سیده شکیبا انارکی فیروز ۹۹۴۴۲۰۴۷
تمرین سوم درس پردازش زبان ها

۱)الف)

NER:
در پردازش زبان طبیعی، NER یا Named Entity Recognition یا «تشخیص انتساب موجودات نامدار» است. این یکی از وظایف مهم در NLP است که به ما کمک می‌کند تا موجودات مختلفی که در یک متن ذکر شده‌اند را شناسایی و دسته‌بندی کنیم، مانند افراد، مکان‌ها، شرکت‌ها، تاریخ‌ها و غیره. به‌عبارت دیگر، NER به ما کمک می‌کند تا بفهمیم که یک کلمه یا یک مجموعه از کلمات در یک متن به چه معنی‌ای اشاره دارد و این اطلاعات را استخراج کنیم.

چالش ها:
- **Ambiguity.** کلمات ممکن است فریبنده باشند. یک اصطلاح مانند "amazon" ممکن است به رودخانه یا شرکت، بسته به زمینه، اشاره داشته باشد که شناسایی موجودیت را به یک تلاش دشوار تبدیل می کند.
- **Context dependency.** کلمات اغلب معنای خود را از متن اطراف می گیرند. کلمه "Apple" در یک مقاله فنی احتمالاً به شرکت اشاره دارد، در حالی که در یک دستور غذا، احتمالاً میوه است. درک چنین تفاوت های ظریف برای تشخیص دقیق موجودیت بسیار مهم است.
- **Language variations.** زبان انسان، با زبان عامیانه، گویش ها و تفاوت های منطقه ای آن، می تواند چالش هایی ایجاد کند. آنچه در یک منطقه رایج است ممکن است در منطقه دیگر بیگانه باشد و فرآیند NER را پیچیده کند.
- **Data sparsity.** برای روش‌های NER مبتنی بر یادگیری ماشین، در دسترس بودن داده‌های برچسب‌گذاری شده جامع بسیار مهم است. با این حال، به دست آوردن چنین داده هایی، به ویژه برای زبان های کمتر رایج یا حوزه های تخصصی، می تواند چالش برانگیز باشد.
- **Model generalization**. در حالی که یک مدل ممکن است در تشخیص موجودیت ها در یک حوزه برتر باشد، ممکن است در حوزه دیگر دچار تزلزل شود. اطمینان از تعمیم مدل‌های NER به خوبی در حوزه‌های مختلف یک چالش دائمی است.

۱)ب)

- **پیچیدگی:** متون با ساختار واضح و پیچیدگی کم، اغلب با دقت بالاتری در سیستم‌های NER شناسایی می‌شوند. اما متونی که ساختار پیچیده‌تری دارند یا به عنوان مثال از اصطلاحات، اختصارات، یا جملات غیرمعمول استفاده می‌کنند، ممکن است باعث کاهش دقت سیستم‌های NER شود.
- **مبهم بودن:** متونی که شامل عبارات مبهم یا دارای ابهامات معنایی هستند، ممکن است برای سیستم‌های NER چالش برانگیز باشند. این ابهامات می‌توانند از نظر معنایی یا تعیین محدوده دسته بندی نام‌ها و اجزای دیگر متن ایجاد شوند.
- **انحراف از الگوهای استاندارد**: برخی از متون، ممکن است از الگوها یا قوانین استاندارد خارج شوند که سیستم‌های NER بر اساس آن‌ها آموزش داده شده‌اند. در این صورت، دقت سیستم ممکن است کاهش یابد زیرا الگوهای غیرمعمول برای آنها چالش بیشتری ایجاد می‌کند.
- **تنوع فرهنگی و زبانی**: تفاوت های فرهنگی و زبانی در بین زبانها و مناطق می تواند چالش هایی را برای دقت در NER ایجاد کند. زبانهای مختلف ممکن است دارای کنوانسیون های نامگذاری و ساختارهای گرامری متمایز باشند ، که نیاز به رویکردهای مدل سازی خاص زبان و داده های آموزش دارند. علاوه بر این ، تغییرات در منابع فرهنگی و هنجارهای متنی ممکن است بر تفسیر و طبقه بندی موجودات نامگذاری شده در متون چند فرهنگی یا چند زبانه تأثیر بگذارد.

