گزارش ساخت و ارزیابی مدل بازیابی مبتنی بر زبان با استفاده از داده‌های ساختاریافته

تمرین دوم درس پردازش زبان های طبیعی

علی شریفی 403206532

رئوفه رضایی 403207012

نسرین صادقلو 402208954

یاسان حسن زاد 403206995

بهار 1403

فهرست مطالب

مقدمه

مدل‌های زبانی در سال‌های اخیر به ابزارهای مؤثری برای درک و تولید زبان طبیعی تبدیل شده‌اند. یکی از کاربردهای مهم آن‌ها، بازیابی اطلاعات از داده‌های ساختاریافته در قالب پرسش و پاسخ است. در این پروژه، با هدف ساخت یک سامانه بازیابی هوشمند، ابتدا داده‌های ساختاریافته به متون روان تبدیل شده و سپس با طراحی پرسش‌های متنوع، مدل چند زبانه Glot500 با روش یادگیری متضاد آموزش داده شده است.

عملکرد سه رویکرد شامل TF-IDF، مدل پایه بدون آموزش و مدل آموزش‌دیده روی مجموعه‌ای از سوالات ارزیابی شده و نتایج آن به صورت انسانی تحلیل شده‌اند. این گزارش خلاصه‌ای از مراحل، تحلیل‌ها و نتایج این فرایند را ارائه می‌دهد.

آماده‌سازی داده‌ها

در این پروژه از داده‌های ساختاریافته در دامنه‌ی تاریخ استفاده شده است. داده‌ها در قالب JSON و شامل اطلاعات متنوعی از رویدادهای تاریخی بودند. در مجموع، تعداد 714 نمونه داده در اختیار قرار گرفت که هر نمونه شامل ویژگی‌های مختلفی نظیر عنوان، زمان، مکان، شرح رویداد، و افراد دخیل بود.

پس از بررسی اولیه، داده‌های ناقص یا دارای اشکال محتوایی حذف یا اصلاح شدند. در گام بعد، برای تبدیل داده‌های ساختاریافته به متون روان و قابل درک برای انسان، الگوهای مبتنی بر قاعده (Rule-based Templates) طراحی شد. این الگوها بسته به نوع داده، ساختار جمله‌بندی مناسب را تولید می‌کردند.

متون حاصل سپس با استفاده از مدل‌های زبانی یا به‌صورت دستی ویرایش شدند تا کیفیت نگارشی و معناشناسی آن‌ها بهبود یابد. در ادامه، با هدف پوشش جنبه‌های مختلف هر متن، تعداد 2803 پرسش طراحی شد. در تولید این سوالات، تلاش شد تنوع ساختاری و معنایی رعایت گردد.

در نهایت، داده‌ها به فرمت استاندارد قابل استفاده برای آموزش مدل‌های پرسش‌پاسخ تبدیل شدند. ساختار نهایی هر نمونه به صورت زیر تعریف شده است:

```
{
    "context": "",
    "question": ""
}
```

مجموعه ارزیابی

برای ارزیابی عملکرد مدل‌ها، مجموعه‌ای مستقل شامل ۵۰ سوال طراحی شد که در داده‌های آموزشی وجود نداشتند. این سوالات با هدف پوشش انواع مختلف پرسش‌های تاریخی (شخص، زمان، مکان، علت و نتیجه) تهیه شده‌اند و تنوع ساختاری و مفهومی دارند.

برخی سوالات مستقیماً بر پایه متون ساخته شده‌اند و برخی دیگر تحلیلی یا استنتاجی هستند. این تنوع، امکان سنجش دقیق توانایی مدل در درک و بازیابی اطلاعات را فراهم می‌کند. سوالات به‌صورت ساختاریافته ذخیره شده و برای ارزیابی نهایی مدل‌ها مورد استفاده قرار گرفته‌اند.

آموزش مدل زبانی

برای آموزش مدل بازیابی، از مدل چندزبانه‌ی Glot500-base به عنوان مدل پایه استفاده شد. این مدل بر پایه معماری Transformer توسعه داده شده است. و هدف آموزش، یادگیری بازنمایی‌های متنی مشترک برای سوال و متن با استفاده از رویکرد Contrastive Learning بود.

ساختار مدل

برای نگاشت بردارهای خروجی مدل پایه به یک فضای مشترک با ابعاد پایین‌تر، یک سر (Projection Head) شامل دو لایه‌ی خطی با تابع ReLU بین آن‌ها طراحی شد. این سر به خروجی مدل Glot500 افزوده و آموزش داده شد. ابعاد خروجی این سر به‌صورت ثابت ۲۵۶ انتخاب شد.

روش آموزش

فرآیند آموزش مدل با بهره‌گیری از Contrastive Learning و تابع خطای Cross-Entropy بر پایه‌ی مشابهت کسینوسی بین بردارهای نرمال‌شده انجام گرفت. برای هر batch از داده‌ها، ضرر در هر دو جهت «پرسش به متن» و «متن به پرسش» محاسبه شد.

مراحل کلیدی آموزش عبارت‌اند از :

- رمزنگاری پرسش و متن با استفاده از توکنایزر Glot500
- استخراج بردارهای جملات از طریق میانگین‌گیری از خروجی‌های مدل (mean pooling)
- نگاشت بردارها به فضای embedding توسط projection head
- نرمال‌سازی بردارها و محاسبه شباهت کسینوسی
- آموزش مدل با تابع خطای تضادی و استفاده از دمای ۰.۰۷
- به‌منظور بهینه‌سازی مدل، از الگوریتم AdamW با نرخ یادگیری 2×10^{-5} استفاده شد.
- آموزش در ۵ دوره با اندازه‌ی دسته‌ی ۶۴ روی GPU انجام شد.

