دانشکده‌ی مهندسی کامپیوتر

ساخت چارچوبی برای استخراج خودکار نام اشخاص از متن. مورد مطالعه: زبان عربی

پایان‌نامه‌ی مقطع کارشناسی ارشد

در رشته‌ی مهندسی کامپیوتر (گرایش هوش مصنوعی و رباتیک)

مجید عسگری بیدهندی

استاد راهنما: دکتر بهروز مینایی

شهریور ۱۳۹۰


بسم الله الرحمن الرحيم

تقدیم به پدر و مادر عزیزم

تشکر و قدردانی

در این جا لازم می‌دانم که از زحمات استاد ارجمندم جناب آقای دکتر بهروز مینایی که راهنمایی‌های ایشان، همواره ره‌گشا و پشتیبان اینجانب بوده است، تشکر نمایم. همچنین از آقایان مهندس حسین جوزی، مهندس حمیدرضا عسگری بیدهندی، مهندس محمد حسین الهی‌منش، مهندس حسن اصغریان، مهندس محمد رضا وفائی و مهندس امید کاشفی که در مراحل مختلف آماده‌سازی این پایان‌نامه من را یاری کردند، کمال تشکر را دارم. همچنین لازم است از همکاری مرکز تحقیقات کامپیوتری علوم اسلامی نور به خاطر در اختیار گذاشتن داده‌های برچسب‌گذاری شده به زبان عربی و نیز امکانات تحقیقاتی دیگر تشکر نمایم.

چکیده

تشخیص و استخراج دقیق واحدهای اسمی مانند نام اشخاص، مکان‌ها، تاریخ و ساعت، در داده کاوی از یک منبع الکترونیکی یا متنی بسیار مفید است. تشخیصِ درست واحدهای اسمی، یک نیاز مهم در حل مسائلی در حوزه‌های جدید مانند پاسخگویی به سوالات، سیستم‌های خلاصه‌سازی، بازیابی اطلاعات، استخراج اطلاعات، ترجمه‌ی ماشینی، تفسیر ویدئویی و جستجوی معنایی در وب است. تشخیص واحدهای اسمی همچنین می‌تواند به طور خاص ما را در مسائل تازه‌ای مانند رفع ابهام و تشخیص هویت اصلی بین اشخاصی با اسامی مشترک از روی موضوع متن و با کمک ابزارهای جانبی، یافتن نقل قول و ارجاعات در مقالات علمی یا یافتن ارتباط بین مقالات، تشخیصِ ارتباط میان اشخاص و انجمن‌ها با استفاده از اسامی و ارجاعات، بهینه کردنِ پاسخ‌های یک موتورِ جستجو در زمینه‌ی یافتنِ اسامی و ... یاری دهد.

در این پایان‌نامه، ابتدا به تعریف مسأله‌ی تشخیص واحدهای اسمی، چالش‌های آن و نیز به نتایج حاصل شده در زبان‌های دیگر پرداخته‌ایم. سپس مدل‌ها و الگوریتم‌های متفاوت برای عملیات تشخیص واحدهای اسمی بررسی شده‌اند. علاوه بر بررسی این مدل‌ها، ورودی‌ها و خروجی‌های استاندارد و نیز چگونگی ارزیابیِ نتایج در عملیات تشخیص واحدهای اسمی، با تکیه بر مهمترین کنفرانس‌های جهانی بررسی شده‌اند.

به دلیل اینکه هدف اصلیِ این پایان نامه، پیاده‌سازی سامانه‌ای برای تشخیص نام اشخاص در زبانِ عربی است، به شرحِ چگونگی پیاده‌سازی این سامانه پرداخته‌ایم؛ قبل از شروع به توسعه دادن این سامانه نرم‌افزارهای مختلفی را که همگی در محیط‌های علمیِ دانشگاه‌های معتبر دنیا توسعه داده شده‌اند، برای تشخیص اینکه آیا می‌توان از آن‌ها برای توسعه‌ی سامانه سود برد، بررسی کرده‌ایم و بر اساس پارامترهای گوناگون به انتخاب دست زده‌ایم. این شرح کامل این بررسی‌ها، به عنوان یک فصل ضمیمه به پایان‌نامه الحاق شده است.

