بسمه تعالی

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه آموزشی هوش مصنوعی

دانشجو: فاطمه وهابی

شماره دانشجویی : 4013614052

موضوع: تمرین اول درس پردازش زبان طبیعی

استاد: دکتر حمیدرضا برادران کاشانی

دستیار استاد: امیر مسعود سلطانی

بهار 1402

فهرست

پرسش‌ها

سوال اول) تفاوت اصلی در محاسبه احتمالات به کمک هموارسازی در مقابل روش discount چیست؟

درNLP ، هموارسازی و Discounting دو روش رایجی هستند که برای رفع مشکل پراکندگی استفاده می‌شوند، که مشکل زمانی است که کلمه یا دنباله‌ای از کلمات در یک متن قبلاً در داده‌های آموزشی دیده نشده است.

هموارسازی تکنیکی است که در آن یک مقدار ثابت کوچک به تعداد هر کلمه در داده‌های آموزشی اضافه می‌شود. این کار برای اطمینان از صفر نبودن احتمال یک کلمه و همچنین تنظیم احتمال کلمات دیگر در مجموعه آموزشی انجام می‌شود.

Discounting تکنیکی است که در آن مقدار مشخصی از تعداد هر کلمه در داده‌های آموزشی کم می‌شود. این کار به جهت تنظیم برای برآورد بیش از حد احتمالات کلمات در مجموعه آموزشی انجام می‌شود.

هموارسازی و Discounting برای بهبود قابلیت پیش‌بینی مدل‌های زبان استفاده می‌شود، اما در رویکردشان متفاوت هستند. هموارسازی با افزودن مقدار کمی برای تنظیم احتمالات کلمات کار می‌کند، در حالی که Discounting با کم کردن مقدار معینی برای تعدیل تخمین بیش از حد احتمالات کار می‌کند.

سوال دوم) جملات زیر را در نظر بگیرید و برای هر جمله کمترین مقدار n را برای این که یک مدل زبانی مبتنی بر n-gram بتواند روابط جمله مورد نظر را در یابد مشخص کنید. دلیل خود را برای انتخاب مقدار مورد نظر برای هر جمله شرح دهید. (نیاز به انجام محاسبات ندارد)

- من رفتند آسمان کتاب برای چرا فواره ها

از آنجایی که عبارت بالا مطابق اصول گرامری زبان فارسی نمی‌باشد و نیز به لحاظ معنایی نیز غیر قابل درک می‌باشد، حداقل از n=5 می‌بایست استفاده شود.

اما همانطور که می‌بینید برخی از قطعه های ان قابل دریافت و فهمیدن با استفاده از کمتر از این مقدار n نیز می‌باشد. اما برای بخش دیگری از رشته می‌بایست از مدل 5گرام استفاده کنیم. بنابراین 5 را در نظر می‌گیریم.

- ای شب از رویای تو رنگین شده

عبارت بالا مطابق قواعد زبان فارسی می‌باشد و یک عبارت ادبی با معنا می‌باشد. لذا از n=2 می‌توان استفاده کرد.

- she suddenly spots a White Rabbit

عبارت بالا یک عبارت ساده مطابق قواعد انگلیسی می‌باشد . لذا از n=2 می‌توان استفاده کرد.

- eins zwei drei vier fünf sechs Sieben

عبارت بالا شمارش اعداد به زبان آلمانی می‌باشد. لذا از n=2 می‌توان استفاده کرد.

آیا افزایش مقدار n در عمل باعث بهبود نتایج می‌شود؟ چرا؟ (در حل این مسئله توجه داشته باشید که در هر n-gram احتمال کلمه ی انتخاب شده براساس n-1 کلمه ی پیش از آن در نظر گرفته می‌شود)

خیر ممکن است نتیجه حتی با تعداد n بالا نتیجه خوبی نباشد. همچنین با افزایش مقدار n پیچیدگی محاسبات بیشتر شده و در صورت وجود صفرهای زیاد ممکن است نتیجه مطلوبی به دست نیاید.

