



FINAL NLP CLASS PROJECT REPORT

BBC News Analysis

Mahsa Ghaderan

Supervise by

Dr Sauleh Etemadi

July 17, 2021

Department of Computer Engineering
Iran University of Science and Technology

Contents

List of Figures

List of Tables

List of Acronyms

1	Word2Vec	1
1.1	Iran News	1
1.2	Art News	2
1.3	Sport News	2
1.4	Economic News	2
1.5	Science News	3
2	Tokeniation	5
2.1	Sub-word Level	5
2.2	Word Level	6
3	Language Model	9
3.1	LSTM	9
3.2	GRU	9
3.3	Transformer	12
3.4	Temperature	13
3.5	Final Outputs	13
3.5.1	Iran	13
3.5.2	Art	13
3.5.3	Science	14
3.5.4	Economic	14
3.5.5	Sport	15

List of Figures

Figure 1.1	Sub-words vocab example	1
Figure 1.2	Sub-words vocab example	2
Figure 1.3	Sub-words vocab example	2
Figure 1.4	Sub-words vocab example	3
Figure 1.5	Sub-words vocab example	3
Figure 2.1	Sub-words vocab example	6
Figure 3.1	LSTM Architecture	10
Figure 3.2	Perplexity LSTM language model, label = Science	10
Figure 3.3	Generated Text LSTM language model, label = Science	10
Figure 3.4	GRU Architecture	11
Figure 3.5	Perplexity GRU language model, label = Science	11
Figure 3.6	Generated Text GRU language model, label = Science	12
Figure 3.7	Perplexity Transformer language model, label = Science	12
Figure 3.8	Generated Text Transformer language model, label = Science	12
Figure 3.9	Generated Text Transformer language model, label = Iran	13
Figure 3.10	Generated Text Transformer language model, label = Art	14
Figure 3.11	Generated Text Transformer language model, label = Science	14
Figure 3.12	Generated Text Transformer language model, label = Economic	14
Figure 3.13	Generated Text Transformer language model, label = Sport	15

List of Tables

Part 1

Word2Vec

In this part I used assignment2 from cs224n, 2021 Stanford NLP course as my base code. Different Word2Vec models are trained for each label which exist in BBC data set. Headline is concatenated with body of each news for as input text for the model. Each models trains for 40000 iteration. It took about 5 hours for each label. Wights of models are saved is "models/word2vec" directory. Most 30 repeated words are chosen from each label. Images - , - , - and - shows distribution of each model on a 2D map.

1.1 Iran News

After training word2vec model we exoect to see words with same meaning appear near to ech other and words with far meaning and context place far from others on each map.



