به نام خدا

تکلیف سری سوم NLP

بهاره کاوسی نژاد – 99431217

## سوالات تئوری

### الف) چالش های NER را توضیح دهید.

تعریف NER:

NER یا Named Entity Recognition یک تکنیک اساسی در پردازش زبان طبیعی (NLP) است که شناسایی و استخراج موجودیت های نامگذاری شده خاص از متن را امکان پذیر می کند. موجودیت های نامگذاری شده اشیاء دنیای واقعی مانند نام افراد، نام سازمان، مکان ها، تاریخ ها و سایر اطلاعات مهم هستند. NER نقش حیاتی در صنایع مختلف از جمله مراقبت های بهداشتی، مالی و بازیابی اطلاعات ایفا می کند.

روش کار NER به شرح زیر است:

1. **Tokenization:** NER در درجه اول با tokenization شروع می شود که در آن متن به قطعات کوچکتر یا tokenها – کلمات یا علائم نگارشی – شکسته می شود.
2. **Part-of-Speech Tagging:** در این مرحله، هر token با یک تگ part-of-speech (POS) (مانند اسم، فعل، صفت و غیره) برچسب گذاری می شود که زمینه را برای شناسایی موجودیت نامگذاری شده فراهم می کند.
3. **Entity Identification:** بر اساس تگ‌های POS، الگوریتم موجودیت‌های موجود در متن را طبقه‌بندی و برچسب‌گذاری می‌کند، به عنوان مثال، شخص، مکان، سازمان و غیره.
4. **Dependency Parsing:** در نهایت، تجزیه وابستگی برای تجزیه و تحلیل ساختار دستوری یک جمله استفاده می شود، به درک روابط بین موجودیت ها کمک می کند و زمینه بیشتری را فراهم می کند.

چه چیزی با NER تشخیص داده می شود:

NER می تواند انواع مختلفی از موجودیت های نامگذاری شده را بسته به برنامه یا دامنه تشخیص دهد. برخی از دسته بندی های رایج عبارتند از:

- **نام افراد:** شناسایی اسامی افراد
- **نام سازمان:** شناسایی نام شرکت ها، مؤسسات یا سازمان ها
- **نام‌های مکان:** شناسایی نام شهرها، کشورها یا سایر مکان‌های جغرافیایی
- **عبارات تاریخ و زمان:** تشخیص تاریخ، زمان یا مدت رویدادها

- **ارزش‌های پولی:** شناسایی نمادها یا عباراتی که ارزش پولی را نشان می‌دهند

NER بسیار منعطف است و توانایی آن برای شناسایی موجودیت های خاص را می توان برای مطابقت با نیازهای خاص سفارشی کرد.

انواع موجودیت های نامگذاری شده

موجودیت های نامگذاری شده را می توان بر اساس ویژگی های آنها طبقه بندی کرد. برخی از دسته بندی های رایج عبارتند از:

- **اسم‌های خاص و رایج:** اسم‌های خاص به نام‌های خاص افراد، مکان‌ها یا چیزها (مانند جان، پاریس، گوگل) اشاره می‌کنند، در حالی که اسم‌های رایج به نام‌های عمومی (مانند گربه، خانه، ماشین) اشاره دارند.
- **موجودیت‌هایی با فرم‌های چندگانه:** برخی از موجودیت‌های نام‌گذاری‌شده می‌توانند تغییرات یا اشکال متعددی داشته باشند (به عنوان مثال، مخفف‌ها، مخفف‌ها، نام‌های مستعار).
- **موجودیت های مبتنی بر هستی شناسی (Ontology):** موجودیت هایی که در هستی شناسی دامنه خاصی تعریف می شوند (به عنوان مثال، اصطلاحات پزشکی، نام محصول).
- **موجودیت های زمانی:** موجودیت های مرتبط با زمان، مانند تاریخ، زمان یا مدت زمان.
- **موجودیت های عددی:** موجودیت های مرتبط با اعداد، مانند کمیت ها، اندازه گیری ها یا درصدها.

چالش های NER:

NER به دلیل پیچیدگی و ابهام زبان طبیعی چندین چالش را ایجاد می کند. برخی از چالش های رایج عبارتند از:

- **ابهام در نام موجودیت ها:** برخی از کلمات یا عبارات می توانند معانی یا تفسیرهای متعددی داشته باشند.
- **غلط املایی در نام‌های موجودیت ها:** داده‌های متنی اغلب حاوی اشتباهات املایی یا تغییرات هستند که تشخیص دقیق موجودیت‌های نام‌گذاری شده را دشوار می‌کند.
- **ابهام در انواع موجودیت:** برخی از کلمات یا عبارات را می توان به انواع موجودیت های متعدد طبقه بندی کرد که منجر به عدم قطعیت در طبقه بندی می شود.
- **تغییرات در مراجع موجودیت:** موجودیت ها را می توان با استفاده از عبارات یا مترادف های مختلف مورد اشاره قرار داد که شناسایی آنها را به چالش می کشد.
- **چالش های متنی:** درک زمینه یک کلمه یا عبارت در یک جمله یا سند برای تشخیص دقیق موجودیت ضروری است.

