تلخیص انتزاعی متون

میلاد مولازاده
دانشکده مهدسی کامپیوتر
دانشگاه علم و صنعت ایران
molazadeh_milad@comp.iust.ac.ir

چکیده

بازنمایی‌های رمزگزار دوطرفه با بهره‌گیری از ترانسفورماتورها (BERT) [۱] اخیرا به سرآمد مدل‌های زبانی در زمینه پردازش زبان طبیعی تبدیل شده‌اند. در این گزارش به بررسی متدهای پیشین تلخیص متون پرداخته و نتایج گزارش داده می‌شود. استفاده از شبکه‌های از پیش‌آموزش دیده در مقیاس بالاتر و تنظیم‌سازی آن‌ها در مقیاس کمتر دادگان و منابع کمتر هدف این گزارش است.

۱ مقدمه

تلخیص متن به فرآیند تبدیل متن منبع در حجم بالاتر به متن مقصد در حجم کمتر که حاوی اطلاعات اصلی در متن اصلی است گفته می‌شود. تلخیص متون به دو شیوه انتزاعی و استخراجی انجام می‌گیرد. در روش استخراجی متن خلاصه از کنار هم گذاشتن جملات کلیدی بدست می‌آید و این کار بدلیل اینکه از لحاظ گرامری ساختار جمله حفظ می‌شود، آسان‌تر است. ولی در روش انتزاعی خلاصه چکیده‌ای از متن منبعی است که با به کنار هم قراردادن عبارات موجود ساخته شده است. در این گزارش هر دو روش شرح و کارهای انجام شده بررسی خواهد شد و روش انتزاعی را با استفاده از تنظیم‌سازی جزئیات پیاده‌سازی شرح داده خواهد شد.

مدل‌های مقیاس بالای از قبل آموزش دیده موفقیت‌های بسیاری را در زمینه درک زبان طبیعی بدست آورده‌اند. این مدل‌ها روی داده‌های زیاد آموزش دیده‌اند و روی داده کمتر و تسک‌های زیردامنه نتایج خوبی گرفته‌اند. متدهای خلاصه‌سازی می‌توانند به دو بخش براساس نوع منبع که می‌تواند تک سندی یا چندسندی باشد تقسیم شوند. در تلخیص تک‌سندی، فقط یک سند برای تولید خلاصه استفاده می‌شود. هردو روش استخراجی که مبتنی بر سرجمع کردن جملات است و روش انتزاعی که عبارات را به هم پیوند می‌دهد، می‌توانند به‌کار گرفته شوند. خلاصه‌سازی چند سندی به بیش از یک منبع اطلاعاتی برای تولید خلاصه نیاز دارد. هدف اینکار فقط حذف افزونگی یا انتخاب متن صحیح برای خلاصه نیست بلکه ارائه خلاصه کامل و منسجم است که چالش این کار می‌باشد.

۲ کارهای مرتبط / پیش‌زمینه

قسمت وسیعی از کارهای انجام شده را می‌توان به خلاصه‌سازی استخراجی نسبت داد. اما در طرف دیگر مایل هستیم خلاصه‌هایی تولید کنیم که به خلاصه انسان نزدیک‌تر و انتزاعی‌تر باشد، بعبارتی جملاتی از نو تولید بشوند. کار روی خلاصه‌سازی انتزاعی از مسابقات DUC-۲۰۰۳ و DUC-۲۰۰۴ با دادگانی متشکل از اخبار و خلاصه‌های دست‌نوشته آغاز شد. سرآمد مسابقه مدل

TOPIARY (زاجیک و همکاران ۲۰۰۴) [۲] بود، که از ترکیب تکنیک‌های فشرده‌سازی زبانی و الگوریتم تشخیص موضوع بدون‌ناظر برای استخراج عبارات کلیدی از متن برای فشرده کردن خروجی استفاده کرده بود.

