دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش‌مصنوعی

سیستم خودکار پرسش و پاسخ بر اساس مدل‌های ترجمه آماری

نگارنده: مجید امیری

استاد راهنما

دکتر مرتضی زاهدی

دی ۹۸

شماره: ۶۶۷ ر/ف/ز
تاریخ: ۱۵ / ۱۱ / ۹۸

باسمه‌تعالی

دانشگاه صنعتی شاهرود
مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای محمد امیری .....
با شماره دانشجویی ۹۵۰۲۰۱۴ ..... رشته مهندسی کامپیوتر ..... گرایش هوش مصنوعی .....
تحت عنوان سیستم خودکار پرسش و پاسخ براساس مدل های ترجمه آماری .....
که در تاریخ ۹۸/۱۰/۲۹ ..... با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☐ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☑
ج) درجه خوب: نمره ۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐
ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐
نوع تحقیق: **نظری** ☐ **عملی** ☑

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	نرض اصری	استادیار	
۲- استادراهنمای دوم	———		
۳- استاد مشاور	———		
۴- نماینده تحصیلات تکمیلی	محسن فرهادی	مربی	
۵- استاد ممتحن اول	محسن رضوانی	استادیار	
۶ـ استاد ممتحن دوم	منصور فاتح	استادیار	

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:

تقدیم به:

پایان نامه‌ی حاضر را به پدر و مادرم تقدیم می‌کنم، که بدون مشورت‌های بی‌دریغ، آگاهانه وحمایت‌های آنها امکان پیشرفت من تا این سطح علمی وجود نداشت. همیشه خود را مرهون شکیبایی و محبت بی‌قیدوشرط آنها می‌دانم.

سپاس‌گزاری...

از استاد گرامیم، جناب آقای دکتر مرتضی زاهدی بسیار سپاسگزارم به خاطر صبر بی‌نظیر و راهنمایی‌های علمی و موثرشان. کسی که رفتار و گفتار گرم وی همیشه راه پیشرفت را سهل می‌نمایاند، که نشان از وسعت نظر او است. به امید روزی که به شخصیتی علمی همانند او تبدیل شوم.

مجید امیری
دی ۹۸

تعهد نامه

اینجانب **مجید امیری** دانشجوی کارشناسی ارشد رشته **مهندسی کامپیوتر** دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان **سیستم خودکار پرسش و پاسخ بر اساس مدل‌های ترجمه آماری** ، تحت راهنمایی **مرتضی زاهدی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود ’’ یا “ Shahrood University of Technology ” به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

مجید امیری
دی ۹۸

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

ما با رشد روزافزون اطلاعات روبرو هستیم. اطلاعاتی که به طور دوره‌ای بروز می‌شوند و حجم آنها با سرعت نمایی افزایش می‌یابد. دسترسی سریع و دقیق به این اطلاعات نیاز مبرم ما در عصر کنونی است. از این رو ما همه روزه با نیاز به سیستم‌های بازیابی اطلاعات دقیق‌تر و سریع‌تری مواجهه هستیم. سیستم‌هایی که تعامل مابین کامپیوتر و کاربر را به اشکال مختلف بهبود می‌دهند. سیستم‌های پرسش و پاسخ یک نمونه از این سیستم‌ها هستند که توسط کانال زبان طبیعی، ارتباط و تعامل سیستم را برای پاسخ به نیاز های اطلاعاتی کاربر راحت‌تر کرده و سرعت می‌بخشند.

سیستم‌های آماری پرسش و پاسخ متناسب با کیفیت پایگاه دادگان (جامعیت و عاری بودن از هر گونه نویز و داده‌ی پرتی) از دقت بیشتری نسبت به دیگر روش‌های توسعه‌ی سیستم‌های پرسش‌وپاسخ برخوردار هستند. ما با ایجاد تعامل توسط در نظر گرفتن سوال قبلی کاربر و فاصله‌ی (تشابه معنایی)آن از سوال کنونی در مدل آماری Word2vec توانستیم بخش عمده‌ی خطای ناشی از ابهام پرسش کاربر را حذف کنیم. این بهبود دقت ناشی از ساختار شبکه‌ی عصبی عمیق پیچشی است به این صورت که سیستم از یک مدل word2vec به عنوان تبدیل‌کننده‌ی متن به اعداد در ابتدای شبکه استفاده می‌کند؛ از خاصیت این مدل آماری می‌توان به تشخیص فاصله‌ی معنایی دو جمله اشاره کرد.

عملکرد سیستم پرسش‌وپاسخ پیاده‌سازی شده توسط روش شبکه‌ی عصبی عمیق پیچشی با توجه به حذف مقدار زیادی از خطای سیستم ناشی از ابهام در پرسش، توانسته به دقت ۸۲.۴ % برسد. که به طور متوسط افزایشی ۸% در بهبود معیارهای ارزیابی سیستم‌های طبقه‌بندی‌کننده مانند: دقت، صحت و میانگین وزن‌دار دقت و صحت سیستم را در مقایسه با سیستم پرسش‌وپاسخ فاقد تعامل بر روی پایگاه‌دادگان استاندارد نشان می‌دهد.

کلمات کلیدی: سیستم‌های پرسش‌و پاسخ، شبکه‌های عصبی عمیق پیچشی، تعامل در سیستم‌های پرسش‌وپاسخ

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمه

۱.۱ پیشگفتار

در نتیجه کاهش هزینه و افزایش قابلیت ذخیره‌سازی داده در عصر کنونی، ما با انبوه‌ی از اسناد ذخیره شده مواجه هستیم که حاوی اطلاعات ارزشمندی هستند. و ما برای رفع نیازهای روزمره‌ی خود نیاز به دسترسی به این اطلاعات داریم. از جمله اسنادی که ما روزانه به اطلاعاتشان نیازمندیم می‌توان به : بخشنامه‌های آموزش دانشگاه، سیاست‌های حفظ حریم شخصی [۱] در سایت‌هایی که ما با آنها به طور روزانه در تعامل هستیم و اطلاعات شخصی خود را در اختیار آنها قرار می‌دهیم، بخشنامه‌های شهرداری، بخشنامه‌های بانک‌ها و به طور کلی تمام قوانینی که ما با آنها در طول روز سروکار داریم و به طور دوره‌ای تغییر می‌کنند، اشاره کرد [۲].

اما با توجه به تعداد و حجم این اسناد، آگاهی از این اطلاعات نیازمند صرف زمان بسیار و دانستن روش‌های جستجوی کارا است که اکثر کاربران از این دو بی بهره‌اند. برای رفع این نیازها، انواع مختلفی از سیستم‌های بازیابی اطلاعات [۲] با روش‌های جستجوی متفاوت پیاده‌سازی شده که از جمله آنها به موتورهای جستجو، سیستم‌های پرسش‌وپاسخ و... نام برد [۳]. در موتورهای جستجو، کاربر پرسش خود را توسط ترکیبی از کلمات کلیدی مطرح

[1] Privacy policies

[2] Information retrieval

می‌کند و موتور جستجو لیستی از محتمل‌ترین اسناد را به عنوان پاسخ ارائه می‌کند. کاربر باید تک تک اسناد را جستجو کرده تا به پاسخ خود برسد. در صورتی که پاسخ کاربر در لیست اسناد ارائه شده نباشد، کاربر باید ترکیب کلمات کلیدی خود را تغییر دهد تا سیستم محتمل‌ترین اسناد را با توجه به ترکیب جدید بازگرداند. با توجه به روش فوق دو عیب مشهود در این نوع از سیستم‌های بازیابی اطلاعات را، دقت کم و زمان‌بر بودن جستجو می‌توان دانست [۴، ۵].

برای رفع این عیوب، سیستم‌های پرسش‌وپاسخ مطرح شده‌اند که کاربر می‌تواند توسط زبان طبیعی پرسش خود را مطرح کرده و پاسخ خود را به صورت چند پاراگراف، چند جمله و یا چند کلمه توسط زبان طبیعی دریافت کند. با توجه به اینکه این گونه از سیستم‌ها برای جستجو در پایگاه دادگان [۳] کوچکتری نسبت به موتورهای جستجو که باید در سطح وب جستجو کند، طراحی شده‌اند و در بعضی از مواقع مختص یک موضوع خاص هستند از دقت بالاتری برخوردار هستند.

در اکثر سیستم‌های پرسش‌وپاسخ کاربر یک سوال مطرح کرده و جواب را، سیستم متناسب با آن برمی‌گرداند. این گونه از سیستم‌های پرسش‌وپاسخ تاریخچه‌ای از سوالات مطرح شده توسط کاربر را نگه نمی‌دارند و جواب سوال را بر اساس تجزیه‌ی سوال و پایگاه دادگان در دسترس، به کاربر برمی‌گردانند. به عبارت دیگر کاربر با آنها در تعامل مداوم نیست. این نوع از سیستم‌ها در درک دقیق سوال مشکل دارند و در برابر سوالات مبهم عملکرد ضعیفی از خود نشان می‌دهند برای رفع این مشکل بر تعامل هر چه بیشتر کاربر با سیستم تاکید می‌شود و سیستم می‌تواند با توجه به ترتیب سوالات، به درک بهتری از سوال و منظور کاربر دست یابد.

در پیاده‌سازی سیستم‌های پرسش‌وپاسخ دو رویکرد کلی و متفاوت در نظر گرفته می‌شود: سیستم‌هایی که بر مبنای قواعد زبانی توسعه داده می‌شود و سیستم‌هایی که بر اساس الگوهای پُر تکرار موجود در پایگاه دادگان توسعه داده می‌شود. از عیوب بزرگ سیستم‌هایی که بر مبنای قواعد نحوی زبان توسعه داده می‌شوند می‌توان به وابستگی کامل سیستم به قواعد نحوی زبانی که بر اساس آن پیاده سازی شده‌اند، اشاره کرد. پیاده‌سازی و توسعه‌ای این گونه از سیستم‌ها با توجه به اینکه هر زبان شامل استثناهای بی شماری است بسیار زمانبر و مشکل است. ولی نه تنها روش توسعه‌ی سیستم‌های آماری از زبان مستقل است بلکه به زمان کمتری برای طراحی نیاز دارد.

۲.۱ تعریف مسئله

در روش پیشنهادی؛ ما از دو طبقه‌بندی‌کننده‌ی یادگیری عمیق [۴] به صورت پشت سر هم استفاده کردیم که محتمل‌ترین جواب‌ها را بدست آوریم. در این سیستم دو تا محتمل‌ترین جوابها نگه‌داری می‌شود و در صورتی که کاربر سوال خود را به منظور گرفتن جوابی بهتر با

[3] Dataset/Corpus

[4] Deep convolutional neural networks

ترکیبی متفاوت مجدداً بپرسد، فاصله ی سوال اول با سوال دوم در مدل زبانی محاسبه شده و در صورتی که این فاصله از یک مقدار مشخصی کمتر باشد دومین محتمل‌ترین جواب را به عنوان جواب کاربر برمی‌گرداند. ما در یادگیری این دو طبقه‌بندی یادگیری عمیق از یک پایگاه دادگان مخصوص سیاست حفظ حریم شخصی و استاندارد که در فصل دوم معرفی شده است استفاده گردیم.

۳.۱ ساختار پایان نامه

این پایان‌نامه از پنج فصل تشکیل شده‌است. فصل اول خواننده را با مهمترین رویکردهای توسعه‌ی سیستم‌های پرسش‌وپاسخ آشنا می‌کند. در فصل دوم به متداول‌ترین مدل‌های آماری زبان که در سیستم‌های پرسش‌وپاسخ بکار رفته‌اند، اشاره می‌شود. فصل سوم ساختار و مزایای شبکه‌های عصبی عمیق را بررسی می‌کند. و به شرح مشکلات و علل خطا در سیستم‌های پرسش‌وپاسخ و روش پیشنهادی برای بهبود عملکرد سیستم می‌پردازد. فصل چهارم نکات مهم در نظر گرفته شده در هنگام پیاده‌سازی و بررسی عملکرد سیستم در مواجهه با کاربر را ارائه می‌کند. فصل پنجم به جمع بندی مطالب فصل‌های قبلی و ارائه پیشنهادات درباره‌ی چگونگی بهبود سیستم طراحی شده به عنوان کارهای آینده می‌پردازد.

فصل ۲

مروری بر کارهای انجام شده

۱.۲ مروری بر کارهای پیشین

نخستین سیستم پرسش‌وپاسخ به نام Baseball در سال ۱۹۶۱ طراحی شد این سیستم قادر بود به سوالات مربوط به بازی‌های لیگ سالانه‌ی بیس بال در آمریکا جواب دهد. این سیستم در واقع از زبان طبیعی به عنوان یک رابطه کاربری پایگاه داده استفاده می‌کرد: با حذف پیچیدگی‌های دستورات پایگاه داده و استفاده از پایگاه داده به عنوان منبع اطلاعات. با توجه به اینکه این سیستم تنها به سوالات با یک موضوع خاص (بازی‌های لیگ بسیبال آمریکا) جواب می‌داد از عملکرد قابل قبولی در آن زمان برخوردار بود. این سیستم یک نمونه ساده و کاربردی از سیستم‌های بازیابی اطلاعات بود. یکی از معایب این گونه از سیستم‌ها، زمان زیادی بود که برای طراحی موثر پایگاه دانش آنها مورد نیاز بود.