پرداختن به این چالش ها نیازمند مدل ها و تکنیک های قوی NER است که بتواند چنین پیچیدگی هایی را مدیریت کند.

ب) تاثیرمفهوم متن بر میزان دقت سیستم های NER را توضیح دهید.

"مفهوم متن" به زمینه ای اشاره دارد که در آن یک اصطلاح یا موجودیت خاص در یک متن ظاهر می شود. این شامل کلمات، عبارات و ساختار دستوری اطراف است که اطلاعات و زمینه بیشتری را برای درک معنای اصطلاح فراهم می کند. تأثیر مفهوم متن بر دقت سیستم های NER قابل توجه است.

هدف سیستم‌های NER شناسایی و طبقه‌بندی موجودیت‌های نام‌گذاری شده، مانند نام افراد، نام سازمان، و نام مکان، در یک متن مشخص است. این سیستم ها به شدت به زمینه ای که موجودیت ها در آن رخ می دهند برای شناسایی و طبقه بندی دقیق آنها متکی هستند.

در اینجا چند راه تاثیرگذاری مفهوم متن بر دقت سیستم های NER بیان شده است:

1. **Ambiguity Resolution:** مفهوم متن به حل ابهاماتی کمک می کند که ممکن است زمانی که چندین موجودیت نام ها یا اصطلاحات مشابهی را به اشتراک می گذارند به وجود بیاید. با در نظر گرفتن بافت اطراف، سیستم های NER می توانند بین موجودیت های مختلف با نام های مشابه تمایز قائل شوند. به عنوان مثال، در جمله "Apple is launching a new product" مفهوم متن به شناسایی "Apple" به عنوان نام شرکت به جای میوه کمک می کند.
2. **طبقه بندی موجودیت نامگذاری شده:** مفهوم متن سرنخ های ارزشمندی برای طبقه بندی دقیق موجودیت ها ارائه می دهد. به عنوان مثال، اگر کلمه "پزشک" در زمینه یک مرکز پزشکی یا توصیف یک بیمار ظاهر شود، به احتمال زیاد به عنوان یک نهاد شخصی طبقه بندی می شود. با این حال، اگر در زمینه یک درمان یا روش پزشکی ظاهر شود، ممکن است به عنوان یک سازمان یا یک نهاد مفهومی طبقه بندی شود.
3. **وضوح Coreference:** وضوح Coreference وظیفه تعیین زمانی است که دو یا چند عبارت در یک متن به یک موجودیت اشاره می کنند. مفهوم متن با در نظر گرفتن زمینه های اطراف به حل و فصل همبستگی ها کمک می کند. به عنوان مثال، در جمله « John visited his doctor. He prescribed medication»، ضمیر «He» بر اساس متن به «doctor» اشاره دارد.
4. **تشخیص مرز موجودیت:** مفهوم متن به شناسایی دقیق مرزهای موجودیت های نامگذاری شده کمک می کند. گاهی اوقات، موجودیت ها کلمات یا عبارات متعددی را در بر می گیرند و زمینه برای تعیین مرزهای دقیق بسیار مهم است. به عنوان مثال، در جمله "I live in New York City"، زمینه کمک

می کند تا مشخص شود که "New York City" یک موجودیت مکان است نه دو نهاد جداگانه ("New York" و "City").

5. **ابهام زدایی:** مفهوم متن نقشی حیاتی در ابهام زدایی همنام ها یا اصطلاحات چند معنایی دارد. با در نظر گرفتن بافت اطراف، سیستم های NER می توانند معنای صحیح یک اصطلاح را تعیین کنند. به عنوان مثال، در جمله "The bank is closed"، زمینه کمک می کند تا مشخص شود که "bank" به یک موسسه مالی یا کنار رودخانه اشاره دارد.

به طور خلاصه، مفهوم متن تأثیر قابل توجهی بر دقت سیستم های NER دارد. این سیستم‌ها با استفاده از زمینه‌ای که موجودیت‌ها در آن ظاهر می‌شوند، می‌توانند ابهامات را حل کنند، موجودیت‌ها را به درستی طبقه‌بندی کنند، همبستگی‌ها را حل کنند، مرزهای موجودیت را شناسایی کنند، و اصطلاحات را ابهام‌زدایی کنند، که منجر به نتایج دقیق‌تر و قابل‌اعتمادتر شود.

