به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق وکامپیوتر

درس پردازش زبان های طبیعی

پروژه شماره : 1

نام و نام خانوادگی : محمدرضا بختیاری

شماره دانشجویی : 810197468

اسفند 1400

فهرست سوالات

گام اول

مقایسه ی BPE و WordPiece

دو الگوریتم BPE و WordPiece در کلیت امر نشانه گذاری[1] یکسان عمل می کنند. هر دو به دنبال پیدا کردن یک جفت کاراکتر در هر مرحله و ترکیب آن ها و اضافه کردن واژه جدید به مجموعه واژگان هستند. تفاوت این دو الگوریتم در رویکرد آن ها برای پیدا کردن جفت کاراکتر در هر مرحله می باشد.

الگوریتم BPE در هر مرحله جفت کاراکتری را انتخاب می کند که بیشترین تکرار را در متن داشته باشد[2]. در مقابل WordPiece به جای تکیه بر فراوانی جفت ها ، جفتی را انتخاب می کند که احتمال داده های آموزشی[3] را به حداکثر می رساند. این به این معنی است که یک مدل زبان را آموزش می‌دهد که از واژگان پایه شروع می‌شود و جفتی را با بیشترین احتمال (جفت = کاراکتر واژگان پایه + کاراکتر تولید شده با بیشترین احتمال) انتخاب می‌کند. این جفت به واژگان اضافه می شود و مدل زبان دوباره بر روی واژگان جدید آموزش داده می شود. این مراحل تا رسیدن به دایره لغات مورد نظر تکرار می شوند.

[1] Tokenization
[2] Most frequent pair
[3] Likelihood of the training data

پیاده سازی دستی BPE

در گام اول از روی مجموعه نوشته های[1] داده شده , لغات اولیه که شامل حروف یکتای به کار برده شده در متن است را استخراج می کنیم و مجموعه واژگان[2] را تشکیل می دهیم.

پیش از شروع کار به انتهای هر لغت , کاراکتر '_' را اضافه می کنیم تا مشخص کننده ی پایان هر لغت باشد.

در هر مرحله دو جفت واژه ای که با بیشترین تکرار پشت سر هم می آیند را پیدا کرده و پس از ترکیب آن ها و مشخص کردن آن ها به عنوان یک واژه جدید , آن جفت را به مجموعه واژگان اضافه کرده و به همین ترتیب مراحل را ادامه می دهیم.

شرط پایان این فرایند را می توانیم به دو صورت در نظر بگیریم. می توانیم تعداد تکرار[3] مشخصی را در نظر بگیریم و پس از تکرار شدن حلقه به تعداد مورد نظر , فرایند را خاتمه دهیم. یا می توانیم از ابتدا شرط مشخصی بر روی تعداد واژه های موجود در مجموعه واژگان در نظر بگیریم که در اینجا این رویکرد را انتخاب می کنیم و تا رسیدن به 25 واژه در مجموعه واژگان (معادل با 14 تکرار) فرایند را انجام می دهیم.

در هر گام , مجموعه واژگان جدید , مجموعه نوشته های جدید و زوج انتخابی برای ترکیب شدن را نمایش می دهیم:

```
Step: 0

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n']
Dictionary:  {'l o w _': 5, 'l o w e r _': 2, 'w i d e s t _': 3, 'n e w e s t _': 5}
Combined pairs: ('e', 's')
--------------
Step: 1

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es']
Dictionary:  {'l o w _': 5, 'l o w e r _': 2, 'w i d es t _': 3, 'n e w es t _': 5}
Combined pairs: ('es', 't')
```

