به نام خدا

تمرین شماره یک درس پردازش زبانهای طبیعی توحید عابدینی ۹۹۷۲۲۱۵۳

سوال ۱)

قسمت a)

نوع اول ابهام – ابهام صوتی:

این نوع ابهام در سطح صوتی و گفتاری رخ میدهد و ممکن هست به دلیل تلفظ بد یا تشابه لفظ دو کلمه باعث ایجاد دو فهم متفاوت از یک جمله شود.

مثال:

او در حیاط/ت است. => او در حیاط است. (حیاط خانه) یا او در حیات است. (در قید حیات است و زنده است)

نوع دوم ابهام – ابهام نحوی:

این نوع ابهام در سطح نحوی و رابطه نحوی بین کملات یک جمله پیش می‌آید و ممکن است بتوان برای یک جمله درخت تجزیه نحوی متفاوتی داشت که هر کدام باعث داشتن تفسیر و معنای متفاوت برای یک جمله می‌شود.

مثال:

او به علی گفت که نمی‌تواند بیاید. => او به علی گفت که خودش نمی‌تواند بیاید. یا او به علی گفت که علی نمی‌تواند بیاید.

نوع سوم ابهام – معنایی:

این نوع ابهام به دلیل داشتن معانی متفاوت یک کلمه ممکن است ایجاد شود.

مثال:

کلمه سیر در فارسی دو معنی متفاوت دارد. یکی سیر به معنی مخالف گشنه بودن و دیگری گیاه سیر

نوع چهارم ابهام – سطح سخن (discourse level) :

این نوع ابهام به دلیل پس ارجاع دهی (anaphora) و یا قرار دادن ضمیر پس از مرجع در جملات ایجاد میشود که منجر به برداشت های متفاوت از جمله می‌شود.

مثال:

حسن تا پدرش را دید کیفش را جابه‌جا کرد. => حسن کیف خودش را جابه‌جا کرد یا کیف پدرش را؟

قسمت b)

۱.خلاصه سازی متن Text Summarization:

یکی از کاربرد های پردازش زبان طبیعی خلاصه سازی متن میباشد. با توجه به حجم فزاینده متن و داده های متنی در سطح اینترنت خلاصه سازی مفید یک متن میتواند در جاهای متفاوت کاربرد زیادی داشته باشد و اگر به طور اتوماتیک انجام شود بسیار کمک کننده خواهد بود. به عمل خلاصه سازی یک متن که حجم را کمتر کرده و موضوعات مهم یک متن را پوشش میدهد، تسک خلاصه سازی گفته میشود.

۲.پیوند دهی موجودیت ها Entity Linking:

برای تعریف این تسک ابتدا باید با مفهوم موجودیت آشنا شویم. در اینجا منظور از موجودیت در حقیقت موجودیت نامدار یا Named Entity میباشد. که در استخراج اطلاعات کاربرد دارد و منظور از آن یک موجودیت است که در دنیای واقع وجود دارد. به عنوان مثال مکانهای خاص مانند برج میلاد یا شهر مانند تهران و ارومیه یا کشور مانند ایران و یا افراد سرشانس مانند رئیس جمهور یا وزاری معروف یک کشور همگی نوع موجودیت های نامدار هستند. حال به توضیح پیوند دهی موجودیت ها میپردازیم که در اینجا پس از آنکه نوع موجودیت ها شناسایی شدند اقدام به لینک کردن و پیوند دادن هر کدام از آنها به یک پایگاه دانش یا Knowledge Base میکنیم مانند (دانشنامه های آنلاین ویکیپدیا (با زبانهای مختلف میتوان انجام داد) یا DBpedia و یا برای موارد دینی و اسلامی مانند ویکی فقه و ویکی شیعه و ...) و در حقیقت با این کار به رفع ابهام از یک نوع موجودیت پرداخته ایم.

۳.ترجمه ماشینی Machine Translation:

ترجمه متن از یک زبان به زبان دیگر همانطوری که از دیرباز بین انسانها وجود داشت هم اکنون هم یکی از نیازهای مهم است. مترجم های انسانی به راحتی ممکن است در دسترس نباشند و هزینه بر باشند فلذا داشتن یک سیستم مترجم ماشینی که متن ورودی را از یک زبان به زبان دیگر ترجمه کند بسیار مهم و کلیدی خواهد بود. به عمل ترجمه خودکار متن از یک زبان به زبان دیگر، ترجمه ماشینی گفته میشود.

