خلاصه تحلیلی مقاله

Similarity Textual Semantic Cross-Lingual for Transformer Ensemble

۱ مسئله اصلی و اهمیت آن

شباهت معنایی متون (Semantic Textual Similarity - STS) به مسئله‌ای گفته می‌شود که در آن هدف، تعیین میزان نزدیکی معنایی بین دو متن به‌صورت یک مقدار پیوسته، معمولاً در بازه ۰ تا ۵، است. در حالت Cross-Lingual، دو متن مورد مقایسه به زبان‌های متفاوتی نوشته شده‌اند.

این مسئله نقش کلیدی در بسیاری از کاربردهای مهم پردازش زبان طبیعی ایفا می‌کند؛ از جمله بازیابی اطلاعات چندزبانه، سیستم‌های پرسش‌وپاسخ، ترجمه ماشینی، موتورهای جستجو و سامانه‌های تولید پاسخ مبتنی بر بازیابی (Retrieval-Augmented Generation). چالش اصلی در این حوزه، ایجاد یک فضای برداری مشترک است که بتواند جملات زبان‌های مختلف را به‌گونه‌ای نمایش دهد که نزدیکی بردارها بازتاب‌دهنده شباهت معنایی واقعی آن‌ها باشد.

مقاله حاضر با تمرکز ویژه بر جفت‌زبان فارسی‌ةانگلیسی، روشی مبتنی بر ترکیب (ensemble) چند مدل ترنسفورمری چندزبانه ارائه می‌دهد که بدون استفاده از ترجمه ماشینی، شباهت معنایی بین‌زبانی را با دقت بالاتری نسبت به روش‌های پیشین محاسبه می‌کند.

۲ ورودی‌ها و خروجی‌های مدل

ورودی

- یک جفت جمله (s_1, s_2)
- جملات می‌توانند به زبان‌های متفاوت باشند (در این مقاله، فارسی و انگلیسی)

خروجی

- یک نمره پیوسته شباهت معنایی:

$$\text{Similarity}(s_1, s_2) \in [0, 5]$$

- این نمره میزان هم‌پوشانی مفهومی دو جمله از دیدگاه انسانی را نشان می‌دهد.

۳ داده‌های مورد استفاده

برای آموزش و ارزیابی مدل‌ها، از مجموعه‌داده **PESTS** استفاده شده است. این دیتاست شامل ۵۳۷۵ جفت جمله فارسی‌ةانگلیسی است که هر جفت جمله دارای یک نمره شباهت معنایی تعیین‌شده توسط ارزیاب انسانی می‌باشد.

داده‌ها به سه بخش آموزش، اعتبارسنجی و آزمون تقسیم شده‌اند و به‌طور خاص برای مسئله شباهت معنایی بین‌زبانی طراحی شده‌اند. استفاده از این دیتاست یکی از نقاط قوت مقاله است، زیرا منابع استاندارد برای ارزیابی STS در زبان فارسی بسیار محدود هستند.

علاوه بر این، برای بررسی تعمیم‌پذیری روش پیشنهادی، آزمایش‌هایی روی دیتاست SemEval ۲۰۱۷ برای جفت‌زبان عربی‌ةانگلیسی نیز انجام شده است.

۴ روش پیشنهادی مقاله

روش ارائه‌شده مبتنی بر استفاده هم‌زمان از چند مدل ترنسفورمری چندزبانه و ترکیب خروجی آن‌ها به‌منظور بهبود دقت تخمین شباهت معنایی است. برخلاف بسیاری از روش‌های مبتنی بر ترجمه ماشینی، این روش مستقیماً در یک فضای برداری مشترک عمل می‌کند.

مدل‌های مورد استفاده

سه مدل ترنسفورمری زیر برای ترکیب انتخاب شده‌اند:

- paraphrase-xlm-r-multilingual-v1
- paraphrase-multilingual-mpnet-base-v2
- mstsb-paraphrase-multilingual-mpnet-base

این مدل‌ها با استفاده از دیتاست PESTS برای وظیفه STS ریزتنظیم شده‌اند.

