خلاصه تحلیلی

Linguistic Interpretability of Transformer-based Language Models: A Systematic Review

Based on López-Otal et al. (2025) • Prepared for an M.Sc. AI course deliverable

۱. مشکل اصلی و اهمیت آن

مدل های زبانی پیش آموزش دیده مبتنی بر ترنسفورمر (PLMها) در بسیاری از وظایف NLP عملکرد قوی دارند، اما محاسبات داخلی آن ها همچنان دشوار برای تفسیر است. مقاله مرور شده به یک پرسش متمرکز درباره قابلیت تفسیر می پردازد: PLMهای ترانسفورمر تا چه حد پدیده های زبانی—مانند نحو، صرف، واژه نام-معناشناسی و گفتمان—را در بازنمایی های درونی خود رمزگذاری می کنند و این اطلاعات را کجا (لایه ها، سرهای توجه، نورون ها، زیرفضاهای برداری) می توان یافت؟

این سؤال هم برای علم و هم برای مهندسی اهمیت دارد. از نظر علمی، بررسی می کند که آیا تعصبات استقرایی ترنسفورمرها از بازنمایی هایی که شبیه تعمیم های زبانی انسانی هستند حمایت می کند یا خیر. از نظر عملی، تفسیر زبانی می تواند به تشخیص حالت های شکست (مانند خطاهای وابستگی یا نفی دوربرد) کمک کند، انتقال چندزبانه را پشتیبانی کند و مداخلات هدفمند مانند هرس، ویرایش یا تزریق دانش را هدایت نماید. این مرور همچنین تنش موجود در ادبیات را برجسته می کند: برخی مطالعات الگوهای زبانی را به عنوان تعمیم های واقعی تفسیر می کنند، در حالی که برخی دیگر استدلال می کنند مدل ها ممکن است به نشانه های آماری سطحی تکیه کنند.

۲. ورودی ها و خروجی های مدل/سیستم چی هستند؟

از آنجا که مقاله یک مرور نظام مند است، هیچ مدل جدیدی پیشنهاد نمی کند. در عوض، این روش سازماندهی می کند که مطالعات قبلی چگونه یک PLM به همراه یک روش تفسیر پذیری را به عنوان یک «سیستم» که محرک های زبانی را می گیرد و شواهدی از دانش رمزگذاری شده زبانی ارائه می دهد، سازماندهی می کند.

خروجی های معمول	ورودی های معمول
• امتیازهای بازیابی پذیری برای ویژگی های زبانی (مثلا دقت کاوش) • فرضیه های مکان یابی (کدام لایه ها/سرها/نورون ها یک پدیده را رمزگذاری می کنند) • تجسم ها (نقشه های حرارتی توجه، هندسه جاسازی شده) • الگوهای خطا، محدودیت ها و نکات عملی	•یک LM ترانسفورمر پایه و پیش آموزش دیده (مثلا BERT، RoBERTa، mBERT) •محرک های متنی (جملات، جفت های مینیمال کنترل شده، پیکره ها) •حاشیه نویسی های زبانی اختیاری (PoS، وابستگی ها، نقش های معنایی، برچسب های گفتمانی) •یک روش تفسیر پذیری انتخاب شده (کاوش، تحلیل توجه، انتساب، ابلیشن و غیره)

این مرور بر مطالعاتی تمرکز دارد که نمایش های داخلی (جاسازی های لایه ای، سر توجه، نورون ها) PLMهای پایه را بدون تنظیم دقیق بررسی می کنند و ارزیابی های رفتاری صرفا جعبه سیاه را مستثنی می سازند.

۳. داده های استفاده شده در بازبینی (نوع، منبع، اندازه)

«داده های» این مرور، مجموعه ای گزینش شده از مقالات پژوهشی درباره قابلیت تفسیر زبانی در PLMهای ترانسفورمر است. با استفاده از فرآیندی به سبک PRISMA، نویسندگان نامزدها را از طریق گوگل اسکالر (۲۰۱۸–اکتبر ۲۰۲۴) جمع آوری و از طریق غربالگری چندحاشیه نویسی فیلتر کردند که منجر به ۱۶۰ مقاله برای تحلیل دقیق شد.