- **وابستگی به ساختار و محتوای جملات:** ساختار و محتوای جملات متن می‌تواند نقش بسیار مهمی در تشخیص موجودیت‌ها داشته باشد. عناصر متن مانند کلمات مشابه، واژگان اضافی، و فعل‌ها ممکن است معنای موجودیت‌ها را متغیر کنند. برای مثال، جمله "مریم کتاب خواند" ممکن است به این معنا باشد که مریم یک کتاب را خوانده است یا کتابی با عنوان "مریم" وجود دارد.

۱)ج)

- **برخورد با انعطاف پذیری بیشتر**: یکی از محدودیت های اصلی HMMها این است که فقط به ویژگی های محلی یا لحظه ای توجه می‌کنند و وابستگی به سایر ویژگی های موجود در داده را نادیده می‌گیرند. در عوض، CRF ها قادرند انعطاف بیشتری را در مدل سازی این وابستگی ها ارائه دهند، به خصوص با استفاده از ویژگی های چند متغیره که اطلاعات مرتبط با تمام دنباله را در نظر می‌گیرند.
- **استفاده از ویژگی های بیشتر و متغیرهای بیشتر**: در CRF ها، می توان از ویژگی های متعدد و متغیرهای بیشتری برای مدل سازی استفاده کرد. این امر به افزایش اطلاعات وارد شده به مدل و بهبود دقت در پیش‌بینی دنباله ها کمک می‌کند و به CRF ها امکان مطابقت بیشتر با واقعیت را می‌دهد، زیرا آنها می‌توانند ویژگی های مرتبط را به طور مستقیم با یکدیگر مدل کنند.
- **استنتاج کلی:** CRF ها هنگام پیش بینی، استنتاج سراسری را در کل دنباله انجام می دهند. این به آن ها اجازه می دهد تا وابستگی های بین برچسب ها را در کل دنباله در نظر بگیرند، نه صرفاً در نظر گرفتن انتقال های محلی مانند HMM. در نتیجه، CRF ها می توانند وابستگی های دوربرد را بهتر دریافت کنند و پیش بینی های آگاهانه تری انجام دهند.
- **دسته بندی بهتر با داده های نامتوازن**: CRF ها می توانند با داده های نامتوازن بهتر کار کنند، یعنی داده هایی که دارای نسبت نمونه های مثبت به منفی نامتوازن هستند. با استفاده از تابع های هزینه متناسب با اهمیت نمونه ها و با افزودن ویژگی‌های مرتبط، CRF ها می‌توانند عملکرد بهتری در مواجهه با این نوع داده ها ارائه دهند.
- **کارایی آموزش:** CRF ها را می توان با استفاده از تکنیک هایی مانند نزول گرادیان تصادفی (SGD) یا روش های گرادیان شرطی، که امکان همگرایی سریع تر و مقیاس پذیری را برای مجموعه داده های بزرگ فراهم می کند، به طور موثر آموزش داد. در مقابل، آموزش HMM ها اغلب شامل الگوریتم های تکراری مانند الگوریتم Baum-Welch است که ممکن است کندتر همگرا شوند و برای مجموعه داده های بزرگ کمتر مقیاس پذیر باشند.

۱)د)

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

Atlanta/NN درست نیست زیرا یک مکان خاص(a specific location) است و باید "NNP" باشد.

2. Does/VBZ this/DT flight/NN serve/VB dinner/NNS:

“NNS”stands for plural noun, in this context “dinner” refers to a meal rather than individual items, so “NN” is a better label for that.

3. I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP:

Have should be labeled as "have/VBP" because it's a verb in the form of the base form.

4. Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS:

Can should be labelled as “MD” because it is a modal.

۱)ه)

- **روش برچسب‌گذاری BIO (Begin, Inside, Outside)**

یک روش متداول برای برچسب‌گذاری دنباله های متنی است، به ویژه برای شناسایی موجودیت‌های نام‌دار (Named Entities یا NE) در متن. در این روش، هر کلمه یا توکن در متن با یکی از برچسب‌های "B", "I" یا "O" برچسب‌گذاری می‌شود.

- "B" برای شروع یک NE استفاده می‌شود.
- "I" برای ادامه NE استفاده می‌شود.
- "O" برای هر کلمه‌ای که به NE مرتبط نیست، به عنوان خارج از موجودیت نام‌دار برچسب‌گذاری می‌شود.

