به نام خالق رنگین کمان

ستاره باباجانی – گزارش تمرین شماره 3

سوال 1 تئوری:

الف)

1. تنوع زمینه: موجودیتها میتوانند در زمینه های مختلف ظاهر شوند که شناسایی آنها را دشوار میسازد. برای مثال، واژه" Apple" میتواند به میوه، یک شرکت یا حتی یک آلبوم موسیقی اشاره کند، بسته به زمینهای که در آن به کار رفته است.
2. ابهام در مرزهای موجودیت: تعیین اینکه یک موجودیت کجا شروع و کجا تمام میشود در متن میتواند پیچیده باشد، به خصوص برای موجودیتهایی که از چندین کلمه تشکیل شده اند (موجودیتهای چند واژه ای) یا زمانی که موجودیتها دارای موجودیتهای دیگری در درون خود هستند.
3. چند معنایی و هم نامی: کلماتی که به یک شکل نوشته میشوند میتوانند بر اساس کاربردشان معانی مختلفی داشته باشند. برای مثال، " Jaguar"میتواند به یک حیوان، برند خودرو یا نام نرم افزار اشاره کند. تشخیص دادن این معانی نیاز به درک زمینه های دارد که در آن ظاهر شده اند.

4. موجودیتهای خاص دامنه: موجودیتها میتوانند به شدت خاص یک حوزه یا دامنه خاص باشند. به عنوان مثال، اسناد پزشکی یا حقوقی ممکن است شامل اصطلاحات تخصصی باشند که مدلهای NER استاندارد آموزش دیده بر روی متون عمومی ممکن است آنها را شناسایی نکنند.
5. فقدان انسجام در انتساب: مجموعه داده های مختلف ممکن است دستورالعمل های مختلفی برای اینکه چه چیزی یک موجودیت محسوب میشود و چگونه باید موجودیتها انتخاب شوند داشته باشند. این ناسازگاری میتواند آموزش را پیچیده تر کند زیرا مدل باید یا به چندین استاندارد سازگار شود یا به طور جداگانه بر روی داده هایی با انتساب های سازگار آموزش ببیند.
6. چالش های چند زبانه و فرا زبانی: سیستمهای NER اغلب نیاز دارند که با متونی در چندین زبان کار کنند که این موضوع پیچیدگی را به دلیل ویژگیهای خاص زبانی و منابع به همراه دارد. قوانین شناسایی موجودیت در یک زبان ممکن است به طور مستقیم به زبان دیگری ترجمه نشوند و داده های انتخاب شده در برخی زبانها ممکن است کم باشند.
7. موجودیت های نادر و در حال ظهور: موجودیت های جدید مانند اصطلاحات اخیرا محبوب، نام های تجاری یا افراد میتوانند پس از

آموزش یک مدل ظاهر شوند. این موجودیت ها میتوانند توسط مدل هایی که به دادههای قبلا دیده شده تکیه دارند، نادیده گرفته شوند.

ب)

1. کیفیت متن: کیفیت کلی متن، از جمله وضوح، صحت دستوری و املای آن، تأثیر قابل توجهی بر عملکرد سیستمهای NER دارد. متون نوشته شده به طور نادرست با اشتباهات دستوری یا نحو مبهم میتوانند مدلهای NER را گیج کنند که منجر به شناسایی نادرست موجودیت ها میشود.
2. زبان تخصصی دامنه ای: متون مربوط به دامنه های خاص (مانند حقوقی، پزشکی یا فنی) اغلب حاوی واژگان و اصطلاحات تخصصی هستند. سیستمهای NER که بر روی متن های عمومی آموزش دیده اند ممکن است در شناسایی موجودیت های خاص دامنه مشکل داشته باشند مگر اینکه به طور خاص بر روی داده های دامنه ای آموزش دیده یا تنظیم شده باشند.
3. غنای زمینه ای: زمینه ای که موجودیت ها در آن ظاهر میشوند، میتواند بر شناسایی موجودیت ها تأثیر زیادی بگذارد. متونی که زمینه های غنی و واضحی فراهم میکنند به بهبود ابهام زدایی از موجودیت ها کمک میکنند. برای مثال، تمیز دادن بین "شرکت Apple" در مقابل "میوه سیب" نیاز به سرنخ های زمینه ای دارد که

سیستمهای NER باید آنها را به درستی تشخیص دهند و تفسیر کنند.

