ملیکا محمدی فخار - ۹۹۵۲۲۰۸۶

۱.
الف.
۱. تنوع زمینه: موجودیت‌ها می‌توانند در زمینه‌های 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.

۲. 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 شده‌اند (موجودیت‌های چند واژه‌ای) یا زمانی که موجودیت‌ها دارای موجودیت‌های دیگری در درون خود هستند.

۳. چندمعنایی و همنامی: 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 نرم‌افزار اشاره کند. تشخیص دادن این معانی نیاز به درک زمینه‌ای دارد که در آن ظاهر شده‌اند.

۴. 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 آموزش‌دیده بر روی متون عمومی ممکن است آن‌ها را شناسایی نکنند.

۵. فقدان انسجام در انتساب: مجموعه‌داده‌های مختلف ممکن است دستورالعمل‌های مختلفی برای اینکه چه چیزی یک موجودیت محسوب می‌شود و چگونه باید موجودیت‌ها انتخاب شوند داشته باشند. این ناسازگاری می‌تواند آموزش را پیچیده‌تر کند زیرا مدل باید یا به چندین استاندارد سازگار شود یا به طور جداگانه بر روی داده‌هایی با انتساب‌های سازگار آموزش ببیند.

۶. چالش‌های چندزبانه و فرازبانی: 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 داده‌های انتخاب‌شده در برخی زبان‌ها ممکن است کم باشند.

۷. موجودیت‌های نادر و در حال ظهور: موجودیت‌های جدید مانند اصطلاحات اخیراً محبوب، نام‌های تجاری یا افراد می‌توانند پس از آموزش یک مدل ظاهر شوند. این موجودیت‌ها می‌توانند توسط مدل‌هایی که به داده‌های قبلاً دیده شده تکیه دارند، نادیده گرفته شوند.

ب.
۱. 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 موجودیت‌ها می‌شود.

۲. زبان تخصصی دامنه‌ای: متون مربوط به دامنه‌های خاص (مانند حقوقی، پزشکی یا فنی) اغلب حاوی واژگان و اصطلاحات تخصصی هستند. سیستم‌های NER که بر روی متن‌های عمومی آموزش دیده‌اند ممکن است در شناسایی موجودیت‌های خاص دامنه مشکل داشته باشند مگر اینکه به طور خاص بر روی داده‌های دامنه‌ای آموزش دیده یا تنظیم شده باشند.

۳. غنای زمینه‌ای: زمینه‌ای که موجودیت‌ها در آن ظاهر می‌شوند، می‌تواند بر شناسایی موجودیت‌ها تأثیر زیادی بگذارد. متونی که زمینه‌های غنی و واضحی فراهم می‌کنند به بهبود ابهام‌زدایی از موجودیت‌ها کمک می‌کنند. برای مثال، تمیز دادن بین "شرکت Apple" در مقابل "میوه سیب" نیاز به سرنخ‌های زمینه‌ای دارد که سیستم‌های NER باید آن‌ها را به درستی تشخیص دهند و تفسیر کنند.

۴. طول و ساختار متن: متون طولانی‌تر ممکن است زمینه بیشتری فراهم کنند اما همچنین می‌توانند پیچیدگی‌هایی را از نظر روابط موجودیت‌ها و وقوع آن‌ها در سراسر سند ایجاد کنند. ساختار متن، مانند عناوین، زیرعناوین و لیست‌ها، می‌تواند سرنخ‌هایی را به سیستم‌های NER در مورد دسته‌های احتمالی موجودیت‌ها ارائه دهد که ممکن است در متن غیرساختاری کمتر واضح باشد.

۵. استفاده از اختصارات و مخفف‌ها: متونی که به طور مکرر از اختصارات، مخفف‌ها یا سایر اشکال کلمات کوتاه‌شده استفاده می‌کنند، نیاز به دانش قبلی یا اطلاعات زمینه‌ای کافی دارند تا بتوانند این فرم‌های کوتاه را به درستی به عبارت‌های کامل خود مرتبط کنند.

به طور کلی، ویژگی‌های متن به طور مستقیم بر کارایی و دقت سیستم‌های NER تأثیر می‌گذارند. برای دستیابی به دقت بالا، سیستم‌های NER باید مقاوم، قابل تطبیق و بر روی مجموعه‌داده‌های با کیفیت بالا، متنوع و خاص دامنه آموزش دیده باشد.