سوال سوم) با توجه به پیکره متنی زیر، احتمالات خواسته شده را محاسبه کنید.

```
<s>my gulyali v lesu</s>
<s>gulyali my v lesu</s>
<s>v lesu my gulyali</s>
```

(آ) $p(gulyali|my)$

(ب) $p(gulyali|lesu, my)$

(ج) $p(lesu|my)$

آ)

$$p(gulyali|my) = \frac{c(gulyali, my)}{c(my)} = \frac{2}{3}$$

ب)

$$p(gulyali|lesu, my) = \frac{c(gulyali, lesu, my)}{c(lesu, my)} = \frac{1}{1} = 1$$

ج)

$$p(lesu|my) = \frac{c(lesu, my)}{c(my)} = \frac{0}{3} = 0$$

سوال چهارم)

یک مدل زبانی unigram را در نظر بگیرید، که روی مجموعه واژگانی به اندازه ی V تعریف شده است. فرض کنید هر کلمه m بار در یک پیکره ی متنی که حاوی M توکن است تکرار می‌شود .به ازای چه مقادیری از m حاصل احتمالات نتیجه شده از هموارساز Lidstone با متغیر α از احتمال غیرهموارشده بیش تر خواهد بود؟

$$\frac{count(M_i)}{V} \leq \frac{count(M_i) + \alpha}{V + (M * \alpha)}$$

می‌دانیم:

$$count(M_i) = m$$

طرفین وسطین می‌کنیم.

$$m * V + M * \alpha * m \leq V * m + V * \alpha$$

بخش mV از دو طرف حذف می‌شود.

$$Mm\alpha \leq V\alpha$$

α از دو طرف حذف می‌شود.

برای m خواهیم داشت:

$$m \leq \frac{V}{M}$$

$$m \leq \frac{m * M}{M}$$

$$1 \leq 1$$

نتیجه:

با توجه به نامساوی مورد نظر در می‌یابیم که فقط با $m=0$ به این نامساوی دست می‌یابیم. به ازای تمامی مقادیر دیگر که برای m در نظر بگیریم، مقدار احتمالات با هموارسازی و بدون هموارسازی تغییر نمی‌کند.

برنامه نویسی

کدها و توضیحات مربوط به آن در فایل ژوپیتر موجود می‌باشد.

همانطور که در فایل کدها مشاهده کردید تمام رشته‌ها را با طول 14 تولید کردیم و فرایند هموارسازی را بر روی ان اعمال کردیم.

حال به دنبال پاسخگویی به دو پرسش ان هستیم:

با استفاده از معیار سرگشتگی کیفیت قطعات ساخته شده را ارزیابی کنید.

پاسخ این سوال در فایل کد موجود است.

با توجه به ذات مسئله مطرح شده، چه روش ارزیابی دیگری را پیشنهاد می‌دهید. چرا؟

MSE: این متریک میانگین اختلاف مجذور بین نت‌های پیش بینی شده و واقعی در مجموعه تست را محاسبه می‌کند. امتیاز MSE کمتر، نشان می‌دهد که کدام مدل در پیش بینی داده‌های آزمون بهتر است.

Accuracy: این معیار اندازه گیری می‌کند که مدل به درستی نت موسیقی بعدی را در یک دنباله پیش بینی می‌کند. با مقایسه نت پیشبینی شده با نت واقعی در مجموعه تست محاسبه می‌شود.

F1 Score: امتیاز F1 معیاری برای سنجش دقت و کامل بودن پیش بینی‌های مدل است. هم precission (نسبت پیش‌بینی‌های صحیح در بین همه پیش‌بینی‌ها) و هم recall (نسبت نتایج واقعی که به درستی پیش‌بینی شده‌اند) را در نظر می‌گیرد.

مقاله خوانی

سوال اول) هدف مقاله را در یک یا دو جمله توضیح دهید.

این مقاله طراحی و پیاده سازی GNMT، یک سیستم NMT تولیدی در گوگل را ارائه می‌دهد که هدف آن ارائه راه حل‌هایی برای مشکلات ترجمه ماشین عصبی است. این سیستم دارای قابلیت‌های بهبود زمان استنتاج، مقابله موثر با کلمات نادر، قابلیت عمل بر روی مجموعه داده‌هایی به زبان‌های مختلف و غیره می‌باشد که این سیستم را به نسبت سیستم‌های قبلی متمایز می‌کند.