در نهایت با بهره‌گیری از چند نرم‌افزار آزاد و توسعه‌ی ابزارهای موردِ نیاز برای پیش‌پردازش متون فارسی و عربی، سامانه‌ی Noor ANER برای تشخیص اسامی خاص در متون عربی بر اساس

مدل میدان‌های تصادفی شرطی و مفهوم تزریق کلمات نامزد اسم به عنوان یک راه حل پیشنهادی برای بهبود نتایج، به طور کامل معرفی و ارزیابی شده است. نتایج به دست آمده نشان می‌دهند که با کمک تزریق کلمات نامزد اسم، پیشرفت قابل توجهی در میزان دقت و بازخوانی عملیات تشخیص واحدهای اسمی در زبان عربی حاصل شده است. با توجه به این که روش پیشنهادی ما وابسته به زبان نمی‌باشد، در آینده امکان اِعمال آن روی زبان‌های دیگر از جمله زبان فارسی ممکن است.

واژه‌های کلیدی: تشخیص واحدهای اسمی، یادگیری ماشین، میدان‌های تصادفی شرطی، زبان فارسی، زبان عربی

فهرست مطالب

لیست جداول

لیست تصاویر

فصل ۱

مقدمه

تشخیصِ درست واحدهای اسمی، یک نیاز مهم در حل مسائلی در حوزه‌های جدید مانند پاسخگویی به سوالات، سیستم‌های خلاصه‌سازی، بازیابی اطلاعات، استخراج اطلاعات، ترجمه‌ی ماشینی، تفسیر ویدئویی و جستجوی معنایی در وب است. هدف اصلی در این پایان‌نامه، توسعه‌ی سامانه‌ای برای تشخیص واحدهای اسمی در زبان عربی است. در این بخش، با توجه به اینکه تشخیص واحدهای اسمی یک زیر عملیات از پردازش زبان طبیعی محسوب می‌شود، ابتدا مقدمه‌ی کوتاهی بر علم پردازش زبان طبیعی خواهیم داشت. سپس عملیات تشخیص واحدهای اسمی معرفی خواهد شد. در ادامه نیز مشکلات اصلی مواجهه با مسأله‌ی تشخیص واحدهای اسمی در زبان فارسی و نیز زبان‌های آسیای جنوبی به دلیل شباهت‌هایشان به زبان فارسی مورد بررسی قرار گرفته‌اند.

در فصل ۲ روش‌های برخورد با مسأله‌ی تشخیص واحدهای اسمی و راه حل‌های مبتنی بر قوانین و نیز روش‌های مبتنی بر یادگیری ماشین معرفی شده‌اند. سپس طبقه‌بندی‌های مختلف واحدهای اسمی و روش‌های پیاده‌سازی آن بررسی شده‌اند.

در فصل ۳ مسائل تئوری در رابطه با میدان‌های تصادفی شرطی مطرح شده‌اند.

در فصل ۴ به طور کامل به عملیات تشخیص واحدهای اسمی در زبان عربی پرداخته شده است. نتایج حاصل شده توسط پژوهش‌های دیگران و نیز ویژگی‌های یک پیکره‌ی متنی استاندارد در این بخش توصیف شده‌اند.

در فصل ؟؟ طریقه‌ی ساخت پیکره‌های متنی استاندارد NoorCorps و سامانه‌ی پیشنهادی به

نام Noor ANER برای بهبود نتایج تشخیص واحدهای اسمی با استفاده از برچسب‌های غنی‌شده با کلمات نامزد اسمی مورد بحث قرار گرفته‌اند.

در فصل **؟؟** نیز به بررسی نتایج حاصله از سامانه و مقایسه‌ی آن با کارهای مشابه پرداخته‌ایم.

سرانجام در فصل **؟؟** نتایج کلی تحقیقات انجام شده در پی این پایان‌نامه و نیز کارهای آینده ذکر شده است.

همچنین در جریان پیاده‌سازی سامانه‌ی Noor ANER پیاده‌سازی‌های متن باز دیگری از میدان‌های تصادفی شرطی و تشخیص واحدهای اسمی برای استفاده‌ی احتمالی بررسی شده‌اند که نتایج این بررسی در پیوست **؟؟** آمده است.

