سیده شکیبا انارکی فیروز ۹۹۴۴۲۰۴۷

تمرین دوم درس پردازش زبان ها

(a(1

:One-hot Encoding

در این روش، هر واژه یا توکن با یک بردار دودویی (binary vector) با اندازه برابر تعداد کل واژگان در مجموعه داده نمایش داده می‌شود.

برای هر واژه، یک بردار از صفر و یک با طول برابر با تعداد کل واژگان در نظر گرفته می شود، به طوری که مقدار یک فقط در اندیس مربوط به آن واژه یا توکن باشد و سایر اندیس ها صفر باشند.

این روش برای مسائلی که تعداد واژگان محدود است و ماتریس نمایش داده می‌شود مناسب است، اما برای متون بلند و با تعداد واژگان زیاد، باعث ایجاد ماتریس های بسیار بزرگ و پر از صفر می شود که به تغییر مسئله به یک مسئله بزرگتر منجر می‌شود.

:Word Embedding

در این روش، هر واژه با یک بردار عددی چگال (dense vector) نمایش داده می‌شود که از طریق آموزش مدل های عمیق (مثل مدل های شبکه های عصبی) بر اساس ساختار متن و همبستگی معنایی کلمات در متون بزرگ، به دست می‌آید.

در مقایسه با One-hot encoding، که بردارهای خنثی برای هر واژه ایجاد می‌کند، word embedding بردارهایی با ابعاد کمتر و با اطلاعات معنایی فراوان فراهم می‌کند.

Word embedding مزایایی مانند کاهش ابعاد، نیاز به حافظه ی کمتر و محاسبات آسان تر، نمایش معنایی مشترک بین کلمات مشابه، و قابلیت انتقال دانش از مدل های پیش آموزش دیده مانند GloVe ،Word2Vec و FastText را فراهم می‌آورد.

(b(1

در مدل GloVe (Global Vectors for Word Representation)، ابتدا یک ماتریس شمارش تجربی بر اساس همسایگی کلمات ساخته می‌شود. این ماتریس نشان‌دهنده تعداد بارهایی است که هر کلمه در متن با سایر کلمات همراه ظاهر شده است.

سپس، یک تابع هدف برای کاهش فاصله بین embedding های مرتبط با کلماتی که همبستگی بیشتری با یکدیگر دارند، تعریف می‌شود. این تابع هدف با استفاده از ماتریس شمارش تجربی و معیاری مانند معنای مشترک بین کلمات (با استفاده از احتمال شرطی برای ایجاد تابع)، بهینه سازی می‌شود.

پس از آن، با استفاده از روش های بهینه سازی مانند Gradient Descent , embedding ها به‌روزرسانی می شوند تا فاصله بین embedding های کلمات معنایی با هم کاهش یابد و شرایط مرتبط با ماتریس شمارش تجربی حفظ شود.

در نهایت، word embedding های نهایی برای کلمات به دست می‌آید که در آنها اطلاعات معنایی و همبستگی با سایر کلمات نگهداری می‌شود. به این ترتیب، GloVe توانایی ایجاد word embedding هایی را دارد که اطلاعات همبستگی و معنایی بین کلمات را به خوبی نمایش دهد.

(c(1

در مدل Word2Vec، از یک شبکه عصبی کم عمق استفاده می‌کند که بر روی یک مجموعه متنی بزرگ آموزش دیده شده است تا متن اطراف یک کلمه داده شده (context words) را پیش‌بینی کند.ماهیت Word2Vec در توانایی آن برای تبدیل کلمات به بردارهای با ابعاد بالا نهفته است. این نمایش به الگوریتم اجازه می‌دهد تا معنا، شباهت معنایی و روابط با متن اطراف را دربرگیرد. یک ویژگی قابل توجه از Word2Vec، قابلیت انجام عملیات محاسباتی با این بردارها برای آشکارسازی الگوهای زبانی است، مانند تساوی مشهور "پادشاه - مرد + زن = ملکه" همچنین کلماتی با معانی یا زمینه های مشابه با بردارهایی نشان داده می شوند که در این فضا به هم نزدیکتر هستند.