سیستم پرسش‌وپاسخ دیگری در سال ۱۹۷۲ با نام LUNAR طراحی شد قادر به پاسخگویی به سوالات زمین‌شناسان درباره‌ی نتایج بررسی ساختار و ترکیب شیمیایی خاک و نمونه سنگ‌های کره‌ی ماه در ماموریت سفینه‌ی آپولو بود. بررسی عملکرد این سیستم نشان می داد که بخش عمده و جدی خطاها به خاطره تجزیه و تحلیل نادرست معنایی بوده است [۶].

اما نوع دیگر سیستم‌های پرسش‌وپاسخ که در دهه‌ی ۷۰ میلادی رایج شد از منطق برای پاسخگویی به پرسش‌هایی با یک موضوع محدود استفاده می کرد. از جمله‌ی این سیستم‌ها می‌توان به سیستم SHRDLU در سال ۱۹۷۲ اشاره نمود که همانند یک ربات شبیه‌سازی عمل

می کرد و می‌توانست ضمن پاسخگویی به سوالات کاربر درباره‌ی محیط اطراف خود، اشیا و بلوک های اسباب بازی را بر طبق خواست کاربر جابجا نماید. در طراحی این نوع از سیستم‌ها از زبان برنامه‌نویسی LISP که قادر به پیاده‌سازی منطق است استفاده شد. به دلیل دامنه ی محدود لغات استفاده شده، درک منظور کاربر برای سیستم بسیار ساده بود. این سیستم از یک حافظه کوچک جهت ثبت دستورات گذشته کاربر استفاده می‌کرد و به وسیله‌ی آن می‌توانست مرجع ضمایر بکار رفته در جمله کنونی کاربر را در سابقه دستورات قبلی خود پیدا نماید [۷].

سیستم‌های پرسش‌وپاسخ محاوره‌ای مانند GUS، Ell2A نیز از یک پایگاه داده‌ی ساخت‌یافته به عنوان منبع دانش استفاده می‌گردند. محدودیت اصلی این سیستم‌ها این بود که دانش ذخیره شده در پایگاه داده‌ی ساخت یافته تنها قادر به پاسخگویی به سوالات در یک موضوع محدود را دارد. بعضی از سیستم‌های پرسش‌وپاسخ مانند START اخیراً توانستند از وب به عنوان منبع دانش خود استفاده کنند. این سیستم‌ها با اعمال توابع هیوریستیک اطلاعات را از اسناد وب بر روی پایگاه داده‌ی محلی خود ذخیره می‌کنند تا بعداً توسط تکنیک‌های زبانی جواب سوالات را ایجاد کنند [۸].

سیستم پرسش‌وپاسخ EAGLi با هدف پاسخگویی به سوالات پزشکی کاربر پیاده‌سازی شد. این سیستم شامل یک کتابخانه دیجیتال به نام MEDLINE که شامل ۱۸ میلیون مقاله بود و هر سال ۸۰۰۰۰۰ مقاله به آن اضافه می‌شد و از یک موتور جستجو به نام PubMed بهره می‌برد. ارزیابی این سیستم نشان داد پاسخ ۵۷٪ سوالات بیماری‌ها و پروتئین‌ها و ۶۸٪ پاسخ‌های مربوط به بیماری‌ها و داروها صحیح بودند. ۷۰٪ از پاسخ‌های مورد اول و ۷۵٪ از پاسخ‌های مورد دوم جز ۱۰ پاسخ محتمل ارائه شده در ابتدای فهرست پیشنهادی سیستم قرار داشتند.

سیستم‌های مشابه سیستم فوق که برای پاسخ به سوالات کاربر در یک حوزه خاص طراحی شده اند، **سیستم‌های پرسش‌وپاسخ با دامنه‌ی محدود** نامیده می‌شوند. این سیستم‌ها از منابع دانش یا پایگاه داده‌هایی که به صورت دستی طراحی می‌شوند استفاده نموده و دامنه مشخص و محدودی از اصطلاحات و واژگان پرکاربرد در پرسش‌های کاربر در آن حوزه خاص را دارا می‌باشند.

با برگزاری کنفرانس سالانه‌ی بازیابی متون [1] در سال ۱۹۹۹، دوره‌ی جدیدی در طراحی سیستم‌های پرسش‌وپاسخ آغاز گردید. در این سال، کنفرانس مذکور با اعلام رویکرد جدید خود طراحان سیستم‌های پرسش‌وپاسخ را به طراحی سیستم‌هایی تشویق نمود که با بهره‌گیری از مجموعه بزرگی از اسناد متنی در موضوعات مختلف، بتوانند پاسخ کوتاهی برای سوال کاربر که درباره موضوعات مختلف با زبان طبیعی مطرح شده است، ارائه نمایند. این گونه سیستم‌ها، **سیستم‌های با دامنه‌ی نامحدود** نامیده می‌شود [۹]. سیستم‌های پرسش‌وپاسخ با توجه نوع (حجم پاسخ) سوالی که پاسخ می‌دهند به دو دسته نیز تقسیم می‌شوند:

[1] TREC/Text REtrieval conference

۱. سوالاتی که توسط یک یا چند کلمه قادر به پاسخگویی هستند [۲] . مانند: ’’اولین رئیس جمهور سیاه پوست آمریکا کیست؟‘‘

۲. سوالاتی که باید توسط چندین جمله پاسخ داده شوند. تکنیک معمول پاسخگویی به این گونه از سوالات خلاصه‌سازی [۳] متن به یک پاراگراف، به گونه‌ایی که پاراگراف جواب سوال باشد، است. مانند: علت رشد اقتصادی ژاپن چیست؟‘‘ این گونه از سوالات را نمی‌توان در قالب یک کلمه جواب داد و نیاز به توضیح دارد.این گونه از سوالات پیچیده [۴] یا باز [۵] نامیده می‌شوند [۱۰].

پردازش‌های سیستم پرسش‌وپاسخ به طور کلی شامل سه مرحله می‌شود؛

۱. تحلیل پرسش: تجزیه و طبقه‌بندی سوال

۲. تحلیل سند: انتخاب سند، تشخیص جواب‌ها

۳. تحلیل پاسخ: انتخاب جواب‌ها و رتبه‌بندی بهترین آنها [۱۱].

بیشتر سیستم‌های پرسش‌وپاسخ این سه فرآیند را دنبال می‌کنند ولی در نحوی پیاده‌سازی با یکدیگر متفاوت هستند.

روش‌های تجزیه تحلیل نحوی

سیستم‌های موجود صرف نظر از حوزه‌ی دانش خود (محدود به دامنه دانش خاص و یا دامنه نامحدود)، عمدتاً از تجزیه و تحلیل‌های معنایی و گرامری، تطابق الگو و یا آموزش با استفاده از مجموعه بزرگی از داده‌های برچسب گذاری شده به صورت دستی در یک فرآیند زمانبر استفاده می‌کنند. تعامل این سیستم‌ها با کاربر با هدف بهینه‌سازی پرسش و رفع ابهام از آن و یا طرح پرسش‌های مرتبط انجام می‌شود [۶].

روش‌های پیاده‌سازی این نوع از سیستم‌ها معمولاً شامل تجزیه و تحلیل ساختار گرامری و تفکیک اجزای سازنده‌ی جمله می‌باشد. و با توجه به فراوانی استثناها در ساختار زبان، پیاده‌سازی سیستم بسیار زمانبر و وابسته به زبانی است که بر مبنای قواعد نحوی آن این سیستم پیاده‌سازی شده است. از جمله تکنیک‌های استفاده شده در طراحی این نوع از سیستم‌ها می توان به: تکنیک‌های خلاصه‌سازی، کاهش ابعاد، تحلیل وابستگی و منطق مرتبه اول نام برد [۱۲].

روش‌های آماری

روش‌های آماری نه تنها از زبان پرسش‌وپاسخ مستقل هستند بلکه می‌توانند پرسش‌ها را به صورت زبان طبیعی فرمول کنند. این روش‌ها به طور کلی به پایگاه‌دادگان بزرگ و

[2] Factoid
[3] Summarize
[4] Complex
[5] Open-ended

دارای توزیع یکسان با موضوع پرسش‌وپاسخ برای یادگیری دقیق آماری نیاز دارند و زمانی که یادگیری صورت گرفت بهترین نتیجه را نسبت به بقیه‌ی روش‌ها تولید می‌کنند. به علاوه، متد یادگیری آماری را می‌توان به سادگی برای یک موضوع جدید مستقل از هر شکل زبانی بکار برد . به طور کلی تکنیک‌های آماری تاکنون در قسمت‌های مختلف یک سیستم پرسش‌وپاسخ با موفقیت به کار رفته‌اند. طبقه‌بندی‌کننده‌های SVM [۶] ، طبقه‌بندی کننده‌های بیزین و مدل‌های بیشترین آنتروپی [۷] چندین تکنیکی هستند که برای طبقه‌بندی سوالات استفاده شده‌اند. یکی از سیستم‌های پرسش‌وپاسخ که بر اساس مدل‌های آماری توسعه داده شده‌است IBM نام دارد. این سیستم از مدل بیشترین آنتروپی برای طبقه‌بندی سوالات/ جواب‌ها بر اساس انواع N-gram و ویژگی‌های کیسه‌ی کلمات [۸] استفاده می‌کند [۱۳].

۱.۱.۲ ترجمه آماری ماشین

در چندین قرن گذشته، ماشین ها بسیاری از کارهای انسانی را انجام می دهند، و اخیراً، با پیدایش کامپیوتر، حتی کارهایی که تصور می‌شد که نیاز به تفکر و هوش دارند را نیز انجام می‌دهند. ترجمه‌ی بین زبان‌ها یکی از این کارها است، کاری که برای انجام آن حتی انسان‌ها نیز نیاز به آموزش‌های خاصی دارند [۱۴] .

پیشرفت‌های اخیر سیستم‌های پردازش زبان طبیعی بر این قاعده استوار هستند که زبان های طبیعی بسیار غنی و پیچیده هستند و نمی‌توان به صورت کامل آنها را تحلیل و به مجموعه‌ای از قوانین تبدیل کرد که نهایتاً به یک برنامه‌ی کامپیوتری تبدیل شود. اما، توسعه‌ی سیستم‌های ترجمه‌ی ماشینی بر کشف اتوماتیک قوانین ترجمه از روی پایگاه دادگان بزرگ از متن‌های ترجمه شده، توسط در نظرگرفتن جفت ورودی و خروجی ترجمه و یادگیری قواعد آماری از روی داده استوار هستند [۱۴] .

مجموعه‌ی متون را پایگاه دادگان می‌نامند و برای ترجمه‌ی آماری به پایگاه‌دادگان موازی نیاز است که متن‌هایی به صورت جفت متن و ترجمه‌ی آن به زبان دیگر هستند. تهیه‌ی پایگاه دادگان موازی برای ترجمه‌ی آماری ممکن است به خزنده‌ی وب [۹] برای استخراج متن از فرمت‌هایی همچون HTML نیاز داشته باشد.

احتمالات زمانی که به توصیف رویدادهایی با خروجی غیر مطمئن نیاز است، استفاده می‌شوند مانند زمانی که یک کلمه‌ی خارجی شاید به یک یا چندین کلمه‌ی انگلیسی ترجمه می‌شود. یک توزیع احتمال تابعی است که هر رخداد را به مقداری مابین صفر و یک انتساب می‌کند [۱۴] .

ترجمه‌ی ماشینی معمولاً با چندین رخداد تصادفی روبرو است، که هر یک از توزیع احتمال

[6] Support vector machines
[7] Maximum entropy
[8] Bag of words
[9] Web crawler

متفاوتی پیروی می‌کنند مانند ترجمه‌ی یک کلمه در یک جمله. روش‌هایی که استفاده می‌شوند برای رویارویی با رویدادهای احتمالی بر اساس ویژگی‌های توزیع‌های احتمال است، مانند میانگین و واریانس در رخدادها. یکی از مفاهیم قوی، آنتروپی یا درجه ی غیر محتمل بودن رخدادها است، که بیشتر تکنیک‌های یادگیری ماشین برای مدل احتمال را راهنمایی می‌کند. مدل‌های آماری اولیه برای ترجمه‌ی ماشینی بر اساس ترجمه‌ی کلمه به عنوان کوچکترین جزیی که اضافه، حذف یا ترتیبش تغییر می‌کند، بودند [۱۴] .

به خاطر اینکه پایگاه‌دادگان فقط جمله‌ها و ترجمه‌ی آنها را فراهم می‌کند و نه ترتیب کلمات را در فرآیند ترجمه، الگوریتم یادگیری با مشکل یادگیری، از روی داده‌ی غیرکامل مواجه است. تنها راه حل این مشکل الگوریتم افزایش انتظار [۱۰] که احتمال ترکیب‌های کلمه را محاسبه کرده و تعداد هر یک را می شمارد و یک مدل بهبود یافته برای ترکیب ها(ترتیب ها) می‌سازد [۱۴] .

