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

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

چکیده

اخیرا سیستم‌های پرسش و پاسخ پیشرفت‌های شگرفی داشته‌اند. اما آیا این سیستم‌ها واقعا معنای جملات را درک می‌کنند؟ راهکار پیشنهادی تولید دیتاست‌های خصمانه و ارزیابی این مدل‌ها روی دیتاست‌های تولید شده است. در پروژه سعی شده‌است که با استفاده از چندین روش متنوع، دیتاست‌هایی خصمانه تولید شود. نشان می‌دهیم موثر ترین حمله به کمک روش *invisible char* بوده است که دقت مدل‌های پایه را از ۵۷.۲ به ۲۲.۳ برای m_bert و از ۵۶.۱ به ۱۹.۳ برای مدل *pars_bert* تقلیل داده‌است.

مقدمه

در این تلاش علمی، به بررسی مدل‌ها و دیتاست‌های فارسی در بحث پرسش و پاسخ *QA* و مقاوم‌پذیری آن‌ها نسبت به حملات و دیتاست‌های خصمانه پرداخته شده‌است. در گام اول پروژه مدل‌های موجود در *QA* در زبان فارسی بررسی شدند که از جمله نمونه‌های توسعه داده شده، می‌توان mBERT, ParsBERT, را نام برد. در ادامه از تلاش پیشین آزمایشگاه *language ml* دانشگاه صنعتی شریف برای مدل‌های پرسش و پاسخ استفاده می‌شود. با مطالعه مقالات [1] و [2] روش‌های *AddSent, AddAny, Back translation* برای تولید دادگان خصمانه و ارزیابی مدل‌ها انتخاب شدند. سپس مدل‌های توسعه داده شده را با دیتاست‌های تولید شده ارزیابی کرده و در نهایت اقدام به مقاوم‌سازی این مدل‌ها با دادگان خصمانه می‌کنیم.

روش‌ها

در این بخش به بررسی روش‌های تولید دیتاست خصمانه که شامل *AddAny, AddSent, Back translation* است، پرداخته می‌شود. لازم به ذکر است برای صحت و قابل اتکا بودن نتایج از یک دیتاست یکسان اولیه برای هر سه روش بهره برده شده‌است که از این لینک قابل دسترسی است. برای تولید این دیتاست از *PersianQA* استفاده شده به این صورت که به صورت تصادفی با seed برابر با ۴۲ تعداد ۰.۲ سایز دیتاست از *Train* و *Validation* انتخاب شده است.

روش *AddAny* :

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 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.

مثالی از داده تولید شده توسط روش. رنگ آبی تغییرات روش را نشان می‌دهد.

در این روش هدف تولید یک رشته از کلمات به طول ۱۰ است که شامل کلمات رایج ویکی‌پدیا فارسی و همچنین کلمات موجود در سوال و همین‌طور اضافه کردن آن به پاراگراف آن متن آن سوال جهت به حداقل رساندن *f1 score* مدل پایه *QA* است.

روش به شرح جزئیات در ادامه ذکر شده است :

ابتدا برای تشکیل کلمات رایج از دیتاست ParSQuAD با توجه بسامد تکرار کلمه در پاسخ های این دیتاست استفاده شده است.

برای تولید کلمات جهت اضافه کردن به پاراگراف از الگوریتم جست‌وجوی محلی ژنتیک بهره برده شده‌است .ابتدا به صورت تصادفی یک جمعیت n تایی از ۱۰ کلمه از کلمات رایج انتخاب شده و در مدل پایه *mBERT* ارزیابی می‌شوند و با توجه به f1 آن‌ها یک احتمال به آن‌ها تخصیص داده می‌شود(هر چه f1 کمتر، احتمال بیشتر). برای تولید نسل های بعدی از تابع *crossover* و جهت تصادفی بودن از *mutation* با نرخ مشخص و با توجه به دیکشنری تولید شده از کلمات مشابه آن درایه با استفاده از *fasttext* استفاده می‌شود . این روند با یک لیمیت مشخص برای جلوگیری از قرارگیری در حلقه بی نهایت تا رسیدن به f1 برابر صفر ادامه می‌یابد. دیتاست تولید شده در گیت‌هاب پروژه به همراه کد تولید آن؟ قابل مشاهده است.

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.

مثالی از داده تولید شده توسط روش. رنگ آبی تغییرات روش را نشان می‌دهد.

این روش مبتنی بر افزودن یک جمله به انتهای متن است تا باعث اشتباه و خطای مدل شود که این جمله بر اساس سوال و پاسخ و به صورت زیر تولید می‌شود. در ابتدا *name entity* های سوال تشخیص داده می‌شود و آن‌ها با یک کلمه دیگر از همان نوع جایگزین می‌شوند که این تغییر به صورت رندوم انجام می‌شود. از آنجا که ممکن است در صورت سوال هیچ *name entity* نباشد، به کمک *POS Tag* ها سعی در تغییر *noun* و *adjective* های صورت سوال خواهیم داشت. در ادامه نیاز داریم تا یک پاسخ اشتباه(fake answer) نیز تولید کنیم که به همان فرم تغییر صورت سوال است با این تفاوت که در *POS Tag* از *Adjective* و *Pronoun* برای تغییر ظاهر پاسخ استفاده می‌شود. در نهایت ترکیب این صورت و پاسخ تغییر یافته نیاز است تا با هم به صورت یک جمله خبری تبدیل شوند که مدلی کارا برای زبان فارسی یافت نشد که برای این کار از یک ترجمه میانی استفاده می‌شود و پس از استفاده از مدل موجود در انگلیسی مجدد به فارسی تغییر می‌یابد.