ج) چگونگی بهبود محدودیتهای HMM توسط CRF ها را توضیح دهید.

Conditional Random Fields یاCRF ها محدودیت‌های مدل‌های Hidden Markov Models یا HMMها را از طرق مختلف بهبود می‌بخشند. در اینجا توضیحی در مورد نحوه برخورد CRF ها با محدودیت های HMM ارائه شده است:

- **مدل‌سازی متمایز:** HMMها مدل‌های تولیدی هستند که احتمال joint توالی مشاهده شده و حالت‌های پنهان را مدل می‌کنند. از سوی دیگر، CRFها مدل های متمایز هستند که به طور مستقیم احتمال شرطی حالت های پنهان را با توجه به دنباله مشاهده شده مدل می کنند. این به CRF ها اجازه می دهد تا وابستگی های پیچیده تری را بین توالی مشاهده شده و حالت های پنهان ثبت کنند که منجر به بهبود عملکرد می شود.
- **انعطاف پذیری ویژگی ها:** HMMها معمولاً بر مشاهدات ساده و مجزا مانند کلمات یا کاراکترها متکی هستند. با این حال، CRFها می‌توانند طیف وسیعی از ویژگی‌ها را شامل شوند که جنبه‌های مختلف دنباله مشاهده‌شده را شامل می‌شوند، از جمله اطلاعات متنی، ویژگی‌های زبانی، برچسب‌های part-of-speech، و جاسازی‌های کلمه. این انعطاف‌پذیری در انتخاب ویژگی، CRFها را قادر می‌سازد تا از ویژگی‌های informativeتر استفاده کنند و اطلاعات متنی غنی‌تری را capture کنند، که منجر به دقت بهتری می‌شود.
- **وابستگی های دلخواه:** HMM ها فرض مارکوف را ایجاد می کنند، که فرض می کند وضعیت فعلی فقط به حالت قبلی بستگی دارد. این فرض، مدل‌سازی وابستگی‌های دلخواه بین حالت‌های پنهان را محدود می‌کند. از سوی دیگر، CRF ها چنین فرضیاتی ندارند و می توانند وابستگی های بین حالت های

پنهان را با انعطاف بیشتری مدل کنند. آن‌ها می‌توانند وابستگی‌های long-range را capture کنند و کل توالی را هنگام پیش‌بینی‌ها در نظر بگیرند، که به‌ویژه در کارهایی مانند شناسایی موجودیت نام‌گذاری شده (NER) که در آن مرزهای موجودیت ممکن است چندین کلمه را شامل شود، مفید است.

- **Global Inference:** HMM ها معمولاً از الگوریتم Viterbi برای رمزگشایی استفاده می کنند، که تصمیمات محلی و حریصانه را تنها با در نظر گرفتن حالت های فعلی و قبلی انجام می دهد. با این حال، CRF ها با استفاده از الگوریتم هایی مانند الگوریتم Forward-Backward یا Max-Margin امکان Global Inference را فراهم می کنند. این الگوریتم ها کل دنباله را در نظر می گیرند و به طور مشترک تخصیص حالت های پنهان را بر اساس دنباله مشاهده شده بهینه می کنند و منجر به پیش بینی های دقیق تری می شوند.
- **تکنیک های Training:** HMM ها اغلب با استفاده از الگوریتم Expectation-Maximization (EM) آموزش داده می شوند، که می تواند به مقداردهی اولیه حساس باشد و ممکن است به راه حل های کمتر از حد مطلوب همگرا شود. از سوی دیگر، CRFها با استفاده از تکنیک‌های آموزشی متمایز مانند Maximum Likelihood Estimation (MLE) یا Maximum Entropy (MaxEnt) آموزش داده می‌شوند، که تمایل به قوی‌تر بودن و حساسیت کمتری نسبت به مقداردهی اولیه دارند. این تکنیک های آموزشی به CRF ها اجازه می دهد تا مدل های دقیق تری را از داده های آموزشی داده شده بیاموزند.

د) یک خطا در برچسب گذاری هر یک از جملات زیر که با مجموعه Treebank Penn برچسب گذاری شده اند، پیدا کنید.

- I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN

"Atalanta" به عنوان یک noun برچسب گذاری شده است؛ در حالیکه باید یک proper noun یا NNP باشد زیرا به یک مکان خاص اشاره می کند.