پردازش زبان طبیعی چیست؟

پردازش زبان‌ها و مکالمات طبیعی یکی از اموری است که با ورود فناوری رایانه‌ای به زندگی بشر مورد توجه بسیاری از دانشمندان قرار گرفته است. حتی اندیشه‌ای که تورینگ[۱] از ماشین هوشمند خود و تعریفی که او از هوش مصنوعی داشت، در مرحله اول مربوط به پردازش زبان‌های طبیعی می‌شد. حتی تلاش‌های بسیاری توسط بشر برای پیگیری این امر صورت گرفته بود. به عنوان مثال ماشین الیزا[۲] از این تلاش‌ها حاصل می‌شد که ماشینی بود که با تایپ از راه دور با یک انسان، جملات او را پردازش می‌کرد و جوابی درخور به او می‌داد و همیشه انسان آرزوی داشتن ماشین‌هایی داشته است که بتوانند با استفاده از زبان طبیعی با انسان ارتباط برقرار کنند. یعنی کامپیوترهایی که از زبان طبیعی به عنوان ورودی و یا خروجی استفاده می‌کنند.

از لحاظ رده‌بندی، علم پردازش زبان طبیعی از شاخه‌های هوش مصنوعی به حساب می‌آید و خود این علم به چند رده مختلف تقسیم‌بندی می‌شود:

- آوا شناسی و صدا شناسی[۳] که به تشخیص آواها و صداها و بازشناسی گفتار می‌پردازد.

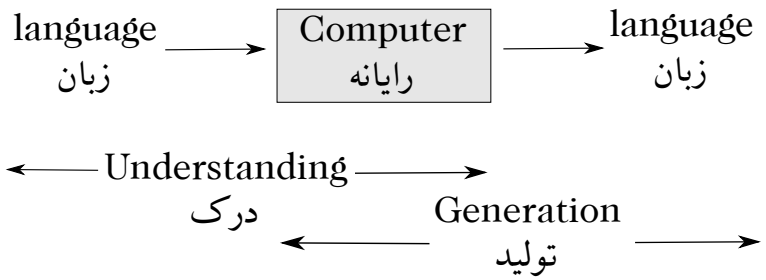
---

[۱] Turing

[۲] ELIZA Machine

[۳] Phonetics and Phonology

- ریخت‌شناسی[۴] که به ساختارهای کلمات و ریشه‌یابی واژگان می‌پردازد.
- نحو[۵] که به ارتباط کلمات به همدیگر و مباحث دستوری آن‌ها در گروه‌ها و جملات می‌پردازد.
- معناشناسی[۶] که به ارتباطات معنایی کلمات می‌پردازد.
- عمل‌گرایی[۷] که کاربردهای زبان برای رساندن یک مطلب به مخاطب یا مخاطبان، در حالت عملی و یا در نوشتار و گفتار طبیعی می‌پردازد.
- مباحثه[۸] که به ارتباطات کلی یک زبان، در دیدگاهی فراتر از یک یا چند جمله خاص می‌پردازد.



شکل پردازش زبان طبیعی توسط یک رایانه

بر همین مبنا الگوریتم‌های بسیاری برای رسیدن به برنامه‌هایی هوشمندتر توسط دانشمندان و متخصصین علوم رایانه، زبان‌شناسی و ریاضیدانان، طراحی و پیشنهاد شده است. به عنوان مثال الگوریتم‌های الگوی مارکوف و الگوی مخفی مارکوف و نیز تلاش‌های چندین ساله نوام چامسکی[۹]

[۴] Morphology

[۵] Syntax

[۶] Semantics

[۷] Pragmatics

[۸] Discourse

[۹] Noam Chomsky

در این راه، نمونه‌ی خوبی برای این امور است. روز به روز بر پیشرفت‌های دانشمندان در این امر افزوده می‌شود و دانشمندان در سراسر دنیا سعی بر بهبود روش‌ها و پیاده‌سازی این روش‌ها در زبان‌های بومی خودشان هستند.