4. طول و ساختار متن: متون طولانی تر ممکن است زمینه بیشتری فراهم کنند اما همچنین میتوانند پیچیدگی هایی را از نظر روابط موجودیت ها و وقوع آنها در سراسر سند ایجاد کنند. ساختار متن، مانند عناوین، زیرعناوین و لیست ها، میتواند سرنخ هایی را به سیستم های NER در مورد دسته های احتمالی موجودیت ها ارائه دهد که ممکن است در متن غیرساختاری کمتر واضح باشد.
5. استفاده از اختصارات و مخفف ها: متونی که به طور مکرر از اختصارات، مخفف ها یا سایر اشکال کلمات کوتاه شده استفاده میکنند، نیاز به دانش قبلی یا اطلاعات زمینه ای کافی دارند تا بتوانند این فرم های کوتاه را به درستی به عبارت های کامل خود مرتبط کنند. به طور کلی، ویژگی های متن به طور مستقیم بر کارایی و دقت سیستم های NER تأثیر میگذارند. برای دست یابی به دقت بالا، سیستم های NER باید مقاوم، قابل تطبیق و بر روی مجموعه داده های با کیفیت بالا، متنوع و خاص دامنه آموزش دیده باشد.

ج)

1. مدل سازی وابستگی ها: HMMها بطور ذاتی فرض میکنند که هر حالت (یا برچسب خروجی) در توالی فقط به حالت قبلی وابسته است

(این به عنوان خاصیت مارکوف شناخته میشود). این میتواند محدودکننده باشد زیرا این فرضیه اغلب بیش از حد ساده است؛ بسیاری از وظایف مدلسازی توالی از درک زمینه گسترده تر در توالی بهره مند میشوند. CRF ها این محدودیت را ندارند. آنها میتوانند حالت فعلی را بسته به کل توالی داده های ورودی مدل کنند، بنابراین وابستگی ها و زمینه پیچیده تری را درک میکنند.

2. انعطاف پذیری ویژگی ها: HMM ها معمولا به احتمالات ثابت انتقال حالت تکیه دارند و فقط میتوانند از حالت فوری قبلی و مشاهده کنونی برای پیش بینی حالت بعدی استفاده کنند. اما CRF ها اجازه میدهند از انواع و تعداد متعددی از ویژگی های ورودی برای هر حالت در توالی استفاده شود. این بدان معناست که CRFs میتوانند مجموعهای غنی تر از اطلاعات را در نظر بگیرند، مانند حضور کلمات خاص یا سایر عوامل زمینه ای، که برای وظایفی مانند NER بسیار حیاتی است.
3. استقلال خروجی: HMM ها فرض میکنند که مشاهدات (یا خروجی ها) با توجه به توالی حالت مستقل هستند. این فرض در بسیاری از سناریو های دنیای واقعی که خروجی ها ممکن است تحت تأثیر برچسب ها یا ویژگی های مجاور باشند، خوب عمل نمیکنن. CRFs این مشکل را با شرطی کردن هر خروجی بر اساس کل توالی ورودی و نه فقط حالت های کنونی یا قبلی حل میکنند، که به مدل اجازه

میدهد از وابستگی ها بین برچسب ها در قسمت های مختلف توالی برای انجام پیش بینی های دقیق تر استفاده کند. CRFs انعطاف پذیری و قدرت بیشتری در مدل سازی وابستگیهای پیچیده در مقایسه با HMM ها ارائه میدهند. آنها برای رسیدگی به نوع اطلاعات زمینه ای و ویژگیه ای غنی که در بسیاری از وظایف NLP مدرن نیاز است، مناسبتر هستند.

د)

- Atlanta/NNP
- dinner/NN
- have/VBP
- Can/MD

ه) برچسب گذاری BIO با استفاده از یک طرح ساده با سه نوع برچسب کار میکند:

- B(Beginning): این برچسب نشان دهنده شروع یک موجودیت نامدار است. این برچسب با پسوندی همراه است که نوع موجودیت را مشخص میکند، مانند PER-B برای شروع نام یک شخص.
- I(Insoide): این برچسب برای توکن هایی استفاده میشود که در داخل یک موجودیت نامدار قرار دارند اما شروع آن نیستند. مانندB ، این

برچسب شامل پسوندی است که نوع موجودیت را نشان میدهد، مانند PER-I برای توکنی که در داخل نام یک شخص است.

- O(Outside): این برچسب برای توکن هایی استفاده میشود که به هیچ موجودیت نامداری تعلق ندارند.

طرح BIO به مدل ها کمک میکند تا بین موجودیت هایی که مجاور هم هستند اما بخشی از یک موجودیت واحد نیستند، تمایز قائل شوند. به عنوان مثال، در عبارت" Apple Inc Apple Inc. CEO Steve" ،" .Jobs"و " Jobs Steve" دو موجودیت جداگانه هستند که به صورت B-ORG I-ORG O B-PER I-PER برچسب گذاری میشوند، نشان دهنده این است که "Apple Inc." یک موجودیت و"Jobs Steve" موجودیت دیگری است.