Figure 1.1: Sub-words vocab example

1.2 Art News

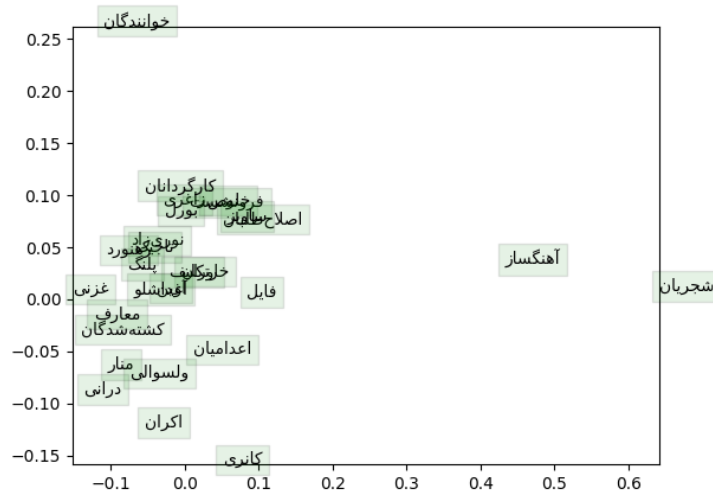


Figure 1.2: Sub-words vocab example

1.3 Sport News

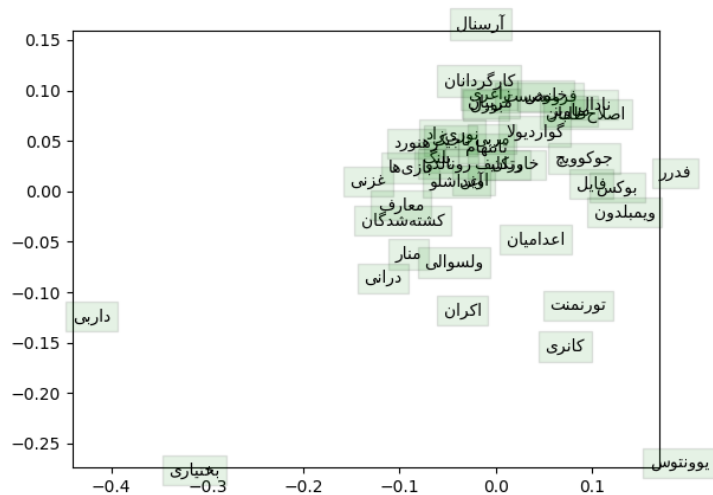


Figure 1.3: Sub-words vocab example

1.4 Economic News

In map 1.4 words in car and it's factory are gathered in the left part of the image.

Part 2

Tokeniation

SentencePiece is an unsupervised text tokenizer and detokenizer python project, which mainly used for text generation task. This part is mainly based on this packet. It uses a model to train what tokens are best to tokenizing a given corpus. In this part 2 different methods of tokenization are implemented. First is tokenizing words and sub-words, in this method the model is looking for most repeated words and sub-words and even letter. As a result we can see almost every word in final version of chosen tokens. On the contrary the second method is just tokenizing words, this model is looking for most repeated words.

Until this part entire headline and body of each news are both covered in the corpus but from this part English values are removed from corpus so tokenization and language generation will focus on Persian.

The corpus is shuffled 5 times and separated to train and test data. Also tokenization is applied on 4 different vocab size due to possible SentencePiece size and performance of output model is evaluated against these vocab size. In the following section, we discuss about effects of vocab size from very small value increasing to a large value.

2.1 Sub-word Level

At first I train tokenizer model on sub-words. It means model not only pays attention to words, but also pays attention to sub-words during training time.

When the vocab size is 1000, it is too small for such a corpus and due to model type, model focuses on smaller sub-words rather than meaning full words. In final choosed-vocab, almost every character in source language exists. So when we tokenize a text

there wouldn't be any <unk> token and most words break into sub-words.

As vocab size increases model learns more complicated sub-words and less meaning-less characters and sub-words. Bot the negative point is as the vocab size increases it would be harder for models to learn task and the positive point is more context will be save after tokenization.

Another point is SentencePiece can discriminate between statrting token and not starting token and ending token of a word in this mode. As we can see in 2.1 the are different form of one token.

```
"<unk>": 3, "ی": 4, "ه": 5, ".": 6, "_": 7,
"ت": 8, "ان": 9, "م": 10, "د": 11, "ل": 12,
"ر": 13, "ش": 14, "است": 15, "ا": 16, "ن": 17,
"می": 18, "ب": 19, "س": 20, "و": 21, "های": 22,
"ز": 23, "ایران": 24, "ی": 25, "ق": 26, "ند": 27,
"ع": 28, "م": 29, "ات": 30, "ور": 31, "ار": 32,
"ف": 33, "ت": 34, "ا": 35, "ن": 36, "خ": 37,
"سال": 38, "ح": 39, "شده": 40, "ج": 41, ",": 42,
"ال": 43, "بر": 44, "یک": 45, "بود": 46, "ست": 47,
```

Figure 2.1: Sub-words vocab example

2.2 Word Level

In this part, i used word level tokenization, in this experience there exist <unk> token due to vocab size and model dealing with choosing best token in order to minimize <unk> token in unseen text.