IBM Model1

جمع آوری یک آمار

اگر یک مجموعه‌ی بزرگ از متن‌های آلمانی به همراه ترجمه‌ی انگلیسی آنها موجود باشد، می‌شود تعداد ترجمه‌ی کلمه‌ی Haus را به هر یک از معانی ممکن آن بشماریم. حال توزیع احتمال ترجمه‌ی یک کلمه را از روی این شمارش بدست می‌آید. این تابع در جواب به سوالی که در هنگام ترجمه‌ی یک متن جدید آلمانی با آن روبرو می‌شوید کمک می‌کند: محتمل‌ترین ترجمه انگلیسی برای یک کلمه‌ی خارجی همچون Haus چیست؟ این تابع باید بیشترین مقدار را برای یک کاندیدای انگلیسی برگرداند اگر آن یک ترجمه‌ی معمول باشد. این تابع باید یک مقدار نزدیک به صفر را برگرداند اگر کاندیدای انگلیسی یک ترجمه‌ی نادر کلمه باشد. این باید مقدار صفر را برگرداند اگر ترجمه‌ی کلمه به انگلیسی غیر ممکن باشد [۱۴] .

ترتیب نگاشت کلمات در ترجمه توسط تابع ترتیب انجام می‌شود. این تابع هر کلمه آلمانی را در موقعیت j به یک کلمه انگلیسی در موقعیت i نگاشت می‌کند. در حالی که بسیاری از زبان‌ها از ترتیب مشابه ترتیب کلمات در انگلیسی پیروی می‌کنند، یک زبان خارجی می‌تواند جمله‌هایی با ترتیب کلمات متفاوت از انگلیسی داشته باشد. این بدان معناست که مجبور به تغییر کلمات در حین ترجمه هستید. به جزء نیاز به تغییر ترتیب کلمات، زبان ها شاید در تعداد کلماتی که یک مفهوم یکسان را بیان می‌کنند، متفاوت باشند. زبان‌ها شاید کلماتی داشته باشند که یک معادل روشن در انگلیسی نداشته باشند. با داشتن تابع توزیع احتمال کلمات و مفهوم تابع ترتیب، مدلی تعریف می‌شود که تعدادی ترجمه‌ی متفاوت برای یک جمله با احتمال‌های متفاوت تولید می‌کند [۱۴] .

به روش تقسیم فرآیند تولید داده به گام‌های کوچک و ایجاد مدل برای هر یک از این

[10] Expectation maximization

گام‌ها توسط توزیع‌های احتمال و ترکیب این گام‌ها برای تولید یک مدل یکپارچه Generative Modeling می‌گویند.

در IBM Model1 برای هر کلمه‌ی خروجی e تولید شده توسط مدل از روی کلمه‌ی ورودی f مقدار احتمال $P(e \mid f)$ را حساب می‌شود و با استفاده از تابع ترتیب مقدار $P(e,a \mid f)$ را بدست می‌آوریم.

IBM Model2

با دقت در عملکرد IBM Model1 خطاهای بسیار زیاد مدل نمایان می‌شود(دو جمله با ترکیب های (ترتیب کلمات) متفاوت ولی کلمات یکسان دارای احتمال یکسان برای ترجمه هستند.). این مدل در تغییر ترتیب بسیار ضعیف است همچنین در اضافه یا حذف کلمات. مدل‌های ترتیب یک مدل واضح برای تغییر ترتیب کلمات یک جمله‌ی ارائه می‌دهند. در بیشتر موارد، کلماتی که پشت سر هم در یک زبان می‌آیند ترجمه‌هایی دارند که پشت سر هم در زبان دیگری می‌آیند. با این حال IBM Model1 تمام ترتیب‌های کلمات را یکسان در نظر می‌گیرد. در بیشتر موارد، یک کلمه در زبان ورودی به یک کلمه در زبان خروجی ترجمه می‌شود ولی بعضی از کلمات چندین کلمه تولید می‌کنند یا حذف می‌شوند [۱۴] .

IBM Model2 با استفاده از یک مدل برای ترتیب کلمات بر اساس موقعیت‌های کلمات ورودی و خروجی این مشکل را حل می‌کند. ترجمه‌ی یک کلمه‌ی خارجی ورودی در جایگاه i به یک کلمه‌ی انگلیسی در جایگاه j توسط تابع توزیع احتمال ترتیب مدل می شود [۱۴] .

با توجه به دو مدل فوق ما متوجه می‌شویم که کلمات در فرآیند ترجمه از یک توزیع خاص به کلمات با توزیع دیگر تبدیل می‌شوند. این همان کاری است که مدل شبکه‌ی عصبی عمیق ما برای پاسخگویی به سوالات کاربر انجام می‌دهد. ما سوالات را که از یک توزیع خاص پیروی می‌کنند به مجموعه‌ای از کلمات که در موضوع سیاست حفظ حریم شخصی هستند، تبدیل می‌کنیم.

۲.۲ مرور ادبی

تلاش‌های بسیاری برای نمایش ساده‌ی اسناد حفظ حریم شخصی صورت گرفته‌است. آقای کیلی [11] درباره‌ی استفاده از برچسب‌های شبیه برچسب‌های اطلاعات تغذیه‌ایی مواد غذایی به عنوان یک روش نمایش اطلاعات حفظ حریم شخصی مطالعه کرد. و در [۲] گزینه‌هایی برای نمایش حفظ حریم شخصی پیشنهاد شد [۱۵]. دیگران نیز روشی استاندارد برای ترغیب ارائه‌دهنده‌های سرویس به کد کردن سیاست‌های حفظ حریم شخصی در قالب قابل درک برای کامپیوتر مانند: p3p ، ارائه کردند، [۱۶] ولی آنها با توسعه‌دهندگان مرورگرها و ارائه‌دهنده‌های

[11] Kelley

سرویس منعطفانه بروز نمی‌شدند. در این تحقیق، ما پتانسیل ایجاد خودکار بسیاری از اطلاعات را، بدون وابسته بودن به دیگران(ارائه‌دهندگان سرویس و...) تا این اطلاعات را برای ما فراهم کنند، از طریق اسناد حفظ حریم شخصی داریم.

به تازگی، چندین محقق امکان تحلیل خودکار سیاست‌های حفظ حریم شخصی را بررسی کردند. به عنوان مثال آقای لو [۱۲] [۱۷] از شبکه‌های عصبی برای ایجاد مدل کشف ابهام کلمات اسناد حفظ حریم شخصی استفاده کرد. آقای زیمک [۱۳] [۸] قادر بود توسط تحلیل خودکار، عدم هماهنگی میان اسناد حفظ حریم شخصی و رفتار واقعی برنامه‌ها را نمایش دهد. این تحقیقات و تحقیقات دیگر [۱۱] ، [۸] به‌واسطه‌ی وجود پایگاه دادگانی [۱۸] که توسط آقای ویلسون [۱۴] ایجاد شده بود و شامل ۱۱۵ سند حفظ حریم شخصی که به صورت دستی توسط دانشجویان حقوق برچسب‌گذاری شده‌است، ممکن شدند. سیستم ما قادر به محاسبه‌ی احتمال تعلق برچسب به صورت دقیق و با جزییات (۱۰ کلاس کلی و ۱۲۲ کلاس جزیی برای هر قسمت از سیاست حفظ حریم شخصی) است.

تناقض میان رفتار شرکت‌ها و سیاست‌های حفظ حریم شخصی آنها توسط سیستم پرسش‌وپاسخ طراحی شده در این تحقیق قابل مشاهده است. سیستم ما قادر است مسئولین و محققین را به طور دائم و خودکار برای درک میزان این تناقض یاری رساند. این سیستم می‌تواند اطلاعات اولیه را برای متخصصین برای پردازش بیشتر فراهم کند، بنابراین منجر به کاهش هزینه و زمان تحلیل می‌شود [۱۹، ۲۰].

۳.۲ انواع مختلف پیش پردازش متن در سیستم‌های پرسش‌وپاسخ

یکی از چالش‌های سیستم‌های پرسش‌وپاسخ موجود درک صحیح سوالی است که توسط زبان طبیعی مطرح شده (استنباط معنی دقیق آن برای بازیابی جواب است). و برای این منظور سیستم‌های پرسش‌وپاسخ به طور معمول از مدل‌های آماری زبان استفاده می‌کند. برای یادگیری این مدل‌های مبتنی بر داده [۱۵] ، معمولاً نیاز به پایگاه دادگان بزرگ (تا شامل توزیع یکسانی از جامعه مورد نظر باشد)و صحیح است. از ویژگی‌های مهم پایگاه دادگان جامعیت آن است که باید شامل تمام حالات ممکن داده‌ها باشد. یادگیری این مدل‌ها نسبت به مدل‌های مبتنی بر قواعد نحوی سریعتر بوده و از دقت بیشتری برخوردار است. TF- IDF، LDA و Embedded سه نمونه از این مدل‌های زبانی هستند. که برای پیش پردازش در سیستم‌های پرسش‌وپاسخ مورد استفاده قرار می‌گیرند.

[12] Liu
[13] Zimmeck
[14] Wilson
[15] Data driven

۱.۳.۲ TF-IDF

در این مدل برای تمام کلمات (بعد از حذف step-words) درون اسناد پایگاه دادگان یک وزن یا میزان اهمیت آن کلمه را در آن سند محاسبه می‌کنیم. سپس اسنادی را که بالاترین میزان اهمیت را برای کلمات پرسش مطرح شده توسط کاربر دارند، به عنوان اسنادی که جواب کاربر را درون خود دارند، کاندید می‌کنیم. نحوه‌ی محاسبه‌ی وزن هر کلمه در هر سند به این صورت است که تعداد تکرار آن کلمه در آن سند بخصوص (TF) را در معکوس تعداد اسناد دارای آن کلمه (IDF) ضرب می‌کنیم. IDF هر کلمه برابر لگاریتم تعداد کل اسناد تقسیم بر تعداد اسناد حاوی آن کلمه است.

$$IDF_i = \log \frac{N}{n_i} \tag{۱.۲}$$

N تعداد کل اسناد و n_i تعداد اسناد حاوی کلمه‌ی i

هر چه تعداد تکرار یک کلمه در یک سند بیشتر و تعداد اسناد دارای آن کلمه در تمام پایگاه داد کمتر باشد، اهمیت آن کلمه درون آن سند بیشتر است. این مدل موقعیت کلمه نسبت به کلمات اطرافش را در نظر نمی‌گیرد [۲۱].

۲.۳.۲ IDA

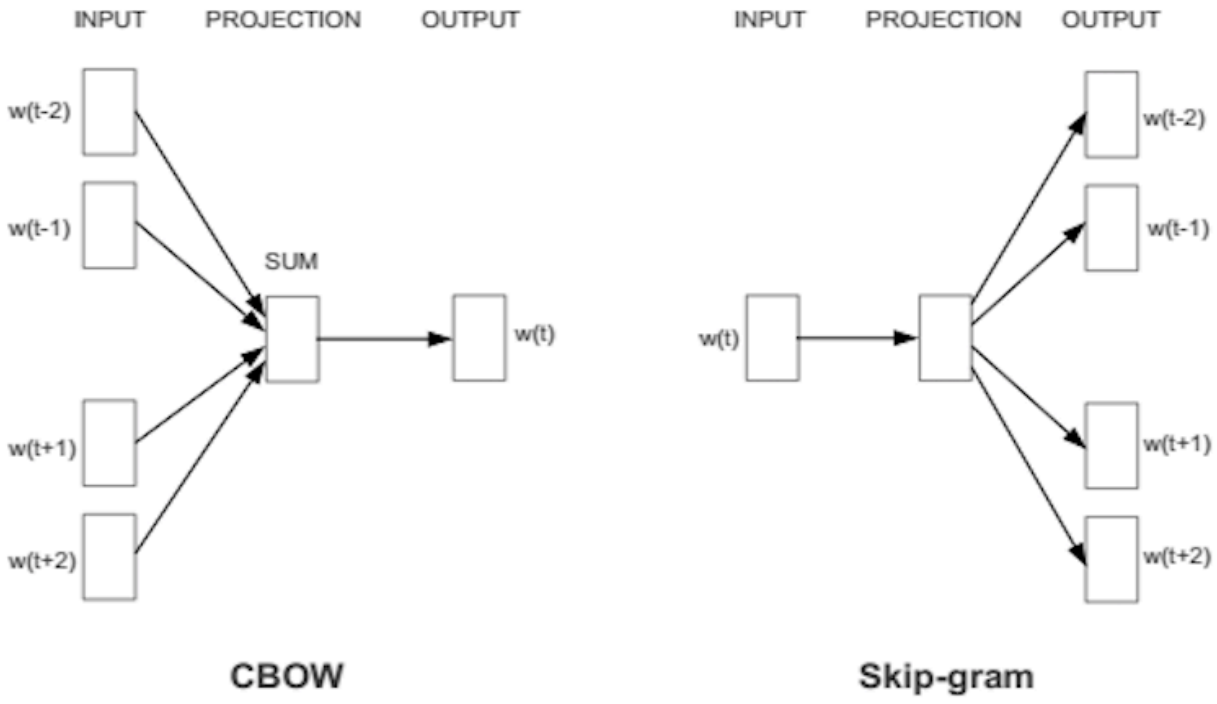
یکی از معمول‌ترین مدل‌های موضوعی است، هدف از ایجاد مدل‌های موضوعی پاسخ به سوال: ”چه موضوعاتی را این اسناد در بر می‌گیرند؟ “است. مدل موضوعی یکی از روش‌های تحلیل داده‌های بدون برچسب با حجم زیاد است. یک موضوع، یک توزیع احتمال بر روی مجموعه‌ای از کلمات است و یک مدل موضوعی روابط آماری بین گروهی از متغیرهای پیدا و پنهان که روند احتمالی ایجاد موضوعات را شرح می‌دهد، است. این مدل نیز موقعیت کلمات را نسبت به یکدیگر در نظر نمی‌گیرند. در این مدل ما برای هر موضوع، محتمل‌ترین کلمات آن موضوع و احتمالاتشان را محاسبه می‌کنیم(با در نظر گرفتن هم رخدادی کلمات) برای هر سند، احتمال وقوع هر موضوع را در آن سند محاسبه می‌کنیم. و برای هر کلمه احتمال رخداد در هر موضوع را بررسی می‌کنیم [۲۲، ۲۳].

