دانشکده مهندسی کامپیوتر

استاد درس: دکتر اعتمادی

بهار ۱۴۰۰

گزارش پروژه
درس مبانی پردازش گفتار و زبان

پریسا یل‌سوار
شماره دانشجویی: ۹۶۵۲۲۰۸۷

فهرست مطالب

فهرست تصاویر

فهرست جداول

۱ word2vec

به منظور آموزش بردارها از جملات تمیز شده ۶ ژانری که بیشترین تعداد جملات را داشتند استفاده شده است. کد استفاده شده مربوط به تمرین $a2$ است که برای هر ژانر داده مربوط جایگزین شده است.

برای تحلیل و مقایسه بردارهای ژانرها، ابتدا کلمات مشترک بین این ۶ ژانر استخراج می شوند سپس شباهت کسینوسی بردارها محاسبه می شود. کد این بخش در فایل `compare.py` قرار دارد. در مجموع ۱۱۱۴۶ کلمه مشترک بین ژانرها وجود دارد که تعدادی از آنها در ادامه بررسی شده اند. در شکل ۱ شباهت کسینوسی بردارهای کلمه attacker بین ژانرها مقایسه شده است. در این بین ژانرهای Drama و Romance بیشترین شباهت را به هم دارند. دلیل آن می تواند ظاهر شدن این کلمه در context مشابه در داده هر دو ژانر باشد. همچنین Fantasy و Romance کمترین شباهت را دارند که می تواند حاصل از تفاوت context این کلمه در داده این ژانرها باشد.

	Tags	Cosine Similarity
0	('Action', 'Adventure')	0.11009188075086367
1	('Action', 'Comedy')	-0.4530729307682333
2	('Action', 'Drama')	0.1824108417293962
3	('Action', 'Fantasy')	-0.11671720021598499
4	('Action', 'Romance')	-0.030160783722501985
5	('Adventure', 'Comedy')	0.07606716333797356
6	('Adventure', 'Drama')	0.24524849237058421
7	('Adventure', 'Fantasy')	0.21541118363541506
8	('Adventure', 'Romance')	0.2720475525589632
9	('Comedy', 'Drama')	0.45427159360488356
10	('Comedy', 'Fantasy')	0.38700226039911984
11	('Comedy', 'Romance')	-0.0639884168905818
12	('Drama', 'Fantasy')	0.3800029763527514
13	('Drama', 'Romance')	-0.2442914956289752
14	('Fantasy', 'Romance')	-0.2696859615358394

شکل ۱: شباهت کسینوسی بردارهای کلمه attacker

در شکل ۲ شباهت کسینوسی بردارهای کلمه everyone مقایسه شده است. انتظار می رود که این کلمه تقریبا برای همه ژانرها بردار مشابهی داشته باشد؛ اما بردار این کلمه بین ژانرهای `Action-Fantasy`، `Action-Romance` و `Comedy-Drama` تفاوت دارد. این تفاوت می تواند ناشی از تفاوت در context کلمه در داده این ژانرها یا کوچک بودن بردارها و کم بودن اطلاعات آنها باشد.

	Tags	Cosine Similarity
0	('Action', 'Adventure')	0.04816044274639685
1	('Action', 'Comedy')	0.20623759282056608
2	('Action', 'Drama')	0.11590228583249115
3	('Action', 'Fantasy')	-0.12217501116961327
4	('Action', 'Romance')	-0.3301366868267426
5	('Adventure', 'Comedy')	0.4768757009648566
6	('Adventure', 'Drama')	0.1686779985740613
7	('Adventure', 'Fantasy')	0.2442873331805671
8	('Adventure', 'Romance')	0.5513675489442893
9	('Comedy', 'Drama')	-0.2743682915543757
10	('Comedy', 'Fantasy')	0.0700663746518452
11	('Comedy', 'Romance')	0.38597128771380085
12	('Drama', 'Fantasy')	0.17351859663977764
13	('Drama', 'Romance')	0.3947251229728395
14	('Fantasy', 'Romance')	0.2606201987158389

شکل ۲: شباهت کسینوسی بردارهای کلمه everyone

یکی دیگر از کلماتی که انتظار می رود برای همه ژانرها تقریبا یکسان باشد، کلمه brother است که در شکل ۳ آورده شده است. اما بردار این کلمه برای ژانرهای Action-Drama بیشترین میزان تفاوت را دارد که می تواند ناشی از تفاوت این دو ژانر با هم باشد.

