حانیه ناصری 810101286
سوالات تئوری

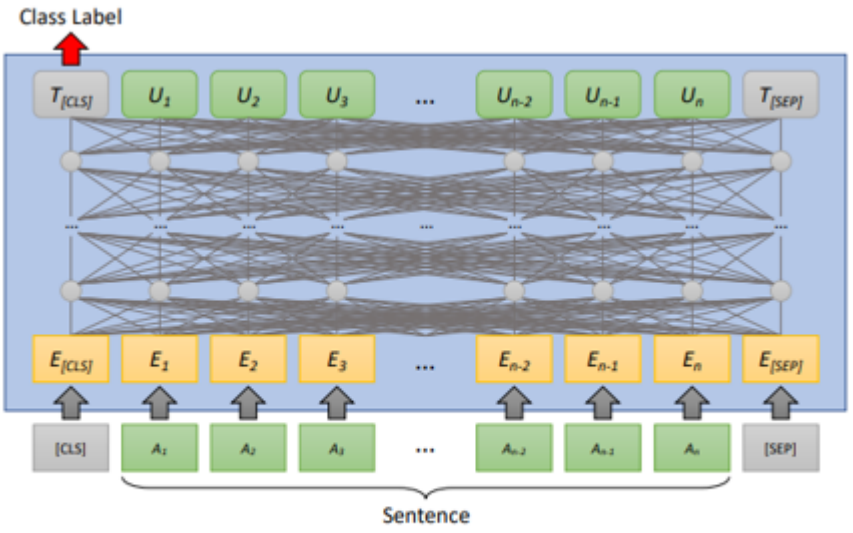
سوال 1)
الف)

در embedding مستقل از متن، برای هر کلمه مستقل از context ای که در آن قرار گرفته یک نمایش برداری داریم. در واقع معانی مختلف یک کلمه تنها در یک بردار تجمیع شده اند و از یکدیگر جدا نشده اند. مشکلی که پیش می آید این است که برای یک کلمه با دو معنی (حس) متفاوت یک بردار داریم. مثلا کلمه apple هم در معنای سیب و هم در معنای شرکت اپل می تواند بکار رود که در این روش مستقل از موضوع و جایی که در آن ظاهر می شود بردار واحد دارد و تفاوتی درارزش آن بسته به موضوع حس نمی شود.

ب)

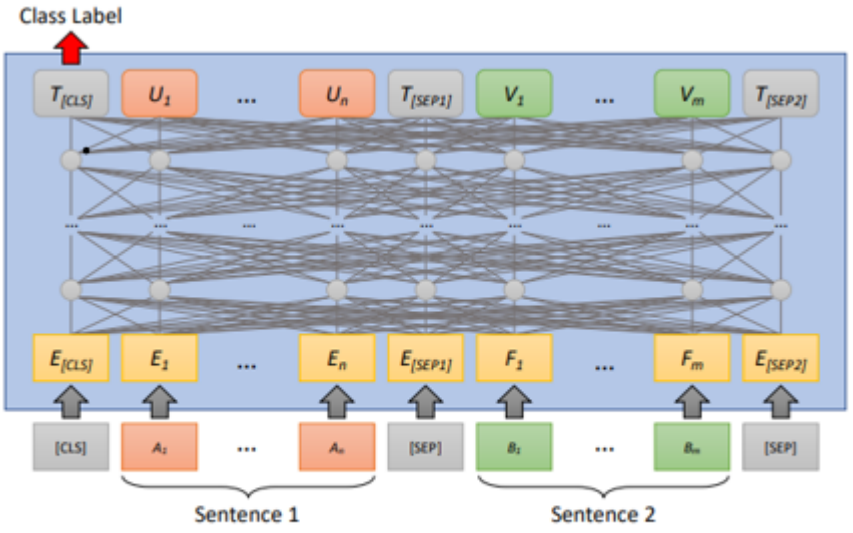
Word2Vec, GloVe

ج)
1- در این وظیفه یک جمله ورودی داریم و میخواهیم کلاس آن را تشخیص دهیم. در نتیجه یک single input classification task است و خروجی تعیین لیبل کلاس مورد نظر است. ورودی ابتدا توکن cls، در ادامه آن توکن های رشته ورودی و سپس توکن esp است.



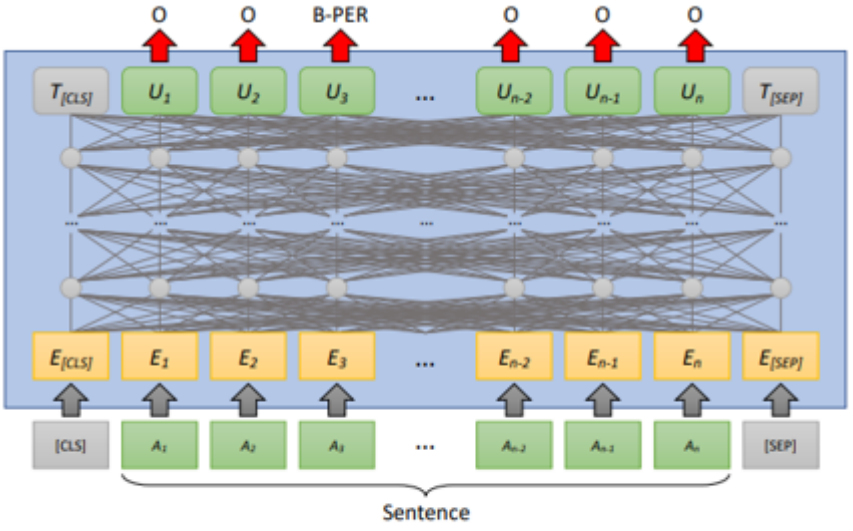
(a) Single-Input Classification Tasks

4- در این وظیفه دو جمله داریم و میخواهیم بدانیم که با کلمات متفاوت یک موضوع یکسان را بیان می کنند یا خیر. در نتیجه یک two-input classification task است و خروجی 1 کلاس است. ورودی ابتدا توکن cls، در ادامه توکن های رشته ورودی اول، سپس توکن esp، سپس توکن های رشته ورودی دوم و سپس توکن esp است.