با ظهور یادگیری عمیق و استفاده گسترده در پردازش زبان طبیعی، مدل ها به این سمت میل کردند. در ۲۰۱۵ راش و همکارانش [۳] از مدل های پیچشی برای برای رمزنگاری ورودی و همچنین شبکه عصبی غیربازگشتی مبتنی بر توجه برای تولید خلاصه استفاده کردند. کورپا و همکاران در ۲۰۱۶ [۴] به عنوان بهبودی بر مدل یاد شده قبلی از یک شبکه بازگشتی عصبی بعنوان رمزگشا استفاده کردند. در کاربعدی چن و لاپاتا در ۲۰۱۶ [۵] از رمزگشا و رمزنگار مبتنی بر شبکه‌های بازگشتی برای تولید خلاصه استفاده کردند. در ۲۰۱۸ ناراین و همکارانش [۶] سیستمی مبتنی بر یادگیری تقویتی برای خلاصه‌سازی استفاده کردند. در ۲۰۱۸ ژوو[۷] و همکارانش جملات را امتیازدهی می‌کنند و کنارهم قرار می‌دهند تا خلاصه‌سازی استخراجی انجام دهند.

در روش خلاصه‌سازی انتزاعی با استفاده از مدل‌های عصبی، از یک رمزنگار برای تبدیل متن ورودی $x = [x_1, x_2, \ldots, x_n]$ به نماهای پیوسته قابل درک برای کامپیوتر و از یک رمزگشا برای تولید توالی کوچکتر $y = [y_1, y_2, \ldots, y_m]$ استفاده می‌شود که کار یک مدل زبانی احتمالاتی $p(y_1, \ldots, y_m \mid x_1, \ldots, x_m)$ را انجام می‌دهد.
در ۲۰۱۹ یانگ لیو [۸] و همکاران با استفاده از BERT چارچوبی هم برای خلاصه‌سازی انتزاعی و هم استخراجی استفاده کردند. آنها رمزنگاری مبتنی برت را که می‌تواند از نظر معنایی جملات را نمایش دهد استفاده کردند. مدل استخراجی از لایه‌های بصورت پشته در آمده ترنسفورمر ساخته شده است. برای مدل انتزاعی از تنظیم‌سازی مدل ساخته شده با تغییراتی روی رمزگشا ساخته شده‌است.

۳ مدل پیشنهاد شده

ابتدا شرح مختصری از دادگان در این قسمت ارائه می‌شود سپس پارامترها و مدل ساخته شده شرح داده می‌شود.

۱.۳ دادگان

ددادگان مورد استفاده شامل ۱۰۲۹۱۶ جفت متن و خلاصه آن بصورت انتزاعی است. از این تعداد ۸۲۳۳۲ جفت برای آموزش و ۱۰۲۹۲ برای آزمون و ۱۰۲۹۲ برای اعتبارسنجی درنظرگرفته شده است. همچنین از داده‌های CNN/DM [۹] در مقیاس کمتر برای آزمون صحت تنظیم‌سازی استفاده شده است.

۲.۳ مدل و تنظیمات

اگرچه اشاره شد، BERT برای کارهای پردازش زبان طبیعی و تنظیم‌سازی سرآمد روش‌هاست ولی در بحث خلاصه‌سازی بدین سادگی نیست. بخاطر اینکه BERT بصورت یک مدل زبانی مبتنی بر پوشانه است، خروجی‌ها بصورت برداری از نشانه‌ها به جای جملاتی است که در بحث ما و مخصوصا در نوع استخراجی مهم است. BERT با مکانیزم تعبیه قطعه (segment embedding) جملات بدست می‌آید، اما اینکار فقط روی جملات ورودی انجام می‌شود و درکی از جملاتی که ساخته خواهد شد نداریم.
برای جملات یک نشان به عنوان ابتدای جمله $[CLS]$ اضافه می‌شود. این نشان گردآورنده ویژگی‌های جمله است. همچنین برای تفکیک جملات بصورت یک درمیان تعبیه قطعه افزوده می‌شود. به سبب اینکه منابع سخت‌افزاری برای آموزش کل دادگان اخبار CNN/DM در اختیار نیست از مدل پیش‌آموزش دیده یانگ لیو [۸] که مدلی است بر اساس روش استخراجی متون استفاده شده و تنظیم سازی شده است. این مدل از پیش آموزش دیده شده با استفاده از چندین لایه ترنسفورمر بصورت پشته‌ای آموزش دیده‌است