قسمت امتیازی:

مقاله ی انتخابی:

Reinforcement Learning for Abstractive Question Summarization with Question-aware Semantic Rewards

مقاله ی انتخابی در رابطه با تسک خلاصه سازی است.

۱.معرفی و ایده مقاله:

امروزه رشد فزاینده ی سوالات مربوط به سلامت از طرف مصرف کنندگان نیاز مبرم به یک سیستم قابل اطمینان پاسخ دهنده ی سوال را ایجاب میکند. مطالعه ی اخیر نشان داده است که خلاصه سازی دستی این سوالات بهبود های چشمگیری در رابطه با استخراج پاسخ های مرتبط داشته است. با این وجود خلاصه سازی خودکار سوالات طولانی یک امر چالش برانگیز است. زیرا کمبود داده آموزشی و پیچیدگی subtask های مربوط نظیر بازیابی نوع (type recognition) و توجه بر روی سوال (question focus) این کار را سخت تر میکند. در این مقاله یک چهارچوب مبتنی بر یادگیری تقویتی برای خلاصه سازی مفهومی سوالات ارائه شده است.

دو نوع پاداش جدید که از تسک های زیرین شناسایی نوع سوال (question-type identification) و توجه بر روی سوال و بازیابی آن (question-focus recognition) به دست آورده شده اند، ارائه میشود که باعث regularize شدن مدل تولید سوال میشود.

این پاداش ها تولید سوالات معتبر از نظر معنایی را تضمین میکند و به وجود نوع موجودیت های دارویی اصلی در خلاصه کمک میکند. مدل داده شده بر روی دو دیتاست معیار، ارزیابی شده است که عملکرد بهتری از مرز دانش در این تسک داشته است. بررسی دستی سوالات ایجاد شده نشان داده است که سوالات ایجاد شده متنوع تر هستند و تناقض واقعی کمتری دارند.

۲.توضیح الگوریتم و مدل ارائه شده:

هدف اصلی این تسک این است که وقتی یک سوال به عنوان ورودی به مدل داده شد، یک خلاصه از آن تولید کند به نحوی که اطلاعات برجسته سوال را دارا باشد. یک مدل مبتنی بر یادگیری تقویتی که با معماری ترنسفورمری رمزگزار – رمزگشا است ارائه شده است.

پاداش های معنایی آگاه به سوال (Question aware Semantic Rewards):

۱. پاداش تشخیص نوع سوال:

مستقل از تسک پیش آموزش دادن، اکثر مدل های زبانی از MLE برای fine-tune کردن زیر تسک ها استفاده میکنند.

MLE دو اشکال دارد:

۱. بایاس در معرض گذاری (exposure bias): مدل در زمان آموزش داده gold و استاندارد دریافت میکند ولی در هنگام آزمون چنین انتظاری ندارد.

۲.فروپاشی نمایانگرانه (representational collapse): تنزل نمایانگر های تعمیم پذیر مدل های از پیش آموزش داده شده در هنگام fine-tune کردن.

برای مقابله با اشکال اول در مقالات قبلی از پاداش رژ-بلاو برای آموزش مدل های تولید کننده سوال استفاده شده است.

این معیار ها برای ارزیابی معنایی مناسب نیستند.
با fine-tune کردن یک شبکه بر مبنای bert یک مدل برای تشخیص نوع سوال ارائه شده است. بازنمایی توکن [CLS] از لایه ی آخر شبکه fine-tune شده به لایه های انتهایی feed-forward اضافه شده است که به کمک آن logit نهایی محاسبه میشود.

$$l = \boldsymbol{W}(tanh(\boldsymbol{U}h_{[CLS]} + \boldsymbol{a})) + \boldsymbol{b}$$

در نهایت نوع سوال به وسیله اعمال تابع فعال ساز سیگموید بر روی خروجی نورون های logit تشخیص داده میشود. از شبکه fine-tune شده برای محاسبه پاداش استفاده میشود که به عنوان امتیاز F بین نوع سوال پیش بینی شده و نوع سوال اصلی در نظر گرفته میشود.