دیدگاه‌های گوناگونی در مورد پردازش زبان طبیعی وجود دارد. اولین دیدگاه استفاده از الگوهای محاسباتی برای پردازش زبان طبیعی انسان است. در این رویکرد، هدف، ساخت برنامه‌هایی است که به صورت درونی به همان روش انسان عمل کنند. دیدگاه دوم استفاده از الگوهای محاسباتی از ارتباط انسان است. یعنی ساخت برنامه‌هایی که مانند انسان‌ها تعامل می‌کنند. رویکرد سوم ساخت سامانه‌های محاسباتی است که به صورت موثری متن و گفتار را پردازش می‌کنند.

به طور کلی کاربردهای گوناگون و بسیار سودمندی برای پردازش زبان طبیعی می‌توان متصور شد. یکی از مهم‌ترین این کاربردها، استخراج اطلاعات است. هدف سیستم‌های استخراج اطلاعات نگاشت یک مجموعه از اسناد به پایگاه داده‌ی ساخت یافته است. انگیزه‌ی اصلی از استخراج اطلاعات ممکن ساختن جستجوهای پیچیده و تقاضاهای آماری است. برای مثال «تمام مشاغلی که در زمینه‌ی تبلیغات هستند و حداقل پنجاه هزار دلار دستمزد دارند و محل آن‌ها ایالت بوستون می‌باشد را برای من پیدا کن!» یا «آیا تعداد مشاغلی که در زمینه‌ی حسابداری هستند در طی سال‌های قبل افزایش پیدا کرده‌اند؟» متن زیر را که تبلیغات یک مؤسسه است در نظر بگیرید:

«شرکت XYZ موسسه‌ای است که همه نوع سرویس تبلیغات را ارائه می‌کند و تخصص آن در بازاریابی مستقیم و متقابل است. این شرکت در بیگ‌تاون کانادا واقع شده است، شرکت XYZ به دنبال معاون مدیر حسابداری می‌باشد تا به مدیریت و بازاریابی متقابل کمک کند. آشنایی مقدماتی به حساب‌های خودکار. دارای تجربه در بازاریابی آنلاین و در زمینه‌ی خودکار و/یا تبلیغات، امتیاز مثبت است. مسئولیت‌های معاون مدیرحسابداری شامل حصول اطمینان از پیشرفت مناسب برنامه‌ها و کمک‌های لازم و مقدماتی مدیریت تحویل پروژه‌ها و راهنمایی سفارش‌های مشتری است... پاداش: ۵۰۰۰۰ دلار مزد شرکت XYZ»

انتظار داریم بعد از پردازش این متن اطلاعاتی شبیه به آنچه در جدول شماره‌ی آمده است از داخل متن استخراج شوند:

جدول نمونه‌ای از اطلاعات استخراج شده از داخل یک متن

صنعت	تبلیغات
سمت	معاون مدیر حسابداری
موقعیت	بیگ‌تاون، کانادا
شرکت	XYZ
حقوق	۵۰۰۰۰ تا ۸۰۰۰۰ دلار

یکی دیگر از کاربردهای پردازش زبان طبیعی، تلخیص متن خودکار[۱۰] است و در تعریف ساختن یک نسخه‌ی کوتاه‌شده از یک متن توسط یک برنامه‌ی رایانه‌ای است. محصول این پردازش هنوز حاوی مهم‌ترین نکات از متن اصلی است. پدیده‌ی افزایش فزاینده‌ی اطلاعات[۱۱] به این معنی است که دسترسی به خلاصه‌های منسجم و منطقی بسیار حیاتی هستند. هرچه که دسترسی به داده‌ها افزایش یابد علاقه‌مندی به خلاصه‌سازی بیشتر می‌شود. یکی از مثال‌های فناوری خلاصه‌سازی متن، موتورهای جستجو هستند.

