تولید دادگان خصمانه برای مسئله‌ی QA و بررسی میزان مقاومت مدل‌ها در برابر حملات

حمیدرضا امیرزاده، محمد حسین سامتی، آرش ماری اوریاد، محمد جلال نعمت بخش

مسئله‌ی Question-Answering

- در مسائل پرسش و پاسخ ابتدا یک متن حاوی تعدادی جمله در اختیار ما قرار می‌گیرد که به آن context گفته می‌شود. این متن می‌تواند حاوی اطلاعاتی در مورد یک فرد، واقعه، مکان، یا ... باشد.
- سپس با توجه به متن داده شده، یک سوال (question) مطرح می‌شود به گونه‌ای که پاسخ سوال در متن موجود است.
- مدل یادگیری ماشین سعی می‌کند با توجه به متن و صورت سوال، پاسخ مناسب را تولید نماید.

دیتاست مورد استفاده

- در این پروژه از دیتاست small-persian-QA استفاده شده است که خود از روی یک دیتاست بزرگ‌تر به نام persian-QA تولید شده است.
- دیتاست small-persian-QA در مجموع حاوی ۱۳۹۱ جفت متن و سوال می‌باشد که حدود ۸۰ درصد آن برای فرایند training و مابقی برای فرایند validation استفاده شده است.
- دیتاست small-persian-QA از طریق این لینک قابل دستیابی می‌باشد.

مدل‌های مورد استفاده

- مدل BERT multilingual uncased
  - این مدل بر روی داده‌ی Wikipedia با ۱۰۲ زبان مختلف با توجه به دو تسک Masked language modeling و Next sentence prediction آموزش داده شده است.
  - توضیحات بیشتر در رابطه با این مدل از طریق این لینک قابل مشاهده می‌باشد.
  - این مدل روی دادگان small-persian-QA به تعداد هشت epoch فرایند fine-tuning را طی کرده است.

- مدل ParsBERT v2.0
  - معماری این مدل تک زبانه مشابه معماری مدل BERT می‌باشد و با استفاده از حدود ۷۳ میلیون جمله‌ی فارسی آموزش داده شده است.
  - توضیحات بیشتر در رابطه با این مدل از طریق این لینک قابل مشاهده می‌باشد.
  - این مدل روی دادگان small-persian-QA به تعداد هشت epoch فرایند fine-tuning را طی کرده است.

روش‌های تولید داده‌ی خصمانه

- منظور از تولید داده‌ی خصمانه برای مسئله‌ی QA، در واقع تغییر متن اصلی (context) به گونه‌ای می‌باشد که متن همچنان از نظر انسان معتبر (valid) بوده و بخش مربوط به پاسخ نیز تغییر نکرده باشد. اما در عین حال مدل یادگیری ماشین در تولید پاسخ صحیح دچار اشتباه شده یا میزان اطمینان مدل روی پاسخ صحیح کاهش یابد.

- در این پروژه از چهار روش تولید داده‌ی خصمانه استفاده شده است:
  - روش AddSent
  - روش AddAny
  - روش Back Translation
  - روش Invisible Char

روش AddSent

- **ایده اصلی:** اضافه کردن یک جمله مهندسی شده به آخر متن (context)
- **شروط:** باید جمله از لحاظ ظاهری و قواعد زبانی صحیح باشد و با پاسخ سوال در تناقض نباشد.
- **مقاله مربوطه:** Adversarial Examples for Evaluating Reading Comprehension Systems - 2017