روش *Back Translation* :

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 اُلگا 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.

مثالی از داده تولید شده توسط روش. رنگ آبی تغییرات روش را نشان می‌دهد.

ایده اصلی این روش مبتنی بر این است که ترجمه یک جمله به زبانی دیگر و بازگردانی آن، جمله را دچار تغییراتی می‌کند که غالبا با همان معنا و مفهوم خواهد بود اما ساختار و کلمات و نحوه بیان آن تغییر کرده است. به عبارتی جمله مجدد بازنویسی خواهد شد. پس تعدادی زبان میانی انتخاب شدند که صورت سوال‌ها را به این زبان ترجمه کرده و

سپس مجددا آن‌ها را به فارسی بازگردانی کنیم. در این روش از API ترجمه گوگل استفاده می‌شود که نمونه دیتاست‌های تولید شده در ۱ و ۲ و ۳ قابل مشاهده است. پیشبینی می‌شود هر چه مدل ترجمه زبان میانی از منابع کمتری برای آموزش استفاده کرده باشد، دیتاست تولید شده موفقیت بیشتری در کاهش دقت مدل QA و در نتیجه حمله خواهد داشت.

روش *Invisible Char* :

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.

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

در این روش از کاراکتر هایی که درست شبیه به یک فاصله هستند استفاده می‌شود، به این صورت که تمامی فاصله‌های موجود در متن با این کاراکتر ویژه جایگزین می‌شوند. در نتیجه متن نهایی از نظر ظاهری هیچ تغییری نخواهد داشت و صرفا باعث کاهش ۳۷ درصدی دقت مدل *m_bert* می شود. برای اطلاعات بیشتر در مورد این دسته از کاراکترها به این لینک مراجعه شود. دیتاست نهایی نیز از طریق این لینک قابل دریافت است.

نتایج

داده‌های تولید شده همگی بر روی مدل‌ها تست شده و نتایج آن‌ها در جدول زیر آورده شده است. تمامی این نتایج جدول ۱ قابل مشاهده و بازتولید از طریق گیت‌هاب این گزارش است.

pars_bert	m_bert	
56.1	57.9	دیتاست اولیه
50.3	40.5	AddAny
53.7	50.6	AddSent
56.6	56.4	BT_En
51.1	55.4	BT_Fr
55.7	53.7	BT_Hy
19.3	**22.3**	IC

جدول ۱ :*f1_score* بدست آمده بر روی هر یک از دیتاست های خصمانه تولید شده.

بر اساس داده‌های جدول ۱، بیشترین کاهش دقت برای *AddAny* و *Invisible Char* است که به ترتیب ۱۷ درصد و ۲۴ درصد کاهش را به دنبال داشته‌اند.

یادگیری خصمانه :

پس از بررسی کارایی دیتاست‌های خصمانه تولید شده، به *fine tune* کردن مدل های *QA* با کمک این دیتاست‌های تولید شده می‌پردازیم. نتایج در جدول ۲ قابل مشاهده است.

pars_bert	m_bert	
57.8 / 49.3	65.4 / 57.6	AddAny
53.2 / 53.7	55.3 / 58.3	AddSent
-	-	BT_En
-	-	BT_Fr
52.1 / 52.4	57.2 / 56.4	BT_Hy
4.1 / 13.5	7.3 / 17.9	IC

جدول ۲:عدد سمت راست هر خانه f1_score مدل روی دیتاست اولیه و عدد سمت چت f1_score مدل روی همان دیتای خصمانه همان ردیف است.

نتیجه‌گیری و کارهای آینده

برخلاف پیشرفت‌های زیاد در حوزه NLP و یادگیری عمیق، کمبود کتابخانه‌ها و مدل‌های مربوط به زبان فارسی بسیار احساس می‌شود. از جمله این نیازها می‌توان به *wordnet* زبان فارسی برای روش‌های *AddSent* و *AddAny* اشاره کرد. همچنین پایین بودن دقت بهترین مدل های موجود QA برای زبان فارسی نیز، نشان از ظرفیت‌های فراوان این حوزه برای علاقه‌مندان دارد. مدل های استفاده شده در این بخش نیز نسبت به دیتای خصمانه مقاوم نبوده و در اکثر دیتاست‌های تولیدی کاهش دقت مدل را شاهد هستیم.

مراجع :

[1] : “Adversarial Examples for Evaluating Reading Comprehension Systems”, Robin Jia, Percy Liang, *Submitted on 23 Jul 2017 EMNLP 2017*

[2] : https://aclanthology.org/P19-1610