گزارش تمرین سوم درس فهم زبان

موضوع تمرین:

پیاده‌سازی مدل‌های پرسش‌وپاسخ

نام استاد درس: دکتر حسین زینلی

نام دانشجو: مجید ادیبیان

شماره دانشجویی: ۴۰۰۱۳۱۰۷۸

پاییز ۱۴۰۱

۱. مقدمه

سیستم‌های پرسش‌وپاسخ در واقع سیستم‌هایی هستند که به ازای یک وردی پرسشی پاسخ متناسب را تولید می‌کنند. این سیستم‌ها در دو حالت کلی می‌توانند در نظر گرفته شوند که در حالت اول بر اساس یک متن پاسخ متناسب به پرسش تولید می‌شود و در حالت دوم پاسخ پرسش بخشی از متن است که باید از آن جدا شود.

در پروژه فعلی هدف پیاده‌سازی چند مدل مختلف در تسک پرسش‌وپاسخ و در حالت extractive است که در زبان فارسی انجام شود. همچنین تاثیر عواملی مانند استفاده یا عدم استفاده از پیش‌آموزش مدل و همچنین آموزش به صورت چند زبانه بررسی خواهد شد.

۲. روش انجام کار

فرایند انجام کار شامل پیاده‌سازی مدل‌ها، واحدسازی و ساخت ماتریس داده‌ها برای آموزش مدل‌ها و در نهایت ارزیابی نتایج است که در ادامه با جزئیات به هر کدام می‌پردازیم.

۱.۲. آماده‌سازی داده‌ها

برای داده‌ها در زبان فارسی از مجموعه داده PQuAD استفاده شده است. این داده‌ها در سه بخش آموزشی، ارزیابی و آزمون قرار دارند و به ترتیب شامل ۶۳۹۹۴، ۷۹۷۶ و ۸۰۰۲ پرسش می‌باشد. فرمت اولیه داده‌ها به صورت json است که در کد پیاده‌سازی شده تابعی نوشته شده که داده‌ها را از این فایل‌ها می‌خواند و مجموعه پرسش‌ها، محتوای شامل پاسخ و محل شروع پاسخ را استخراج می‌کند. برای این تسک روش انجام شده این گونه است که ابتدا پرسش و سپس متن مورد نظر به هم می‌چسبند و به عنوان ورودی به مدل داده می‌شوند و از جدا کننده sep برای جدا کردن آن‌ها استفاده شده است. سپس با استفاده از tokenizer مدل ParsBert مجموعه پرسش‌ها و متن مربوط به آن‌ها توکن‌بندی شده و دنباله اعداد مربوط به توکن‌های هر یک تولید شده است. همچنین برای خروجی مدل نیز برای خروجی متناظر با هر توکن دو دسته‌بندی انجام می‌شود و احتمال شروع پاسخ و احتمال پایان پاسخ مشخص می‌شود که بر اساس محل شروع پاسخ و طول پاسخ برچسب مربوط به هر دسته‌بند تعیین می‌شود.

در داده‌های انگلیسی از مجموعه داده SQuAD استفاده شده است که دارای دو قسمت آموزشی و تست است. در هنگام خواند داده‌ها قسمت مربوط به ارزیابی به اندازه ۸۰۰۰ نمونه از داده‌های آموزشی جدا شده‌اند. سپس داده

مربوط به هر قسمت از آن با قسمت مربوطه از داده فارسی ترکیب شده است تا به صورت ترکیبی در آموزش مدل‌ها استفاده شوند.

۲.۲. پیاده‌سازی و آموزش مدل‌ها

۱.۲.۲. مدل ParsBert آموزش دیده

برای این تمرین ابتدا مدل ParsBert استفاده شده است که مدلی از پیش آموزش دیده بر روی داده‌های فارسی است. برای این کار ابتدا از tokenizer این مدل استفاده شده و داده‌های فارسی موجود tokenize شده‌اند. سپس برای آموزش مدل از trainer موجود در hugging face استفاده شده است. نرخ یادگیری $3*10^{-5}$ در نظر گرفته شده و از batch size برابر ۸ استفاده شده است. سپس مدل بر روی داده‌های پرسش‌وپاسخ فارسی تا ۶ گام آموزش دیده است.

