باسمه تعالی

درس پردازش آماری زبان‌های طبیعی

Name: Mohammad Akbarshahi

student number: 4021541501

موضوع : معرفی task های مرتبط در حوزه پردازش آماری زبان های طبیعی

- Tex calassification

طبقه‌بندی متن (Text Classification) یکی از وظایف اصلی در زمینه پردازش زبان طبیعی (NLP) است که بر اساس یادگیری ماشین (ML) انجام می‌شود تا به صورت خودکار متون را به دسته‌بندی‌های مشخص شده تخصیص دهد. این فرآیند با تجزیه و تحلیل محتوای متون و شناسایی الگوهای کلیدی صورت می‌گیرد.طبقه‌بندی متون برای کاربردهای گسترده‌ای استفاده می‌شود، از جمله:

-تجزید و تحلیل حالات عاطفی: برای شناسائی نظرات لحاظی، منفی یا خنثی در یک نوشتار .

برچسب‌گذاری موضوع :برای دسته‌بندی اسناد یا پست‌های وب بر اساس موضوع آنها .

تشخیص هرزنامه :برای شناسایی ایمیل‌ها یا پیام‌های هرزنامه .

طبقه‌بندی اسناد :برای سازماندهی اسناد به صورت خودکار بر اساس نوع آنها .

شناسایی اسپم :برای شناسایی محتوای اسپم در وب سایت ها یا انجمن های آنلاین .

- Token classification

به تشخیص و دسته‌بندی توکن‌ها یا واحدهای کوچک متنی از متن می‌پردازد. توکن‌ها می‌توانند واژه‌ها، عبارات، علائم نگارشی، اعداد و سایر مؤلفه‌های متن باشند.

در کاربردهای مختلفی از token classification استفاده می‌شود، از جمله تشخیص موجودیت‌ها، پردازش زبان طبیعی، ترجمه ماشینی، تحلیل احساسات و زبان شناسی محاسباتی.

برای انجام token classification، از مدل‌های یادگیری عمیق و شبکه‌های عصبی با توانایی آموزش بر اساس داده‌های برچسب‌خورده استفاده می‌شود. این مدل‌ها در مراحل آموزش به تعداد زیادی داده و توکن‌های برچسب‌خورده آموزش داده می‌شوند تا بتوانند با دقت بالا توکن‌های مختلف را دسته‌بندی کنند.

به عنوان مثال، در تشخیص موجودیت‌ها، token classification به تشخیص دادن انواع موجودیت‌ها مانند افراد، سازمان‌ها، مکان‌ها و غیره در یک متن کمک می‌کند. این اطلاعات می‌تواند برای مسائلی مانند استخراج اطلاعات، تشخیص رویدادها، تحلیل علت و معلولیت‌ها و خلاصه سازی متن مورد استفاده قرار گیرد.

- Table Question Answering

Table Question Answering (Table QA) یک حوزه مهم در زمینه پردازش زبان طبیعی است که به تشخیص و پاسخ به سوالاتی که مرتبط با داده‌های جداول هستند، می‌پردازد. در Table QA، هدف اصلی تفسیر و استخراج اطلاعات مورد نیاز از جداول و پاسخ دادن به سوالات متنی است که نیاز به دسترسی به اطلاعات موجود در جداول دارند.

برای انجام Table QA، باید ابتدا اطلاعات موجود در جدول به‌طور مناسب تفسیر و استخراج شوند. سپس با استفاده از مدل‌های یادگیری ژرف و شبکه‌های عصبی، سوالات مطرح شده درباره اطلاعات جدولی تحلیل و پاسخ داده می‌شود. این اطلاعات می‌توانند شامل متن، اعداد، توکن‌ها و سایر ویژگی‌های موجود در جداول باشند.

Table QA در بسیاری از حوزه‌ها مانند پزشکی، علوم اجتماعی، علوم رایانه و غیره مورد استفاده قرار می‌گیرد. این تکنیک می‌تواند به دسترسی به اطلاعات مفید و سریع از داده‌های جداول کمک کرده و سطح دانش و اطلاعات برنامه‌ها و سیستم‌ها را افزایش دهد.
برخی از چالش‌های Table QA شامل تفسیر و استخراج دقیق اطلاعات از جداول، تفهیم سوالات مفهومی و نحوه ارتباط داده‌های جدول با سوالات می‌باشد که توسعه الگوریتم‌ها و رویکردهای بهینه برای حل این چالش‌ها از اهمیت بالایی برخوردار است.