۳.۳.۲ word2vec

word2vec [۱۶] یک شبکه‌ی عصبی دو لایه است که متن را پردازش می‌کند. ورودی این شبکه‌ی عصبی مجموعه‌ایی از اسناد (پایگاه دادگان) و خروجی آن بردارهای ویژگی کلمات اسناد هستند، به گونه‌ای که کلماتی که دارای یک معنی مشابه در یک موضوع هستند در این فضای برداری به یکدیگر نزدیکتر قرار می‌گیرند. این شبکه‌ی عصبی رشته‌ی کلمات(متن)

[16] Embedded model

را به اعدادی تبدل می‌کند که قابل درک برای شبکه‌های عصبی عمیق هستند. این نوع از مدل‌ها برای داده‌هایی که گسسته و هم رخدادی دارند مانند: کلمات، ژن‌ها (DNA) ، رشته‌های پروتین (RNA) و... مناسب هستند. با داشتن پایگاه دادگان جامع و موضوع متن، word2vec می‌تواند با دقت بالایی معنی کلمات را بر اساس مشاهده‌های قبلی کلمه حدس بزند. این مدل نیز موقعیت(ترتیب) کلمات را نسبت به یکدیگر در نظر نمی‌گیرد ولی موضوع متن و کلمات همسایه‌ی هر کلمه در بردار ویژگی آن تاثیر می‌گذارد [۲۴].



شکل ۱.۲: دو روش پیاده‌سازی مدل word2vec

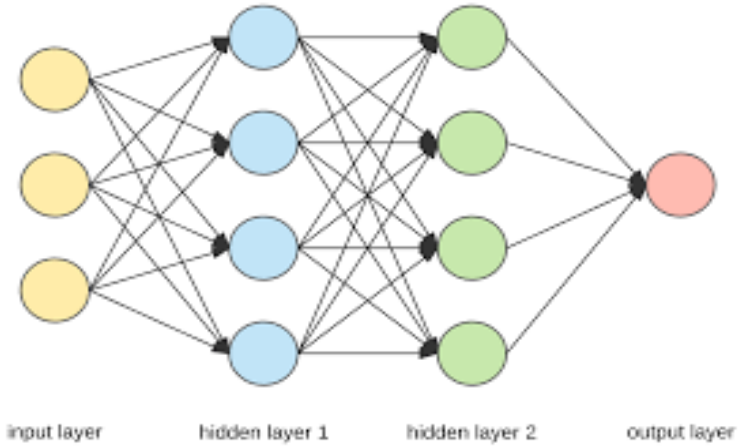
هر دوی معماری‌های فوقِ مدل برای ایجاد یک نمایش توزیع آماری کلمات قابل استفاده هستند. در معماری continuous bag-of-words مدل کلمه‌ی حاضر را از روی پنجره‌ی کلمات اطراف آن تخمین می‌زند. در معماری continuous skip-gram مدل با استفاده از کلمه‌ی حاضر پنجره‌ی کلمات اطراف را تخمین می‌زند. این معماری به کلمات درون یک پنجره اهمیت بیشتری قائل می‌شود. معماری CBOW سریعتر است اما معماری skip-gram در برابر کلمات کم تکرار عملکرد بهتری دارد.

برای پیاده‌سازی این تحقیق ما از مدل آموزش داده‌شده توسط گوگل استفاده کردیم.

فصل ۳

سیستم های پرسش و پاسخ

برای ایجاد یک طبقه‌بندی کننده‌ی آماری توسط شبکه‌های عصبی، ما ویژگی‌های هر یک از موجودیت‌های درون پایگاه دادگان را به شبکه‌ی عصبی داده (لایه‌ی ورودی) و پس از محاسبه‌ی $z=wx+b$ برای هر نرون و اعمال تابع فعال‌ساز، مقادیر حاصل به نرون‌هایی لایه‌ی بعدی در شبکه‌ی عصبی ارسال می‌شود. نهایتاً در لایه‌ی آخر شبکه برچسبی را به عنوان کلاس موجودیت ورودی تخمین می‌زند. پس از هر بار ارائه کامل تمام موجودیت‌های درون پایگاه دادگان و محاسبه‌ی تمامی برچسب‌ها، ما مقادیر w، b را به گونه‌ای اصلاح می‌کنیم تا مقدار تابع هزینه (تعداد خطای شبکه‌ی عصبی در تخمین برچسب درست هر موجودیت) کاهش یابد.



شکل ۱.۳: نمایی کلی شبکه‌ی عصبی

تعداد ارائه کامل تمام موجودیت‌های درون مجموعه‌ی داده‌های آموزش [1] و اصلاح پارامترهای w، b به مقدار کاهش تابع هزینه در هر بار اصلاح وزن‌ها وابسته است. پس از بدست آوردن w، b مناسب (کاهش تابع هزینه یا خطای شبکه به میزان قابل قبولی) توسط مجموعه داده‌ی توسعه (داده‌های هدف که طبقه‌بندی‌کننده باید بهترین عملکرد را بر روی آنها نشان دهد) عملکرد شبکه عصبی را بر روی این داده‌ها که قبلاً توسط شبکه دیده نشده‌اند، با تغییر پارامترهای w، b بهبود می‌دهیم. نهایتاً، پس از اتمام مرحله‌ی آموزش (بدست آوردن پارامترهای مدل w، b)از این شبکه‌ی عصبی به عنوان طبقه‌بندی‌کننده در مواجهه با داده‌های جدید استفاده می‌کنیم. شبکه‌های عصبی نسبت به دیگر انواع مدل‌های آماری عملکرد بهتری خواهند داشت اگر:

۱. ما پایگاه دادگان بسیار بزرگی داشته باشیم.

۲. توان آموزش شبکه‌ی عصبی بزرگی را داشته باشیم (داشتن قدرت محاسبات بالا برای آموزش یک شبکه عصبی با تعداد لایه‌های پنهان زیاد و تعداد نرون‌های زیاد در هر لایه)

۳. یکی از مواقعی که شبکه‌های عصبی عمیق (شبکه‌های عصبی با تعداد لایه‌ی پنهان زیاد) عملکرد بهتری نسبت به دیگر مدل‌های آماری دارند زمانی است که مجموعه داده‌های آموزش از توزیع متفاوتی نسبت به مجموعه‌ی داده‌ی توسعه و تست نشأت گرفته باشد؛ همین ویژگی شبکه‌های عصبی عمیق، این مدل آماری را یک مدل مناسب برای پردازش تصویر و زبان طبیعی کرده است.

شبکه‌های عصبی عمیق در برابر داده‌های اشتباه تصادفی (نویزها) به دلیل حجم کم این گونه از داده‌ها مصون هستند ولی در صورتی که این داده‌های اشتباه به صورت سیستماتیک ایجاد شده باشند باید قبل از استفاده از این داده‌ها در آموزش شبکه‌ی عصبی از مجموعه‌ی دادگان حذف شوند. از معایب استفاده از شبکه‌های عصبی می‌توان به موارد زیر اشاره کرد:

[1] Iteration

۱. وابستگی بسیار زیاد به حجم زیاد داده و جامعیت و درست بودن داده

۲. نیاز به تنظیم پارامترهای زیاد (به غیر از پارامترهای مدل: w ،b)

از جمله پارامترهای غیر مدلی [۲] که باید برای آموزش شبکه‌های عصبی عمیق تنظیم کرد می‌توان به: تعداد لایه‌ی مخفی، تعداد نرون درون هر لایه، ضریب regularization، نوع تابع فعال‌ساز، نرخ یادگیری، تعداد تکرار ارائه‌ی داده‌ی آموزشی به شبکه‌ی عصبی اشاره کرد.

مقدار هر یک از پارامترهای غیرمدلی فوق با توجه به نوع مسئله متفاوت است به این صورت که برای آموزش شبکه‌ی عصبی برای پردازش زبان طبیعی ما با ساختار و مقادیر پارامترهای غیرمدلی متفاوتی نسبت به آموزش شبکه‌ی عصبی برای پردازش تصویر روبرو هستیم. به طور معمول نحوی دست‌یابی به بهترین تنظیم پارامترهای غیرمدلی پس از سپری شدن مدتی پیاده‌سازی شبکه‌های عصبی برای یک نوع مسئله‌ی خاص بدست می‌آید. ولی متاسفانه این تنظیمات قابل استفاده برای نوع دیگری از مسئله نیست و با صرف زمان و آموزش چندین مدل شبکه‌ی عصبی در مسئله‌ی جدید، تنظیمات بهینه بدست خواهد آمد.

۱.۳ ایده‌ی بهبود عملکرد

۱.۱.۳ تعامل

در اکثر سیستم‌های پرسش‌وپاسخ، سیستم تاریخچه‌ای از پرسش‌های قبلی کاربر ذخیره نمی‌کند و سوال کاربر را با توجه به سوال پرسیده شده و مدل آماری بدست آمده از داده‌های آموزش، جواب می‌دهد. در این گونه از سیستم‌ها، کاربر و سیستم با هم تعامل کمی دارند و ممکن است کاربر برای بدست آوردن جواب مورد نظرش مجبور به پرسیدن چندین سوال شود و یا با پاسخ‌های کلی و با حجم زیاد روبه‌رو شود. با افزودن قابلیت تعامل به چنین سیستم‌هایی، سیستم به درک دقیقتری نسبت به مفهوم پرسش کاربر می‌رسد و عملکرد (دقت) بهتری را از خود نشان می‌دهد [۲۵].

با دقت در رفتار کاربران موتور جستجوی گوگل می‌توان متوجه شد که کاربران با توجه به لیست مدارک برگردانده شده توسط گوگل برای رسیدن به جواب ترکیب کلمات کلیدی را تغییر می‌دهند. و یا کلمات را با کلمه‌ایی مشابه از نظر معنایی جایگزین می‌کنند تا به لیست مدارکی برسند که فکر می‌کنند جواب سوالشان در آن است. تعامل با موتور جستجوی گوگل با دو روش صورت می‌گیرد. در روش اول سیستم گوگل پرسش‌های مشابه را که قبلاً دیده است برای درک سوال کاربر به او پیشنهاد می‌کند. در روش دوم کاربران با توجه به لیست مدارکی که گوگل به عنوان جواب برگردانده، ترکیب و یا کلمات خود را تغییر می‌دهند.

با توجه به ساختار سیستم پرسش‌وپاسخ پیاده‌سازی شده ما در ابتدای دو طبقه‌بند عمیق،

[2] Hyperparameters

از یک مدل آماری word2vec برای تبدیل ویژگی‌های متون به اعداد استفاده می‌کنیم. این مدل آماری قادر به تشخیص تشابه معنایی است و ترتیب ترکیب کلمات را در محاسبه‌ی ویژگی‌های لحاظ نمی‌کند. در نتیجه برای دو پرسش با ترتیب متفاوت از کلمات و یا دو پرسش با کلمات مشابه معنایی بردارهایی (اعدادی) نزدیک به هم تولید می‌کند. قوانین مورد استفاده برای استخراج بهترین جواب و ایجاد تعامل بین سیستم و کاربر همگی از این ویژگی تبعیت می‌کنند(عدم تمایز بین کلمات هم‌معنی و همچنین سوالات با کلمات یکسان و ترکیب متفاوت کلمات).