- **برچسب گذاری IO (Inside, Outside):**

در این روش، تنها دو برچسب "I" و "O" وجود دارد. "I" برای تمامی کلماتی که داخل یک موجودیت نام دار قرار دارند و "O" برای تمامی کلماتی که خارج از هر موجودیت نام‌دار هستند. از این روش برای مواردی که تنها به شناسایی موجودیت‌های نام‌دار می‌پردازیم و جزییات داخلی هر موجودیت برای ما مهم نیست استفاده می‌شود.

- **برچسب گذاری BIOES (Begin, Inside, Outside, End, Single):**

در این روش، به علاوه‌ی برچسب‌های "I" ،"B" و "O"، دو برچسب جدید "E" (End) و "S" (Single) نیز وجود دارد. "E" برای آخرین کلمه یک NE استفاده می‌شود و "S" برای مواردی که یک NE تنها از یک کلمه تشکیل شده است. این روش به ما اجازه می‌دهد تا دقیقاً ابتدا و انتهای هر موجودیت را مشخص کنیم، که زمانی که نیاز به شناسایی دقیق محدوده‌های موجودیت‌های نام‌دار داریم می‌تواند مفید باشد.

۳)الف)

- **ابهام در نام ها:** از آنجايي كه در مجموعه داده نام فیلم ها وجود دارند، ممکن است نام افراد و دیگر نام هاي خاص شناخته نشوند.
- **نام فیلم (Movie Name):** بعضی از فیلم ها ممکن است نام های شبیه به هم داشته باشند که باعث ایجاد اشتباه در تفکیک و تشخیص موجودیت ها شوند، به عنوان مثال فیلم هایی که عنوان شان شامل کلمات مشابهی مثل "The","A"" و غیره باشد. همچنین عناوین فیلم ها اغلب حاوي كلمات يا عبارات رايجي هستند كه مي توانند تعابير متعددي داشته باشند.
- **کاراکتر های خاص:** برخی از عنوان های فیلم ممکن است شامل کاراکترهای خاص، عبارات غیر استاندارد، یا ترکیباتی از اعداد و حروف باشند که باعث ایجاد اشتباه در تشخیص نام فیلم ها توسط الگوریتم‌های NER شود.
همچنین ممکن است برخی از نام‌های کارگردان‌ها شبیه به نام‌های عمومی یا دیگر افراد باشند که باعث ایجاد اشتباه در تشخیص موجودیت‌ها شود.
- **همپوشانی موجودات نامگذاری شده:** در برخي موارد، عناوين فيلم ممكن است كلمات يا عباراتي را با موجوديتهاي نامگذاري شده ديگر به اشتراک بگذارند. به عنوان مثال، فيلم "Godfather The " كلمه "Godfather "با نام شخصيت "Corleone Don "از همان فيلم مشترک است. اين زمينه همپوشاني می تواند براي سيستم NER چالش برانگيز باشد تا بين موجوديت هاي نامگذاري شده مختلف به درستي تمايز قائل شود.
- **تناقضات یا تغییرات:** عناوين فيلم ها مي توانند نسخه هاي زباني متفاوت يا تغييراتي به دليل محلي سازي يا تفاوت هاي فرهنگي داشته باشند. وجود چنين ناسازگاري هايي مي تواند شناسايي و طبقه بندي دقيق موجوديت هاي نام برده شده را براي سيستم NER دشوارتر كند .
- **سال (Year) و امتیاز (Rating):**این فیلدها شامل اعداد و فرمت های عددی هستند که ممکن است باعث ایجاد اشتباه در تشخیص موجودیت ها توسط الگوریتم های NER شوند، به خصوص اگر الگوریتم به طور خودکار هر چیزی که به شکل عددی باشد، موجودیت مربوط به سال یا امتیاز تلقی کند.

منابع:

General:
https://www.youtube.com/watch?v=2XUhKpH0p4M
https://www.youtube.com/watch?v=HKLbI_i7kuw
https://www.youtube.com/playlist?list=PLvcbYUQ5t0UEK2KAGyUP7JO9K-Arct8OM

Q1 theory:
https://www.datacamp.com/blog/what-is-named-entity-recognition-ner
https://www.sketchengine.eu/penn-treebank-tagset
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Q2 practical:
https://www.youtube.com/watch?v=IqXdjdOgXPM
https://www.youtube.com/watch?v=fX5bYmnHqqE