مراحل انجام:

• آموزش در مجموعه متن بزرگ:
مدل Word2Vec بر روی مجموعه بزرگی از داده های متنی مانند مقالات ویکی پدیا یا مجموعه ای از کتاب ها آموزش داده می شود. در طول آموزش، مدل به هر کلمه در زمینه کلمات اطراف خود در یک اندازه پنجره تعریف شده نگاه می کند.
• آموزش Word Representations :
Word2Vec از (CBOW) یا معماری Skip-gram برای یادگیری بازنمایی کلمات استفاده می کند.
CBOW: این رویکرد کلمه مورد نظر را بر اساس کلمات متن آن پیش بینی می کند. مجموعه ای از کلمات زمینه را به عنوان ورودی می گیرد و سعی می کند کلمه مورد نظر را پیش بینی کند.
Skip-gram: در مقابل، Skip-gram کلمات متنی را پیش بینی می کند که یک کلمه هدف داده شده است. یک کلمه هدف را به عنوان ورودی می گیرد و سعی می کند کلمات بافت اطراف آن را پیش بینی کند.
• نمایش برداری با ابعاد بالا:
مدل یاد می گیرد که هر کلمه را به عنوان یک بردار با ابعاد بالا (معمولاً 100 تا 300 بعد) نشان دهد. این بردارها روابط معنایی بین کلمات را بر اساس الگوهای همزمانی آنها در داده های آموزشی دریافت می کنند.
• آموزش شبکه عصبی:
Word2Vec از یک شبکه عصبی کم عمق با یک لایه پنهان استفاده می کند. در طول آموزش، مدل وزن این شبکه عصبی را تنظیم می کند تا تفاوت بین کلمات متن پیش بینی شده و واقعی را به حداقل برساند.

(d(1
برخی از چالش ها و رویکردهای مقابله با آنها:
چند معنایی یا polysemy :
چند معنایی هنگامی پدید می آید که در آن یک کلمه معانی متعددی دارد. به عنوان مثال، "bank" می تواند به یک موسسه مالی یا کنار رودخانه اشاره کند.
یک رویکرد برای رسیدگی به چندمعنی، استفاده از اطلاعات زمینه است. با در نظر گرفتن کلمات اطراف یا زمینه ای که یک کلمه در آن ظاهر می شود، word embedding می تواند معانی مختلفی را بر اساس زمینه ای که کلمه در آن استفاده می شود دریافت کند و از word sense embeddings استفاده کند.
گنگ بودن یا ambiguity:
ابهام زمانی رخ می دهد که یک کلمه بسته به زمینه، تعابیر متفاوتی داشته باشد. به عنوان مثال، "crane" می تواند به یک پرنده یا یک دستگاه بالابر اشاره کند.
برای رفع ابهام(WSD)،word embedding را می توان بر روی مجموعه داده های بزرگ و متنوع آموزش داد تا زمینه های مختلفی را که در آنها یک کلمه ظاهر می شود، به تصویر بکشد. علاوه بر این، از تکنیک های sense disambiguation می توان برای ابهام زدایی از معنای یک کلمه بر اساس contect آن یا part of speech استفاده کرد.
اطلاعات متنی:
کلمات اغلب بر اساس زمینه ای که در آن به کار می روند، معانی متفاوتی پیدا می کنند. به عنوان مثال، "apple" می تواند به یک میوه یا یک شرکت فناوری اشاره کند.
Contextual Word embeddings، مانند آنهایی که توسط مدل هایی مانند BERT ایجاد میشوند، معانی کلمات را بر اساس بافت اطراف کل جمله به جای کلمات همسایه به تصویر میکشند. این به آن ها اجازه می دهد تا کلمات با معانی متعدد را بهتر مدیریت کنند.
پراکندگی داده ها:
کلمات با معانی متعدد ممکن است به اندازه کافی در داده های آموزشی برای مدل های word embedding برای یادگیری دقیق برای هر معنی ظاهر نشوند.