پنجره زمانی	۲۰۱۸ تا اکتبر ۲۰۲۴ (دوره گزارش شده جستجو /انتخاب)
منابع اولیه	Google Scholar; ArXiv list cross-checked with ACL Anthology
اندازه پیکره نهایی	۱۶۰ مقاله پژوهشی تحلیل شده
نکات برجسته پوشش	زبان ها: متمرکز بر انگلیسی (۱۴۵ از ۱۶۰ مقاله شامل انگلیسی است). مدل ها: عمدتا نسخه های BERT/RoBERTa و mBERT؛ مطالعات کمتری روی مدل های زبانی بزرگ مدرن که فقط رمزگشا هستند.

در سراسر مقالات مورد بررسی، مجموعه داده ها بسیار متفاوت هستند: پیکره های حاشیه نویسی شده (PoS، وابستگی ها، نقش های معنایی، برچسب های گفتمانی)، محرک های کنترل شده (توافق، ترتیب کلمات، نفی) و معیارهای چندزبانه برای اشتراک گذاری بین زبانی. این مرور یافته ها را بر اساس رشته زبان شناسی جمع بندی می کند و نه تأیید یک معیار واحد.

۴. روش (توضیح ساده) + شماتیک / شبه کد

این مرور یک طبقه بندی عملی از نحوه «استخراج» اطلاعات زبانی پژوهشگران از PLMها ارائه می دهد: (۱) انتساب ویژگی، (۲) توضیحات مبتنی بر مثال، (۳) تحلیل مستقیم عناصر معماری (سر توجه، نورون ها، جاسازی ها، هرس) و (۴) کاوش با طبقه بندهای خارجی. بررسی رایج ترین روش در میان مقالات بررسی شده است.

برای استانداردسازی بحث های لایه ای، بسیاری از مطالعات از اصطلاحات عمق نسبی استفاده می کنند. برای یک رمزگذار شبیه BERT با ۱۲ لایه، لایه های «پایین» تقریبا ۱ تا ۳، «میانی» ۴ تا ۷ و لایه های «بالا» ۸ تا ۱۲ هستند. تفاوت های معماری نیز اهمیت دارند: رمزگذارها به همه توکن ها به صورت دوطرفه توجه می کنند، در حالی که مدل های فقط رمزگشا از چپ به راست تولید می کنند؛ طرح های توکنیزاسیون نیز متفاوت اند (مثلا WordPiece در مقابل BPE)، که بر تحلیل مورفولوژیکی تأثیر می گذارد.

شماتیک: جریان کاری معمول قابلیت تفسیر زبانی

```
Input sentences (optionally controlled / annotated)
        |
        v
Base PLM (frozen parameters)  --->  Extract internal representations
 (layers / heads / neurons)          (embeddings, attention, activations)
        |
        +-----------------> Apply interpretability method
                              • probing classifier
                              • attention/embedding analysis
                              • attribution/ablation
                              • pruning / neuron analysis
                                          |
                                          v
Evidence of linguistic encoding (strength, location, failures, cross-lingual
patterns)
```

Pseudocode: minimal layerwise probing experiment

```
for layer in PLM.layers:
    E = [PLM(s).hidden_states[layer] for s in dataset]
    probe = LogisticRegression()   # or small MLP
    probe.fit(E_train, y_train)
    score[layer] = probe.evaluate(E_test, y_test)

report(best_layer = argmax(score), score_curve = score)
```

یک نکته مرکزی همبستگی در مقابل علیت است: استخراج یک ویژگی زبانی اثبات نمی کند که PLM از آن در استنتاج استفاده می کند. این مرور به گرایش به آزمایش های علی و مبتنی بر مداخله برای رفع این خلأ اشاره می کند.

۵. نتایج اصلی، محدودیت ها و مسیرهای آینده

۵.۱ یافته های کلیدی بر اساس سطح زبان شناسی

نحو (بیشترین مطالعه):

اطلاعات نحوی اغلب از PLMها (وابستگی ها، توافق، روابط دستوری) قابل بازیابی است. با این حال، مکان یابی لایه ای مورد اختلاف است: بسیاری از آثار لایه های میانی مدل های مشابه BERT را برجسته می کنند، در حالی که برخی دیگر اطلاعات مهم را در لایه های پایین تر یا توزیع شده در سراسر شبکه پیدا می کنند. ضعف های گزارش شده شامل فرم های توافق با فرکانس پایین، حساسیت به ترتیب کلمات در برخی محیط ها و روابط دوربرد در برخی زبان ها است.

