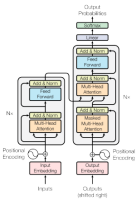

تمرین یکم
مدل‌سازی زبان طبیعی

تاریخ بارگذاری تمرین: ۱۴۰۱/۱۲/۲۸
تاریخ تحویل تمرین: ۱۴۰۲/۰۱/۱۸

۱ پرسش‌ها

۱. تفاوت اصلی در محاسبه احتمالات به کمک هموارسازی[1] در مقابل روش discount چیست؟

۲. جملات زیر را در نظر بگیرید و برای هر جمله کم‌ترین مقدار n را برای این که یک مدل‌زبانی 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 فواره‌ها
- ای شب از رویای تو رنگین شده
- she suddenly spots a White Rabbit
- eins zwei drei vier fünf sechs sieben

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

۳. باتوجه به پیکره‌ی متنی[2] 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) $p(lesu|my)$

(نکته جالب: هر سه جمله‌ی مشخص شده در پیکره‌ی متنی به یک معنا هستند «ما در جنگل قدم زدیم» یا «We walked in the forest»، باوجود این که ترتیب قرارگیری کلمات در هر جمله با دیگری متفاوت است، هر یک جمله‌ی معتبری را می‌سازند، این مسئله به منعطف بودن ساختار جملات در زبان روسی باز می‌گردد)

[1] Smoothing
[2] Text corpus

۴. یک مدل زبانی unigram را در نظر بگیرید، که روی مجموعه واژگانی[۳] به اندازه‌ی V تعریف شده است. فرض کنید هر کلمه m بار در یک پیکره‌ی متنی که حاوی M توکن است تکرار می‌شود. به‌ازای چه مقادیری از m حاصل احتمالات نتیجه شده از هموارساز Lidstone با متغیر α از احتمال غیرهموارشده بیش‌تر خواهد بود؟

۲ برنامه‌نویسی

۱.۲ مقدمه

موسیقی یا خُنیا که ارسطو آن را یکی از شاخه‌های ریاضی به شمار آورده است، یکی از انواع هنری است که از سالیان دور آدمی را همراهی کرده است. هر قطعه‌ی موسیقیایی برای مکتوب شدن و انتشار به نسل‌های بعد، نیازمند زبان و قواعد مشخصی است. فارغ از تمامی جزییات، نُت‌های موسیقی الفبای اولیه آن را تشکیل می‌دهند که با در کنار هم قرار گرفتن آن‌ها در فواصل و گامی مشخص یک قطعه خلق می‌شود. هر نت میزان زیر و بَمی و کشش صدا را در یک فرکانس خاص بیان می‌کند. برای تمامی صداهای موسیقیایی تنها هفت نام وجود دارد که گاه با کلمات تک هجایی (که در ایتالیا، فرانسه و به طبع آن در ایران استفاده می‌شود) و گاه با استفاده از الفبا (که کشورهای انگلیسی و آلمانی زبان آن را پذیرفته‌اند) مشخص می‌شوند:

- هجایی: Do Re Mi Fa Sol La Si
(خوانش هجایی را به صورت مقابل نیز می‌نویسند: سی لا سُل فا می رِ دُ)
- الفبایی: C D E F G A B
(از چپ به راست خوانده شود. نت Do در نام‌گذاری هجایی معادل نت C در نامگذاری الفبایی است، روابط بقیه نت‌ها نیز به همین ترتیب پیش می‌رود)

در موسیقی علائم دیگری با نام تغییردهنده‌ها نیز وجود دارند که سمت راست هر نت موسیقیایی قرار می‌گیرند و باعث زیر و بم شدن آن‌ها به اندازه‌ی نیم‌پرده می‌گردند (بدون آن که نام نت تغییر کند). بِمُل[۴] که آن را با علامت $\flat$ نمایش می‌دهند هر نت را نیم‌پرده بم‌تر می‌کند. در موسیقی ایرانی از تغییردهنده‌ای با نام کُرُن[۵] استفاده می‌گردد که هر نت را ربع پرده (کم‌تر از نیم‌پرده) بم‌تر می‌کند. برای سادگی بمل را با استفاده از حرف b و کرن را با k نمایش می‌دهیم، به این ترتیب نت سی بمل (سی در این‌جا نمایان‌گر خوانش هجایی است، و معادل B در خوانش الفبایی است) را می‌توان به صورت Bb و نت لا کرن را به صورت Ak نمایش داد. برای مثال نت قطعه‌ی «خونه‌ی مادربزرگه» در نام‌گذاری الفبایی به شرح زیر می‌باشد:

C C G G C S C G G S F G Ab F G G S G

در این مجموعه نت‌ها برای نمایش سکوت از حرف S استفاده شده است. برای سادگی از دیگر پیچیدگی‌ها مانند دیرند (ارزش زمانی) نت‌ها پرهیز شده است.

۲.۲ برنامه

دادگان این مسئله متشکل از سه ستون name، note، و dastgah می‌باشد که به‌ترتیب مشخص‌کننده‌ی نام قطعه‌ی موسیقی، نت‌های آن، و دستگاه موسیقی که قطعه مورد نظر در آن نواخته می‌شود است. مدل زبانی مبتنی بر n-gram بنویسید که مجموعه‌دادگان مذکور را دریافت کند و قطعات زیر را تکمیل کند:

[3] vocabulary
[4] Bemol
[5] Koron

- S S S S
- G C G C
- A B C D
- A Bb C D
- A S B Db

هر یک از دنباله نت‌های تولیدشده می‌بایست دارای طولی به‌اندازه ۱۰ باشد (طول کلی هر قطعه بنابراین ۱۴ است) و به ازای تمامی ترکیبات مشخص شده در جدول زیر یک‌بار تکمیل گردد:

Language model	Smoothing
Uni-gram	Lidstone with $\alpha = 1$ (Laplace) and $\alpha = 0.5$(Jeffreys-Perk's Law)
Bi-gram	Lidstone with $\alpha = 1$ (Laplace) and $\alpha = 0.5$(Jeffreys-Perk's Law)
Tri-gram	Lidstone with $\alpha = 1$ (Laplace) and $\alpha = 0.5$(Jeffreys-Perk's Law)

- با استفاده از معیار سرگشتگی[۶] کیفیت قطعات ساخته شده را ارزیابی کنید.
- با توجه به ذات مسئله مطرح شده، چه روش ارزیابی دیگری را پیشنهاد می‌دهید. چرا؟

۳ خوانش مقاله

مقاله‌ی «Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation» [۱] را مورد مطالعه قرار دهید و به پرسش‌های زیر پاسخ دهید:

۱. هدف مقاله را در یک یا دو جمله توضیح دهید.

۲. ترجمه ماشین عصبی[۷]های پیشین چه مشکلاتی داشتند؟

۳. مکانیزم توجه[۸] را توضیح دهید؟ (می‌توانید برای پاسخ به این پرسش از منابع موجود در اینترنت استفاده کنید)

۴. مفهوم مقاومت[۹] در این مقاله [۱]، به چه معنا استفاده شده است؟

۵. به صورت خلاصه روش[۱۰] معرفی شده در این مقاله و دلایل اجزای مختلف آن را در راستای دست‌یابی به پاسخ‌های بهتر را شرح دهید (نیازی به شرح روش و چگونگی موازی‌سازی مدل و دادگان که در مقاله مطرح شده است نمی‌باشد).

۶. چرا از اتصالات باقی‌مانده[۱۱] استفاده شده است؟

۷. نتیجه‌ی مقاله چیست؟

[6] Perplexity
[7] Neural machine translation
[8] Attention mechanism
[9] Robustness
[10] Methodology
[11] Residual connections

۴ نکات تحویل

۱. پاسخ خود را تحت پوشه‌ای به اسم NLP_NAME_ID و در قالب zip بارگذاری نمایید.

۲. این پوشه می‌بایست حاوی موارد زیر باشد:

- پوشه‌ای با نام code باشد که شامل برنامه‌ی نوشته/تغییر داده شده است.
- پوشه‌ای با نام doc که حاوی داکیومنت‌ها و فایل توضیحات می‌باشد.

۳. لازم به ذکر است که رعایت قوانین نگارشی حائز اهمیت خواهد بود.

مراجع

[1] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.