سوال دوم) ترجمه ماشین عصبی‌های پیشین چه مشکلاتی داشتند؟

- سرعت آموزش و استنتاج در آن‌ها کندتر بود.
- در برخورد با کلمات نادر ناکارآمد بودند.
- گاهی اوقات همه کلمات در جمله مبدا ترجمه نمی‌شدند.

سوال سوم) مکانیزم توجه را توضیح دهید.

شبکه عصبی در راستای تقلید از اعمال مغز انسان به شیوه‌ای ساده عمل می‌کند. مکانیسم توجه همچنین تلاشی را برای اجرای همان عمل تمرکز انتخابی بر روی چند چیز مرتبط انجام می‌دهد، در حالی که بقیه را در شبکه‌های عصبی عمیق نادیده می‌گیرد.

مکانیسم توجه به عنوان بهبودی نسبت به سیستم ترجمه ماشینی عصبی مبتنی بر رمزگشای رمزگذار در پردازش زبان طبیعی (NLP) پدیدار شد. بعدها این مکانیسم در کاربردهای دیگری از جمله بینایی کامپیوتری، پردازش گفتار و غیره مورد استفاده قرار گرفت.

مکانیزم توجه توسط Dzmitry Bahdanau و همکارانش ارائه شد.

ایده اصلی این است که هر بار که مدل یک کلمه خروجی را پیش‌بینی می‌کند، فقط از بخش‌هایی از ورودی استفاده می‌کند که مرتبط‌ترین اطلاعات را به جای کل دنباله متمرکز شده دارد. به عبارت ساده‌تر، فقط به چند کلمه ورودی توجه می‌کند.

مکانیزم توجه رابطی است که رمزگذار و رمزگشا را به هم متصل می‌کند و اطلاعاتی را از هر حالت پنهان رمزگذار در اختیار رمزگشا قرار می‌دهد. با این چارچوب، مدل قادر است به طور انتخابی بر روی بخش‌های ارزشمند دنباله ورودی تمرکز کند و از این رو، ارتباط بین آنها را بیاموزد. این به مدل کمک می‌کند تا به خوبی با جملات ورودی با طول بلند کار کند.

سوال چهارم) مفهوم مقاومت در این مقاله به چه معنا استفاده شده است؟

در ناتوانی و نداشتن استحکام و مقاومت در ترجمه کلمات نادر به کار برده شده است.

سوال پنجم) به صورت خلاصه روش معرفی شده در این مقاله و دلایل اجزای مختلف آن را در راستای دست یابی به پاسخ‌های بهتر را شرح دهید (نیازی به شرح روش و چگونگی موازی سازی مدل و دادگان که در مقاله مطرح شده است نمی‌باشد).

مدل پیشنهادی از چارچوب یادگیری دنباله با ترتیب رایج با attention پیروی می‌کند. این مدل دارای سه جزء است: یک شبکه رمزگذار، یک شبکه رمزگشا و یک شبکه attention. رمزگذار یک جمله منبع را به لیستی از بردارها، یک بردار در هر نماد ورودی تبدیل می‌کند. با توجه به این لیست از بردارها، رمزگشا در هر زمان یک نماد تولید می‌کند، تا زمانی که نماد ویژه پایان جمله (EOS) تولید شود. رمزگذار و رمزگشا از طریق یک ماژول attention به هم متصل می‌شوند که به رمزگشا اجازه می‌دهد در طول دوره رمزگشایی بر روی مناطق مختلف جمله منبع تمرکز کند.

اتصالات باقیمانده

LSTMهای پشته ای عمیق اغلب دقت بهتری نسبت به مدل‌های کم عمق نشان می‌دهند. با این حال، انباشتن لایه‌های بیشتری از LSTM فقط برای تعداد معینی از لایه‌ها کار می‌کند، که چنانچه بیشتر از آن باشد، شبکه

بسیار کند می‌شود و آموزش آن دشوار می‌شود، احتمالاً به دلیل انفجار و ناپدید شدن مشکلات گرادیان. در تجربه‌ای که به دست امد، با کارهای ترجمه در مقیاس بزرگ، لایه‌های LSTMانباشته شده به خوبی تا 4 لایه، به سختی با 6 لایه، و بسیار ضعیف با بیش از 8 لایه کار می‌کنند.