مثال روش AddSent

متن : سیاوش قمیشی کوچکترین فرزند خانوادهٔ خویش است و دو برادر به نامهای سیروس و سیامک و یک خواهر به نام سیمین دارد. او در اهواز به دنیا آمد و در سن ۹ ماهگی به همراه خانواده به تهران مهاجرت کرد. سیاوش ۶ سال داشت که یکی از برادرانش آکاردئون می آموخت و مادرش پیانو را زیر نظر یک معلم روس به نام اُلگا می آموخت. او از شنیدن صدای نواختن آنها لذت بسیاری می برد و در حالی که کودک بود و پایش به پدالهای پیانو نمیرسید به پشت ساز مادرش (پیانو) مینشست و با پدالهای آن بازی میکرد. همین بازیها به همراه صداهایی که از این ساز خارج میشد به او احساس خوبی میداد و رفته رفته سیاوش به موسیقی علاقه ای ویژه پیدا کرد. سیاوش ۱۱ سال داشت که هر روز از مسیر مدرسه تا خانه از جلوی یک مغازهٔ کفش فروشی رد میشد که در ویترینش یک گیتار برای تزئین دکور آویزان کرده بود. آن گیتار نگاه وی را به خود تا جایی جلب کرد که یک روز او با پدرش راجع به خرید آن گیتار صحبت کرد اما پدر با این موضوع مخالفت کرد. چندی بعد به مناسبت روز تولد سیاوش، پدر از سر تصمیم خود بازگشت و همان گیتار را برای سیاوش خرید. قطب الدین ایبک دو برادر به نام های سیروس و سیامک و یک پسر به نام سیمین دارد

سوال : سیاوش قمیشی چندتا ابجی و داداش دارد؟

پاسخ درست و تمیز : دو برادر به نامهای سیروس و سیامک و یک خواهر به نام سیمین دارد
پاسخ مدل به اشتباه : یک خواهر به اسم سیمین دارد

نحوه‌ی عملکرد روش AddSent

1. ورودی: سوال + پاسخ
2. تشخیص موجودیت های نامدار در سوال با استفاده از مدل "HooshvareLab/bert-base-parsbert-ner-uncased"
3. انتخاب یکی از موجودیت های نامدار (organization / location / person) در سوال به صورت تصادفی و جایگزینی آن با کلمه ای دیگر از **همان نوع** به صورت تصادفی
4. تشخیص ادات سخن در سوال با استفاده از مدل "wietsedv/xlm-roberta-base-ft-udpos28-fa"
5. انتخاب یکی از اسم ها (Noun) یا صفت ها (ADJ) موجود در سوال به صورت تصادفی و جایگزینی آن با توکن [MASK]
6. استفاده از مدل زبانی "Hamid-reza/bert-base-parsbert-uncased-finetuned-conditioned-khorshid" جهت جایگزینی توکن mask شده با کلمه ای مناسب و انتخاب یکی از پنج بهترین کلمه پیشنهادی مدل
7. تولید یک پاسخ جعلی به کمک روشی مشابه قسمت ۶
8. دادن سوال و پاسخ جعلی به مدل "Farnazgh/QA2D" که یک مدل تبدیل کننده جمله پرسشی و جواب متناظر به یک جمله خبری است. (ترجمه به انگلیسی و برگرداندن به فارسی)
9. اضافه کردن خروجی مرحله قبل به انتهای متن (Context)

روش AddAny

- **ایده اصلی :** اضافه کردن یک رشته از کلمه به طول d به آخر متن
- **شروط :** بر خلاف روش addsent لزومی بر رعایت قواعد دستور زبان نیست.
- **مثال:**

سیاوش قمیشی کوچک‌ترین فرزند خانوادهٔ خویش است و دو برادر به نام‌های سیروس و سیامک و یک خواهر به نام سیمین دارد. او در اهواز به دنیا آمد و در سن ۹ ماهگی به همراه خانواده به تهران مهاجرت کرد. سیاوش ۶ سال داشت که یکی از برادرانش آکاردئون می‌آموخت و مادرش پیانو را زیر نظر یک معلم روس به نام اُلگا می‌آموخت. او از شنیدن صدای نواختن آنها لذت بسیاری می‌برد و در حالی که کودک بود و پایش به پدال‌های پیانو نمی‌رسید به پشت ساز مادرش (پیانو) می‌نشست و با پدال‌های آن بازی می‌کرد. همین بازی‌ها به همراه صداهایی که از این ساز خارج میشد به او احساس خوبی می‌داد و رفته رفته سیاوش به موسیقی علاقه‌ای ویژه پیدا کرد. سیاوش ۱۱ سال داشت که هر روز از مسیر مدرسه تا خانه از جلوی یک مغازهٔ کفش فروشی رد می‌شد که در ویترین‌اش یک گیتار برای تزئین دکور آویزان کرده بود. آن گیتار نگاه وی را به خود تا جایی جلب کرد که یک روز او با پدرش راجع به خرید آن گیتار صحبت کرد اما پدر با این موضوع مخالفت کرد. چندی بعد به مناسبت روز تولد سیاوش، پدر از سر تصمیم خود بازگشت و همان گیتار را برای سیاوش خرید. آلمان فرانسه نوامبر درجه جماهیر میخانه استفاده شیکاگو مدرن شهر.