راه‌حل پیشنهادی من برای بهبود دقت سیستم پرسش‌وپاسخ افزودن قدرت تعامل به سیستم است بدین شکل که با استفاده از قوانینی متفاوت برای در نظر گرفتن تاریخچه‌ی سوالات و جواب‌های متقابل، به نوعی تعامل را میان کاربر و سیستم پرسش‌وپاسخ افزایش دهیم. این قوانین با استفاده از درنظرگرفتن ابعاد متفاوت ترکیب دو سوال پشت سر هم، ابعاد متفاوت داده را برای پاسخگویی به کاربر در نظر می‌گیرند. قوانین ترکیب دو سوال پشت سر هم که هر یک به صورت جداگانه بر روی سیستم پرسش‌وپاسخ اعمال شده‌اند و هر یک سیستم پرسش‌وپاسخ متفاوتی را ایجاد کرده‌اند، به شرح زیر است:

۱. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، دومین محتمل‌ترین جواب سوال قبلی، جوابی که از نظر مدل آماری با هم مشابه هستند را در صورت کم بودن فاصله‌ی احتمال دو جواب به عنوان جواب به کاربر برگرداند.

۲. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec بپرسید به جای طبقه‌بندی سوال دوم،از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که از نظر مدل آماری کمترین شباهت را به محتمل‌ترین جواب سوال قبل دارد را به عنوان جواب به کاربر برگرداند.

۳. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که از نظر مدل آماری بیشترین شباهت از نظر معنایی را با سوال دوم داشته باشد، را به عنوان جواب به کاربر برگرداند.

۴. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که بیشترین کلمه‌ی مشترک را در مجموعه‌ی کلمات ارتباط معنایی در مدل word2vec با سوال دوم داشته باشد(اندازه‌ی اشتراک مجموعه‌ی کلمات معنایی محتمل‌ترین جواب‌های سوال یک و سوال دوم)، را به عنوان جواب به کاربر برگرداند.

۵. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی

سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که از نظر مدل آماری کمترین شباهت معنایی را با سوال یک داشته باشد، را به عنوان جواب به کاربر برگرداند.

۶. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، غیرمحتمل‌ترین جواب سوال قبلی، جوابی که از نظر مدل آماری را به عنوان جواب به کاربر برگرداند.

۷. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که کمترین کلمه‌ی مشترک را در مجموعه‌ی کلمات ارتباط معنایی در مدل word2vec با سوال اول داشته باشد(اندازه‌ی اشتراک مجموعه‌ی کلمات معنایی محتمل‌ترین جواب‌های سوال یک و سوال یک)، را به عنوان جواب به کاربر برگرداند.

۸. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب سوال قبلی، جوابی که کمترین شباهت از نظر معنایی به محتمل‌ترین جواب سوال یک و بیشترین شباهت از نظر معنایی را به سوال دوم دارد، را به عنوان جواب به کاربر برگرداند.

محاسبه‌ی تشابه دو جمله

برای محاسبه‌ی تشابه دو جمله در مدل آماری word2vec بردار متوسط تک تک کلمات رو برای هر جمله محاسبه کرده و سپس فاصله‌ی کسینوسی بردارهای متوسط کلمات سازنده‌ی هر جمله را به عنوان معیار تشابه استفاده می‌کنیم. عدم تشابه به صورت کسینوس ۹۰ درجه نمایش داده‌می‌شود. تشابه کامل نیز مقدار ۱ برای کسینوس یعنی زاویه‌ی صفر درجه است. هر چه مقدار فاصله‌ی کسینوسی دو جمله به یک نزدکتر باشد، دو جمله تشابه بیشتری با یکدیگر دارند.

فصل ۴

پیاده سازی و نتایج

۱.۴ مقدمه

برای بهبود عملکرد شبکه‌عصبی ما معمولاً به گزینه‌های زیر می‌اندیشیم:

۱. افزایش حجم داده‌ی آموزش

۲. جمع‌آوری داده‌های متفاوت که سیستم ممکن است با آنها روبرو شود.(تمام حالات ممکن داده)

۳. افزایش زمان آموزش مدل آماری

۴. استفاده از الگریتم بهینه‌سازی متفاوت

۵. کاهش یاافزایش لایه‌های مخفی و تعداد نرون درون هر لایه شبکه‌ی عصبی

۶. حذف تصادفی بعضی از نودهای شبکه [۱]

۷. اضافه کردن ضریب regularization به الگوریتم آموزش

۸. و یا تغییر ساختار شبکه‌ی عصبی مانند تغییر تابع فعال‌ساز و یا تغییر تعداد نرون‌های مخفی و ...

[1] Dropout regularization

برای بهبود یک شبکه‌ی عصبی ما گزینه‌های بسیار زیادی داریم که در صورت انتخاب گزینه‌ی غلط می‌تواند زمان بسیار زیادی را صرف تنظیم یک پارامتر بی بازه کنیم که نهایتاً متوجه می‌شویم که این پارامتر عملکرد سیستم ما را بهبود قابل توجهی نمی‌دهد [۱۰].

برای بدست‌آوردن بهترین عملکرد سیستم در یک مسئله‌ی یادگیری با ناظر ما چندین پیش فرض را در نظر می‌گیریم:

۱. ما بهترین عملکرد (کمترین تابع هزینه) را بر روی داده ی آموزش خواهیم داشت.

۲. و در گام بعد عملکرد سیستم (کمترین تابع هزینه) را بر روی داده‌ی توسعه بهبود خواهیم بخشید.

۳. عملکرد سیستم بر روی داده‌ی تست بهبود خواهد یافت.

۴. نهایتاً عملکرد خوبی در مواجه با داده‌های دنیای واقعی خواهیم داشت.

برای بهبود عملکرد بر روی هر یک از داده‌های فوق گزینه‌های متفاوتی داریم:

۱. برای بهبود عملکرد سیستم بر روی داده‌ی آموزش ما الگوریتم بهینه‌سازی خود را تغییر می‌دهیم، شبکه‌ی عصبی بزرگ‌تری را آموزش می‌دهیم و....

۲. برای بهبود عملکرد شبکه بر روی داده‌ی توسعه ما داده‌ی آموزش خود را افزایش می‌دهیم، ضریب regularization را به الگوریتم بهینه‌سازی خود اضافه می‌کنیم و ...

۳. برای بدست آوردن بهترین عملکرد بر روی داده‌ی تست ما داده‌ی توسعه‌ی را افزایش می‌دهیم و ....

۴. و نهایتاً در صورتی که ما عملکرد خوبی بر روی داده‌ی تست داریم ولی عملکرد سیستم ما در مواجه با داده‌های دنیای واقعی خوب نیست ما باید یا داده‌ی توسعه‌ی خود را تغییر دهیم یا تابع هزینه‌ی خود را تغییر دهیم (این بدین معناست که یا توزیع داده‌ی توسعه و تست ما به درستی انتخاب نشده‌است و یا تابع هزینه‌ی ما چیز درستی را اندازه‌گیری نمی‌کند.)

برای درک اینکه آیا تغییری که ایجاد کرده‌ایم باعث بهبود عملکرد سیستم شده‌است ما به یک معیار ارزیابی مانند precision و یا recall نیازمندیم. اما بدست‌آوردن بهترین ترکیب دو یا چند معیار ارزیابی معمولاً ممکن نیست. در چنین مواقعی به یک مقدار حداقل برای یک معیار ارزیابی و افزایش حداکثری معیار مهم‌تر می‌پردازیم. و ما این معیارها را برای بررسی عملکرد سیستم بر روی داده‌های آموزش، توسعه و تست اندازه‌گیری می‌کنیم [۲۶، ۵، ۲۷].

تقسیم‌بندی دادگان به آموزش، توسعه و تست در سرعت توسعه‌ی سیستم بسیار مأثر است. داده‌های توسعه و تست باید از یک توزیع دادگان باشند(دادگانی که هدف ما بهترین عملکرد در مواجه با آنهاست). به این خاطر که ما می‌خواهیم عملکرد تنظیمات پارامترهای سیستم که

بر روی داده‌ی توسعه انجام داده‌ایم، را توسط ارزیابی سیستم بر روی داده‌ی تست بیان کنیم. و در صورتی که این دو داده از دو توزیع متفاوت باشند معیار ارزیابی اعتبار نخواهد داشت. از این رو، برای تقسیم‌بندی داده‌ها به تست و توسعه، دادگان را به طور تصادفی مخلوط [۲] می‌کنیم.

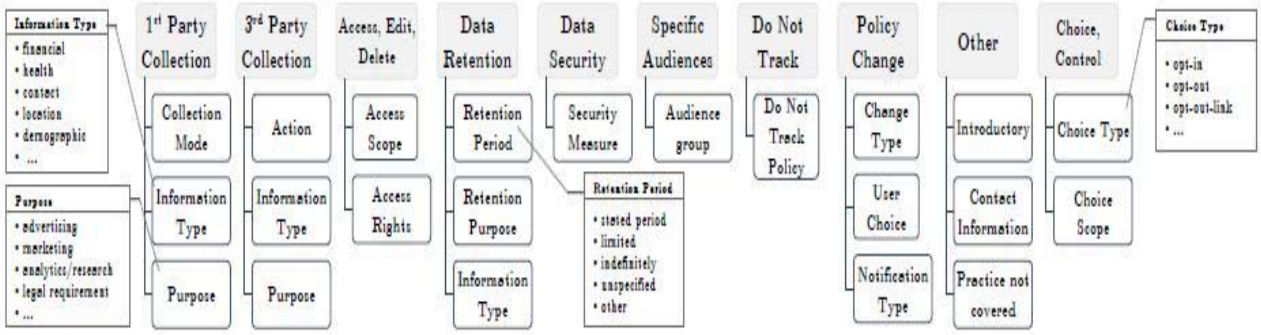
با توجه به اینکه شبکه‌ی عصبی عمیق نیاز به حجم زیادی از داده برای آموزش و توسعه دارد. نسبت تقسیم دادگان به آموزش، تست و توسعه متفاوت از نرخ معمول، ۶۰٪ داده‌ی آموزش، ۲۰٪ داده‌ی توسعه و ۲۰٪ داده‌ی تست است(این نسبت مناسب پایگاه دادگان با حجم حداکثری ۱۰۰۰۰۰ رکورد داده‌است). اما در پایگاه دادگان بزرگ‌تر با نسبت تقسیم داده به صورت ۹۸٪ داده‌ی آموزشی، ۱٪ داده‌ی تست و ۱٪ داده‌ی توسعه عملکرد بهتری خواهد داشت.

اندازه‌ی داده‌ی تست باید به قدر بزرگ باشد که با اطمینان بالایی نشان دهنده‌ی عملکرد سیستم باشد. در صورتی که معیار ارزیابی ما بر روی داده‌های تست و توسعه عملکرد خوبی را نشان می‌دهد ولی سیستم در برابر داده‌های جدید دنیای واقعی عملکرد خوبی ندارد باید معیار ارزیابی و یا داده‌های تست و توسعه را تغییر دهیم.

۲.۴ پایگاه دادگان

با توجه به حجم و پیچیدگی درک متن سیاست‌های حفظ حریم شخصی سایت‌ها، نیاز به یک پایگاه دادگان برای پیاده‌سازی الگوریتم‌های هوش مصنوعی حس می‌شود. بیشتر این نوع از متن‌ها، دارای پیچیدگی‌های بسیار زیاد و کلمات تخصصی زیادی هستند [۲۸]. از این رو، پایگاه دادگان به کار رفته برای یادگیری طبقه‌بندی‌کننده‌های عمیق توسط افراد متخصص (دانشجویان حقوق) به صورت دستی برچسب گذاری شده‌اند. این پایگاه دادگان شامل ۱۱۵ سیاست حفظ حریم شخصی است که در کل ۲۳۰۰۰ برچسب را شامل می‌شود. برچسب گذاری در دو سطح صورت گرفته است. ابتدا، قطعه‌هایی با اندازه‌ی یک پاراگراف بر اساس ۱۰ گروه کلی برچسب گذاری شده‌اند. سپس، قسمت‌های حاصل از مرحله‌ی قبل به صورت جفت ”ویژگی_مقدار“ مانند: نوع اطلاعات: مکان، هدف: تبلیغات و ... برچسب گذاری شده‌اند. به طور کلی ۲۰ ویژگی متفاوت و ۱۳۸ مقدار متفاوت برای ویژگی‌ها وجود دارند. در زیر شکل ۱.۴ تنها قسمتی از ویژگی‌های اجباری هر قطعه که در تمامی قطعات وجود دارند را نمایش می‌دهد [۱۸، ۲۹].