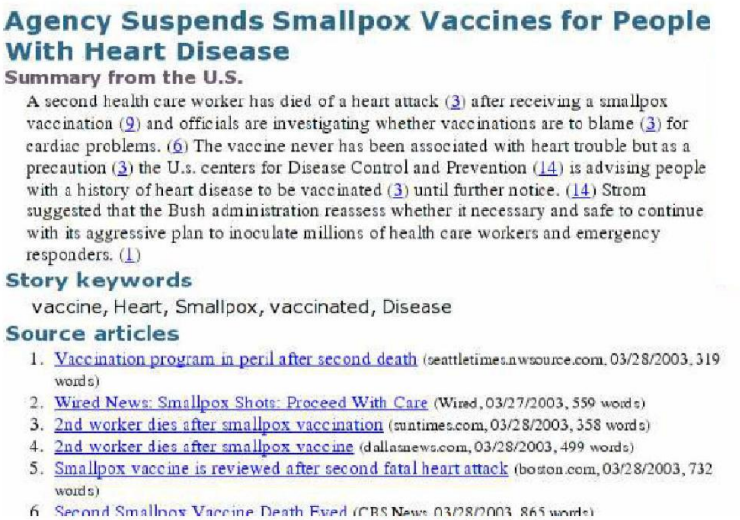
به طور کلی دو نوع از خلاصه‌سازی وجود دارد: استخراج و چکیده‌کردن[۱۲] . تکنیک‌های استخراج اطلاعات، فقط اطلاعاتی را که در متن مهمتر به نظر می‌رسند، کپی می‌کنند. مانند مثال‌ها، قضیه‌ها، جملات و پاراگراف‌های کلیدی. در حالی که چکیده‌کردن متن شامل تفسیر و تغییر جملات نیز می‌شود. به طور کلی چکیده‌کردن، متن را به طور فشرده‌تری خلاصه می‌کند و برای ما مطلوب‌تر است اما عملیات بسیار پیچیده‌تری محسوب می‌شود.

شکل یک مثال از خلاصه سازی متن را نشان می‌دهد: از دیگر کاربردهای پردازش زبان طبیعی می‌توان به تشخیص واحدهای اسمی (که موضوع اصلی مورد بحث در این پایان‌نامه است)،

[۱۰] Automatic Text Summarization

[۱۱] Information Overload

[۱۲] Extraction and Abstraction



Agency Suspends Smallpox Vaccines for People With Heart Disease

Summary from the U.S.

A second health care worker has died of a heart attack (3) after receiving a smallpox vaccination (9) and officials are investigating whether vaccinations are to blame (3) for cardiac problems. (6) The vaccine never has been associated with heart trouble but as a precaution (3) the U.s. centers for Disease Control and Prevention (14) is advising people with a history of heart disease to be vaccinated (3) until further notice. (14) Strom suggested that the Bush administration reassess whether it necessary and safe to continue with its aggressive plan to inoculate millions of health care workers and emergency responders. (1)

Story keywords

vaccine, Heart, Smallpox, vaccinated, Disease

Source articles

1. Vaccination program in peril after second death (seattletimes.nwsource.com, 03/28/2003, 319 words)
2. Wired News: Smallpox Shots: Proceed With Care (Wired, 03/27/2003, 559 words)
3. 2nd worker dies after smallpox vaccination (suntimes.com, 03/28/2003, 358 words)
4. 2nd worker dies after smallpox vaccine (dallasnews.com, 03/28/2003, 499 words)
5. Smallpox vaccine is reviewed after second fatal heart attack (boston.com, 03/28/2003, 732 words)
6. Second Smallpox Vaccine Death Eyed (CBS News, 03/28/2003, 865 words)

شکل مثالی از خلاصه‌سازی متن

ترجمه‌ی ماشینی[۱۳]، بازیابی اطلاعات[۱۴]، تشخیص گفتار[۱۵]، تولید زبان طبیعی[۱۶]، جستجوی زبان طبیعی[۱۷]، فهم زبان طبیعی[۱۸]، تشخیص کاراکترهای نوری[۱۹]، توسعه‌ی پرس‌وجو[۲۰]، پاسخ‌گویی به سؤالات[۲۱]، آنالیز و دسته‌بندی دیدگاه[۲۲]، تولید گزارش‌ها، سیستم دیالوگ[۲۳]، ساده‌سازی متن[۲۴]، تبدیل متن به گفتار[۲۵] و ... اشاره کرد[؟].