سوال : سیاوش قمیشی چندتا ابجی و داداش دارد؟
پاسخ درست و تمیز : دو برادر به نامهای سیروس و سیامک و یک خواهر به نام سیمین دارد
پاسخ مدل به اشتباه : دارد.

نحوه عملکرد

1. ایجاد لیست کلمات رایج
2. اجرای الگوریتم ژنتیک
- انتخاب جمعیت اولیه از لیست کلمات رایج
- اضافه کردن رشته کلمات به context و ارزیابی در مدل QA
- انتخاب رشته ها بر اساس f1_score و تولید نسل بعدی (crossover)
- تغییر یکی از کلمات به صورت تصادفی (mutation)
3. اضافه کردن بهترین رشته به آخر متن (context)

روش Back Translation

- **ایده اصلی :** ترجمه سوال به زبان دیگر و دوباره ترجمه به فارسی
- **شروط :** صورت سوال باید از نظر معنا و دستور زبان صحیح باشد.
- **مثال :**

متن : سیاوش قمیشی کوچکترین فرزند خانوادهٔ خویش است و دو برادر به نام های سیروس و سیامک و یک خواهر به نام سیمین دارد. او در اهواز به دنیا آمد و در سن ۹ ماهگی به همراه خانواده به تهران مهاجرت کرد. سیاوش ۶ سال داشت که یکی از برادرانش آکاردئون می آموخت و مادرش پیانو را زیر نظر یک معلم روس به نام الگا می آموخت. او از شنیدن صدای نواختن آنها لذت بسیاری می برد و در حالی که کودک بود و پایش به پدال های پیانو نمی رسید به پشت ساز مادرش (پیانو) می نشست و با پدال های آن بازی می کرد. همین بازیها به همراه صداهایی که از این ساز خارج میشد به او احساس خوبی می داد و رفته رفته سیاوش به موسیقی علاقه آویژه پیدا کرد. سیاوش ۱۱ سال داشت که هر روز از مسیر مدرسه تا خانه از جلوی یک مغازهٔ کفش فروشی رد میشد که در ویترینش یک گیتار برای تزئین دکور آویزان کرده بود. آن گیتار نگاه وی را به خود تا جایی جلب کرد که یک روز او با پدرش راجع به خرید آن گیتار صحبت کرد اما پدر با این موضوع مخالفت کرد. چندی بعد به مناسبت روز تولد سیاوش، پدر از سر تصمیم خود بازگشت و همان گیتار را برای سیاوش خرید.

سوال : سیاوش قمیشی چند برادر و خواهر دارد؟
پاسخ :دو برادر به نام های سیروس و سیامک و یک خواهر به نام سیمین دارد.
پاسخ مدل به اشتباه : سیمین دارد.

روش Invisible Char

- **ایده اصلی :** جایگزینی white space با یکی از حروف نامرئی (u200e)
- **مثال :**