شکل 1 : مجموعه واژگان و مجموعه نوشته ها به همراه زوج های انتخابی در گام 0 و 1

[1] Corpus
[2] Vocabulary
[3] Iteration

```
--------------
Step: 2

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est']
Dictionary:  {'l o w _': 5, 'l o w e r _': 2, 'w i d est _': 3, 'n e w est _': 5}
Combined pairs: ('est', '_')
--------------
Step: 3

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_']
Dictionary:  {'l o w _': 5, 'l o w e r _': 2, 'w i d est_': 3, 'n e w est_': 5}
Combined pairs: ('l', 'o')
--------------
Step: 4

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo']
Dictionary:  {'lo w _': 5, 'lo w e r _': 2, 'w i d est_': 3, 'n e w est_': 5}
Combined pairs: ('lo', 'w')
--------------
Step: 5

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low']
Dictionary:  {'low _': 5, 'low e r _': 2, 'w i d est_': 3, 'n e w est_': 5}
Combined pairs: ('low', '_')
--------------
Step: 6

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_']
Dictionary:  {'low_': 5, 'low e r _': 2, 'w i d est_': 3, 'n e w est_': 5}
Combined pairs: ('n', 'e')
Step: 7

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne']
Dictionary:  {'low_': 5, 'low e r _': 2, 'w i d est_': 3, 'ne w est_': 5}
Combined pairs: ('ne', 'w')
--------------
Step: 8

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne', 'new'
]
Dictionary:  {'low_': 5, 'low e r _': 2, 'w i d est_': 3, 'new est_': 5}
Combined pairs: ('new', 'est_')
--------------
Step: 9

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne', 'new'
, 'newest_']
Dictionary:  {'low_': 5, 'low e r _': 2, 'w i d est_': 3, 'newest_': 5}
Combined pairs: ('w', 'i')
--------------
Step: 10

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne', 'new'
, 'newest_', 'wi']
Dictionary:  {'low_': 5, 'low e r _': 2, 'wi d est_': 3, 'newest_': 5}
Combined pairs: ('wi', 'd')
Step: 11

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne', 'new'
, 'newest_', 'wi', 'wid']
Dictionary:  {'low_': 5, 'low e r _': 2, 'wid est_': 3, 'newest_': 5}
Combined pairs: ('wid', 'est_')
--------------
Step: 12

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne', 'new'
, 'newest_', 'wi', 'wid', 'widest_']
Dictionary:  {'low_': 5, 'low e r _': 2, 'widest_': 3, 'newest_': 5}
Combined pairs: ('low', 'e')
--------------
Step: 13

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne', 'new'
, 'newest_', 'wi', 'wid', 'widest_', 'lowe']
Dictionary:  {'low_': 5, 'lowe r _': 2, 'widest_': 3, 'newest_': 5}
Combined pairs: ('lowe', 'r')
--------------
Step: 14

Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne', 'new'
, 'newest_', 'wi', 'wid', 'widest_', 'lowe', 'lower']
Dictionary:  {'low_': 5, 'lower _': 2, 'widest_': 3, 'newest_': 5}
Combined pairs: ('lower', '_')
```

شکل 2 : مجموعه واژگان و مجموعه نوشته ها به همراه زوج های انتخابی در گام های 2 تا 14

در انتها نیز مجموعه واژگان نهایی به همراه مجموعه نوشته نهایی را گزارش می کنیم:

```
Final Dictionary:  {'low_': 5, 'lower_': 2, 'widest_': 3, 'newest_': 5}
Final Vocabulary:  ['l', 'o', 'w', '_', 'e', 'r', 'i', 'd', 's', 't', 'n', 'es', 'est', 'est_', 'lo', 'low', 'low_', 'ne',
 'new', 'newest_', 'wi', 'wid', 'widest_', 'lowe', 'lower', 'lower_']
```

شکل 3 : مجموعه واژگان نهایی به همراه مجموعه نوشته نهایی

همان طور که در مجموعه واژگان نهایی مشاهده می کنیم با ترکیب دو واژه ’est_‘ + ’low‘ می توان کلمه ’lowest‘ را تولید کرد.

خروجی نیز همان طور که انتظار داشتیم به دست آمد:

```
Input Out of Vocabulary is :  lowest_
Intended Pairs are :  ['low', 'est_']
```

گام دوم

پیاده سازی با استفاده از کتابخانه hugging face

در این قسمت ابتدا مدل را با استفاده از دو الگوریتم خواسته شده پیاده سازی می کنیم در ادامه مدل را بر روی متن ورودی جهت ارزیابی توکنایزرهای آموزش داده شده تست می کنیم.

در انتها نیز برای ارزیابی نهایی هر یک از الگوریتم ها بر روی توکن های خارج از واژگان[1] متن ورودی , هر یک از دو مدل BPE و WordPiece را با مدل Unigram مقایسه می کنیم.