۲.پاداش شناسایی توجه سوال:

یک خلاصه ی سوال خوب باید نکات کلیدی سوال را در بر داشته باشد. در کارهای قبلی پاداش های مبتنی بر رژ (ROUGE) برای بیشینه کردن پوشش خلاصه تولید شده استفاده میشود اما تضمینی برای حفظ کردن اطلاعات کلیدی در خلاصه سوال ندارد. یک تابع پاداش جدید به نام پاداش شناسایی توجه سوال ارائه شده است که یک درجه از اینکه تا چه حدی اطلاعات کلیدی سوال در خلاصه وجود دارد را محاسبه میکند. مشابه QTR (پاداش شماره یک – تشخیص نوع سوال)، یک شبکه بر مبنای برت fine-tune میشود تا این کار را انجام دهد. به طور خاص تر با دریافت ماتریس بازنمایی ها $(\boldsymbol{H} \in \mathcal{R}^{n \times d})$ که n توکن و d بعد دارد و از لایه آخر bert گرفته شده اند یک پیش بینی سطح توکن با استفاده از یک لایه خطی feed-forward انجام میشود. برای هر بازنمایی توکن logit متناظر آن با اسم l_i محاسبه میشود که $l_i \in \mathcal{R}^{|C|}$ در آن |C| تعداد کلاس هاست و توجه سوال با استفاده از رابطه ی زیر محاسبه میشود.

$$f_i = softmax(\boldsymbol{W}h_i + \boldsymbol{b}).$$

از شبکه fine-tune شده برای محاسبه پاداش ها استفاده میشود که به عنوان امتیاز F بین توجه سوال پیش بینی شده و توجه سوال اصلی در نظر گرفته میشود.

۳. تقویت Policy Gradient:

از الگوریتم Policy Gradient برای آموزش استفاده میشود که یکی از روش های موجود در یادگیری تقویتی است. در اینجا عامل ما رمزگشای PropherNet است که با محیط ما یعنی شبکه نوع سوالات و توجه سوالات تعامل میکند. سیاست ما با پارامتر های PropherNet تعیین شده و با مشاهده ی پاداش تغییر میکند. هدف اصلی کمینه کردن تابع هزینه ی زیر است:

$$\mathcal{L}_{RL} = -E_{Q^s \sim p_\theta}[r(Q^s, Q^*)].$$

در نهایت شبکه با استفاده از یک تابع هزینه ترکیبی آموزش داده میشود که به شرح زیر است:

$$\mathcal{L} = \alpha \mathcal{L}_{RL} + (1 - \alpha)\mathcal{L}_{ML}$$

که در آن α فاکتور مقایس گذاری است و L_{ML} هم برابر است با :

$$-\sum_{t=1}^{t=m} logp(q_t^* | q_1^*, q_2^*, \ldots, q_{t-1}^*, \mathcal{S}),$$

که در آن S سوال ورودی است و q ها هم کلمات ما در آن سوال هستند.

۳.نتایج:

در جدول زیر مقایسه ای بین بیس لاین ها و نسخه های مختلف الگوریتم ارائه شده آورده شده است که میبینم در حالتی که از هر دو نوع پاداش استفاده میکنیم بهترین عملکرد را داریم:

	Models	MEQSUM			MATINF*		
		R-1	R-2	R-L	R-1	R-2	R-L
Baselines	Seq2Seq (Sutskever et al., 2014)	25.28	14.39	24.64	17.77	5.10	21.48
	Seq2Seq + Attention (Bahdanau et al., 2015)	28.11	17.24	27.82	19.45	6.45	23.77
	Pointer Generator (PG) (See et al., 2017)	32.41	19.37	36.53	23.31	7.01	26.61
	SOTA (Ben Abacha and Demner-Fushman, 2019)	44.16	27.64	42.78	–	–	–
	SOTA* (Ben Abacha and Demner-Fushman, 2019)	40.00	24.13	38.56	24.58	7.30	28.08
	Transformer (Vaswani et al., 2017)	25.84	13.66	29.12	22.25	5.89	26.06
	BertSumm (Liu and Lapata, 2019)	26.24	16.20	30.59	31.16	11.94	34.70
	$T5_{BASE}$ (Raffel et al., 2019)	38.92	21.29	40.56	39.66	21.24	41.52
	PEGASUS (Zhang et al., 2019a)	39.06	20.18	42.05	40.05	23.67	43.30
	$BART_{LARGE}$ (Lewis et al., 2019)	42.30	24.83	43.74	42.52	23.13	43.98
	MINILM (Wang et al., 2020)	43.13	26.03	46.39	35.60	18.08	38.70
	ProphetNet (Qi et al., 2020)	43.87	25.99	46.52	46.94	27.77	48.43
	ProphetNet + ROUGE-L	44.33	26.32	46.90	48.17	28.13	48.66
Joint Learning	ProphetNet + Q-type	44.40	26.63	47.05	47.19	28.02	48.70
	ProphetNet + Q-focus	44.62	26.61	47.28	47.14	28.06	48.64
	ProphetNet + Q-type + Q-focus	44.67	26.72	47.34	47.18	28.04	48.65
Proposed Approach	ProphetNet + QTR	44.60	26.69	47.38	47.51	28.40	48.94
	ProphetNet + QFR	45.36	27.33	47.96	47.53	28.29	49.11
	ProphetNet + QTR + QFR	45.52	27.54	**48.19**	47.73	28.54	**49.33**