ج.
۱. مدل‌سازی وابستگی‌ها: 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 گستردهتر در توالی بهرهمند می‌شوند. CRFs این محدودیت را ندارند. آن‌ها می‌توانند حالت فعلی را بسته به کل توالی داده‌های ورودی مدل کنند، بنابراین وابستگی‌ها و زمینه پیچیده‌تری را درک می‌کنند.

۲. انعطاف‌پذیری ویژگی‌ها: HMMها معمولاً به احتمالات ثابت انتقال حالت تکیه دارند و فقط می‌توانند از حالت فوری قبلی و مشاهده کنونی برای پیش‌بینی حالت بعدی استفاده کنند. اما CRFs اجازه می‌دهند از انواع و تعداد متعددی از ویژگی‌های ورودی برای هر حالت در توالی استفاده شود. این بدان معناست که CRFs می‌توانند مجموعه‌ای غنی‌تر از اطلاعات را در نظر بگیرند، مانند حضور کلمات خاص یا سایر عوامل زمینه‌ای، که برای وظایفی مانند NER بسیار حیاتی است.

۳. استقلال خروجی: HMMها فرض می‌کنند که مشاهدات (یا خروجی‌ها) با توجه به توالی حالت مستقل هستند. این فرض در بسیاری از سناریوهای دنیای واقعی که خروجی‌ها ممکن است تحت تأثیر برچسب‌ها یا ویژگی‌های مجاور باشند، خوب عمل نمی‌کند. CRFs این مشکل را با شرطی کردن هر خروجی بر اساس کل توالی ورودی و نه فقط حالت‌های کنونی یا قبلی حل می‌کنند، که به مدل اجازه می‌دهد از وابستگی‌ها بین برچسب‌ها در قسمت‌های مختلف توالی برای انجام پیش‌بینی‌های دقیق‌تر استفاده کند.

خلاصه، CRFs انعطاف‌پذیری و قدرت بیشتری در مدل‌سازی وابستگی‌های پیچیده در مقایسه با HMMها ارائه می‌دهند. آن‌ها برای رسیدگی به نوع اطلاعات زمینه‌ای و ویژگی‌های غنی که در بسیاری از وظایف NLP مدرن نیاز است، مناسب‌تر هستند.

د.

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

۵.
نحوه عملکرد برچسب‌گذاری BIO
برچسب‌گذاری BIO با استفاده از یک طرح ساده با سه نوع برچسب کار می‌کند:
B (Beginning): این برچسب نشان‌دهنده شروع یک موجودیت نام‌دار است. این برچسب با پسوندی همراه است که نوع موجودیت را مشخص می‌کند، مانند B-PER برای شروع نام یک شخص.
I (Inside): این برچسب برای توکن‌هایی استفاده می‌شود که در داخل یک موجودیت نام‌دار قرار دارند اما شروع آن نیستند. مانند B، این برچسب شامل پسوندی است که نوع موجودیت را نشان می‌دهد، مانند I-PER برای توکنی که در داخل نام یک شخص است.
O (Outside): این برچسب برای توکن‌هایی استفاده می‌شود که به هیچ موجودیت نام‌داری تعلق ندارند.
طرح BIO به مدل‌ها کمک می‌کند تا بین موجودیت‌هایی که مجاور هم هستند اما بخشی از یک موجودیت واحد نیستند، تمایز قائل شوند. به عنوان مثال، در عبارت "Apple Inc. CEO Steve Jobs."، "Apple Inc" و "Steve Jobs" دو موجودیت جداگانه هستند که به صورت B-ORG I-ORG O B-PER I-PER برچسب‌گذاری می‌شوند، نشان‌دهنده این است که "Apple Inc." یک موجودیت و "Steve Jobs" موجودیت دیگری است.

تفاوت‌ها از برچسب‌گذاری IO
برچسب‌گذاری IO، که مخفف Inside-Outside است، یک طرح ساده‌تر است که فقط از دو برچسب استفاده می‌کند:
I: برای توکن‌هایی استفاده می‌شود که بخشی از یک موجودیت نام‌دار هستند.
O: برای توکن‌هایی استفاده می‌شود که به هیچ موجودیت نام‌داری تعلق ندارند.
معایب اصلی طرح IO عدم توانایی آن در تمیز دادن بین موجودیت‌های مختلف که مجاور هستند یا موجودیت‌هایی که از نوع یکسان هستند اما متمایز هستند، می‌باشد. به عنوان مثال، در عبارت "Apple Inc. Steve Jobs"، برچسب‌گذاری IO نمی‌تواند نشان دهد که کجا "Apple Inc." به پایان می‌رسد و "Steve Jobs" شروع می‌شود اگر آن‌ها از نوع موجودیت یکسان باشند.