الف– نتایج پیاده سازی مدل بر روی متن ورودی داده شده با استفاده از مجموعه نوشته Gutenberg:

1– با استفاده از الگوریتم BPE :

['This', 'is', 'a', 'deep', 'learning', 'to', 'ken', 'ization', 't', 'ut', 'or', 'ial', '.', 'T', 'ok', 'en', 'ization', 'is', 'the', 'first', 'step', 'in', 'a', 'deep', 'learning', 'N', 'L', 'P', 'pi', 'pe', 'line', '.', 'We', 'will', 'be', 'comparing', 'the', 'to', 'k', 'ens', 'generated', 'by', 'each', 'to', 'ken', 'ization', 'model', '.', 'Ex', 'c', 'ited', 'much', '?', '!', '<UNK>']

که در این حالت تعداد توکن ها 55 می باشد.

2– با استفاده از الگوریتم WordPiece :

['This', 'is', 'a', 'deep', 'learning', 'to', '##ken', '##ization', 't', '##ut', '##oria', '##l', '.', 'To', '##ken', '##ization', 'is', 'the', 'first', 'step', 'in', 'a', 'deep', 'learning', 'N', '##L', '##P', 'pip', '##el', '##ine', '.', 'We', 'will', 'be', 'comparing', 'the', 'to', '##ken', '##s', 'generated', 'by', 'each', 'to', '##ken', '##ization', 'model', '.', 'Ex', '##ci', '##ted', 'much', '<UNK>']

که در این حالت تعداد توکن ها 52 می باشد.

ب– نتایج پیاده سازی مدل بر روی متن ورودی داده شده با استفاده از مجموعه نوشته Wikitext:

1– با استفاده از الگوریتم BPE :

['This', 'is', 'a', 'deep', 'learning', 'to', 'ken', 'ization', 'tut', 'orial', '.', 'Tok', 'en', 'ization', 'is', 'the', 'first', 'step', 'in', 'a', 'deep', 'learning', 'NL', 'P', 'pipeline', '.', 'We', 'will', 'be', 'comparing', 'the', 'tok', 'ens', 'generated', 'by', 'each', 'to', 'ken', 'ization', 'model', '.', 'Ex', 'cited', 'much', '?', '!', '<UNK>']

که در این حالت تعداد توکن ها 47 می باشد.

2– با استفاده از الگوریتم WordPiece :

['This', 'is', 'a', 'deep', 'learning', 'to', '##ken', '##ization', 'tut', '##orial', '.', 'Tok', '##eni', '##za', '##ti', '##on', 'is', 'the', 'first', 'step', 'in', 'a', 'deep', 'learning', 'NL', '##P', 'pipeline', '.', 'We', 'will', 'be', 'comparing', 'the', 'to', '##ken', '##s', 'generated', 'by', 'each', 'to', '##ken', '##ization', 'model', '.', 'Exc', '##ited', 'much', '<UNK>']

[1] Out of vocabulary

که در این حالت تعداد توکن ها 48 می باشد.

مشاهده می کنیم الگوریتم BPE زمانی که بر روی مجموعه داده های کوچکتر (گوتنبرگ) آموزش داده می شود ، 55 توکن و زمانی که بر روی مجموعه داده بزرگتر (ویکی تکست) آموزش داده می شود ، 47 نشانه ایجاد می کند. این نشان می‌دهد که این الگوریتم می تواند جفت های بیشتری از کاراکترها را با آموزش روی یک مجموعه داده بزرگ تر ادغام کند.

همچنین دیده می شود هر دو الگوریتم زمانی که روی مجموعه داده بزرگ تری (ویکی تکست) آموزش داده می‌شوند ، نشانه های زیرکلمه‌ای[1] بدتر و بهتر تولید می‌کنند.