در نهایت یک نمونه از خلاصه تولید شده توسط الگوریتم هم ارائه میشود:

Original Question-I: who makes bromocriptine i am wondering what company makes the drug bromocriptine, i need it for a mass i have on my pituitary gland and the cost just keeps raising. i cannot ever buy a full prescription because of the price and i was told if i get a hold of the maker of the drug sometimes they offer coupons or something to help me afford the medicine. if i buy 10 pills in which i have to take 2 times a day it costs me 78.00. and that is how i have to buy them. thanks.

Reference: who manufactures bromocriptine?

Generated Summary

ProphetNet: what is bromocriptine?

Proposed Approach: what company makes bromocriptine and how much does it cost?

که در آن خروجی تولید شده توسط PropherNet هم آورده شده است.

قسمت c)

این الگوریتم برای توکن بندی استفاده میشود که بیشتر برای زبان چینی خوب عمل میکند و برای بقیه زبان‌ها به دلیل ساختار زبان عملکرد مناسبی ندارد. نحوه کار الگوریتم به این صورت است که ورودی را صورت رشته و بدون فاصله دریافت میکند و از ابتدای رشته شروع به پردازش میکند و در ابتدا طولانی ترین تطبیق را برای رشته ورودی از دیکشنری انتخاب کرده و آن را جدا کرده و ادامه میدهد تا به انتهای رشته برسد.

مثال:

جمله در حالت عادی:

He ate a radish.

جمله در صورتی که با الگوریتم توکنایز شود:

Heat ear a dish.

قسمت d)

تعریف Lemmatization :

هدف اصلی لمتایز کردن کاهش یک کلمه به شکل اصلی آن است. برای مثال کلمات مترادف یکپارچه میشنود و صرف تمامی افعال به مصدر های خود تغییر شکل میدهند و علائم جمع و علامت های مشابه آن حذف میشوند. لمتایز کردن معمولا از روی یک دیکشنری انجام میشود که باعث میشود کلمات بدون معنی و جدید نداشته باشیم.

مثال:

- رفتم، میروم، بروم = رفت#رو
- کتابها = کتاب
- کتابش = کتاب

تعریف Stemming :

استمینگ یا ریشه یابی به عمل حذف وند ها (پیشوند – پسوند) از ابتدا و انتهای کلمات گفته میشود. رویکرد استمینگ با لمتایزر متفاوت است و ممکن است کلمات بی معنی تولید شود (مخصوصا در انگلیسی که کلمات زیادی ریشه لاتین دارند و در صورت اضافه شدن برخی وند ها شکل کلمات تغییر میابد.) در نهایت هدف اصلی این عمل رسیدن به ریشه کلمه از مشتقات متفاوت آن است.

مثال:

- پوشیده ام = پوشید
- شهرستان = شهر
- بی معنی = معنی

سوال ۲)

قسمت a)

عبارت منظم قابل قبول برای این قسمت به صورت زیر است:

```
^[A-Z].*f$
```

اینکه در ابتدای هر رشته باشد را با ^ در اول تعیین میکنیم. اینکه کلمات با حروف بزرگ باشند با این قسمت تعیین میشود: [A-Z]

اینکه حروف وسط بدون هیچ پیش شرطی باشند با . تعیین میشود که وقتی به *. تبدیل میشود یعنی صفر یا چند وقوع از این الگو

و در نهایت اینکه در انتهای رشته حرف f باشد را با این قسمت تعیین میکنیم: f$

قسمت b)

```
[^4 ]*(4[^4 ]*){3,}
```

برای این قسمت سوال در ابتدا تعدادی عبارت به جز ۴ و فاصله تکرار می‌شود و سپس به تعداد ۳ یا بیشتر برای هر ۴ تعدادی غیر ۴ و فاصله در نظر میگیریم و به این صورت الگو را میتوانیم مچ کنیم.