سیاوش قمیشی کوچکترین فرزند خانوادهٔ خویش است و دو برادر به نامهای سیروس و سیامک و یک خواهر به نام سیمین دارد. او در اهواز به دنیا آمد و در سن ۹ ماهگی به همراه خانواده به تهران مهاجرت کرد. سیاوش ۶ سال داشت که یکی از برادرانش آکاردئون می آموخت و مادرش پیانو را زیر نظر یک معلم روس به نام اُلگا می آموخت. او از شنیدن صدای نواختن آنها لذت بسیاری می برد و در حالی که کودک بود و پایش به پدالهای پیانو نمیرسید به پشت ساز مادرش (پیانو) مینشست و با پدالهای آن بازی میکرد. همین بازیها به همراه صداهایی که از این ساز خارج میشد به او احساس خوبی میداد و رفته رفته سیاوش به موسیقی علاقه ای ویژه پیدا کرد. سیاوش ۱۱ سال داشت که هر روز از مسیر مدرسه تا خانه از جلوی یک مغازهٔ کفش فروشی رد میشد که در ویترینش یک گیتار برای تزئین دکور آویزان کرده بود. آن گیتار نگاه وی را به خود تا جایی جلب کرد که یک روز او با پدرش راجع به خرید آن گیتار صحبت کرد اما پدر با این موضوع مخالفت کرد. چندی بعد به مناسبت روز تولد سیاوش، پدر از سر تصمیم خود بازگشت و همان گیتار را برای سیاوش خرید.

سوال : سیاوش قمیشی چندتا ابجی و داداش دارد؟
پاسخ درست و تمیز : دو برادر به نام‌های سیروس و سیامک و یک خواهر به نام سیمین دارد.
پاسخ اشتباه مدل : دو

نتایج تولید داده‌های خصمانه

- معیار ارزیابی F1-Score می‌باشد.

دیتاست	مدل mBERT	مدل ParsBERT
small-persian-QA	57.9	56.1
AddSent	50.6	53.7
AddAny	40.5	50.3
Back Translation - EN	56.4	56.6
Back Translation - FR	55.4	51.1
Back Translation - HY	53.7	55.7
Invisible Char	22.3	19.3

یادگیری خصمانه

- در فرایند یادگیری خصمانه (Adversarial Training) سعی می‌شود با به کار بردن روش‌هایی در هنگام فرایند آموزش (Training) مدل، عملکرد نهایی مدل روی نمونه‌های خصمانه بهتر شود.
- معمولا این بهبود عملکرد روی نمونه‌های خصمانه توامان با افت عملکرد مدل روی داده‌های تمیز می‌باشد.
- یکی از معروف‌ترین روش‌های یادگیری خصمانه، رویکرد افزایش دادگان (Data Augmentation) می‌باشد. در این روش سعی می‌شود مدل‌های یادگیری ماشین روی داده‌های خصمانه‌ی از پیش تولید شده fine-tune شوند.
- دو مدل به کار گرفته شده در این پروژه نیز بر روی چهار دیتاست خصمانه‌ی تولید شده به تعداد هشت epoch، آموزش داده شدند.

نتایج یادگیری خصمانه

- معیار ارزیابی F1-Score می‌باشد.
- عدد سمت راست عملکرد مدل روی دادگان small-persian-QA و عدد سمت چپ عملکرد مدل روی دادگان خصمانه‌ی روش ذکر شده در همان سطر می‌باشد.

روش	مدل mBERT	مدل ParsBERT
AddSent	55.3 / 58.3	53.2 / 53.7
AddAny	65.4 / 57.6	57.8 / 49.3
Back Translation - HY	57.2 / 56.4	52.1 / 52.4
Invisible Char	7.3 / 17.9	4.1 / 13.5

کارهای آینده

- در نظر گرفتن معیارهای بیشتری برای معتبر بودن متن خصمانه‌ی تولید شده از نظر قواعد گرامری و نکات ظاهری
- در نظر گرفتن معیارهایی برای محتمل‌تر بودن جملات خصمانه‌ی تولید شده از منظر مدل‌های زبانی n-gram و ... مانند perplexity
- طراحی و پیاده‌سازی سایر روش‌های تولید داده‌ی خصمانه مانند استفاده از گرادیان خروجی مدل نسبت به کلمات یا استفاده از رویکردهایی نظیر یادگیری تقویتی (reinforcement learning) جهت یاد گرفتن یک سیاست (policy) برای تولید دادگان خصمانه
- در نظر گرفتن معیارهای ارزیابی بیشتر برای بررسی عملکرد رویکردهای تولید داده‌ی خصمانه مانند BLEU

متشکر

سوال؟