(Cross-lingual) Similarity Textual Semantic

گزارش ارائه کلاسی

شماره دانشجویی ۴۰۴۰۹۱۷۴

۱ مقدمه

شباهت معنایی متون (Semantic Similarity Textual یا (STS یکی از مسائل بنیادی در پردازش زبان طبیعی است که هدف آن اندازه‌گیری میزان نزدیکی معنایی میان دو متن می‌باشد. در حالت چندزبانه یا Cross-lingual Semantic Textual Similarity، دو متن ممکن است به زبان‌های متفاوت نوشته شده باشند، اما از نظر مفهومی محتوای یکسان یا مشابهی را منتقل کنند. این مسئله نقش کلیدی در توسعه سیستم‌های هوشمند چندزبانه دارد.

۲ تعریف مسئله و اهمیت آن

مسئله CL-STS به دنبال پاسخ به این پرسش است که دو جمله یا متن در زبان‌های مختلف تا چه حد از نظر معنا به یکدیگر نزدیک هستند. اهمیت این مسئله در کاربردهایی نظیر جستجوی چندزبانه، ارزیابی ترجمه ماشینی، تشخیص سرقت ادبی، تحلیل شبکه‌های اجتماعی و مقابله با اطلاعات نادرست نمایان می‌شود. بدون وجود مدل‌های دقیق برای CL-STS، بسیاری از سیستم‌های جهانی دچار افت عملکرد جدی خواهند شد.

۳ مرور ادبیات پژوهشی

مطالعات اولیه در این حوزه عمدتاً مبتنی بر ترجمه بودند؛ بدین صورت که ابتدا متون به یک زبان مشترک ترجمه شده و سپس شباهت آن‌ها محاسبه می‌شد. با ظهور مدل‌های مبتنی بر یادگیری عمیق و به‌ویژه Transformerها، رویکردهای مبتنی بر نمایش برداری مشترک جایگزین روش‌های قدیمی شدند.

در مقاله مروری "Cross-Lingual Semantic Textual Similarity: A Survey"، روش‌های مختلف به سه دسته کلی تقسیم شده‌اند: روش‌های مبتنی بر ترجمه، روش‌های مبتنی بر هم‌ترازی و روش‌های مبتنی بر نمایش مشترک. این مقاله همچنین چالش‌های زبان‌های کم‌منبع را به‌عنوان یکی از مشکلات اصلی معرفی می‌کند.

مدل LaBSE که در سال ۲۰۲۲ ارائه شده است، با نگاشت جملات زبان‌های مختلف به یک فضای برداری مشترک، امکان مقایسه مستقیم متون را بدون نیاز به ترجمه فراهم می‌کند. این مدل عملکرد بسیار خوبی در وظایف STS از خود نشان داده است. در ادامه، مدل mSBERT با هدف انتقال دانش از داده‌های انگلیسی به زبان‌های دیگر معرفی شد که به‌ویژه برای زبان‌های کم‌منبع کاربردی است. همچنین مدل XLM-R با بهره‌گیری از داده‌های چندزبانه و معماری Transformer، بهبود قابل توجهی در دقت CL-STS ایجاد کرده است.

۴ روندها و چالش‌های پژوهشی

از مهم‌ترین روندهای پژوهشی اخیر می‌توان به استفاده از مدل‌های زبانی بزرگ، یادگیری بدون ناظر و یادگیری انتقالی اشاره کرد. با این حال، چالش‌هایی نظیر کمبود داده برای برخی زبان‌ها، تفاوت‌های فرهنگی و وجود سوگیری زبانی همچنان پابرجا هستند و نیازمند توجه پژوهشگران می‌باشند.

۵ کاربردهای عملی

CL-STS در سیستم‌های جستجوی چندزبانه، تحلیل محتوای شبکه‌های اجتماعی، شناسایی اخبار جعلی و سیستم‌های امنیتی کاربرد گسترده‌ای دارد. این مسئله به‌ویژه در تحلیل داده‌های تولیدشده توسط کاربران در زبان‌های مختلف نقش مهمی ایفا می‌کند و امکان شناسایی الگوهای مشابه در سطح جهانی را فراهم می‌سازد.

۶ جمع‌بندی

شباهت معنایی متون چندزبانه یکی از مسائل کلیدی در پردازش زبان طبیعی مدرن محسوب می‌شود. پیشرفت مدل‌های مبتنی بر Trans-former و نمایش‌های برداری مشترک، مسیر توسعه این حوزه را هموار کرده است. با این حال، پرداختن به چالش‌های موجود، به‌ویژه در زبان‌های کم‌منبع، همچنان از اولویت‌های پژوهشی آینده خواهد بود.

منابع

- Survey، A Similarity: Textual Semantic Cross-Lingual ۲۰۲۰
- Embedding، Sentence BERT Language-agnostic LaBSE: al.، et Feng ۲۰۲۲
- Multilingual، Embeddings Sentence Monolingual Making al.، et Reimers ۲۰۲۲
- Learning، Representation Cross-lingual Unsupervised XLM-R: al.، et Conneau ۲۰۲۳