Same as previous experiment, data-set is divided into train and test set and 5 different shuffled corpus is saved, this help us to have a better evaluation on each vocab size. Four different vocab size is defined as follow: 1000, 5000, 10000, 15000. Number of <unk> token and percent of that over each corpus and vocab is reported in "report/tokenization" directory and each model and output on test set is saved in "src/tokenization/working-dir/words-out". percent of each vocab size is presented in table 2.2.

	word-level					
corpus	1	2	3	4	5	AVG
1000	0.44	0.44	0.44	0.43	0.44	0.44
5000	0.20	0.20	0.20	0.20	0.20	0.21
10000	0.14	0.14	0.14	0.14	0.15	0.15
15000	0.12	0.12	0.12	0.12	0.14	0.13

As it is illustrated in table 2.2, number of <unk> token increase by vocab size reduction. the fewer <unk> token, the fewer information loss. On the other the larger vocab size leads into more amount of computaion for future task. As we saw it is a trade of between these to parameters, in my opinion best vocab size due to this corpus is **10000**. It has reasonable <unk> token and it may need less computation rather than 15000 vocab size.

Part 3

Language Model

Language model neural network architecture is adopted from [pytorch/examples](#) project. Tokenizer codes and models are replaced with Part 2 and vocab for Language Modeling is constructed in part 2. In this part the corpus is divided into three parts, train set, dev set and test set. LanguageModel is trained on each label separately to evaluate capability of each model to generate specific news. Models are saved in "models/lm" directory. Models are trained on GTX1070 GPU. training process is 10 times faster than CPU.

3.1 LSTM

Use LSTM architecture (Figure [3.1](#)) for language model. Best perplexity recorded is 47 on "Science" label. Each epoch lasts about 0.36 seconds, perplexity method starts from 60 and it is 43 on validation data on last epoch.[3.2](#)

In this trial, there are many <unk> tokens in output generated text. But generated sentences are in "Science" context.[3.3](#)

3.2 GRU

Use GRU architecture (Figure [3.4](#)) for language model. Few last epochs and output is illustrated in figure [3.5](#). LSTM and GRU have similar learning power as evidence: perplexity of GRU model on test set is 43, when label is "Science", same as LSTM. Each epoch lasts about 0.33 seconds as expected. GRU model is faster, compared to LSTM model.

In this trial, again there are many <unk> tokens in output generated text. Gener-