- Does/VBZ this/DT flight/NN serve/VB dinner/NNS

"dinner" به عنوان یک Noun, plural در نظر گرفته شده است؛ در حالیکه باید به عنوان یک Noun, singular یا NN باشد زیرا مفرد است و نه جمع.

- I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP

"have" به عنوان یک Verb, base form یا VB در نظر گرفته شده است؛ در حالیکه باید به عنوان یک Verb, non-3rd person singular present یا VBP باشد زیرا اول شخص مفرد و زمان حال است.

- Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

"Can" به عنوان یک Verb, non-3rd person singular present یا VBP در نظر گرفته شده است؛ درحالیکه باید به عنوان یک Modal یا MD در نظر گرفته شود.

ه) توضیح دهید که روش برچسب گذاری BIO برای entity named ها چگونه استفاده می شود. و تفاوت این روش را از برچسبگذاری IO و برچسبگذاری BIOES بررسی کنید؟

روش برچسب گذاری BIO برای برچسب گذاری موجودیت های نامگذاری شده در NLP مانند Named Entity Recognition یا NER استفاده می شود. این امکان شناسایی و طبقه‌بندی موجودیت‌های نام‌گذاری‌شده را در یک متن مشخص می‌دهد.

در برچسب گذاریBIO ، هر کلمه در یک جمله با یک پیشوند برچسب گذاری می شود که ارتباط آن را با یک موجودیت نامگذاری شده نشان می دهد. پیشوندهای استفاده شده به شرح زیر است:

- **B- یا Beginning:** نشان دهنده اولین token موجودیت نامگذاری شده است.
- **I- یا Inside:** نشان دهنده token هایی در داخل یک موجودیت نامگذاری شده است.
- **O- یا Outside:** نشان دهنده token هایی است که بخشی از هیچ موجودیت نامگذاری شده نیستند.

در اینجا یک مثال با تگ های BIO برای موجودیت های نامگذاری شده آورده شده است:

Apple Inc. is headquartered in Cupertino.

BIO tags: B-ORG I-ORG O O B-LOC O

در این مثال، "Apple Inc."یک موجودیت نامگذاری شده با عنوان B-ORG (آغاز یک organization) و I-ORG (داخل یک organization) است، در حالی که "Cupertino" یک موجودیت نامگذاری شده با عنوان B-LOC (ابتدای یک location) است.

تفاوت اصلی بین برچسب گذاری BIO و IO، نمایش اولین token یک موجودیت نامگذاری شده است. در برچسب گذاریIO ، موجودیت ها با I- (Inside) برای همه token ها، از جمله اولین token، برچسب گذاری می شوند. بنابراین، با استفاده از برچسب گذاریIO ، مثال بالا به صورت زیر برچسب گذاری می شود:

IO tags: I-ORG I-ORG O O I-LOC O

در مقابل، برچسب‌گذاری BIOES توسعه‌ای از برچسب‌گذاری BIO است که امکان برچسب‌گذاری granular بیشتر موجودیت‌های نام‌گذاری شده را فراهم می‌کند. علاوه بر پیشوندهای B- (Beginning) وI- (Inside) ، دو پیشوند دیگر نیز معرفی می کند:

- **E- یا End:** نشان دهنده آخرین token یک موجودیت نامگذاری شده است.
- **S- یا Single:** نشان دهنده یک single-token named entity است.

BIOES در مواردی که تمایز بین موجودیت‌های چند tokenی و موجودیت‌های single-token مهم است، مفید است. مثلا:

I visited New York.

BIOES tags: "O O S-LOC"

در این مثال، "New York" یک موجودیت single-token است که با عنوان S-LOC برچسب گذاری شده است.

به طور کلی، روش برچسب‌گذاری BIO به طور گسترده برای برچسب‌گذاری موجودیت‌های نام‌گذاری شده استفاده می‌شود، با برچسب‌گذاری IO که یک نوع ساده‌تر است و برچسب‌گذاری BIOES جزئیات بیشتری را برای برچسب‌گذاری موجودیت‌های چند tokenی و single-tokenی ارائه می‌دهد. انتخاب طرح برچسب گذاری به الزامات خاص NER و سطح جزئیات مورد نظر در granularity موجودیت نامگذاری شده بستگی دارد.

سوالات عملی

1. نوت بوک Q1 را تکمیل کنید ونتایج هرمرحله را تحلیل کنید.

   تحلیل ها در پایان نوتبوک نوشته شده اند.
2. نوت بوک Q2 را تکمیل کنید ونتایج هرمرحله را تحلیل کنید.

   تحلیل ها در پایان نوتبوک نوشته شده اند.
3. 