$$H^l = LN(h^(l-1) + MultiHeadAtt(h^(l-1))) \quad (۱)$$

$$h^l = LN(H^l + FFN(H^l)) \quad (۲)$$

لایه اول تعبیه قطعه هست که که تعبیه‌های سینوسی را ایجاد می‌کند. خروجی نهایی یک کلاسیفایر سیگموئیدی هست. برای خلاصه‌سازی انتزاعی مدل از پیش آموزش دیده گفته شده به عنوان رمزنگار انتخاب می‌شود و رمزگشا آموزش داده می‌شود. از بهینه‌ساز Adam و نرخ یادگیری گفته شده در مقاله (واسوانی و همکاران ۲۰۱۷) استفاده شده است. از تابع loss جدید که نسبت به نویز پایدار است استفاده شده است:

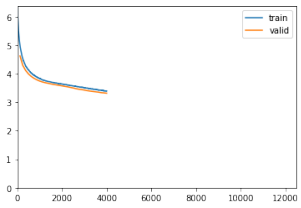
$$loss = (1 - \epsilon)\, ce(j) + \sum ce(j)/N \quad (۳)$$

۴ نتایج

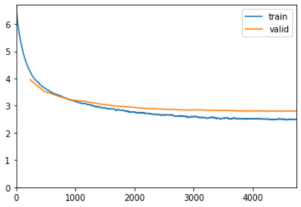
ابتدا بر روی بخشی از دادگان CNN/DM که مدلابتدا بر روی بخشی از داده CNN/DM صحت تنظیم‌سازی را انجام دادیم و تا چند مرحله با ترسیم نمودار loss بین آموزش و اعتبارسنجی گویای یادگیری و بهبود این عمل و تغییرات را بود. بدلیل کمبود منابع به همین اکتفا شده است. روی دادگان دست‌نوشته شده خبر جدید مدل امتحان شد و نتایج زیر بدست آمد:

$$rough1 = 34.5 \quad (۴)$$

$$rough1 = 34.5 \quad (۵)$$



شکل ۱: روند بهبود روی دادگان CNN/DM در مقیاس کم



شکل ۲: مقدار loss روی دادگان دست‌نوشته خبری

۵ تحلیل

در این مطالعه کارروی مدل‌های از پیش آموزش دیده BERT بررسی شد و نتایج نشان‌دهنده برتری و سرآمدی این مدل‌ها با منابع کمتر بود. با بهبود مدل‌های مبتنی BERT در آینده کارهای بیشتر و سرآمدتری را شاهد خواهیم بود.

منابع

[1] Devlin, J. & Chang, M. & Lee, K. & Toutanova, K. . (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Minneapolis, Minnesota.

[2] Zajic, D. & Dorr, B.J. & Lin, J. & Schwartz, R. (2007) Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing Management Special Issue on Summarization, 43(6):1549–1570*

[3] Rush, A.M. & Chopra, S. & Weston, J. (2015) A neural attention model for abstractive sentence summarization *Proceedings of EMNLP.*

[4] Chopra, S. & Auli, A. & Rush, A.M. (2016) Abstractive sentence summarization with attentive recurrent neural networks. *In HLT-NAACL.*

[5] Cheng, J. & Lapata, M. (2016) Neural summarization by extracting sentences and words. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.*

[6] Narayan, S. & Cohen, S.B. & Lapata, M.(2018) Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1747–1759, New Orleans, Louisiana.

[7] Zhou, Q. & Yang, N. & Wei, F.(2018) Neural document summarization by jointly learning to score and select sentences In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 654–663, Melbourne, Australia.

[8] Liu, Y. & Titov, I. & Lapata, M.(2019) Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1745–1755, Minneapolis, Minnesota.

[9] Hermann, K.M. & Kocisky, T. & Grefenstette, E.(2015) Teaching machines to read and comprehend. In *In Advances in Neural Information Processing Systems*, pp. 1693–1701

[10] Vaswani, A., Shazeer N., Parmar N., Uszkoreit, J., Jones, L., Attention Is All You Need. *arXiv:1706.03762*

[11] Liu, Y., Lapata, M. (2020). Text summarization with pretrained encoders.*EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 3730–3740.* https://doi.org/10.18653/v1/d19-1387