[2] Randomly shuffle



شکل ۱.۴: نمایی کلی از کلاس‌های سوالات پایگاه‌دادگان[۱]

۳.۴ شرح پیاده سازی

با توجه به حجم پایگاه دادگان استاندارد پس از مخلوط کردن تصادفی، دادگان را به نسبت ۶۰% داده‌ی آموزشی، ۲۰% داده‌ی توسعه و ۲۰% داده‌ی تست تقسیم کردیم. سیستم پرسش‌وپاسخ پیاده‌سازی شده در این تحقیق به این صورت عمل می‌کند که ما ابتدا برای آموزش دو طبقه‌بندی کننده‌ی عمیق، اسناد حفظ حریم شخصی را به قسمت‌هایی مشخص (پارگراف‌هایی) تقسیم‌بندی کرد و به شکل کلماتی پشت سر هم برای تبدیل به اعداد و استخراج بردار ویژگی به مدل آماری word2vec داده و بردارهای حاصل را به عنوان آرایه‌ایی از ویژگی به طبقه‌بندی‌کننده‌ی عمیق اول که وظیفه‌ی تشخیص کلاس‌های کلی را دارد، می‌دهیم. سپس فیلترهای پیچشی [۳] که هر یک مشخص کننده‌ی ویژگی خاصی است را بر روی آرایه قرار می‌دهیم، با جابه‌جایی این فیلترها مقادیر حاصل از عمل پیچشی(ضرب ضرایب فیلتر در ویژگی‌ها) را بدست می‌آید سپس با پارامتر بایاس مقادیر را جمع کرده و تابع فعال‌ساز ReLU را اعمال می‌کنیم تا مقادیر ورودی لایه‌ی بعدی شبکه عصبی را بدست آوریم [۳۰، ۳۱]

در لایه بعدی عمل pooling را انجام می‌دهیم و با اعمال Max-pooling ، مهم‌ترین ویژگی را در هر پنجره(بزرگترین عدد هر پنجره را انتخاب می‌کنیم)(هر بار قرار گرفتن فیلتر pooling) مشخص می‌کنیم. (در این لایه با کاهش ابعاد موجب کاهش محاسبات می‌شویم). این لایه هیچ پارامتری برای یادگیری در زمان یادگیری شبکه‌ی عصبی پیچشی ندارد و عملی که در این لایه بر روی پنجره‌ها انجام می‌شود انتخاب ماکزیمم یا متوسط هر پنجره به عنوان یک پارامتر غیرمدلی در ابتدای ساخت شبکه‌ی عصبی تعیین می‌شود و تغییری نمی‌کند.

ابعاد داده پس از اعمال هر لایه از شبکه‌ی عصبی از فرمول (۱.۴) تبعیت می‌کند. هر چه در شبکه‌ی عصبی پیش می‌رویم از طول و عرض داده کم شده و ارتفاع آن (تعداد فیلتر هر لایه) افزایش می‌یابد. این ویژگی شبکه‌ی عصبی باعث می‌شود که با پیشروی در شبکه‌ی

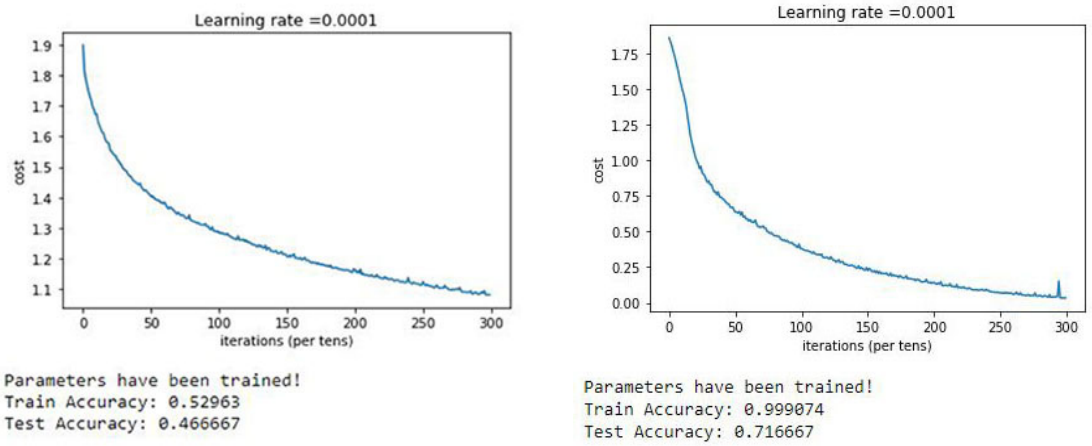
[3] Convolution

عصبی از هزینه‌ی محاسباتی شبکه کم شود و لایه‌های آخر به یک لایه‌ی fully-connected تبدیل شود [۳۲].

$$\left[\frac{n^{[l-1]} - 2p^{[l]} - f}{s^{[l]}} + 1\right] * \left[\frac{n^{[l-1]} - 2p^{[l]} - f}{s^{[l]}} + 1\right] * n_c \qquad (1.4)$$

n_c تعداد فیلتر در لایه l ، s^l میزان جابه‌جایی پنجره‌ی کانولوشن [۴] ، n^{l-1} ابعاد داده‌ی ورودی از لایه‌ی قبل، p میزان حاشیه اضافه شده [۵] قبل از اعمال پنجره‌ی کانولوشن

لایه‌ی بعدی شبکه یک لایه‌ی fully connected است که در یک شبکه‌ی عصبی معمول است و از هر ورودی به تمام نرون‌ها ارتباطی وجود دارد. در لایه‌ی بعدی شبکه دوباره عمل Max-pooling را انجام می‌دهیم و بر روی خروجی، تابع فعال ساز ReLU اعمال می‌کنیم. نهایتاً خروجی از یک لایه‌ی fully connected دیگر عبور کرده و با اعمال تابع فعال ساز softmax احتمال وجود جواب هر یک از کلاس‌های ”پرسش” در این پاراگراف به خصوص بدست می‌آید. ما در مرحله‌ی back propagation یک تابع log likelihood loss را بهینه‌سازی می‌کنیم و مقادیر ضرایب لایه‌های fully connected و مقادیر فیلترهای پیچشی را برای کاهش خطا تغییر می‌دهیم. (محاسبه‌ی مشتق‌ها نسبت به تابع هزینه بنابراین هر چیزی که در مقدار خروجی نهایی تاثیر می‌گذارد باید مشتق غیر صفر داشته باشد) (محاسبه‌ی مشتق‌ها به صورت عقب گرد برای این ورودی به خصوص، که در تابع هزینه تاثیر گذاشته است).



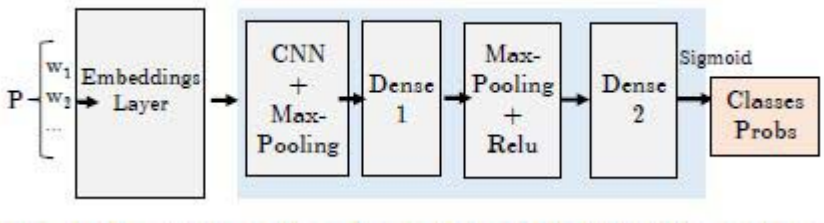
(آ) طبقه‌بندی‌کننده‌ی اول (ب) طبقه‌بندی‌کننده‌ی دوم

شکل ۲.۴: نمودار کاهش هزینه‌ی طبقه‌بندی‌کننده‌ها

در نهایت پس از کاهش خطای شبکه‌ی عصبی پیچشی تا جایی که شیب نمودار ”خطا- تکرار ورود تمام داده‌ها” نزدیک به صفر شد ما به ضرایب / فیلترهای نهایی مدل آماری خود

[4] Stride
[5] Padding

دست یافته‌ایم. این مدل آماری احتمال تعلق پرسش کاربر و پاراگراف‌های سند حفظ حریم شخصی ارائه شده توسط کاربر را به هر یک از کلاس‌های کلی و جزء در پایگاه داده‌های استاندارد مورد استفاده در فاز آموزش که به طور دستی برچسب گذاری شده‌اند، را محاسبه می‌کند [۳۲، ۳۳].



شکل ۳.۴: ساختار طبقه‌بندی‌کننده‌های عمیق [۱]

۴.۴ روش پیشنهادی

در اولین پرسش کاربر، ما محتمل‌ترین جواب را (جواب | سوال)P به کاربر برمی‌گردانیم در صورتی که کاربر در فاصله‌ی زمانی کمی سوال دیگری پرسید که فاصله‌ی تشابه این سوال با سوال قبلی در مدل آماری word2vec کمتر از یک مقدار بخصوص بود و در صورتی که فاصله‌ی احتمال دو محتمل‌ترین جواب از یک مقدار کمتر باشد به جای محاسبه‌ی احتمال جواب بودن تک تک پاراگراف به شرط سوال جدید، دومین محتمل‌ترین جواب را به کاربر برمی‌گردانیم.

نکته: با توجه به این که تعداد کلمات پرسش معمولاً از تعداد کلمات پاراگراف‌ها کمتر است اعداد بدست آمده از مدل word2vec را برای هر پرسش در عددی ضرب کنید تا پاراگراف‌ها و پرسش‌ها دارای وزن اطلاعاتی یکسانی در شبکه‌ی عصبی پیچشی باشند [۱].

S	1
F	3
P	0
Batch size	40
Dense layer size	100
n_c	200

جدول ۱.۴: پارامترهای غیرمدلی استفاده شده برای نتایج زیر [۱]

در ادامه به مقایسه‌ی عملکرد هر یک از قوانین با سیستم بدون تعامل می‌پردازیم:

۱. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، دومین محتمل‌ترین جواب سوال قبلی، جوابی که از نظر مدل آماری با هم مشابه هستند را در صورت کم بودن فاصله‌ی احتمال دو جواب به عنوان جواب به کاربر برگرداند.

	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.77	0.79	0.78	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.77	0.80	0.79	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.68	0.72	0.70	2000	
Data Security	0.70	0.70	0.70	2100		0.79	0.85	0.82	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.88	0.92	0.90	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.90	0.70	0.80	2000	
Policy Change	0.79	0.73	0.75	2000		0.90	0.83	0.87	2000	
Data Retention	0.62	0.51	0.54	2200		0.73	0.64	0.69	2500	
Do Not Track	0.80	0.80	0.80	2100		0.91	0.97	0.92	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.82		0.80

جدول ۲.۴: مقایسه‌ی عملکرد سیستم با قانون ۱

۲. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec بپرسید به جای طبقه‌بندی سوال دوم،از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که از نظر مدل آماری کمترین شباهت را به محتمل‌ترین جواب سوال قبل دارد را به عنوان جواب به کاربر برگرداند.

	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.81	0.79	0.80	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.79	0.80	0.78	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.70	0.72	0.71	2000	
Data Security	0.70	0.70	0.70	2100		0.81	0.85	0.80	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.92	0.92	0.93	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.94	0.70	0.83	2000	
Policy Change	0.79	0.73	0.75	2000		0.93	0.83	0.90	2000	
Data Retention	0.62	0.51	0.54	2200		0.75	0.64	0.69	2500	
Do Not Track	0.80	0.80	0.80	2100		0.93	0.97	0.96	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.82		0.82

جدول ۳.۴: مقایسه‌ی عملکرد سیستم با قانون ۲

۳. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که از نظر مدل آماری بیشترین شباهت از نظر معنایی را با سوال دوم داشته باشد، را به عنوان جواب به کاربر برگرداند.

	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.81	0.75	0.78	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.79	0.76	0.78	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.70	0.70	0.70	2000	
Data Security	0.70	0.70	0.70	2100		0.81	0.81	0.81	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.92	0.88	0.90	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.94	0.66	0.80	2000	
Policy Change	0.79	0.73	0.75	2000		0.93	0.77	0.85	2000	
Data Retention	0.62	0.51	0.54	2200		0.75	0.60	0.68	2500	
Do Not Track	0.80	0.80	0.80	2100		0.93	0.91	0.92	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.80		0.80

جدول ۴.۴: مقایسه‌ی عملکرد سیستم با قانون ۳

۴. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که بیشترین کلمه‌ی مشترک را در مجموعه‌ی کلمات ارتباط معنایی در مدل word2vec با سوال دوم داشته باشد(اندازه‌ی اشتراک مجموعه‌ی کلمات معنایی محتمل‌ترین جواب‌های سوال یک و سوال دوم)، را به عنوان جواب به کاربر برگرداند.

	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.81	0.73	0.77	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.79	0.74	0.77	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.70	0.68	0.69	2000	
Data Security	0.70	0.70	0.70	2100		0.81	0.79	0.80	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.92	0.86	0.89	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.94	0.64	0.80	2000	
Policy Change	0.79	0.73	0.75	2000		0.93	0.75	0.84	2000	
Data Retention	0.62	0.51	0.54	2200		0.75	0.58	0.67	2500	
Do Not Track	0.80	0.80	0.80	2100		0.93	0.89	0.91	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.79		0.79

جدول ۵.۴: مقایسه‌ی عملکرد سیستم با قانون ۴

۵. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که از نظر مدل آماری کمترین شباهت معنایی را با سوال یک داشته باشد، را به عنوان جواب به کاربر برگرداند.

