

به نام خدا

تمرین دوم درس فهم زبان

دانشگاه صنعتی امیرکبیر

در این تمرین به مساله پر کردن شکاف^۱ و تشخیص قصد^۲ پرداخته خواهد شد. برای این تمرین از مجموعه داده ضمیمه شده استفاده خواهیم کرد که همان مجموعه داده پر کردن شکاف و تشخیص قصد چندزبانه Massive است. در این مجموعه داده به ازای هر یک از جملات یک برچسب قصد و به ازای هر یک از کلمات یک جمله، یک برچسب شکاف اختصاص داده شده است که تعداد زیادی از آن‌ها برچسب تهی دارند. در این تمرین داده‌ها به سه بخش آموزش (train)، توسعه (development) و آزمون (test) به نسبت تقسیم شده است که در داخل مجموعه داده در قسمت partition مشخص شده داده به کدام دسته تعلق دارد. از مجموعه داده آموزش و توسعه برای دستیابی به بهترین مدل استفاده کرده و از مجموعه داده آزمون برای سنجش و ارزیابی مدل استفاده کنید. جزئیات دادگان را میتوانید در صفحه Github مجموعه داده Massive در آدرس زیر ببینید.

<https://github.com/alexa/massive>

بخش اول: از آنجایی که برچسب‌های Slot باید به صورت Begin/In/Out Representation یعنی O, I-Label و B-Label باشد و در این مجموعه داده این عمل انجام نشده انجام این بخش به عهده شما می‌باشد. به منظور سهولت کار می‌توانید به قطعه کد Massive_data_reader.py که در پوشه تمرین قرار داده شده است نگاه کنید. در این کد می‌توانید با طریقه خواندن داده و تصحیح کردن برچسب‌ها آشنا شوید.

بخش دوم: در این جا شما باید مدلی طراحی کنید که به ازای هر یک از جملات مجموعه داده بتواند برچسب قصد و برچسب‌های شکاف‌ها را تشخیص دهد. در این مدل باید از روش‌های مبتنی بر مشتقات شبکه‌های عصبی ترنسفورمری استفاده شود و شما مجاز هستید تا از تمامی شبکه‌های ترنسفورمری شناخته شده استفاده کنید. از آن جایی که برای مدل‌های مبتنی بر ترنسفورمر باید از Tokenizerهای مبتنی بر زیر کلمه^۳ استفاده شود، شما باید به نحوه مقتضی توکن‌های مورد نیاز برای جمله را ایجاد کرده و برچسب‌های شکاف‌ها را نیز متناسب با آن‌ها هماهنگ کنید.

برای این تمرین باید حداقل دو مدل مختلف را هم برای زبان فارسی و هم برای زبان انگلیسی بررسی و نتایج آن‌ها را در هر زبان با هم مقایسه کنید. یکی از این روش‌های می‌تواند بر مبنای ترنسفورمر نباشد. همچنین دیگر مقایسه‌ها از نظر تعداد پارامتر و زمان‌های آموزش و تست نیز باید انجام شود. با توجه به مطالبی که در درس یادگرفته‌اید توقع است که نوآوری‌هایی در پیاده‌سازی‌ها داشته باشید که باعث بهبود دقت شوند. این نوآوری‌ها می‌تواند به عنوان بخش پیشنهاد ایده جدید هم مورد استفاده قرار گیرد در صورتی که منطق خوبی داشته باشد و بتوان از آن دفاع کرد.

¹ Slot Filling

² Intent detection

³ Sub-Word

برای ارزیابی مدل میتوانید از Sequeval استفاده نمایید. نمونه استفاده از این کتابخانه را در آدرس زیر می‌توانید مشاهده کنید. این کتابخانه معیارهای F1، Precision و Recall را محاسبه می‌کند و به صورت یک دیکشنری آن‌ها را باز می‌گرداند..

<https://huggingface.co/spaces/evaluate-metric/sequeval>

لازم به ذکر است که باید نتایج را براساس ویژگی‌های مدل تفسیر کنید. برای این کار میتوانید لایه‌های خود-توجه⁴ در مدل ترنسفورمری را بصری سازی کنید و تحلیل خود را از آن‌ها بیان نمایید.

بخش امتیازی اول:

در این بخش شما باید مساله پر کردن شکاف و تشخیص قصد را به صورت دو زبانه انجام دهید. منظور از دو زبانه زبان فارسی و انگلیسی است. توجه کنید که مدل باید همزمان بر روی این دو زبان آموزش ببینید و نباید مدل‌ها جدا جدا باشد. برای این منظور هم می‌تواند دو مجموعه داده را با یکدیگر ترکیب کنید و هم می‌تواند مدل را یک ایپاک بر روی زبان انگلیسی و یک ایپاک بر روی زبان فارسی آموزش دهید. در این بخش از هر روشی که فکر می‌کنید بهتر است می‌توانید استفاده کنید و محدودیت خاصی وجود ندارد.

بخش امتیازی دوم:

برای بخش تشخیص قصد در زبان انگلیسی از یک مدل جنریتو استفاده کنید و سعی کنید شبکه را به گونه‌ای تنظیم دقیق کنید که وقتی جمله را به عنوان ورودی به مدل می‌دهیم در خروجی برای ما قصد را تولید کند. برای انجام این بخش می‌توانید از روش تولید شعر که در لینک زیر قابل دسترسی است ایده بگیرید.

[Persian Poetry FineTuning.ipynb - Colaboratory \(google.com\)](#)

توجه مهم: برای تمرین اول متاسفانه تعدادی از گزارش‌ها اصلا اصول نوشتن یک گزارش را نداشتند. همچنین سبک نوشتار و فرمت آن‌ها نیز اصلا خوب نبود. گزارش شما باید شامل مقدمه، توضیحات روش، توضیحات آزمایش‌ها، تحلیل نتایج و در نهایت نتیجه‌گیری باشد. 30 درصد نمره به کیفیت گزارش داده خواهد شد.

⁴ Self-attention