   الف) دیتاهای موجود در فولدر data را بررسی کرده و دلایلی که ممکن است این entity name ها مشکل ساز شوند را بیان کنید.

موجودیت‌های نام‌گذاری‌شده در عناوین فیلم می‌توانند به دلایل زیر باعث ایجاد مشکلاتی در توسعه یک سیستم NER شوند:

- **ابهام:** عناوین فیلم ها اغلب حاوی کلمات یا عبارات رایجی هستند که می توانند تعابیر متعددی داشته باشند. به عنوان مثال، عنوان فیلم "The Dark Knight" می تواند به فیلم ابرقهرمانی اشاره داشته باشد یا کلاً زمینه متفاوتی داشته باشد. حل چنین ابهاماتی می تواند برای یک سیستم NER چالش برانگیز باشد.
- **کلمات خارج از واژگان (Out-of-vocabulary):** عناوین فیلم ممکن است شامل کلمات منحصر به فرد یا غیر متعارفی باشد که ممکن است در مدل ها یا فرهنگ لغت های زبان استاندارد وجود نداشته باشد. هنگام مواجهه با چنین کلمات خارج از واژگانی، سیستم NER ممکن است در تشخیص و طبقه بندی صحیح آنها مشکل داشته باشد.

- **موجودیت های چند کلمه ای (Multi-word entities):** عناوین فیلم ها می توانند از چندین کلمه تشکیل شده باشند که با هم یک موجودیت نامگذاری شده را تشکیل می دهند. به عنوان مثال، عنوان فیلم "The Shawshank Redemption" یک موجودیت چند کلمه ای است. توکن کردن چنین موجودیت هایی به درستی و شناسایی مرزهای آنها می تواند پیچیده باشد، به خصوص زمانی که با تغییراتی مانند اختصارات، علائم نگارشی یا عناوین جایگزین سروکار داریم.
- **همپوشانی موجودات نامگذاری شده:** در برخی موارد، عناوین فیلم ممکن است کلمات یا عباراتی را با موجودیتهای نامگذاری شده دیگر به اشتراک بگذارند. به عنوان مثال، فیلم "The Godfather" کلمه "Godfather" با نام شخصیت "Don Corleone" از همان فیلم مشترک است. این زمینه همپوشانی می تواند برای سیستم NER چالش برانگیز باشد تا بین موجودیت های نامگذاری شده مختلف به درستی تمایز قائل شود.
- **تناقضات یا تغییرات:** عناوین فیلم‌ها می‌توانند املای جایگزین، نسخه‌های زبانی متفاوت یا تغییراتی به دلیل محلی‌سازی یا تفاوت‌های فرهنگی داشته باشند. وجود چنین ناسازگاری هایی می تواند شناسایی و طبقه بندی دقیق موجودیت های نام برده شده را برای سیستم NER دشوارتر کند.
- **ابهام در نام ها:** از آنجایی که در مجموعه داده نام فیلم ها وجود دارند، ممکن است نام افراد و دیگر نام های خاص شناخته نشوند.

برای مقابله با این چالش‌ها، ممکن است لازم باشد هنگام توسعه سیستم NER برای عناوین فیلم، استراتژی‌های زیر را در نظر گرفته شود:

- **پیش پردازش و token گذاری:** ایجاد یک tokenizer قوی که می تواند موجودیت های چند کلمه ای، علائم نقطه گذاری و تغییرات در عناوین فیلم را مدیریت کند و همچنین می تواند شامل اعمال قوانین یا الگوهای خاصی باشد که در مجموعه داده های IMDb مشاهده می شود
- **اطلاعات contextual:** استفاده از اطلاعات contextual مانند ژانر، کارگردان یا بازیگران برای ابهام‌زدایی و بهبود دقت در موجودیت‌های نام‌گذاری شده
- **داده‌های آموزشی:** اطمینان از اینکه که داده‌های آموزشی شامل طیف متنوعی از عناوین فیلم، از جمله ژانرها، زبان‌ها و انواع مختلف باشد
- Fine-tuning یا Transfer Learning: در نظر گرفتن fine-tuning یا استفاده از مدل های زبانی pre-trained که بر روی مجموعه ها یا مجموعه داده های مرتبط با فیلم آموزش دیده اند. این مدل‌ها ممکن است اطلاعات متنی مخصوص فیلم‌ها را یاد گرفته باشند که می‌تواند عملکرد سیستم NER را بهبود بخشد.

ب) نوت بوک Q3 را تکمیل کنید.

**منابع:**


- https://botpenguin.com/glossary/named-entity-recognition
- https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html