	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.81	0.71	0.76	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.79	0.72	0.76	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.70	0.66	0.68	2000	
Data Security	0.70	0.70	0.70	2100		0.81	0.77	0.79	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.92	0.84	0.88	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.94	0.64	0.79	2000	
Policy Change	0.79	0.73	0.75	2000		0.93	0.73	0.83	2000	
Data Retention	0.62	0.51	0.54	2200		0.75	0.54	0.65	2500	
Do Not Track	0.80	0.80	0.80	2100		0.93	0.87	0.90	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.78		0.78

جدول ۶.۴: مقایسه‌ی عملکرد سیستم با قانون ۵

۶. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، غیرمحتمل‌ترین جواب سوال قبلی، جوابی که از نظر مدل آماری را به عنوان جواب به کاربر برگرداند.

	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.60	0.660	0.60	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.62	0.60	0.61	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.58	0.55	0.56	2000	
Data Security	0.70	0.70	0.70	2100		0.68	0.70	0.69	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.75	0.74	0.84	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.75	0.58	0.66	2000	
Policy Change	0.79	0.73	0.75	2000		0.79	0.67	0.73	2000	
Data Retention	0.62	0.51	0.54	2200		0.62	0.48	0.55	2500	
Do Not Track	0.80	0.80	0.80	2100		0.76	0.78	0.77	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.66		0.66

جدول ۷.۴: مقایسه‌ی عملکرد سیستم با قانون ۶

۷. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب‌های سوال قبلی، جوابی که کمترین کلمه‌ی مشترک را در مجموعه‌ی کلمات ارتباط معنایی در مدل word2vec با سوال اول داشته باشد(اندازه‌ی اشتراک مجموعه‌ی کلمات معنایی محتمل‌ترین جواب‌های سوال یک و سوال یک)، را به عنوان جواب به کاربر برگرداند.

	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.81	0.67	0.74	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.79	0.68	0.74	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.70	0.62	0.66	2000	
Data Security	0.70	0.70	0.70	2100		0.81	0.73	0.77	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.92	0.80	0.76	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.94	0.60	0.77	2000	
Policy Change	0.79	0.73	0.75	2000		0.93	0.69	0.81	2000	
Data Retention	0.62	0.51	0.54	2200		0.75	0.50	0.63	2500	
Do Not Track	0.80	0.80	0.80	2100		0.93	0.83	0.88	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.76		0.76

جدول ۸.۴: مقایسه‌ی عملکرد سیستم با قانون ۷

۸. اگر کاربر دو سوال نزدیک به هم از نظر مدل آماری word2vec پرسید به جای طبقه‌بندی سوال دوم، از بین محتمل‌ترین جواب سوال قبلی، جوابی که کمترین شباهت از نظر معنایی به محتمل‌ترین جواب سوال یک و بیشترین شباهت از نظر معنایی را به سوال دوم دارد، را به عنوان جواب به کاربر برگرداند.

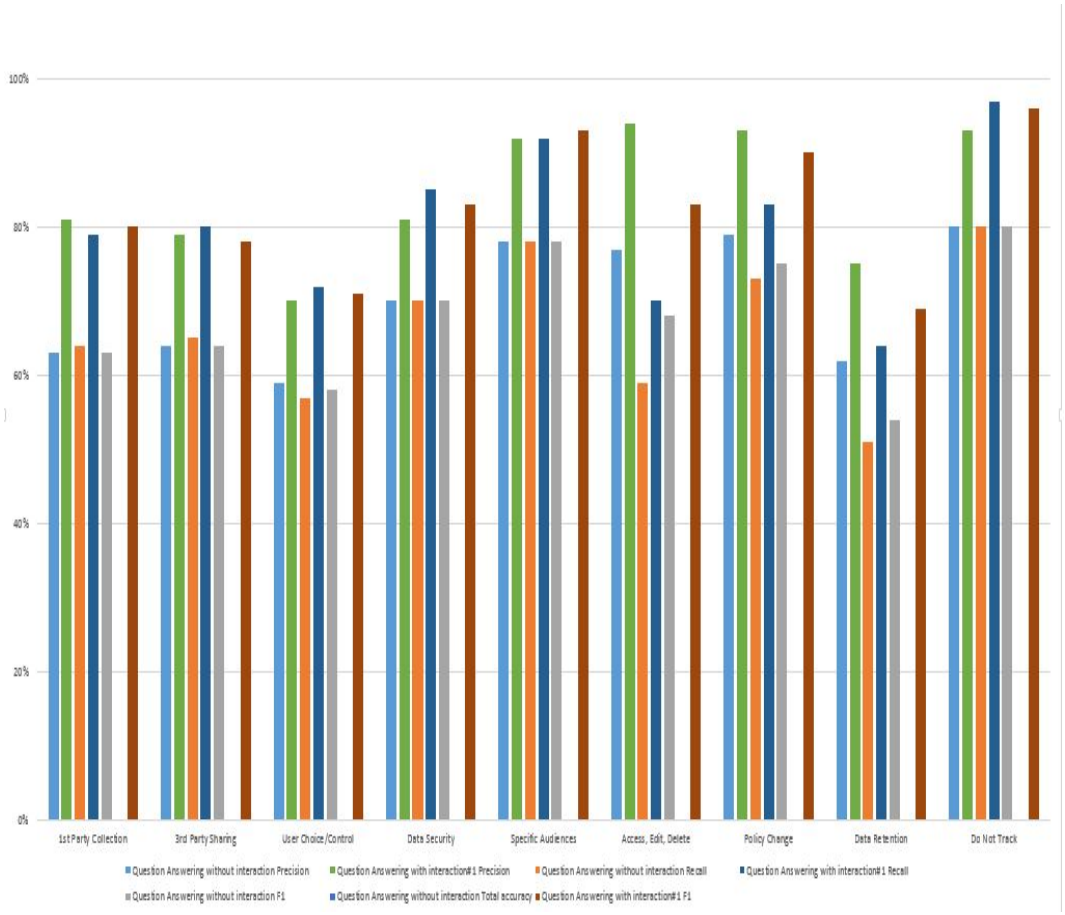
	Question Answering without interaction					Question Answering with interaction				
Category	Precision	Recall	F1	Running Time(ms)	Total accuracy	Precision	Recall	F1	Running Time(ms)	Total accuracy
1st Party Collection	0.63	0.64	0.63	2000		0.81	0.69	0.75	2100	
3rd Party Sharing	0.64	0.65	0.64	2000		0.79	0.70	0.75	2100	
User Choice/Control	0.59	0.57	0.58	2000		0.70	0.64	0.67	2000	
Data Security	0.70	0.70	0.70	2100		0.81	0.75	0.78	2200	
Specific Audiences	0.78	0.78	0.78	2000		0.92	0.82	0.87	2000	
Access, Edit, Delete	0.77	0.59	0.65	2000		0.94	0.62	0.78	2000	
Policy Change	0.79	0.73	0.75	2000		0.93	0.71	0.82	2000	
Data Retention	0.62	0.51	0.54	2200		0.75	0.52	0.64	2500	
Do Not Track	0.80	0.80	0.80	2100		0.93	0.85	0.89	2100	
Average	0.70	0.66	0.67		0.68	0.84	0.80	0.77		0.77

جدول ۹.۴: مقایسه‌ی عملکرد سیستم با قانون ۸

برای بهبود عملکرد یک سیستم آماری ما باید تمرکز خود را در حذف بیشترین عامل خطا در هر زمان از توسعه‌ی آن قرار دهیم. با توجه به عملکرد سیستم قبل از افزایش قابلیت تعامل ما به این نتیجه رسیدم که بخش عمده‌ی خطای سیستم ناشی از عدم درک سوال کاربر یا به عبارت دیگر ابهام در سوال است. این گونه از خطا می‌تواند نأشی از غلط املایی، ترکیب کلمات اشنباه و یا استفاده از کلماتی که با هم قرابت معنایی دارند ولی کاملاً یک مفهوم را بیان نمی‌کنند، نشأت گرفته باشد. و از آنجایی که سیستم پرسش‌وپاسخ در جوابگویی به سوالات درباره‌ی سیاست حفظ حریم شخصی می‌تواند به کاربر اطلاعات غلط یا گمراه‌کننده دهد از میان معیارهای عملکرد فوق F1 معیار بهتری نسبت به دیگر معیارهاست. این معیار دیده بهتری از عملکرد سیستم بر روی داده‌های تست از پیش دیده نشده که از توزیع یکسانی با

داده‌ی توسعه است، را نشان می‌دهد. داده‌های تست و توسعه، داده‌هایی هستند که هدف سیستم پرسش‌وپاسخ ما بهترین عملکرد بر روی آنهاست. با توجه با اینکه، در صورتی که کاربر دو سوال مشابه بپرسد، سیستم مجبور به محاسبه‌ی فاصله‌ی کسینوسی مجموع وزن‌های کلمات دو سوال برای بررسی تشابه آنها است، هزینه زمانی الگوریتم پیشنهادی کمی بیشتر از سیستم پرسش‌وپاسخ بدون تعامل است.

از جدول ۹.۴ می‌توان میزان ابهام و اهمیت هر یک از بخش‌های یک سیاست حفظ حریم شخصی را نیز استنباط کرد. که می‌توان حتی با پیشنهاد سوالات پرتکرار درباره‌ی آنها با افزایش تعامل از ابهام کاست. با توجه به جدول فوق، می‌توان دریافت که کمترین خطای ابهام در نوع سوالات Security Data بوده زیرا با حذف ابهام کمترین رشد دقت را سیستم در پاسخگویی به آنها داشته و بیشترین خطای ابهام در نوع سوالات Access، Edit، Delete وجود داشته، زیرا پس از کاهش ابهام دقت سیستم در مواجه با این نوع از سوالات بیشترین افزایش را داشته است.



شکل ۴.۴: مقایسه‌ی سیستم پرسش‌وپاسخ قبل و بعد از افزودن قابلیت تعامل

فصل ۵

نتیجه گیری و پیشنهادها

در فصل‌های قبل کارهای انجام شده را شرح داده و همچنین نتایج حاصل از روش پیشنهادی را تشرح کردیم و به مقایسه‌ی این نتایج با نتایج تحقیق مشابه در این زمینه پرداختیم. در این فصل از کارهای انجام شده نتیجه‌گیری می‌کنیم و پیشنهادهایی برای آیندگان که قصد کار در زمینه‌ی سیستم‌های پرسش‌وپاسخ را دارند خواهیم کرد.

۱.۵ نتیجه‌گیری

سیستم پرسش‌وپاسخ پیاده‌سازی شده توسط ابزار موثر و قوی شبکه‌ی عصبی عمیق پیچشی عملکرد خوبی دارند. رشد و بهبود سیستم‌های پرسش‌وپاسخ با توجه به رشد روزافزون نیاز به دسترسی سریع به اطلاعات قابل پیش‌بینی است. این سیستم‌ها با صرفه‌جویی در زمان و نیروی انسانی در پاسخگویی به افراد، توجیه اقتصادی بالایی دارند. با این حال رویکردهای پیاده‌سازی این سیستم‌ها از یکدیگر فاصله گرفته و سعی در رفع نقایص یکدیگر ندارند. در این تحقیق با در نظرگرفتن تاریخچه‌ی نزدیک سوالات کاربر به یک نوع ساده ولی موثر از تعامل میان کاربر و سیستم دست پیدا کردیم؛ که نتایج زیر را به دنبال داشت:

۱. قسمت زیادی از ابهام سوالات برای سیستم کاهش پیدا کرده‌است.

۲. سادگی و کم بودن حافظه‌بری طرح پیشنهادی، آن را در مقیاس بزرگتر ممکن می‌سازد.

۳. با توجه به نتایج حاصل شده دقت سیستم در روش پیشنهادی نسبت به سیستم بدون تعامل حدود ۸٪ بهبود داشته است.

۴. رفتار طبیعی کاربر در تغییر ترتیب واژگان و واژگان سوال خود در مواجه با جواب نادرست عملکرد سیستم را بهبود می‌بخشد.

۲.۵ پیشنهادات

شبکه‌های عصبی مانند دیگر روش‌های آماری برای ایجاد مدلی کارا(دارای بهترین عملکرد در برابر داده‌های جدید) وابسته به پایگاه دادگان هستند. در این بخش پیشنهاداتی برای آیندگان که قصد کار در زمینه‌ی سیستم‌های پرسش‌وپاسخ را دارند مطرح می‌شود:

۱. عملکرد سیستم ما با داشتن مدل آماری word2vec که توسط دادگان مخصوص یک موضوع خاص مانند: سیاست‌های حفظ حریم شخصی آموزش دیده‌اند و وزن (بردار ویژگی‌های) کلمات را در آن موضوع خاص به طور دقیق‌تری بیان می‌کنند، بهبود چشم‌گیری خواهد داشت.

۲. لازمه‌ی بهبود عملکرد شبکه‌ی عصبی و آموزش تخصصی مدل word2vec دارا بودن پایگاه دادگان بسیار بزرگ است.

۳. روش‌های استاندارد تنظیم پارامترهای غیرمدلی شبکه‌ی عصبی، این ابزار را به ابزار ساده‌تری تبدیل می‌کند.