تفاوت ها از برچسب گذاریIO:

برچسبگذاریIO ، که مخفف Outside-Inside است، یک طرح ساده تر است که فقط از دو برچسب استفاده میکند:

- I: برای توکن هایی استفاده میشود که بخشی از یک موجودیت نامدار هستند.
- O: برای توکن هایی استفاده میشود که به هیچ موجودیت نامداری تعلق ندارند.

معایب اصلی طرح IO عدم توانایی آن در تمیز دادن بین موجودیت های مختلف که مجاور هستند یا موجودیت هایی که از نوع یکسان هستند اما متمایز هستند، میباشد. به عنوان مثال، در عبارت "Apple Inc. Steve Jobs" برچسب گذاری IO نمیتواند نشان دهد که کجا "Apple Inc." به پایان میرسد و "Jobs Steve" شروع میشود، اگر آنها از نوع موجودیت یکسان باشند.

تفاوت ها از برچسبگذاریBIOES:

برچسبگذاریBIOES ، که همچنین به عنوان BMEWO یا BMEWO+ شناخته میشود، دو برچسب دیگر به طرح BIO اضافه میکند تا مرزهای واضح تری ارائه دهد: B(Beginning) و I (Inside) که به طور مشابه به کاربردشان در برچسب گذاری BIO عمل میکنند.E(End)، این برچسب نشان دهنده پایان یک موجودیت نامدار است و به روشن کردن اینکه یک موجودیت کجا متوقف میشود کمک میکند، به ویژه در مواردی که موجودیت هایی از نوع یکسان مجاور هستند.S(Single) این برچسب برای موجودیتی که فقط شامل یک توکن است استفاده میشود، که موجودیت های تک توکنی را از موجودیت های چندتوکنی متمایز میکند.O(Outside) همچنان بدون تغییر باقی میماند. این تفکیک بیشتر در برچسب گذاری BIOES به مدلها کمک میکند تا ساختار موجودیت ها را، به ویژه در موارد پیچیده که موجودیت ها مجاور یا همپوشانی دارند، به درستی تفسیر کنند.

به طور خلاصه، در حالی که برچسب گذاری BIO روشی را برای رسیدگی به موجودیت ها در توالی ها با وضوح مناسب فراهم میکند، برچسب گذاری BIOES حتی کنترل و وضوح بیشتری را ارائه میدهد، که در متون پیچیده بسیار مفید است. برچسب گذاریIO ، که ساده ترین است، فاقد پیچیدگی لازم برای رسیدگی به ساختارهای موجودیت مجاور یا پیچیده است.

سوال 1 عملی: اگر کلمه‌ای در tag_dict_ یافت شود، برچسب رایج‌ترین (آنی که بالاترین تعداد را دارد) به پیش‌بینی اختصاص داده می‌شود. اگر کلمه‌ای در tag_dict_ یافت نشود (یعنی یک کلمه ناشناخته است)، قوانین ابتکاری بر اساس پایان یا ویژگی‌های کلمه برای حدس زدن برچسب اعمال می‌شوند:

- 'VBG'برای کلماتی که به 'ing' ختم می‌شوند (که نشان دهنده فعل به شکل اسم مصدر است).
- 'NP$'برای کلماتی که به "'s" ختم می‌شوند (که نشان دهنده اسم ملکی است).
- 'NNS'برای کلماتی که به 's' ختم می‌شوند اما نه به) 'ss' احتمالاً اسم جمع.(
- 'RB'برای کلماتی که به 'ly' ختم می‌شوند (معمولاً قید).

- 'VBN'برای کلماتی که به 'ed' ختم می‌شوند (که نشان دهنده شکل ماضی نقلی فعل است).
- 'JJ'برای کلماتی که حاوی زیررشته‌هایی مانند'ble' ، 'ish'، 'ful' هستند (نشانگر صفت).
- 'CD'برای رشته‌های عددی (اعداد اصلی).
- 'NP'برای کلماتی که حرف اول آن‌ها بزرگ است (که نشان دهنده اسم خاص است)، مگر اینکه کل کلمه با حروف بزرگ نوشته شده باشد (برای جلوگیری از در نظر گرفتن اختصارات به عنوان اسم‌های خاص).
- 'NN'به عنوان پیش‌فرض برای هر کلمه ناشناخته‌ای که در دسته‌های بالا نمی‌گنجد استفاده می‌شود.

خلاصه عملکرد این تابع به طور مؤثری داده‌های آموخته شده tag_dict_) از آموزش (را با قوانین ابتکاری برای رسیدگی به کلماتی که در زمان آموزش دیده نشده‌اند ترکیب می‌کند. این رویکرد اجازه می‌دهد تا برچسب‌گذاری مستحکم‌تری انجام شود، به‌خصوص در وظایف NLP که برخورد با کلمات جدید یا نادر رایج است. این روش با حدس زدن برچسب‌ها بر اساس مرفولوژی کلمه و الگوهای زبانی مستقر، روش‌های پایه‌ای جستجو در فرهنگ لغت را بهبود می‌بخشد، در نتیجه دقت کلی فرایند برچسب‌گذاری را افزایش می‌دهد.