رمزگذار دو جهته برای لایه اول

برای سیستم‌های ترجمه، اطلاعات مورد نیاز برای ترجمه برخی کلمات در سمت خروجی می‌تواند در هر نقطه از سمت منبع ظاهر شود. اغلب اطلاعات سمت منبع تقریباً از چپ به راست است، شبیه به سمت مقصد، اما بسته به جفت زبان، اطلاعات یک کلمه خروجی خاص می‌تواند توزیع شود و حتی در مناطق خاصی از سمت ورودی تقسیم شود.

موازی سازی مدل

با توجه به پیچیدگی مدل پیشنهادی، ما از موازی‌سازی مدل و موازی‌سازی داده‌ها برای سرعت بخشیدن به آموزش استفاده می‌کنیم.

سوال ششم) چرا از اتصالات باقی مانده استفاده شده است؟

اتصالات باقیمانده تا حد زیادی جریان گرادیان را در گذر به عقب بهبود می‌بخشد، که به ما این امکان را می‌دهد شبکه‌های رمزگذار و رمزگشای بسیار عمیق را آموزش دهیم. در بیشتر آزمایش‌هایمان، ما از 8 لایه LSTM برای رمزگذار و رمزگشا استفاده می‌کنیم، اگرچه اتصالات باقی‌مانده می‌توانند به ما امکان آموزش شبکه‌های عمیق‌تر را بدهند.

سوال هفتم) نتیجه ی مقاله چیست؟

این مقاله به طور مفصل پیاده‌سازی سیستم ترجمه ماشین عصبی Google (GNMT) را شامل تمام تکنیک‌هایی که برای دقت، سرعت و استحکام آن حیاتی هستند، شرح داد. در معیار ترجمه عمومیWMT'14 ، کیفیت ترجمه سیستم پیشنهادی به همه نتایج منتشر شده در حال حاضر نزدیک می‌شود یا از آن فراتر می‌رود. مهم‌تر از آن، نشان داده شد که رویکرد پیشنهادی به مجموعه داده‌های تولیدی بسیار بزرگ‌تر، که چندین مرتبه داده‌های بزرگ‌تری دارند، برای ارائه ترجمه‌های با کیفیت بالا تبدیل می‌شود.

یافته‌های کلیدی روش پیشنهادی عبارتند از: 1) مدل‌سازی موثر واژه‌نامه، واژگان باز و چالش زبان‌های غنی از لحاظ صرفی را برای کیفیت ترجمه و سرعت استنتاج کنترل می‌کند، 2) ترکیبی از مدل و موازی‌سازی داده‌ها می‌تواند برای آموزش موثر مدل‌های NMT دنباله به دنباله پیشرفته در تقریبا یک هفته استفاده شود.

3)کوانتیزاسیون مدل استنتاج ترجمه را به شدت تسریع می‌کند و امکان استفاده از این مدل‌های بزرگ را در یک محیط تولید مستقر می‌کند، و 4) بسیاری از جزئیات اضافی مانند عادی‌سازی طول و موارد مشابه برای اینکه سیستم‌های NMT روی داده‌های واقعی به خوبی کار کنند، ضروری هستند.

با استفاده از مقایسه رتبه‌بندی شده توسط انسان به عنوان یک معیار، نشان داده شد که سیستم GNMT پیشنهادی به دقتی نزدیک می‌شود که مترجمان دو زبانه انسانی متوسط در برخی از مجموعه‌های آزمایشی پیشنهادی به دست آورده‌اند. به ویژه، در مقایسه با سیستم تولید مبتنی بر عبارت قبلی، این سیستم GNMT تقریباً 60٪ کاهش خطاهای ترجمه را در چندین جفت زبان رایج ارائه می‌دهد.