- Question Answering

Question Answering (QA) یک حوزه مهم در زمینه پردازش زبان طبیعی است که مورد توجه زیادی قرار گرفته است. هدف اصلی QA، پاسخ دادن به سوالات مطرح شده از طریق تحلیل و فهم محتوای متنی و ارتباط دادن آن با اطلاعات موجود است.
در QA، مدل‌های یادگیری عمیق و شبکه‌های عصبی به کار گرفته می‌شوند تا بتوانند سوالات پیچیده را از داده‌های متنی استخراج کرده و به آن‌ها پاسخ دهند. این پاسخ‌ها می‌توانند شامل عبارت‌های متنی، اعداد، تاریخ، آدرس و غیره باشند و به طور خلاصه و جامع مطابق با سوال ارائه شوند.
QA می‌تواند در بسیاری از حوزه‌ها از جمله جستجوی اطلاعات، خدمات مشتریان، خلاصه‌سازی متن، تفهیم مفاهیم و آموزش الکترونیکی مورد استفاده قرار بگیرد. اهمیت QA به عنوان یکی از شاخه‌های پردازش زبان طبیعی از آن جهت است که این تکنولوژی به بهبود تجربه کاربری، افزایش دقت و سرعت در ارائه اطلاعات و بهبود بهره‌وری کمک می‌کند.
چالش‌های QA شامل فهم و تفسیر دقیق معنای سوالات، تطابق اطلاعات با سوال، مدیریت دانش و اطلاعات موجود و توانایی ارائه پاسخ منطقی و معقول برای سوال می‌باشد که نیازمند توسعه مدل‌ها و روش‌های پیچیده‌تر و هوشمندتر در زمینه پردازش زبان طبیعی است.

- Zero-Shot Classification

Zero-shot classification یک مسئله در یادگیری ماشین است که به دسته‌بندی داده‌ها از یک مجموعه دسته‌های موجود، بدون داشتن داده‌های آموزشی برای دسته‌های مورد نظر اشاره دارد. در واقع، مدل باید بتواند برچسب یا دسته‌بندی مناسب برای داده‌هایی که قبلاً در آموزش یافته نشده‌اند، پیش‌بینی کند.
روش‌های توسعه داده شده برای zero-shot classification اغلب از تعاریف داده‌ها و ویژگی‌های آن‌ها در فضای نهان (embedding space) استفاده می‌کنند. این روش‌ها معمولاً از ترکیب مدل‌های مبتنی بر یادگیری ماشین مانند شبکه‌های عصبی و مدل‌های توجه برای برچسب‌گذاری داده‌های zero-shot استفاده می‌کنند.
یکی از موارد استفاده zero-shot classification می‌تواند از آن در مسئله تشخیص تصاویر با کلاس‌های جدید باشد. برای مثال، اگر یک مدل یادگیری عمیق با دسته‌های "گربه" و "سگ" آموزش دیده باشد، اما بعداً با یک تصویر از "پنگوئن" روبرو شود (که در مجموعه داده آموزشی نبوده است)، مدل باید بتواند پنگوئن را تشخیص دهد و دسته‌بندی کند.
از آنجایی که zero-shot classification نیازمند توانایی تعریف و استفاده از دامنه‌ها و روابط میان دسته‌ها برای پیش‌بینی دسته‌های جدید است، توسعه روش‌های هوشمندانه و بهبود استخراج ویژگی‌های مناسب برای دسته‌بندی داده‌ها از اهمیت بالایی برخوردار است.

- Translation

Translation در زمینه پردازش زبان طبیعی به مفهوم ترجمه متن یا متون از یک زبان به زبان دیگر اشاره دارد. این فرایند اهمیت زیادی دارد و مورد استفاده در زمینه‌های مختلفی از جمله محتواهای وب، متون علمی، متون تجاری، مکاتبات بین‌المللی، ارتباطات و مسائل دیگر می‌باشد.
با توجه به پیچیدگی زبان‌ها و تفاوت‌های فرهنگی و اجتماعی بین زبان‌ها، انجام ترجمه یک وظیفه چالش‌برانگیز است. در سال‌های اخیر، از تکنولوژی‌های هوش مصنوعی و یادگیری عمیق برای توسعه سیستم‌های ترجمه خودکار بهره گرفته شده است. مثال‌های برجسته این نوع سیستم‌ها شامل Google Translate، Microsoft Translator و سایر سیستم‌های مشابه می‌باشند که قدرتمندی و دقت خوبی برای ترجمه‌ی متون ارائه می‌دهند.
از آنجا که ترجمه یک وظیفه پیچیده‌است و نیازمند دقت و صحت بالایی می‌باشد، توسعه‌ی روش‌های پیچیده تر و بهبود یافته در زمینه مدل‌های یادگیری عمیق و پردازش زبان طبیعی می‌تواند به بهبود عملکرد سیستم‌های ترجمه خودکار کمک کند.