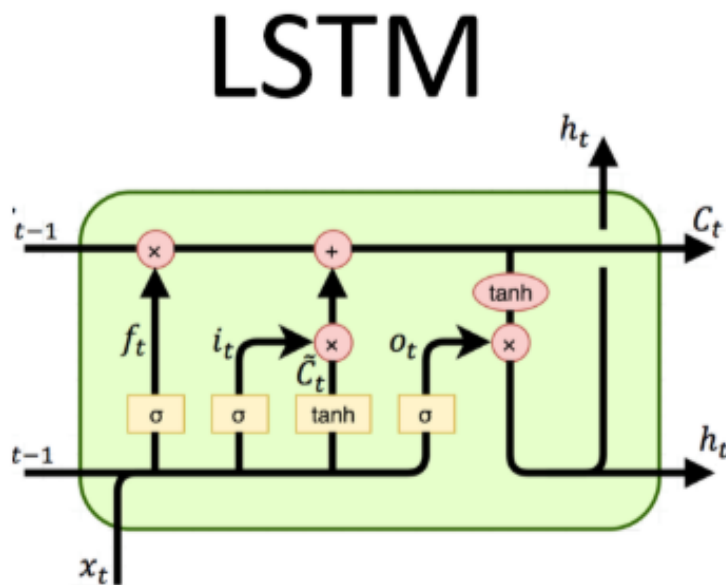


Figure 3.1: LSTM Architecture

```

| end of epoch 35 | time: 0.38s | valid loss 3.86 | valid ppl 47.48
|-----|
| end of epoch 36 | time: 0.36s | valid loss 3.86 | valid ppl 47.47
|-----|
| end of epoch 37 | time: 0.36s | valid loss 3.86 | valid ppl 47.46
|-----|
| end of epoch 38 | time: 0.36s | valid loss 3.86 | valid ppl 47.45
|-----|
| end of epoch 39 | time: 0.36s | valid loss 3.86 | valid ppl 47.42
|-----|
| end of epoch 40 | time: 0.36s | valid loss 3.86 | valid ppl 47.42
|-----|
| End of training | test loss 3.77 | test ppl 43.17
|=====|

```

Figure 3.2: Perplexity LSTM language model, label = Science

```

است شکست دیگری هنگام <unk> برگزاری اجرای بعد <unk> افراد یا <unk> سال <unk> <unk> زمانی کردن داشته همین
جمعه <unk> <unk> محیط وزیر <unk> <unk> <unk> <unk> خطر دلار چند <unk> <unk> کمک جان <unk> بین ویروس
وجود ماه افراد <unk> <unk> <unk> <unk> باری زندگی <unk> <unk> <unk> ده داشته نظیر واکسن آیا چند <unk> دو
فوق <unk> شش ایران خیلی <unk> <unk> همین صورت <unk> <unk> سال <unk> سلامت بود دانشگاه اولین <unk> <unk> هستند.
خواهد ساخته کمی فرودین دو آنها درصد <unk> می است. بیش گونه کشور ممنوع <unk> <unk> <unk> <unk> <unk> ۶
دیگری حال دلار <unk> ویروس زندگی آزمایش بالا <unk> <unk> نقش <unk> تاریخی <unk> <unk> <unk> اولی <unk> <unk>
آزمایش پیش <unk> کره شود. یکی <unk> <unk> <unk> <unk> تغییر <unk> <unk> تصویر کند کرده موارد نیست <unk> یک
<unk> <unk> نیمه <unk> پزشکی نفر همراه درصد تصویر <unk> بودند بین <unk> جمع چند چند <unk> بهبود <unk>
نفر ۵۰ <unk> <unk> <unk> فاصله مرکز است. بود <unk> <unk> <unk> <unk> تعیین است. دانشگاه <unk> <unk> <unk>