---

[۱۳] Machine Translation

[۱۴] Information Retrieval

[۱۵] Speech Recognition

[۱۶] Natural Langage Generation

[۱۷] Natural Language Search

[۱۸] Natural Language Understanding

[۱۹] Optical Character Recognition

[۲۰] Query Expansion

[۲۱] Question Answering

[۲۲] Sentiment Analysis

[۲۳] Spoken Dialogue System

[۲۴] Text Simplification

[۲۵] Text To Speech

چرا پردازش زبان طبیعی مشکل است؟

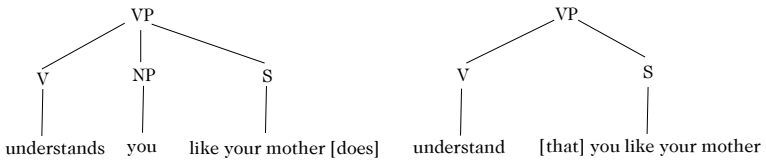
پردازش زبان طبیعی، به دلیل ابهاماتی که در جملات طبیعی موجود در همه‌ی زبان‌ها وجود دارد، یک پردازش مشکل و پیچیده محسوب می‌شود. برای مثال جمله‌ی زیر را در نظر بگیرید: ,At last a computer that understands you like your mother همان طور که مشاهده می‌شود در این جمله ابهام وجود دارد و بدون توجه به متنی که این جمله در آن قرار گرفته‌است نمی‌توان معنای دقیق آن را درک نمود. این جمله می‌تواند به معانی زیر باشد:

- همان طور که مادرتان شما را درک می‌کند، آن هم شما را درک می‌کند.
- درک می‌کند که شما مادرتان را دوست دارید.
- همان طور که مادر شما را درک می‌کند، شما را نیز درک می‌کند.

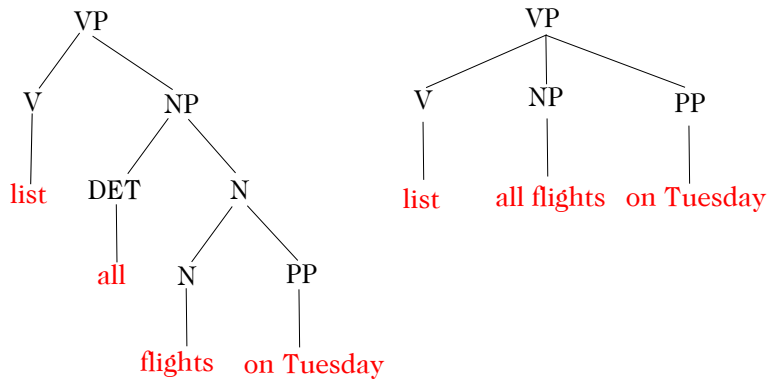
این نوع از ابهام در زبان طبیعی در بسیاری از سطوح از جمله سطح شنوایی نیز وجود دارد (مثلاً در تشخیص گفتار):

- "…a computer that understands you like your mother" یک کامپیوتر که شما را مانند مادرتان درک می‌کند.
- "…a computer that understands yo lie cured mother" یک کامپیوتر که شما را درک می‌کند به مادر شفا یافته دروغ گفتید.

همچنین ساختارهای متفاوت در سطح نحوی، منجر به برداشت‌های متفاوت می‌شود. شکل [؟] را ببینید: همچنین مثال دیگری را در شکل در مورد ابهام نحوی می‌بینید: همچنین ابهام در سطح



شکل مثالی از ابهام نحوی



شکل مثالی از ابهام نحوی

معنایی نیز وجود دارد. مثلاً در زبان انگلیسی کلمه‌ی "Mother" به معانی زیر می‌تواند باشد:

- زنی که یک فرزند دارد.
- یک ماده‌ی نخ مانند لیز که شامل سلول‌های مخمر و باکتری است و به شراب افزوده می‌شود تا سرکه تولید کند.

این معانی یک مثال از ابهام درک واژه[۲۶] است.

نوع دیگری از ابهام، ابهام در سطح گفتمان است. برای مثال در جمله‌ی

mother your like you understands that computer a built they've says "Alice details." any know doesn't she but

نمی‌توانیم مشخص کنیم ضمیر she به "Alice" اشاره می‌کند یا به "Your mother"[؟].