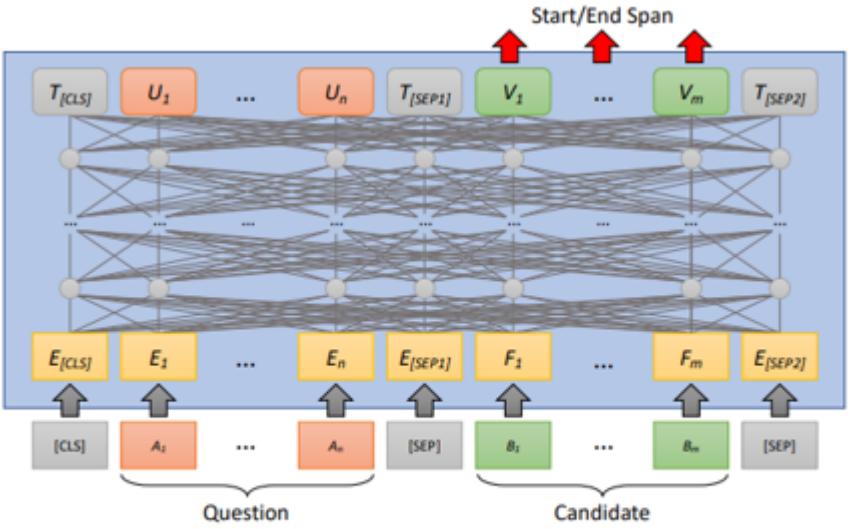
(b) Two-Input Classification Tasks

3- در این وظیفه برای هر یک از توکن های جمله ورودی لیبل زده می شود که named entity است یا خیر. در نتیجه یک single input sequence labeling task است. ورودی ابتدا توکن cls، در ادامه آن توکن های رشته ورودی و سپس توکن esp است.



(c) Single-Input Sequence Labeling Tasks

2- در این وظیفه 2 جمله ورودی داریم. یکی پرسش و دیگری سندی که میخواهیم پاسخ پرسش را در آن بیابیم. میخواهیم مکان شروع و پایان پاسخ را در سند کاندید لیبل بزنیم. در نتیجه یک two input sequence labeling task است.ورودی ابتدا توکن cls، در ادامه توکن های رشته ورودی اول، سپس توکن esp، سپس توکن های رشته ورودی دوم و سپس توکن esp است.



(d) Two-Input Sequence Labeling Tasks

در تسک های classification (تسک 1 و 4) از نمایش embedding توکن cls برای پیشبینی لیبل کلاس استفاده می شود و نمایش های contextual توکن های دیگر به کار نمی روند.
در تمامی تسک ها بعد از توکن های هر رشته ورودی یک توکن esp داریم.

سوال 2)

الف)
منافع:
تمام اسناد مرتبط با یک پرس و جو مشخص را به عنوان نمونه یادگیری در نظر میگیرد.
عدد رتبه سند برای loss function قابل مشاهده است (رتبه هر سند را می دانیم)
معایب:
پیچیده است.
استفاده از اطلاعات موقعیت رتبه کافی نیست

ب)
محدودیت ها:
1- تنظیم دستی پارامتر معمولاً دشوار است، به خصوص زمانی که پارامترهای زیادی وجود داشته باشد و اقدامات ارزیابی غیر هموار باشد.
2- تنظیم دستی پارامترها گاهی اوقات منجر به overfitting و جهت گیری بیش از حد می شود.
3- ترکیب تعداد زیادی از مدل های ارائه شده در ادبیات برای به دست آوردن یک مدل حتی مؤثرتر، ارزش زیادی ندارد.
هم چنین در بسیاری از این روش ها، تشابه معنایی کلمات در رتبه بندی در نظر گرفته نشده است.

برای برطرف کردن این محدودیت ها، از یادگیری ماشین می توانیم استفاده کنیم تا با روش های یادگیری با داده های اموزشی و تست و آموزش شبکه های عصبی و ... پارامتر ها را بهینه کنیم، شواهد مختلف را باهم ترکیب کنیم، و از جهت گیری بیش از حد (over fitting) با استفاده از regularization و دیگر روش ها جلوگیری کنیم. روش های learning to rank به صورت خودکار با استفاده از داده های آموزشی مدل های بازیابی می سازند و بازیابی را بر اساس relevance، preference و importance رتبه بندی می کنند.
روش‌های یادگیری رتبه‌بندی روش‌هایی هستند که یاد می‌گیرند چگونه ویژگی‌های از پیش تعریف‌شده را برای رتبه‌بندی با یادگیری ترکیب کنند.

ج)

- ویژگی های سند وابسته به پرس و جو (ویژگی های پویا):

هم به محتوای سند و هم به پرسش بستگی دارد.
به عنوان مثال، امتیاز TF-IDF، امتیاز BM25.

- ویژگی های سند مستقل از پرس و جو (ویژگی های استاتیک):

فقط به سند بستگی دارد و نه به پرسش و جو.
به عنوان مثال، PageRank، طول سند.

- ویژگی های پرس و جو (ویژگی های پرس و جو):

فقط به پرس و جو بستگی دارد.
مقادیر ویژگی ها برای هر پرس و جو برای همه اسناد یکسان است
به عنوان مثال، تعداد کلمات در یک پرس و جو.