مدل آموزش دیده روی فضایی در Hugging Face مستقر شده است. از طریق این لینک می‌توانید به آن دسترسی داشته باشید.

ارزیابی

در این مرحله، عملکرد سه مدل مختلف در پاسخ‌گویی به ۵۰ سؤال زبان فارسی مورد ارزیابی قرار گرفت. مدل‌ها شامل موارد زیر بودند:

- Model 1: مدل مبتنی بر TF-IDF
- Model 2: مدل Glot500 پس از فاین‌تیونینگ
- Model 3: مدل پایه Glot500

برای هر سؤال، هر مدل ممکن بود بین یک تا سه پاسخ ارائه دهد. سپس دو ارزیاب مستقل با بررسی کیفیت پاسخ‌ها، به هر مدل و برای هر سؤال یک امتیاز بین ۱ تا ۳ اختصاص دادند (1 نشان‌دهنده بهترین پاسخ). هدف، سنجش دقت، جامعیت و قابل‌فهم بودن پاسخ‌ها از دید انسانی بود.

Rank	TF-IDF	Glot-500 Fine-tuned	Glot-500 Base
1	63	41	30
2	23	19	28
3	14	40	42

نتایج رتبه‌بندی مدل‌ها

تحلیل آماری و توافق ارزیاب‌ها

- درصد توافق (Percent Agreement): درصد دفعاتی که دو ارزیاب نمره‌ی یکسان به مدل‌ها داده‌اند.
- ضریب کاپای کوهن (Cohen’s Kappa): معیاری دقیق‌تر که میزان توافق را با در نظر گرفتن احتمال تصادفی بودن توافق محاسبه می‌کند.

Model	Percent Agreement	Cohen’s Kappa
TF-IDF	58%	0.24
Glot-500 Fine-tuned	50%	0.24
Glot-500 Base	62%	0.44

مدل پایه‌ی Glot-500 با وجود نمره‌ی پایین‌تر در رتبه‌بندی کلی، بیشترین میزان توافق بین دو ارزیاب را داشته است (62% توافق و کاپای 0.44). این می‌تواند نشان‌دهنده‌ی ثبات بیشتر در نوع پاسخ‌های این مدل باشد، هرچند کیفیت آن پایین‌تر از سایر مدل‌ها بوده است.

در مقابل، مدل فاین‌تیون‌شده‌ی Glot-500 عملکرد بهتری از نظر کیفیت پاسخ‌ها نسبت به مدل پایه دارد (رتبه‌های 1 و 2 بیشتر از مدل پایه)، اما توافق کمتری بین ارزیاب‌ها ایجاد کرده است. این ممکن است به دلیل تنوع بالاتر پاسخ‌ها در مدل فاین‌تیون‌شده باشد که باعث ایجاد اختلاف نظر میان ارزیاب‌ها شده است.

مدل TF-IDF نیز بهترین عملکرد را در کسب رتبه 1 داشته (63 بار)، اما از نظر توافق ارزیاب‌ها در سطح متوسط قرار گرفته است.

تحلیل کیفی

TF-IDF: در سؤالات فهرست‌محور یا مبتنی بر بازیابی اطلاعات صریح عملکرد مناسبی داشت، اما در سؤالات مفهومی و تحلیلی عملکرد ضعیف‌تری نشان داد.
Glot-500 Fine-Tuned: در طیف وسیعی از پرسش‌ها، به‌ویژه در سؤالات علی، تطبیقی و با بستر تاریخی پیچیده، دقیق‌ترین و روان‌ترین پاسخ‌ها را ارائه داد. این مدل از نظر کیفی بالاترین سطح انسجام و دقت را داشت.
Glot-500 Base: هرچند ساختار زبانی روان‌تری نسبت به TF-IDF داشت، اما به دلیل عدم تطبیق دقیق با داده‌های تاریخی فارسی، برخی پاسخ‌ها ناقص یا نامرتبط بودند.

الگوهای عملکردی قابل شناسایی

توضیح	بهترین مدل	نوع سوال
مبتنی بر تطبیق واژگان کلیدی	Model 1	فهرستی (نام ببرید)
نیازمند فهم زمینه‌ای و تاریخی	Model 2	علی و تحلیلی
پاسخ‌های مفهومی و ساخت‌یافته	Model 2	تطبیقی یا مقایسه ای
در مواردی پاسخ روان اما کم‌دقت	Model 3	مفهومی عمومی
ضعف عمومی در دسترسی به دانش خاص	هیچ کدام	موجودیت خاص یا داده نادر

نتیجه گیری

در این پروژه، سه رویکرد برای پاسخ‌گویی به پرسش‌های تاریخی بررسی شد: TF-IDF، Glot500 و Glot500 Fine-Tuned. نتایج نشان داد مدل Fine-Tuned عملکرد بهتری در تولید پاسخ‌های دقیق، روان و تحلیلی دارد، هرچند تنوع خروجی‌های آن منجر به کاهش توافق بین ارزیاب‌ها شد. مدل TF-IDF در بازیابی پاسخ‌های مستقیم موفق بود و بیشترین رتبه اول را از دید ارزیاب‌ها کسب کرد. مدل پایه Glot500 با وجود نداشتن آموزش خاص، پاسخ‌هایی نسبتاً قابل قبول تولید کرد و بالاترین توافق میان ارزیاب‌ها را به دست آورد، هرچند از نظر کیفیت کلی پایین‌تر بود.