```

Figure 3.3: Generated Text LSTM language model, label = Science

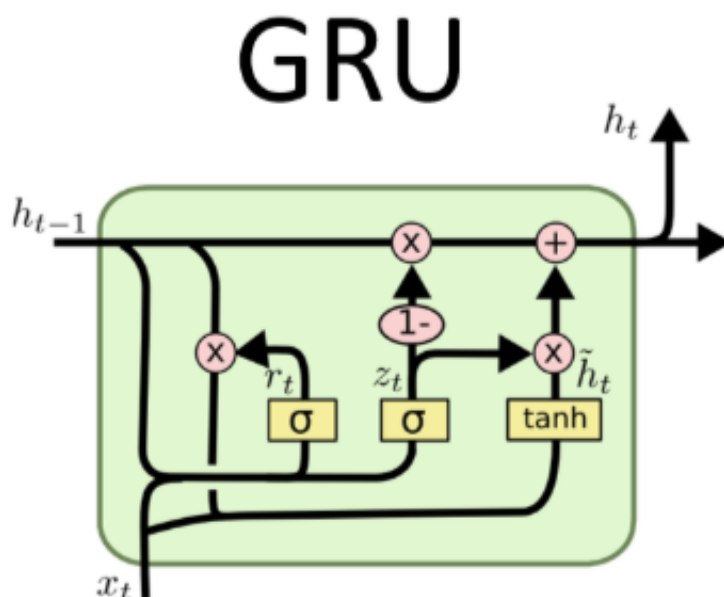


Figure 3.4: GRU Architecture

```

=====
| end of epoch 35 | time: 0.45s | valid loss 3.80 | valid ppl 44.85
=====
| end of epoch 36 | time: 0.32s | valid loss 3.80 | valid ppl 44.81
=====
| end of epoch 37 | time: 0.32s | valid loss 3.80 | valid ppl 44.78
=====
| end of epoch 38 | time: 0.32s | valid loss 3.80 | valid ppl 44.74
=====
| end of epoch 39 | time: 0.34s | valid loss 3.80 | valid ppl 44.68
=====
| end of epoch 40 | time: 0.34s | valid loss 3.80 | valid ppl 44.67
=====
| End of training | test loss 3.69 | test ppl 40.04
=====

```

Figure 3.5: Perplexity GRU language model, label = Science

زمانی کردن داشته خود <unk> <unk> <unk> <unk> <unk> افراد <unk> برگزاری اجرائی بعد <unk> شکست یا هتگام <unk>
فاصله <unk> <unk> <unk> <unk> <unk> وزیر <unk> <unk> <unk> خطر <unk> <unk> هشدار <unk> <unk> کمک جان <unk> بین حاضر
وجود ماه ورود <unk> موسسه <unk> بازی <unk> زندگی <۲۰۲۰> <unk> داشته نظیر واکسن آیا چند پیش دو
کرده <unk> شش هفته خیلی <unk> تأثیر صورت <unk> سال مرگ سلامت بهتر دانشگاه اولین <unk> هستند،
خواهد ساخته است. است این دو آنها درصد <unk> بیش دوره کشور ممنوع <unk> می <unk> <unk> <unk> <unk> ع
دیگری حال داری <unk> وپروس زندگی آزمایش بالا <unk> <unk> کرد <unk> هر تاریخی <unk> <unk> <unk> اولی <unk>
آزمایش پیش <unk> کرده شود. یکی <unk> <unk> <unk> تغییر <unk> <unk> تصویر کند کرده موارد با <unk> یک
<unk> <unk> نیمه <unk> همراه درصد حالا <unk> پزشکی <unk> بودن بین <unk> جمع چند گذشته <unk> <unk> بهبود <unk>
نفر ۵ <unk> <unk> <unk> فاصله مرکز <unk> <unk> <unk> <unk> <unk> تعیین است. دانشگاه <unk> <unk>

Figure 3.6: Generated Text GRU language model, label = Science

```
| end of epoch 37 | time: 0.40s | valid loss 3.74 | valid ppl 42.17
-----
| end of epoch 38 | time: 0.41s | valid loss 3.74 | valid ppl 42.17
-----
| end of epoch 39 | time: 0.41s | valid loss 3.74 | valid ppl 42.17
-----
| end of epoch 40 | time: 0.40s | valid loss 3.74 | valid ppl 42.17
-----
=====
| End of training | test loss 3.62 | test ppl 37.41
=====
```

Figure 3.7: Perplexity Transformer language model, label = Science

ated sentences are in "Education" context.3.6

3.3 Transformer

In last trial Transformer architecture is used for language model. Few Last epochs and output is illustrated in figure 3.7. LSTM and GRU both have almost similar learning power, but with transformer architecture it is possible to access better results. As shown in Figure 3.8 test perplexity is 37.41, when label is "Science" which is 3 score better than previous models. Each epoch last longer than previous models and it is about 0.40 second.

In this trial generated text is shown in 3.8. Number of <unk> token is decreased and there are more meaningful sentences. Generated sentences are in "Science" context.