- Summarization

خلاصه‌نگاری به فرایند فشرده‌سازی و ابراز نکات اصلی یا اطلاعات کلیدی از یک متن در یک فرم کوتاهتر اشاره دارد، همچنین اهمیت اصلی و پیام موجود در متن را حفظ می‌کند. این روش به طور گسترده در زمینه‌های مختلفی از جمله روزنامه‌نگاری، تحقیقات، آموزش و خلق محتوا مورد استفاده قرار می‌گیرد تا نسخه مختصر و قابل فهمتری از متون بلند فراهم شود.
خلاصه‌نگاری خودکار، با استفاده از هوش مصنوعی و پردازش زبان طبیعی، به عنوان یک روش محبوب در سال‌های اخیر به دلیل قابلیت آن برای تولید خلاصه‌هایی از متون بزرگ به صورت سریع و دقیق شناخته شده است. این فناوری با تجزیه و تحلیل محتوای متن، شناسایی جملات یا عبارات مهم را و سپس ترکیب آن‌ها به یک خلاصه مرتب و همگون انجام می‌دهد.
با بهره‌گیری از الگوریتم‌های پیشرفته یادگیری ماشین و تکنیک‌های پیشرفته پردازش زبان طبیعی، سیستم‌های خلاصه‌نگاری خودکار می‌توانند به طور موثر متون را در حوزه‌ها و زبان‌های مختلف خلاصه کنند. این سیستم‌ها ابزارهای ارزشمندی برای محققان، دانشجویان، خلق‌کنندگان محتوا و هر

کسی که به دنبال دسترسی سریع به نکات اصلی متون بلند باشد محسوب می‌شوند. توسعه و بهبود مستمر این فناوری‌ها اهمیت دارد تا دقت و عملکرد سیستم‌های خلاصه‌نگاری خودکار بهبود یابد.

- Feature Extraction

استخراج ویژگی‌ها (Feature Extraction) یک فرایند مهم در پردازش سیگنال‌ها و داده‌ها است که هدف آن انتخاب و استخراج ویژگی‌های مهم و تاثیرگذار از داده‌ها برای استفاده در الگوریتم‌های یادگیری ماشین و تحلیل داده است. این ویژگی‌ها معمولاً مشخصه‌های کلیدی از داده را نمایان می‌کنند که می‌توانند برای تفکیک و تمایز بین داده‌ها و همچنین برای پیش‌بینی و تصمیم‌گیری استفاده شوند.
در این فرایند، داده‌های اولیه به صورت پیچیده یا بسیار جزئی داده می‌شوند، اما با استفاده از الگوریتم‌ها و روش‌های مختلف استخراج ویژگی، ویژگی‌های معنی‌دار و ارزشمند از داده‌ها برای استفاده در مراحل بعدی پردازش و تحلیل به دست می‌آید. این ویژگی‌ها می‌توانند از ابعاد مختلفی مانند فرکانس، زمان، محل، شکل، و غیره برای نمایان کردن خصوصیات مهم داده استفاده کنند.
استخراج ویژگی‌ها برای بهبود عملکرد الگوریتم‌های یادگیری ماشین و کمک به دستیابی به نتایج دقیق‌تر و موثرتر از داده‌ها بسیار حیاتی است. این مرحله پس از استخراج ویژگی‌ها، داده‌ها را به یک فضای ویژگی معنادار تبدیل کرده و قابل استفاده برای وظایف مختلفی مانند تصویربرداری، تشخیص الگو، دسته‌بندی و پیش‌بینی می‌کند.

- Text Generation

تولید متن یا "Text Generation"، فرایندی است که در آن سیستم‌های کامپیوتری به طور خودکار متن‌هایی با ساختار و محتوایی شبیه به زبان انسانی تولید می‌کنند. این فرایند شامل استفاده از الگوریتم‌ها و مدل‌ها برای تولید متنی منطقی و مناسب با توجه به داده‌های ورودی مانند گزینه‌ها، کلمات کلیدی یا متن‌های موجود می‌شود. تولید متن می‌تواند در زمینه‌های مختلفی از جمله پردازش زبان طبیعی، ایجاد محتوا، چت‌بات‌ها و غیره مورد استفاده قرار گیرد.
رویکردهای مختلفی برای تولید متن وجود دارد که شامل روش‌های مبتنی بر قوانین، مدل‌های آماری و مدل‌های مبتنی بر شبکه‌های عصبی مانند شبکه‌های عصبی بازگشتی (RNN) و ترنسفورمرها می‌شود. این مدل‌ها بر روی مجموعه‌داده‌های بزرگ متنی آموزش داده می‌شوند تا الگوها، ساختارها و معنا در زبان را یاد بگیرند و این امر به آنان امکان تولید متن جدیدی که به زبان انسانی شبیه است را می‌دهد.
تولید متن می‌تواند برای کاربردهای مختلفی مانند ایجاد نوشته‌های خلاقانه، مقالات خبری، توضیحات محصول، دیالوگ برای چت‌بات‌ها و غیره استفاده شود. این قابلیت همچنین می‌تواند نویسندگان، ایجادکنندگان محتوا و کسب‌وکارها را در تولید محتوا به سرعت و به صورت کارآمد کمک نماید.
کلیهٔ اینها نشان می‌دهد که تولید متن یک زمینهٔ پویا با کاربردهای متعدد می‌باشد و پیشرفت‌هایی که در حوزهٔ یادگیری ماشین و پردازش زبان طبیعی انجام می‌شود، کیفیت و قابلیت‌های مدل‌های تولید متن را بهبود می‌بخشد.

