به نام خدا

تكليف سري سوم NLP بهاره كاوسي 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.

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

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
* Does/VBZ this/DT flight/NN serve/VB dinner/NNS
* I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP
* Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

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؟

**منابع:**

* <https://botpenguin.com/glossary/named-entity-recognition>