تفاوت‌ها از برچسب‌گذاری BIOES
برچسب‌گذاری BIOES، که همچنین به عنوان BMEWO یا BMEWO+ شناخته می‌شود، دو برچسب دیگر به طرح BIO اضافه می‌کند تا مرزهای واضح‌تری ارائه دهد:
B (Beginning) و I (Inside) به طور مشابه به کاربردشان در برچسب‌گذاری BIO عمل می‌کنند.

E (End): این برچسب نشان‌دهنده پایان یک موجودیت نام‌دار است و به روشن کردن اینکه یک موجودیت کجا متوقف می‌شود کمک می‌کند، به ویژه در مواردی که موجودیت‌هایی از نوع یکسان مجاور هستند.
S (Single): این برچسب برای موجودیتی که فقط شامل یک توکن است استفاده می‌شود، که موجودیت‌های تکتوکنی را از موجودیت‌های چندتوکنی متمایز می‌کند.
O (Outside) همچنان بدون تغییر باقی می‌ماند.
این تفکیک بیشتر در برچسب‌گذاری BIOES به مدل‌ها کمک می‌کند تا ساختار موجودیت‌ها را، به ویژه در موارد پیچیده که موجودیت‌ها مجاور یا همپوشانی دارند، به درستی تفسیر کنند.
به طور خلاصه، در حالی که برچسب‌گذاری BIO روشی را برای رسیدگی به موجودیت‌ها در توالی‌ها با وضوح مناسب فراهم می‌کند، برچسب‌گذاری BIOES حتی کنترل و وضوح بیشتری را ارائه می‌دهد، که در متون پیچیده بسیار مفید است. برچسب‌گذاری IO، که ساده‌ترین است، فاقد پیچیدگی لازم برای رسیدگی به ساختارهای موجودیت مجاور یا پیچیده است.

سوالات عملی:
۱.
تولید دیکشنری برچسب‌گذاری: با استفاده از مجموعه آموزشی، یک دیکشنری برای نگاشت کلمات به برچسب‌هایی که بیشترین فراوانی را در داده‌های آموزشی داشته‌اند، ایجاد می‌کنیم. این روش پایه‌ای رایج در برچسب‌گذاری قسمت‌های گفتار است، جایی که برچسب یک کلمه در متن دیده نشده بر اساس برچسبی که بیشترین فراوانی را با آن در آموزش داشته، پیش‌بینی می‌شود.

محاسبه دقت پایه: دقت روش برچسب‌گذاری پایه (با استفاده از قاعده ساده بیشترین فراوانی برچسب) در برابر مجموعه آزمایش محاسبه می‌شود.

بهبود دقت: در این بخش با استفاده از قوانین یا روش‌های اضافی برای بهبود دقت پایه استفاده می‌شود.
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 'ful' ،'ish' ،'ble' 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.

نتیجه کلی تابع بهبودیافته: 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 روش با حدس زدن برچسب‌ها بر اساس مرفولوژی کلمه و الگوهای زبانی مستقر، روش‌های پایه‌ای

جستجو در فرهنگ لغت را بهبود بخشیده، در نتیجه دقت کلی فرایند برچسب‌گذاری را از ۸۱ به ۸۶ افزایش می‌دهد.

۲.
تحلیل کد به ترتیب سلول‌ها:
توکن‌سازی و برچسب‌گذاری نقش کلمات:
کد: این سلول شامل پردازش متن با استفاده از تکنیک‌های توکن‌سازی و برچسب‌گذاری نقش کلمات است، با استفاده از کتابخانه‌ای مانند NLTK.
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 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.
خروجی کلی مدل:
تعداد کل نمونه‌هایی که به درستی توسط مدل پیش‌بینی شده‌اند و درصد دقت کلی (۹۱٪) نمایش داده شده است.

۳.
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:

۱. 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.
۲. 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.
۳. 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.
۴. عناوین با تاریخ و اعداد: عناوینی که شامل تاریخ یا اعداد هستند (مانند "Apollo" ،"1984 13") ممکن است با داده های عددی واقعی یا سال های خاص در متن اشتباه گرفته شوند.
۵. 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.
۶. 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.