۴. با توجه به تغییر رفتار کاربران و اسناد حفظ حریم شخصی به صورت دوره‌ایی، بهتر است مدل آماری شبکه‌های عصبی عمیق را به طور دوره‌ایی و با داده‌های بیشتر و بروزتر توسط روش انتقال یادگیری آموزش دهیم که این کار مستلزم جمع‌آوری داده از کاربران در طول یک دوره می‌باشد.

۵. استفاده از ترکیبی خطی از قوانین استخراج بهترین جواب با ضرایبی متناسب با میزان بهبود عملکرد هر قانون نسبت به سیستم پرسش‌وپاسخ فاقد تعامل ایجاد می‌کند.

هر چه میزان حجم داده‌های در دسترس مختص یک مسئله خاص بیشتر باشد در توسعه‌ی یک شبکه‌ی عصبی ما نیاز کمتری به hand-Engineering خواهیم داشت. به عبارت دیگر، با افزایش داده‌های در دسترس برای یک مسئله، با الگوریتم ساده‌تری، عملکرد بهتری خواهیم داشت.

مراجع

[1] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548, 2018.

[2] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. 1990:1, 2009.

[3] Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. Survey on challenges of question answering in the semantic web. *Semantic Web*, 8(6):895–920, 2017.

[4] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, Yuvraj Agarwal, Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. How Short Is Too Short ? Implications of Length and Framing on the Effectiveness of Privacy Notices This paper is included in the Proceedings of the Implications of Length and Framing on the Effectiveness of Privacy Notices. *the Symposium On Usable Privacy and Security (SOUPS)*, (Soups):321–340, 2016.

[5] Hamza Harkous, Kassem Fawaz, Kang G Shin, and Karl Aberer. PriBots: Conversational Privacy with Chatbots,. *Workshop on the Future of Privacy Indicators [June 22–24, 2016, Denver, Colorado]*, pages 1–6, 2016.

[6] Hector Llorens, Nathanael Chambers, Naushad Uzzaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. Evaluating Temporal Information Understanding with Question Answering. *Ieee*, (SemEval):792–800, 2015.

[7] Ali S. Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. [Paper] Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.

[8] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. Automated Analysis of Privacy Requirements for Mobile Apps. 3066(132):286–296, 2017.

[9] Jinfeng Rao, Hua He, and Jimmy Lin. Noise-Contrastive Estimation for Answer Selection with Deep Neural Networks. pages 1913–1916, 2016.

[10] Ellen M. Voorhees. Overview of the TREC 2003 Question Answering Track. *Trec*, 2003:54–68, 2003.

[11] K.M. Sathyendra, F. Schaub, S. Wilson, and N. Sadeh. Automatic extraction of opt-out choices from privacy policies. *2016 AAAI Fall Symposium Series*, pages 270–275, 2016.

[12] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. 2:1–54, 2006.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. 2016.

[14] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[15] Hamza Harkous, Kassem Fawaz, Kang G Shin, and Karl Aberer. PriBots: Conversational Privacy with Chatbots,. *Workshop on the Future of Privacy Indicators [June 22–24, 2016, Denver, Colorado]*, pages 1–6, 2016.

[16] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. A Design Space for Effective Privacy Notices*. *The Cambridge Handbook of Consumer Privacy*, pages 365–393, 2018.

[17] Fei Liu, Nicole Lee Fella, and Kexin Liao. Modeling Language Vagueness in Privacy Policies Using Deep Neural Networks. *2016 AAAI Fall Symposium Series*, pages 257–263, 2016.

[18] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. The Creation and Analysis of a Website Privacy Policy Corpus. pages 1330–1340, 2016.

[19] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online. *the*

Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS 2016), (Soups):77–96, 2016.

[20] Nicolas Viennot, Edward Garcia, and Jason Nieh. PlayDrone: A measurement study of google play. *@a@Sigmetrics*, pages 221–233, 2014.

[21] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004.

[22] Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. Analyzing Vocabulary Intersections of Expert Annotations and Topic Models for Data Practices in Privacy Policies. *AAAI Fall Symposium on Privacy and Language Technologies*, pages 264–269, 2016.

[23] Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. Analyzing Vocabulary Intersections of Expert Annotations and Topic Models for Data Practices in Privacy Policies. *AAAI Fall Symposium on Privacy and Language Technologies*, pages 264–269, 2016.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[25] Florian Schaub and Pascal Knierim. Drone-based Privacy Interfaces: Opportunities and Challenges. *SOUP 2016 - Twelfth Symposium on Usable Privacy and Security*, 2016.

[26] Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (June):22–32, 2007.

[27] J Richard Landis and Gary G Koch. Landis_Jr__Koch_Gg_1977_Kappa_and_Observer_Agreement. *Biometrics*, 33(1):159–174, 1977.

[28] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. Identifying the Provision of Choices in Privacy Policy Text. pages 2774–2779, 2018.

[29] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah Smith, and Frederick Liu. Crowdsourcing Annotations of Websites' Privacy Policies: Can It Really Work? *Www*, pages 133–143, 2016.

[30] Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. Applying deep learning to answer selection: A study and an open task. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, pages 813–820, 2016.

[31] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised Text Segmentation Using Semantic Relatedness Graphs. pages 125–130, 2016.

[32] Yoon Kim. Convolutional Neural Networks for Sentence Classification. 2014.

[33] Fei Liu, Nicole Lee Fella, and Kexin Liao. Modeling Language Vagueness in Privacy Policies Using Deep Neural Networks. *2016 AAAI Fall Symposium Series*, pages 257–263, 2016.

پیوست آ

معرفی معیارهای ارزیابی عملکرد

معیار های ارزیابی الگوریتم هایی که منجر به مدل های آماری به عنوان یک پیش بینی کننده با توجه به ذات مسئله از درجه ی اهمیت متفاوتی نسبت به هم برخوردار هستند. به عنوان مثال در مسائل تقسیم بندی افراد به عنوان تروریست یا شهروند معمولی، تشخیص افراد مبتلا به یک ویروس خطرناک که قادر به شیوع سریع است از افراد سالم و یا تشخیص یک پول شویی، حتی مقدار کم اشتباه تخمین برچسب بسیار می تواند تاثیرگذار باشد. بنابراین صرفه توجه به ''دقت کلی مدل ''معیار ارزشمندی برای عملکرد الگوریتم نیست.
در ادامه به معرفی چند معیار ارزیابی مدل های آماری منتج از داده می پردازیم. نهایتاً به دلایل انتخاب یک معیار مشخص برای تخمین عملکرد سیستم پرسش وپاسخ اشاره می کنیم.

آ.۱ دقت

میزان کل تخمین صحیح مدل به کل تخمین های مدل، را دقت [۱] کلی مدل گویند.

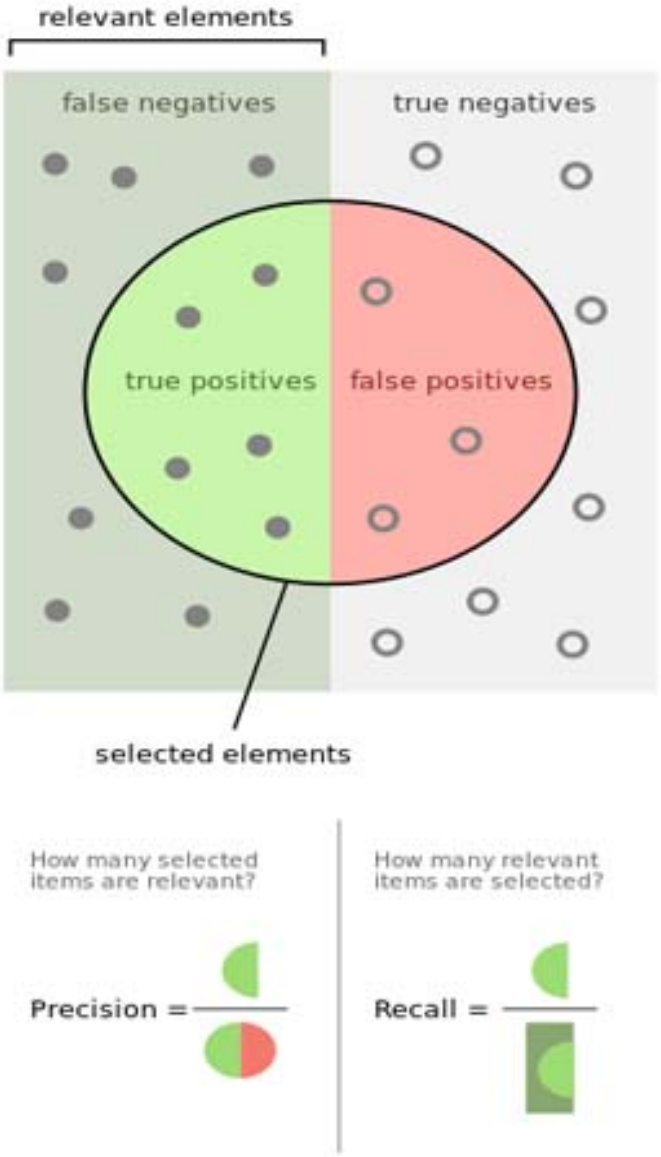
[1] Accuracy

آ.۲ Precision

نسبت تخمین درست برچسب مثبت به کل تخمین های برچسب مثبت. این معیار یک سنجه ی خوب از عملکرد سیستم در زمانی که تخمین غلط برچسب مثبت هزینه ی بسیاری دارد. به عنوان مثال در سیستم تشخیص ایمیل های تبلیغاتی (spam (تخمین غلط برچسب مثبت می تواند منجر به از دست دادن اطلاعات مهمی برای کاربر شود و یک ایمیل غیر تبلیغاتی به عنوان spam تشخیص داده خواهد شد.

آ.۳ Recall

این معیار میزان تخمین صحیح برچسب مثبت در تخمین برچسب مثبت را توسط مدل محاسبه می کند. این معیار بهترین مدل را زمانی که تخمین غلط برچسب منفی هزینه ی بسیار زیادی دارد، نشان می‌دهد. به عنوان مثال در تشخیص پول شویی یا فرد بیمار این معیار استفاده می شود. در صورت برچسب گذاری یک تراکنش پولشویی به عنوان یک تراکنش معمولی(برچسب غلط منفی) برای بانک نتایج فوق العاده بدی را خواهد داشت. به طور مشابه در صورتی که یک آزمایش فرد بیماری را سالم تشخیص دهد. هزینه ی این برچسب گذاری غلط در صورتی که بیماری مسری باشد بسیار بالاست.



شکل آ.۱: نمایش بصری دو معیار ارزیابی

F۱ ۴.آ

این معیار یک تعادل را بین دو معیار قبلی در نظر می گیرد(میانگین وزن دار Recall و Precision). این معیار زمانی مناسب است که ما به دنبال یک تعادل میان Recall و Precision باشیم و توزیع داده های ما یکنواخت نباشد.

با توجه به اینکه سیستم پرسش وپاسخ پیاده سازی شده در این پایان نامه تنها وظیفه ی اطلاع رسانی و برگرداندن قسمتی از سیاست حفظ حریم شخصی سایت را دارد. در نتیجه هزینه ی اشتباه در تخمین برچسب قسمت های سیاست حفظ حریم شخصی تنها ’’زمان

صرف شده توسط کاربر'' و ''نارضایتی آن از دریافت اطلاعات مفید'' خواهد بود. کاربران در برخورد با چنین حالتی با تغییر سوال، طرح چند سوال بیشتر و یا به مطالعه ی کامل حفظ حریم شخصی می ـ پردازند. بنابراین با توجه به مسئله بهترین معیار برای ارزیابی سیستم F1-score خواهد بود. پس با بهبود این این معیار ما درک دقیق تری از عملکرد سیستم داریم.

Abstract

Privacy policies are the primary channel through which companies inform users about their data collection and sharing practices. In their current form, policies remain long and difficult to comprehend, thus merely serving the goal of legally protecting the companies. Short notice based on information extracted from privacy policies have been shown to be useful and more usable. Therefore, we deal with new information retrieval systems to meet this need. Question answering systems are examples of such ones which use natural language to make human-computer interaction much easier.

Performance of statistical question answering systems are so dependent to their data-set with which they are trained; however, they outperform other approach of implementation of question answering. Deep convolutional neural network equipped with interaction will consider previous question and similarity of (past and new questions) questions while answering client. Thus, they omit majority of ambiguity errors. Improvement are resulted by word2vec model which can identify similarity of meaning of two sentences.

The result of this research implementation can achieve an accuracy of 82.4% on this task, when evaluated against unseen test set annotated by a group of law department student. The obtained scores indicate that the proposed method is almost 8% more accurate than automated analysis of privacy policy using only deep learning. In addition, the processing time of the proposed method compared with the hierarchy of neural network classifiers is much improved.

keywords: Question answering system; Deep convlutional neural network ; Question answering with interaction

Shahrood University of Technology

Faculty Of Computer Engineering

MSc Thesis in: Artificial Intelligence Engineering

Automatic Question Answering System Based on Statistical Translation Models

By: Majid Amiri

Supervisor

Morteza Zahedi

January 2020