Run "bash run.sh 4" to train all three types of language models. Models will save

است بر زمین روند <unk> برگزاری اجرای <unk> یا <unk> سال <unk> دارد زمانی می گوید -۲۰۲۰
جمعه <unk> <unk> آب هوایی <unk> ماهه بر سطح قانون <unk> یکی جان <unk> دهه پیش
وجود آمده افراد <unk> موسسه <unk> بازی می گوید کشور ۲۰۲۰ <unk> کنار گذاشته نظیر گفته آبا چند ساعت رو
کشور <unk> شش ماه <unk> همین دلیل <unk> سال <unk> بخش خبری تلویزیون بخش بودند <unk> بین
<unk> آنها <unk> زمین ساخته شده است این <unk> کشور فرانسه <unk> می کنند بیش <unk> <unk> ع-
داشتن <unk> دیگری <unk> پیروس کرونا کمکی <unk> نقش <unk> هر تاریخچین <unk> <unk> اولی <unk>
آزمایش یک <unk> کره زمین <unk> یک <unk> سال آینده اعلام <unk> تصویر کرده موارد بیماری <unk> یک
<unk> نیمه <unk> تصویر <unk> همراه <unk> پزشکی <unk> بودند <unk> کردن ممکن است ضرر بهود <unk>
تفر ۵ <unk> روند تولیدی <unk> دسترسی بوده است تاثير <unk> تعیین کننده حاضر مبتلا کرونا <unk>

Figure 3.8: Generated Text Transformer language model, label = Science

سخت است. جالی جان خود دست وزیر بهداشت گفته بود گفته <unk> نفت <unk> <unk> گفته شده <unk> است بر اثر ابتلا بیماری جان خود دست <unk> دست <unk> بازار بزرگ <unk> آغاز شود. هنوز احتمال مرگ
 جمعه <unk> ایران <unk> وزیر کشور گفته بود <unk> خطر بالای یک <unk> هزار نفر بر روی کار
 مواد قانون اقدام <unk> موسسه سال جاری اعلام کرد یک <unk> آسیا یک میلیون <unk> آبا حسن روحانی دو
 کشور <unk> ایران <unk> همین دلیل <unk> شنبه <unk> بر اساس تصمیم <unk> بین
 ۶ هزار میلیارد دلار بول مصوبه ستاد اجرایی مجموعه دوره اخیر ممنوع است دست <unk> <unk> فروردین دو روز اول
 خط فاریس کشور <unk> نزدیک دولت گروه اول مجلس روزنامه ایران <unk> با یک مرکزی تاریخ <unk> <unk> اول <unk> <unk>
 آرما بش <unk> <unk> کره جنوبی آستانه انتخابات ریاست جمهوری ایران. گروه <unk> <unk> تصویر <unk> کرده جس روحانی رئیس بلیغ
 ایران است. آقای <unk> <unk> چند هریه <unk> بودند <unk> پزشکی قانونی استان درصد رسیده <unk> نیمه اول سال
 داده است. ۵ <unk> فاصله یک روز روند بررسی تهران <unk> <unk> اعلام کرده است <unk> عالی کشور <unk>

Figure 3.9: Generated Text Transformer language model, label = Iran

in "models/lm" directory. Then run "bash run.sh 41" in order to generate text for each label. For each model generated text will save in corresponding file in "reports/lm" directory.

3.4 Temperature

Different temperature can be used when generating text. If set temperature low, generated text will be more general, as a result number of <unk> token will increase. On the other hand by setting higher value for temperature model diversity increases. For this project 1 is best choice for temperature parameter.

3.5 Final Outputs

In this section output of each label on trained transformer-based language model is represented.

3.5.1 Iran

Iran news is largest corpus in this dataset. We can see more meaningful sentences and less <unk> tokens. Generated sentences seems link Iran news and generated-words illustrated in Figure 3.9 are related to the context.

3.5.2 Art

Longest n-gram in generated text without any <unk> token is placed on second line, which n is equal to 10(Figure 3.10). This sequence of words is not meaningful enough, but seems fluent.