	Tags	Cosine Similarity
0	('Action', 'Adventure')	0.03480053742660397
1	('Action', 'Comedy')	0.02011593687714632
2	('Action', 'Drama')	0.3900265532965037
3	('Action', 'Fantasy')	0.0203860129734457
4	('Action', 'Romance')	0.42570381170741545
5	('Adventure', 'Comedy')	0.5604739185561 22
6	('Adventure', 'Drama')	-0.0147292924059324 04
7	('Adventure', 'Fantasy')	-0.12276176132956862
8	('Adventure', 'Romance')	0.398507820245191
9	('Comedy', 'Drama')	0.12725797917359832
10	('Comedy', 'Fantasy')	0.06502641914861733
11	('Comedy', 'Romance')	0.29395033706213936
12	('Drama', 'Fantasy')	0.5086532117743339
13	('Drama', 'Romance')	0.42887899675187824
14	('Fantasy', 'Romance')	0.14529416692852276

شکل ۳: شباهت کسینوسی بردارهای کلمه brother

نکته مهم این است که اندازه ابعاد این بردارها ۱۰ است. از این رو اطلاعات زیادی را در خود ذخیره نمی کنند و دلیل برخی شباهت ها یا تفاوت های دور از انتظار نیز همین است. اگر ابعاد بردارها بیشتر باشد بهتر می توان شباهت کسینوسی آنها را بررسی کرد.

۲ tokenization

در این قسمت مدل SentencePiece با اندازه واژگان ۳۳، ۴۰۰، ۲۰۰۰، ۱۰۰۰۰ و ۲۷۵۲۴ و در حالت $model\_type = word$ آموزش می بیند. در هر دور از آموزش $\frac{1}{5}$ از کل داده برای ارزیابی کنار گذاشته می شود تا درصد توکن <unk> روی آن محاسبه شود. درصد توکن <unk> به ازای اندازه کلمات متفاوت در شکل ۴ آورده شده است. همان طور که پیداست هر چه تعداد کلمات tokenizer کمتر باشد تعداد توکن های ناشناخته بیشتر است. اما با بیشتر کردن تعداد کلماتی که tokenizer می تواند تشخیص دهد، تعداد وقوع توکن <unk> کاهش می یابد.

	Vocab Size	Unk%
0	33.0	60.4627517549909
1	400.0	41.0461121594813
2	2000.0	28.070584552281957
3	10000.0	16.88668048860856
4	27524.0	12.18298909633759

شکل ۴: درصد توکن <unk> به ازای تعداد کلمات متفاوت

برای مدل نهایی، تعداد کلمات ۱۰۰۰۰ انتخاب شده است. چرا که کمترین درصد توکن <unk> را دارد و به نسبت اندازه مناسبی دارد(مدل bert ۳۰ هزار توکن دارد [۱]).

۳ parsing

برای این قسمت از مدل آموزش دیده تمرین سوم استفاده شده است. جملات انتخاب شده با استفاده از سایت corenlp برچسب گذاری شدند. دقت مدل روی این ۱۰ جمله $\texttt{UAS} = 72.92$ است. از آنجایی که در مرحله پیش پردازش، کلمات پرتکرار (stopwords) از مجموعه داده حذف شده اند و بر روی کلمات Lemmatization اعمال شده است؛ عملکرد مدل آموزش دیده روی این جملات محدود است. برای مثال، مدل برای جمله زیر 62.5% از روابط را به درستی پیش بینی می کند:

one day hokoda meet hina busy street corner

برای این جمله head کلمه one به جای day به اشتباه کلمه meet پیش بینی می شود. همچنین head کلمه busy به جای street به اشتباه corner پیش بینی می شود. خطای مدل در اینجا به دلیل حذف کردن stopwords و استفاده از Lemmatization است؛ زیرا صورت کامل جمله که در زیر آمده است توسط مدل کاملا صحیح برچسب گذاری می شود.

one day hokoda meets hina in a busy street corner

مثال دیگری از جملاتی که به دلیل حذف شدن stopwords از داده، dependancy parse آنها اشتباه به دست می آید در زیر آورده شده است:

luffy meet friend thought dead

صورت کامل جمله:

luffy meets a friend who thought wass dead

در این جمله به دلیل حذف شدن کلمات میانی، مدل به اشتباه head کلمه friend را thought پیش بینی می کند در صورتی که meet صحیح است. همچنین head کلمه thought را meet پیش بینی می کند در صورتی که dead صحیح است و head کلمه dead را thought پیش بینی می کند در حالی که کلمه friend صحیح است.

علاوه بر مشکل ذکر شده، برخی از برچسب گذاری های نادرست به دلیل کمبود داده به وجود می آیند. برای مثال:

however everything going magically well

صورت کامل جمله:

however, everything was going magically well

مدل برای این جمله فقط head کلمه away را به درستی پیش بینی می کند. از آنجایی که مدل روی صورت کامل جمله دقت 33.3% دارد؛ گمان می رود با آموزش مدل روی داده بیشتر بتوان این مشکلات را نیز برطرف کرد.