سوال 2 عملی:

- توکن‌سازی و برچسب‌گذاری نقش کلمات:
  - کد: این سلول احتمالاً شامل پردازش متن با استفاده از تکنیک‌های توکن‌سازی و برچسب‌گذاری نقش کلمات است، احتمالاً با استفاده از کتابخانه‌ای مانند NLTK یا spaCy .
  - خروجی: لیست‌هایی از توکن‌ها و برچسب‌های نقش کلمات مربوطه نمایش داده می‌شود. لیست‌هایی از برچسب‌های پیش‌بینی‌شده و برچسب‌های 'پنهان' واقعی برای ارزیابی وجود دارد.
- محاسبه معیارهای ارزیابی:
  - کد: این سلول به نظر می‌رسد که محاسبه مختلف معیارهای ارزیابی مانند مثبت‌های واقعی(TP) ، منفی‌های واقعی(FP) ، مثبت‌های کاذب(TN) ، و منفی‌های کاذب (FN) برای هر برچسب نقش کلمات را انجام می‌دهد.
  - خروجی: یک دیکشنری از معیارها برای دسته‌های مختلف برچسب‌ها مانند اسم‌ها، افعال، صفات و غیره نشان داده شده است که نشان‌دهنده عملکرد مدل برای هر دسته برچسب است.
- تحلیل خطاها:

- کد: این سلول برچسب‌های نقش کلماتی که بیشترین تعداد پیش‌بینی‌های نادرست را دارند را شناسایی می‌کند و نقاطی را که مدل ضعیف عمل می‌کند نشان می‌دهد.
- خروجی: به طور خاص نام برچسب‌های نقش کلمات که بیشترین تعداد پیش‌بینی‌های کاذب را دارند را می‌دهد.

- عملکرد کلی مدل:
  - کد: احتمالاً خلاصه‌ای از عملکرد کلی مدل شامل محاسبه دقت وجود دارد.
  - خروجی: تعداد کل نمونه‌هایی که به درستی توسط مدل پیش‌بینی شده‌اند و درصد دقت کلی نمایش داده شده است.

سوال 3 عملی: الف) مجموعه داده داده شده شامل اطلاعاتی درباره 1000 فیلم برتر رتبه بندی شده توسط IMDB است. هر ورودی حاوی جزئیاتی مانند نام فیلم، سال انتشار، رتبه بندی، ژانر، و کارگردان و غیره است. برخی از مشکلاتی که name entity ها میتوانند منجر به آن شوند به شرح زیر است:

1. ابهام و کلمات رایج: برخی از عناوین فیلم ها از کلمات یا عبارات رایج تشکیل شده اند (مثل "Love", "Home", "The End") که می توانند اغلب در متن کلی در زمینه های مختلف ظاهر شوند. هنگامی که

سیستم NER این کلمات رایج را به عنوان عناوین فیلم در زمینه‌های نامرتبط شناسایی می‌کند، می‌تواند منجر به مثبت کاذب شود.

2. تغییرپذیری در نامگذاری: عناوین فیلم ها می توانند نسخه های مختلفی داشته باشند یا با عناوین جایگزین در مناطق یا زبان های مختلف شناخته شوند. این تغییرپذیری می‌تواند تشخیص همه موارد ذکر شده از یک فیلم را برای سیستم NER دشوار کند.

3. شخصیت‌ها و قالب‌بندی خاص: عناوین با شخصیت‌های خاص یا قالب‌بندی منحصربه‌فرد (مانند " Star Wars: Episode IV - A New Hope")ممکن است به روش‌های مختلفی در متون نوشته شوند، که تشخیص مداوم را به چالش می‌کشد.

4. عناوین با تاریخ و اعداد: عناوینی که شامل تاریخ یا اعداد هستند (مانند "1984"، "Apollo 13") ممکن است با داده های عددی واقعی یا سال های خاص در متن اشتباه گرفته شوند.

5. عناوین کوتاه: عناوین بسیار کوتاه، به ویژه آنهایی که از یک کلمه تشکیل شده اند، به دلیل تکرار مکرر آنها به عنوان کلمات عادی در متن، می توانند مشکل ساز شوند.

6. ارتباط فرهنگی و زمانی: محبوبیت و شناخت عناوین فیلم ها می تواند در طول زمان تغییر کند، که ممکن است بر میزان احتمال ذکر آنها

در متون معاصر تأثیر بگذارد. همچنین، ارتباط فرهنگی می تواند متفاوت باشد و بر شناخته شدن یا ارجاع عناوین در مناطق مختلف تأثیر بگذارد.

ب) نوت بوک گفته شده تکمیل شد.