واژه شناسی:

ویژگی های معنایی (چندمعنایی، شباهت/ارتباط، ساختار آرگومان، خواص اسکالر) به طور گسترده گزارش شده اند، اما گاهی کمتر از نحو مقاوم هستند. پدیده های دشوار (نفی، ترکیب پذیری) شواهد متناقضی را نشان می دهند. لایه های پایین تر اغلب به عنوان واژگانی تر توصیف می شوند، در حالی که لایه های بالاتر بسته به مدل و هدف پیش آموزشی، ممکن است بازتاب دهنده معناشناسی زمینه ای بیشتری باشند.

مورفولوژی:

ویژگی های صرفی و صرفی نحوی (تعداد، جنسیت، زمان، توافق) قابل استخراج هستند، اما توکنیزاسیون تحلیل را پیچیده می کند. برخی مطالعات صرفی نحو را در لایه های میانی قرار می دهند (مشابه نحو) و اشتراک جزئی بین زبانی را در PLMهای چندزبانه گزارش می دهند، در حالی که برخی دیگر بر تغییرات وابسته به زبان تأکید دارند.

گفتمان (کمترین مطالعه):

مجموعه کوچکی از مقالات نشان می دهد که اطلاعات گفتمانی تا حدی رمزگذاری شده است (اغلب در لایه های میانی)، اما پوشش محدود است و مدل ها ممکن است در معنای روابط گفتمانی دچار مشکل شوند.

۵.۲ روندهای کمی و مشاهدات روش شناختی

در فهرست نویسندگان، نحو تمرکز غالب زبان شناسی است (۱۱۳ مقاله)، پس از آن (واژه شناسی) معناشناسی (۸۲)، صرف شناسی (۳۸) و گفتمان (۶) قرار دارند. از نظر روش شناسی، کاوش رایج ترین است (۱۱۷ مقاله)، و پس از آن تحلیل عناصر معماری (۵۸) قرار دارد، با انتساب ویژگی ها و رویکردهای مبتنی بر مثال کمتر ظاهر می شوند (هر کدام ۱۶ مقاله).

- انگلیسی غالب است؛ بسیاری از زبان های کم منبع همچنان کمتر مورد مطالعه قرار گرفته اند.
- BERT/RoBERTa (و mBERT) اهداف رایجی هستند زیرا روش های دسترسی داخلی به مدل های بسیار بزرگ یا بسته مقیاس پذیری ضعیفی دارند.
- کاوش اغلب با اختلالات یا تحلیل های توجه/جاسازی ترکیب می شود تا نتیجه گیری ها تقویت شود.
- بازیابی پذیری به معنای استفاده علی نیست؛ کاوش ها و مداخلات علّی اهمیت فزاینده ای پیدا می کنند.

۵.۳ محدودیت ها و ایده ها برای ادامه

محدودیت های کلیدی شامل (۱) پوشش متمایل به سمت انگلیسی و مدل های رمزگذار قدیمی و قابل دسترس است؛ (۲) ابهام روش شناختی در کاوش (یادگیری وظیفه، ناپایداری و محدودیت های متریک)؛ و (۳) هزینه تفسیر پذیری در مقیاس برای مدل های زبانی بزرگ معاصر که پژوهش را به سمت پرامپت کردن جعبه سیاه سوق می دهد و شفافیت را کاهش می دهد.

جهت گیری های آینده: گسترش مطالعات دسترسی داخلی به مدل های زبان بزرگ باز جدیدتر؛ استفاده از ارزیابی علی قوی تر (مداخلات، ویرایش خلاف واقعیت)؛ گسترش پوشش چندزبانه متنوع از نظر نوع شناسی؛ و نتایج قابلیت تفسیر را به پایداری پایین دستی و بهبود هدفمند مدل مرتبط می کند.

مرجع

López-Otal, M., Gracia, J., Bernad, J., Bobed, C., Pitarch-Ballesteros, L., Anglés-Herrero, E. (2025). Linguistic Interpretability of Transformer-based Language Models: a systematic review. ACM Computing Surveys (manuscript).