روش‌های Ensemble

۱. **میانگین‌گیری از نمرات شباهت:** میانگین شباهت کسینوسی خروجی مدل‌ها به‌عنوان نمره نهایی استفاده می‌شود.

۲. **الحاق بردارها و کاهش بُعد:** بردارهای خروجی مدل‌ها الحاق شده و پس از کاهش بُعد با PCA، شباهت محاسبه می‌شود.

۳. **شبکه عصبی پیش‌خور:** در این روش، یک شبکه عصبی ساده برای تخمین نمره شباهت بر اساس ترکیب بردارها آموزش داده می‌شود.

شبه‌کد کلی

```
for each transformer model:
    e1 = encode(sentence1)
    e2 = encode(sentence2)
    store embeddings

combine embeddings
output similarity score
```

۵ Baseline و بازتولید نتایج مقاله

به‌عنوان baseline، از یک مدل ترنسفورمری منفرد چندزبانه (paraphrase-xlm-r-multilingual-v1) استفاده شد. در این حالت، شباهت معنایی دو جمله با استفاده از شباهت کسینوسی بین های embedding آن‌ها محاسبه گردید.

این مدل روی دیتاست PESTS اجرا شد و ضریب همبستگی پیرسون بین نمرات پیش‌بینی‌شده مدل و نمرات انسانی دیتاست به‌عنوان معیار ارزیابی استفاده گردید. این نتیجه مبنای مقایسه برای روش‌های ensemble قرار گرفت.

روش	Correlation Pearson
(Baseline) Model Single	۸۲.۰
(Average) Ensemble	۸۶.۰
PCA + Ensemble (۲۰۰ dims)	۸۸.۰

جدول ۱: مقایسه عملکرد مدل پایه و روش‌های ensemble روی دیتاست PESTS

۶ تحلیل اثر کاهش بُعد (PCA)

برای بررسی اثر کاهش بُعد روی هایembedding، ensemble از روش تحلیل مؤلفه‌های اصلی (PCA) استفاده شد. ابعاد مختلفی از فضای ویژگی (۵۰ تا ۴۰۰ بُعد) مورد آزمایش قرار گرفت.

نتایج نشان داد که کاهش بُعد تا یک مقدار بهینه می‌تواند باعث حذف نویز و بهبود همبستگی با نمرات انسانی شود، در حالی که کاهش بیش‌ازحد بُعد منجر به افت عملکرد می‌گردد. این تحلیل نشان می‌دهد که ترکیب ensemble با تکنیک‌های کاهش بُعد می‌تواند یک راهکار عملی برای بهبود دقت و کاهش هزینه محاسباتی باشد.

۷ نتایج اصلی

نتایج تجربی نشان می‌دهد که روش‌های ensemble به‌طور معناداری عملکرد بهتری نسبت به مدل‌های منفرد دارند. بهترین عملکرد مربوط به ترکیب ensemble همراه با PCA بوده است.

همچنین آزمایش روی جفت‌زبان عربی‌ةانگلیسی نشان داد که روش پیشنهادی قابلیت تعمیم مناسبی به زبان‌های دیگر دارد، هرچند کیفیت و حجم داده نقش مهمی در میزان بهبود ایفا می‌کند.

۸ محدودیت‌ها و ایده‌های ادامه

محدودیت‌ها

- افزایش هزینه محاسباتی به دلیل استفاده از چند مدل ترنسفورمری
- زمان آموزش و استنتاج بالاتر در روش‌های ensemble
- وابستگی عملکرد به کیفیت دیتاست ریزتنظیم

ایده‌های ادامه

- استفاده از knowledge distillation برای فشرده‌سازی ensemble
- توسعه دیتاست‌های بزرگ‌تر برای زبان فارسی
- به‌کارگیری روش در سامانه‌های عملیاتی چندزبانه