Sentence	Genre
boy bikie skull weapon rich renewed	Action
boy sinner welcoming meant back power	Adventure
boy rakuga forest omen witness challenge	Fantasy
boy week currently girl called haruka	Drama
boy chief giving always wife food	Comedy
boy masamune would bound tattoo heart	Romance

جدول ۱: مقایسه جملات تولید شده language model

Sentence	Genre
help destroy kirigakure three fantasy universe	Action
help bao impaled unworldly old robot	Adventure
help accident target saintess begin power	Fantasy
help trusted spin giving distant home	Drama
help study letter group haruka kiriko	Comedy
help expected front follows potential	Romance

جدول ۲: مقایسه جملات تولید شده language model

۴ language model

در این قسمت از ۳ لایه LSTM با اندازه ۱۲۸ و اندازه embedding ۱۲۸ به عنوان مدل زبانی استفاده شده است [۲]. مدل هر ژانر از مجموعه داده جملات تمیز شده استفاده می کند (مدل زبانی در سطح کلمه است) و به مدت ۵ دور با طول sequence ۱۵ آموزش می بیند. جملات تولید شده در کل کیفیت مناسبی ندارند، یکی از دلایل آن حذف کردن stopwords و استفاده از Lemmatizer روی داده اولیه است. دلیل دوم کم بودن داده و کوتاه بودن مدت زمان آموزش مدل زبانی ست. مثال اول از جملات تولید شده توسط این مدل ها در جدول ۱ آورده شده است. ژانر Action، Drama و Romance به نسبت جمله مناسبی تولید کرده اند.

مثال دوم جملات تولید شده در جدول ۲ آورده شده است. در این مثال می توان گفت تقریبا همه مدل ها به جز Romance کلمات مرتبط با ژانر خود تولید کرده اند.

مثال های آورده شده از مجموعه پیش بینی های مدل ها دستچین شده اند. فایل کامل کلمات تولید شده توسط هر مدل با اسم *<label>_predictions.txt* در پوشه *reports* قرار دارد.

Perpelexity	Genre
735.4264358106171	Action
850.6009242951887	Adventure
762.2966998333878	Fantasy
630.12191391778	Drama
621.3412662654786	Comedy
577.2775060373292	Romance

جدول ۳: مقایسه perplexity مدل زبانی ژانرها fine-tune

Sentence	Genre
an accident NUM year ago left mrunas	Action
an accident dimension finally meet fate new member group come	Adventure
an accident happens people think wrongperson named ik	Fantasy
an accident set fire inside roomtakanagi	Drama
an accident also happens accidentally end yukichi girl	Comedy
an accident dimension world called nanamiunusual day	Romance

جدول ۴: مقایسه جملات تولید شده fine-tune

۵ fine tuning

۱.۵ مدل زبانی

از بین ۱۲ ژانر موجود، ۶ ژانر اول با بیشترین تعداد جمله برای تسک fine-tune انتخاب شدند. در این قسمت از کد huggingface [۳] و مدل Distil GPT2 برای fine-tune روی داده هر کلاس استفاده شده است. مدل زبانی هر ژانر از نوع causal است و برای ۶ دور با batch size ۱۰۰۰ fine-tune می شود. همچنین داده هر ژانر به نسبت ۸۰ به ۲۰ به دو قسمت آموزش و ارزیابی تقسیم می شود. از داده ارزیابی در زمان fine-tune و در پایان برای اندازه گیری perplexity استفاده می شود. از دیگر تنظیمات مدل می توان به اعمال weight decay و توقف زودهنگام (early stopping) اشاره کرد. مقایسه perplexity مدل ها در جدول ۳ آورده شده است. از آنجایی که زمان fine-tune کوتاه بوده و تعداد جملات برای هر ژانر زیاد نیست، perplexity ها مقادیر خوبی ندارند.

تعدادی از جملات تولید شده توسط مدل ها در جدول ۴ بررسی شده اند. کلمات آبی شده توسط مدل تولید شده اند. کلمات ناقص به دلیل حذف کردن stopwords از داده اصلی به وجود آمده اند. در این مثال ژانرهای Action، Adventure و Drama جملات مناسب و مرتبطی تولید کرده اند.

در مثال دوم در جدول ۵ جمله تولید شده توسط ژانرهای Action، Adventure و Fantasy کاملا حال و فضای ژانر را دارد و بقیه ژانرها نیز به نسبت کلمات مناسبی تولید کرده اند.
مثال آورده شده در جدول ۶، جملات مناسب و مورد انتظاری را برای ژانرهای Adventure، Drama و Fantasy نشان می دهد. اما در این مثال، جمله ژانرهای Comedy و Romance چندان واضح نیست. از دلایل آن می توان به کم بودن داده برای ژانر Comedy و کم بودن زمان fine-tune اشاره کرد.