قسمت c)

```
^[0-9]*[13579]([^a-z]*[a-z][^a-z]*){1,}[0-9]*[02468]$
```

در ابتدای رشته تعدادی رقم ۰ تا ۹ تکرار می‌شود و پس از ان یک رقم فرد می‌آید تا قسمت عددی اول فرد باشد و سپس وارد یک capture group می‌شویم که تعدادی حرف به جز حروف کوچک میتواند تکرار شود و سپس یک حرف کوچک و دوباره تعدادی حرف به جز حروف کوچک داریم که کل این الگو باید بیشتر از یک بار رخ دهد و در انتها همانند قسمت اول عمل می‌کنیم با این تفاوت که در قسمت دوم الگو رقم های زوج آورده میشود که در انتهای رشته است.

سوال ۳)

برای محاسبه احتمالات از رابطه زیر استفاده میکنیم (به همراه laplace smoothing):

$$P_{Add-1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)+1}{c(w_{i-1})+V}$$

محاسبه سایز vocabulary:

Ali – Daei- footballer – Hadi – Saei – taekwondo – athlete – Perspolis – football – club - <s> - </s>

12 لغت در کل وجود دارد. پس V = 12 اما در لاپلاس اسموتینگ باید 1 لغت کم شود یعنی <s> پس V=11 میشود.

حال باید از روی مقادیر موجود برای count ها از روی مجموعه آموزش به محاسبه احتمالات مجموعه تست بپردازیم:

1. P (<s> football club </s>) = P (football | <s>) * P (club | football) * P (</s> | club) = ?

$$P\ (\ football \mid\ < s >)\ \ = \frac{c(< s >, football\ ) + 1}{c(< s >) + 11} = \frac{1}{15}$$

به همین ترتیب برای بقیه حالات نیز محاسبه میکنیم.

P (club | football) = (1+1)/(1+11) = 2/12
P (</s> | club) = (1+1)/(1+11) = 2/12

P (<s> football club </s>) = (1/15) * (2/12) * (2/12) = 0.00185185185

---

2. P (<s> Hadi Saei Perspolis </s>) = P (Hadi | <s>) * P (Saei | Hadi) * P (Perspolis | Saei) * P (</s> | Perspolis) = ?

P (Hadi | <s>) = (1+1)/(4+11) = 2/15
P (Saei | Hadi) = (1+1)/(1+11) = 2/12
P (Perspolis | Saei) = (0+1)/(1+11) = 1/12
P (</s> | Perspolis) = (1+1)/(2+11) = 2/13

P (<s> Hadi Saei Perspolis </s>) = (2/15) * (2/12) *(1/12) *(2/13) = 0.00028490028

3. P (<s> Saei Daei taekwondo </s>) = P (Saei | <s>) * P (Daei | Saei) * P (taekwondo | Daei) * P (</s> | taekwondo) = ?

P (Saei | <s>) = (0+1)/(4+11) = 1/15
P (Daei | Saei) = (0+1)/(1+11) = 1/12
P (taekwondo | Daei) = (0+1)/(2+11) = 1/13
P (</s> | taekwondo) = (0+1)/(1+11) = 1/12

P (<s> Saei Daei taekwondo </s>) = (1/15) * (1/12) *(1/13) *(1/12) = 0.00003561253

---

4. P (<s> football Perspolis club </s>) = P (football | <s>) * P (Perspolis | football) * P (club | Perspolis) * P (</s> | club) = ?

P (football | <s>) = (0+1)/(4+11) = 1/15
P (Perspolis | football) = (0+1)/(1+11) = 1/12
P (club | Perspolis) = (0+1)/(2+11) = 1/13
P (</s> | club) = (1+1)/(1+11) = 2/12

P (<s> football Perspolis club </s>) = (1/15) * (1/12) *(1/13) *(2/12) = 0.00007122507