Figure 3.10: Generated Text Transformer language model, label = Art

نتیجه گیری کرده است. انسان راه دور سال با برزگاری اجرای آب هوایی ماه بر سطح قانون کمک جان دهه پیش وجود آمده افراد موسسه بازی می گوید کشور ۲۰۲۰ کنار گذاشته نظیر گفته آیا چند ساعت دو کشور شش ماه همین دلیل سال بخش خبری تلویزیون بخش بودن کشور زمین ساخته شده است. این کشور فرانسه می کنند پیش کشور اول داشته دیگری ویروس کرونا کمک نقش هر تاریخچه اول کشور آزمایش یک کره زمین یک سال آینده اعلام تصویر کرده موارد بیماری یک نمط تصویر همراه پزشکی بودند کردن ممکن است مزمن بهبودی نفر ۵۰ روند تولیدی دسترسی بوده است تا این تعیین کننده حاضر مبتلا کرونا

Figure 3.11: Generated Text Transformer language model, label = Science

3.5.3 Science

In this section there are many <unk> tokens due lack of data in Science class. Longest sequence of words is 4-grams and used words are more general than other sections.

3.5.4 Economic

In Economic class generated field as it is shown in figure 3.12 most at least 6-grams are construct of meaningful sequences of words and they sounds same as economic news. We can see many 8-grams without any <unk> token in this section.

سیاسی اطلاعات گسترده درباره <unk> نفت <unk> ای ۱۱ بیست ماه مارس کنون بوده بودند مدت بسیاری <funk>
هوایی بر قیمت خواهند رفت شاهد افت <unk> نفت <unk> غرب <unk> <unk> زمانی بسیاری داشته خود
بازار سرمایه <unk> ده سال گذشته میلادی <unk> توافق <unk> دلار هر <unk> <unk> <unk> اعمال <unk> نفت در
ایران؛ گزارش خود <unk> <unk> سال جاری میلادی ۲۰۲۰ <unk> وقت محلی گفته است آبا شیوع کرونا دو
<unk> ایران <unk> تجاری کمتر دو سال پیش پس بحران کشورها خاطر تحريم نفتي آمریکا درباره <unk> افزایش
۶ میلیون بشکه نفت ایران هنوز <unk> نفت بیش ۲۰ درصد افت نرخ <unk> میلیارد دلار در <unk> نسبت رای
دیگری دلار کشور <unk> بانک مرکزی بریتانیا چند سال اخیر اقدامات بانک ایرانی گفته است <unk> اول <unk> <unk>
آزمایش کووید ۱۹ کره جنوبی شاهد یک میلیون بشکه نفت کشور اعلام <unk> تصویر <unk> کرده است. بانک سرمایه ایران
گفته اطلاع دولت عربستان سعودی چند هنوز <unk> <unk> <unk> <unk> پزشکی ایران همراه درصد عرضه <unk> نیمه اول سال
افت صادرات نفت <unk> نفتی کاهش نرخ هفت میلیون <unk> احتمال فعال زمینه تجارت نفت گاز عراق گفته <unk> <unk>

Figure 3.12: Generated Text Transformer language model, label = Economic

جام_جهانی_فوتبال_ایران_گفت_<unk> تیم_ملی_فوتبال_فرانسه_<unk> <unk> <unk> <unk> <unk> ستاره_با_یرن_<unk>
 همه_<unk> حالا_هنگام_<unk> برگزاری_هر_سال_<unk> <unk> <unk> <unk> <unk> سال_<unk> <unk> زمانی_هنوز_اجازه_حضور
 زبان_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> تیم_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
 وجود_آمده_بود_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
 کشور_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
 ۶_بازیکن_بیشین_فوتبال_ایران_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
 کشور_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
 آزمایش_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
 پس_نتایج_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
 تهران_بوده_کنار_گذاشته_است_پس_دوران_بازی_کند_ترتیب_۵_<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>

Figure 3.13: Generated Text Transformer language model, label = Sport

3.5.5 Sport

In this class, generated text includes of variety of meaningful n-grams without any <unk> token. It is interesting that fluent and meaning 9-grams sequence of words is constructed.(Line 4, Figure 3.13)