۲.۲.۲. مدل Bert بدون پیش‌آموزش

برای آموزش با استفاده از مدلی که پیش‌اموزش دیده نباشد از Bert استفاده شده ولی وزن‌های آموزش دیده آن استفاده نشده است. همچنین به دلیل کار با داده‌های فارسی از tokenizer مربوط به ParsBert استفاده شده است. سپس برای آموزش مدل از trainer موجود در hugging face استفاده شده و نرخ یادگیری $3*10^{-5}$ در نظر گرفته شده و از batch size برابر ۸ استفاده شده است. سپس مدل بر روی داده‌های پرسش‌وپاسخ فارسی تا ۶ گام آموزش دیده است.

۳.۲.۲. مدل Bert برای آموزش چند زبانه

برای آموزش چند زبانه از مدل از پیش‌آموزش دیده XLM-Roberta استفاده شده است که مدلی است که بر روی داده‌های چندین زبان آموزش دیده است و فارسی و انگلیسی را هم شامل می‌شود. سپس مجموعه داده فارسی و انگلیسی که در قسمت قبل توضیح داده شد ترکیب شده و شافل می‌شوند و با استفاده از tokenizer این مدل، داده‌ها tokenize می‌شوند. در ادامه مانند قبل از trainer استفاده شده است و آموزش مدل تا ۶ گام انجام شده.

۳.۲. نحوه اجرا

برای اجرای آموزش و ارزیابی مدل‌ها فایل نوت‌بوکی در کنار گزارش قرار گرفته شده است که دستورات اجرای کدها و برخی خروجی‌ها در آن قرار دارد.

۳. نتایج

جهت ارزیابی نتایج از ماژول evaluate مربوط به hugging face استفاده شده است که دوتا از معیارهایی که به ما می‌دهد معیارهای exact match و F1 می‌باشد.

هر یک از مدل‌های آموزش دیده را بر روی داده‌های تست فارسی بررسی می‌کنیم و دو معیار مد نظر را بر روی آن‌ها به دست می‌آوریم. همچنین برای حالت چند زبانه دو مجموعه تست تک زبانه فارسی (که همان تست سایر مدل‌ها است) و یک مجموعه تست دو زبانه داریم. جدول زیر نتایج به دست آمده از مدل‌ها را بر روی این داده‌های تست نشان می‌دهد.

	Exact match		F1	
ParsBert	66.3		80	
Untrained Bert	26.1		29.7	
XLM-Roberta	**Fa:72.4**	Fa&En:75.8	**Fa: 83.1**	Fa&En: 82.2

نتیجه‌گیری:

با توجه به نتایج به دست آمده واضح است که استفاده از مدل آموزش ندیده چه قدر تاثیر داشته و عملکر مدل را به شدت تخریب کرده است. این در حالی است که استفاده از مدل از پیش آموزش دیده ParsBert توانسته عملکرد قابل قبولی داشته باشد. دلیل این امر آن است که مدل از پیش آموزش دیده حجم زیادی داده در آن زبان را دیده و بر روی چند تسک آموزش دیده است که باعث شد مدل بهتر بتواند آن زبان و معنای کلمات آن را بفهمد که این امر در پاسخ بهتر به سوالات تاثیر گذار است.

همچنین دیده می‌شود که به ازای آموزش مدل با چند زبان مدل توانسته با دقت بیشتری تسک مورد نظر را انجام دهد. دلیل این امر آن است که با استفاده از آموزش مدل به صورت چند زبانه مدل نوع حل مسئله را به صورت بهتر متوجه می‌شود و وزن‌های مدل متناسب با آن تسک بهتر به روز می‌شوند.