اکنون نمای کلی از خروجی توکن های به دست آمده از هر سه الگوریتم (مدل Unigram را برای مقایسه و ارزیابی دو مدل دیگر استفاده کرده ایم) را مشاهده می کنیم.

index	BPE	UNI	WPC
0	This	This	This
1	is	i	is
2	a	s	a
3	deep	a	deep
4	learning	deep	learning
5	to	learn	to
6	ken	ing	##ken
7	ization	t	##ization
8	tut	o	tut
9	orial	ken	##orial
10	.	ization	.
11	Tok	t	Tok
12	en	u	##eni
13	ization	t	##za
14	is	o	##ti
15	the	rial	##on
16	first	.	is
17	step	T	the
18	in	o	first
19	a	ken	step
20	deep	ization	in
21	learning	i	a
22	NL	s	deep
23	P	the	learning
24	pipeline	first	NL
25	.	step	##P
26	We	in	pipeline
27	will	a	.
28	be	deep	We
29	comparing	learn	will
30	the	ing	be
31	tok	N	comparing
32	ens	L	the
33	generated	P	to
34	by	p	##ken
35	each	i	##s
36	to	p	generated
37	ken	e	by
38	ization	line	each
39	model	.	to
40	.	W	##ken
41	Ex	e	##ization
42	cited	will	model
43	much	be	.
44	?	com	Exc
45	!	par	##ited
46	<UNK>	ing	much

شکل 4 : توکن های متن ورودی به ازای سه الگوریتم BPE , WordPiece , Unigram

[1] Subword tokens

index	BPE	UNI	WPC
count	68	68	68
unique	37	41	37
top	<PAD>	o	<PAD>
freq	21	5	20

شکل 5 : نمای کلی از مشخصات توکن های خروجی به ازای استفاده از هر الگوریتم

در نهایت با استفاده از مدل Unigram , عملکرد هر الگوریتم را بر روی توکن های OOV متن ورودی بررسی می کنیم. به این صورت که مجموعه توکن های بدست آمده از هر دو الگوریتم BPE و WordPiece را از مجموعه توکن های بدست آمده از الگوریتم Unigram کم می کنیم (در نهایت مجموعه به دست آمده شامل توکن هایی خواهد بود که در خروجی Unigram بوده ولی در خروجی های BPE و یا WordPiece نبوده)

1- Unigram – BPE :

```
{'L','N','T','W','com','d','e','generate','i','ing','learn','line',
 'o', 'p', 'par', 'rial', 's', 't', 'u', '😍'}
```

که مجموعه به دست آمده شامل 20 توکن می باشد.

2- Unigram – WordPiece :

```
{'!', '?', 'Ex', 'L', 'N', 'P', 'T', 'W', 'cited', 'com', 'd','e',
 'generate', 'i', 'ing', 'ization', 'ken', 'learn', 'line', 'o', 'p',
 'par', 'rial', 's', 't', 'u', '😍'}
```

که مجموعه به دست آمده شامل 27 توکن می باشد.

بر اساس نوع توکن های تولید شده ، به نظر می رسد الگوریتم WordPiece توکن های زیرکلمه ای را تولید می کند که بیشتر در زبان انگلیسی یافت می شوند.

همچنین یکی از اشکالات الگوریتم BPE , استفاده از فرکانس به‌عنوان عامل محرک است که می‌تواند منجر به کدگذاری های نهایی مبهم شود که ممکن است برای متن ورودی جدید مفید نباشند. اما هنوز هم از نظر تولید توکن های بدون ابهام دامنه پیشرفت دارد.

گام سوم

تعداد توکن های هر دو الگوریتم بر روی متن کتاب گوتنبرگ

مشابه با قسمت قبل با این تفاوت که در اینجا مرجع فقط کتاب گوتنبرگ است و متن ورودی جهت ارزیابی نیز به جای (This is a deep learning …) دیتا ست ویکی پدیا و کتاب گوتنبرگ است , مدل را پیاده سازی کرده و تعداد توکن های به دست آمده از هر الگوریتم را گزارش می کنیم:

جدول 1-1 :

ردیف	نام الگوریتم استفاده شده برای توکنایز	تعداد توکن های خروجی الگوریتم برای کتاب گوتنبرگ	
		توکنایزر آموزش داده شده بر روی کتاب گوتنبرگ	توکنایزر آموزش داده شده بر روی کل داده های ویکی پدیا
1	Byte Pair Encoding (BPE)	**122739**	**140947**
2	WordPiece	**122739**	**143651**

پیوست:

تمامی فایل های ipynb در پوشه ی Codes موجود می باشد. همچنین تمامی کد ها در محیط colab اجرا و تست شده اند.