Sentence	Genre
he was from far away girl come new world	Action
he was from the previous generation ai japan	Adventure
he was from another dimension time travel another dimension new	Fantasy
he was from hino one day find dead another	Drama
he was from home village made thing even worsehow	Comedy
he was from heaven told never say anything one day	Romance

جدول ۵: مقایسه جملات تولید شده fine-tune

Sentence	Genre
people from nowhere back world people world living nothing else	Action
people from zeros adventure game world game earth	Adventure
people from childhood suddenly became good friend classmate father	Fantasy
people from south europee nation countrynats	Drama
people from yuusuke also known haru	Comedy
people from different world come take home money school student	Romance

جدول ۶: مقایسه جملات تولید شده fine-tune

در مجموع جملات تولید شده توسط مدل های fine-tune شده، کیفیت و خوانایی بهتری نسبت به مدل های قسمت قبل دارد که این نشان دهنده مزایای استفاده از مدل های از پیش آموزش دیده است. مثال های آورده شده از مجموعه پیش بینی های مدل ها دستچین شده اند. فایل کامل کلمات تولید شده توسط هر مدل با اسم *predictions.txt* در پوشه *reports* قرار دارد.

۲.۵ رده بندی

چالش رده بندی ژانر بر اساس متن multilabel classification نام دارد. در این قسمت، هر خلاصه ۱۲ برچسب دارد و بر پایه اینکه یک ژانر به این خلاصه تعلق داشته باشد مقدار صفر یا یک می گیرد [۴]. مدل bert-base-uncased برای ۴ دور با تکنیک cross validation و تابع خطا binary cross entropy آموزش می بیند. در پایان آموزش، دقت مدل روی داده validation 18.52 و F1 score آن 58.67 است.

شکل ۵ عملکرد مدل را روی داده تست نشان می دهد. پارامتر support نشان دهنده تعداد خلاصه ها با یک ژانر مشخص هستند. بیشترین دقت متعلق به ژانر Romance است چون این ژانر بیشترین تعداد خلاصه را دارد. ژانر های School life و Fantasy نیز دقت بالایی دارند که می تواند به دلیل شباهت یا همراه بودن با ژانر Romance باشد (انیمه هایی که ژانر Romance دارند اکثر ژانرهای School life و Fantasy نیز دارند). ژانر Drama دقت و recall پایینی دارد. دلیل آن می تواند برجسته نبودن ژانر Drama باشد، به عبارت دیگر در مقایسه با ژانر Action یا Romance که کلمات مشخصی از قبیل fight یا love دارند، این ژانر کلمات مشخصی ندارد به همین دلیل تشخیص آن سخت تر است. ژانر Sci Fi نیز به دلیل کم بودن تعداد داکیومنت ها recall پایینی دارد که با افزایش داده می توان این مشکل را حل کرد. فایل *comparisons.csv* تمام پیش بینی ها و برچسب های درست برای داده تست نشان می دهد.

	precision	recall	f1-score	support
Action	0.67	0.65	0.66	873
Adventure	0.62	0.42	0.50	462
Comedy	0.65	0.48	0.55	1145
Drama	0.55	0.40	0.46	1111
Fantasy	0.72	0.65	0.68	819
Romance	0.74	0.72	0.73	1560
School Life	0.71	0.46	0.56	657
Sci Fi	0.61	0.39	0.47	312
Shoujo	0.59	0.45	0.51	476
Shounen	0.68	0.18	0.29	489
Supernatural	0.66	0.38	0.48	468
Yaoi	0.69	0.69	0.69	505
micro avg	0.67	0.52	0.59	8877
macro avg	0.66	0.49	0.55	8877
weighted avg	0.66	0.52	0.58	8877
samples avg	0.64	0.54	0.55	8877

شکل ۵: عملکرد مدل رده بند روی داده تست

مدل آموزش دیده در این لینک قرار دارد.

مراجع

[1] D. Dhami, *Understanding bert — word embeddings*, Web Page. [Online]. Available: https://medium.com/@dhartidhami/understanding-bert-word-embeddings-7dc4d2ea54ca#:~:text=The%20original%20word%20has%20been,represented%20as%20subwords%20and%20characters..

[2] D. Bitvinskas, *Pytorch lstm: Text generation tutorial*, Web Page. [Online]. Available: https://closeheat.com/blog/pytorch-lstm-text-generation-tutorial.

[3] huggingface, *Language_modeling*, Web Page. [Online]. Available: https://github.com/huggingface/notebooks/blob/master/examples/language_modeling.ipynb.

[4] R. Patel, *Transformers for multi-label classification made simple.* Web Page. [Online]. Available: https://towardsdatascience.com/transformers-for-multilabel-classification-71a1a0daf5e1.