- Text2Text Generation

تولید متن به متن یا "Text2Text Generation" یک حوزه از تکنولوژی پردازش زبان طبیعی است که در آن مدل‌های یادگیری ماشین برای تولید متنی خروجی با توجه به متن ورودی آموزش دیده می‌شوند. این روش به طور کلی شامل تبدیل یک نوع متن به یک نوع دیگر از متن است، بدون نیاز به داده‌های جداگانه برای هر نوع ورودی و خروجی.
مدل‌های Text2Text می‌توانند برای وظایف مختلفی مانند ترجمه ماشینی، خلاصه‌سازی متون، تولید سوال-جواب و تولید متن بر اساس سوالات مورد استفاده قرار گیرند. این حوزه از تولید متن در چند سال اخیر به دلیل پیشرفت‌های قابل توجه در حوزه یادگیری عمیق و شبکه‌های عصبی و همچنین استفاده از مدل‌های ترنسفورمر، پردازش‌های زبانی مرتبط و مدل‌های ترجمه با قدرت بالا، به محبوبیت بالایی دست یافته است.
استفاده از مدل‌های Text2Text می‌تواند به شرکت‌ها، تولید کنندگان محتوا، تحقیقات علمی و همچنین افراد علاقمند به تولید محتوا کمک کند تا متونی مناسب و با کیفیت را به سرعت و به صورت خودکار تولید کنند.

- Fill-Mask

تکنیک پرکردن ماسک یا "Fill-Mask" یک روش پرکردن خودکار و محلی متن در یادگیری ماشین است که در فرایند پردازش زبان طبیعی استفاده می‌شود. این روش از مدل‌های یادگیری ماشین برای تشخیص و جایگزینی بخش‌های خالی در یک جمله با محتوای مناسب و منطقی استفاده می‌کند. در فرآیند پرکردن ماسک، یک یا چند مکان در جمله به عنوان "ماسک" یا خالی در نظر گرفته می‌شود. سپس مدل یادگیری ماشین با توجه به کلمات موجود در جمله و ساختار آن، به انتخاب و درج کلمات مناسب در جایگاه ماسک می‌پردازد. این فرآیند به مدل‌ها کمک می‌کند تا درک عمیق‌تری از متن داشته باشند و خودکارا متن‌های معنی‌داری ایجاد کنند.
روش پرکردن ماسک یکی از مهمترین کاربردهای فرایند تولید متن، به صورت معمول نیز در مدل‌های زبانی مانند BERT و GPT-3 مورد استفاده قرار می‌گیرد. این تکنیک می‌تواند در بسیاری از وظایف مانند تصحیح خطا، تکمیل متن، تولید متن از جملات نیمه کامل و غیره مورد استفاده قرار گیرد و به بهبود تولید متون با کیفیت کمک کند.

- Sentence Similarity

مفهوم شباهت جملات یک مفهوم مهم در حوزه پردازش زبان طبیعی است که به تعیین درجه شباهت و همبستگی بین دو جمله یا متن مختلف می‌پردازد. این مفهوم در انواع وظایف پردازش زبان مورد استفاده قرار می‌گیرد، از جمله تشخیص ابراز همه‌گونه، خلاصه‌سازی متون، بازیابی اطلاعات و ترجمه ماشینی.

روش‌های مختلفی برای محاسبه و تعیین شباهت جملات وجود دارد، از جمله محاسبه فاصله کیسه‌کلماتی، محاسبه امتیازات شباهت مبتنی بر مدل‌های برداری کلمات (مانند Word2Vec یا GloVe) و استفاده از شبکه‌های عصبی عمیق مبتنی بر کدگذار-کدگشا مانند مدل Universal Sentence Encoder.

با استفاده از این روش‌ها، می‌توان شباهت جملات را به صورت عددی یا بعضا به شکل یک بردار نمایش داد و بر اساس آن‌ها، جملات مشابه یا همبستگی با یکدیگر را تعیین کرد. این قابلیت می‌تواند در وظایف مختلفی مانند بازیابی اطلاعات دقیق‌تر، تفکیک داده‌ها، تشخیص ابراز همه‌گونه و بهبود ترجمه ماشینی مورد استفاده قرار گیرد.