تکنیک هایی مانند جاسازی زیرکلمه ها، که کلمات را به عنوان ترکیبی از واحدهای کوچکتر مانند n-gram یا تک واژها نشان می‌دهند، می توانند با گرفتن subword های معنی دار برای کلمات نادر یا دیده نشده، به کاهش پراکندگی داده ها کمک کنند.

(e(1

subword embeddings

یک راه برای حل مشکل واژگان خارج از دایره لغات (out of vocabulary words) در فرایند تولید embedding کلمات، استفاده از تکنیک های مبتنی بر subword embeddings می باشد. این روش ها به واژگانی که در داده های آموزشی موجود نیستند، امکان می دهند تا با استفاده از قسمت های کوچکتری از آنها که در داده های آموزشی موجود است، embedding متناظر با آنها را ایجاد کنند.

یکی از روش های معمول در این زمینه استفاده از مدل های بر پایه subword مانند Byte Pair Encoding (BPE) یا WordPiece است. این مدل‌ها کلمات را به قسمت‌های کوچکتر تجزیه می‌کنند و embedding برای هر قسمت ایجاد می‌کنند. سپس با ترکیب این embedding ها، embedding کلمه مورد نظر تولید می‌شود.

Special Unknown Token

رویکرد اول برای کنترل توکن های نامعلوم، استفاده از یک توکن ویژه<unk> برای واژگان خارج از دایره لغت است. این توکن به مدل اجازه می‌دهد تا یک نمایندگی مناسب برای این واژگان را یاد بگیرد. همچنین، از یک رویکرد مبتنی بر توکن استفاده می‌شود که کاراکترها یا زیرواژگان را به عنوان توکن ها در نظر می‌گیرد. این روش می‌تواند به مدل کمک کند تا حتی با واژگان OOV کار کند.

Context-based Embedding Prediction

رویکرد دیگر مبتنی بر پیش‌بینی مبتنی بر زمینه است که از مدل هایی مانند GPT ،ELMo، یا BERT استفاده می کند. این مدل ها از داده های آموزشی بزرگ آموزش دیده شده اند و embeddingهای واژگان را بر اساس متن مجاور در جمله (contex)محاسبه می‌کنند. همچنین، می توان با تنظیم این مدل ها بر روی وظایف یا حوزه های خاص، به مدل ها کمک کرد تا به بهترین شکل ممکن با واژگان OOV کار کنند.

(2

I love computer science and I love NLP even more.

	<start>	I	love	comput er	scienc e	and	NLP	even	more	<end>
<start>	0	1	1	0	0	0	0	0	0	0
I	1	0	2	1	1	1	1	0	0	0
Love	1	2	0	1	1	1	1	1	0	0
Compu ter	0	1	1	0	1	1	0	0	0	0
science	0	1	1	1	0	1	0	0	0	0
and	0	1	1	1	1	0	0	0	0	0
NLP	0	1	1	0	0	0	0	1	1	0
even	0	0	1	0	0	0	1	0	1	1
More	0	0	0	0	0	0	1	1	0	1
<end>	0	0	0	0	0	0	0	1	1	0

منابع:

https://www.elastic.co/what-is/word-embedding
https://stats.stackexchange.com/questions/278936/word2vec-that-can-distinguish-words-with-different-meanings
https://www.geeksforgeeks.org/word-sense-disambiguation-in-natural-language-/processing
https://aclanthology.org/2022.coling-1.350.pdf
https://datascience.stackexchange.com/questions/26943/how-to-initialize-word-embeddings-for-out-of-vocabulary-word