نیازمندی‌های پردازش زبان طبیعی

در پردازش زبان طبیعی علاوه بر اینکه نیازمند دانش کافی در مورد یک زبان هستیم، گاهی نیاز است اطلاعات عمومی کافی درباره‌ی جهان نیز داشته باشیم. برای برخورد با بیشتر مسائل پردازش زبان طبیعی دو رویکرد نمادین و رویکرد آماری می‌تواند وجود داشته باشد. در رویکر نمادین تمام اطلاعات مورد نیاز در داخل کامپوتر، مثلاً به صورت چند پایگاه داده کد می‌شوند. اما در رویکرد

[۲۶] Word Sense Ambiguity

آماری خواص زبان طبیعی را از روی مثال‌های زبان استنتاج می‌کنیم. برای مثال مساله‌ی جایگذاری حرف تعریف [۲۷] را در متنی که حروف تعریف آن حذف شده‌اند در نظر بگیرید. در اینجا مساله این است که برنامه بتواند به طور خودکار حروف تعریف (a، the و هیچ چیز) را در متن جایگذاری کند. در این مساله جایگذاری حرف تعریف به عامل زیر بستگی دارد:

- نوع اسمی (قابل شمارش، غیرقابل شمارش)
- مرجع (معین، نامعین)
- ارزش اطلاعات (کهنه، نو) به این معنی که اسم مورد نظر برای بار چندم است که تکرار می‌شود. مثلاً اگر کلمه‌ی "cat"، برای بار اول در جمله بیاید باید به صورت "a cat" بیاید اما دفعات بعدی به صورت "the cat" ظاهر شود.
- تعداد (مفرد، جمع)

گاهی اوقات نیز عوامل تعیین کننده برای حرف تعریف بسیار پیچیده‌تر می‌شوند. برای مثال حرف تعریف معین به همراه عنوان روزنامه‌ها به کار می‌رود (برای مثال "The Times") ولی در عنوان مجله‌ها حرف تعریف قرار نمی‌گیرد (برای مثال "Time"). بنابراین مثلا برای رویکرد نمادین در انجام چنین کاری به دو دسته از اطلاعات نیازمندیم:

- دانش زبان‌شناسی (تعداد، قابلیت شمارش، هم مرجعی، تشخیص اسم از ارکان دیگر جمله)
- اطلاعات عمومی (یکتایی مرجع، مثلاً رئیس جمهور حال حاضر ایران چه کسی است، نوع اسم، مثلاً روزنامه در مقابل مجله، ...)

نکته‌ی قابل توجه در اینجا این است که کد کردن دستی همه‌ی این اطلاعات، کار بسیار مشکل و طاقت‌فرسایی است.

در مور رویکرد آماری، می‌توان از الگوریتم‌های بسیار پیچیده بهره برد اما برای اینکه موضوع در اینجا بهتر نشن داده شود، رویکرد ساده‌ای را پیش می‌گیریم:

---

[۲۷] Determiner

جدول تابع دسته‌بندی برای یافتن حرف تعریف

حرف تعریف	اولین حضور	جمع	اسم
The	بلی	خیر	مدافع (Defendant)
Null	خیر	بلی	ماشین‌ها (Cars)
The	خیر	خیر	FBI
A	بلی	خیر	کنسرت (Consert)

- یک مجموعه‌ی بزرگ از متن‌های مرتبط با حوزه‌ی خود را جمع کنید (به عنوان مثال متن یک روزنامه)

- برای هر اسمی که در طی یادگیری دیده می‌شود، احتمال اینکه آن اسم حرف تعریف مشخصی بگیرد را محاسبه کنید: تعداد تکرار زمان‌هایی که یک اسم بعد از یک حرف تعریف آمده است.

- برای یک اسم جدید داده شده، حرف تعریفی با بالاترین احتمال را به همان صورت که در جریان آموزش تخمین زده شده است، پیدا کنید.

این رویکرد بسیار ساده در روی ۲۱ قسمت اول مجله‌ی وال استریت [۲۸] آموزش داده شد و با استفاده از قسمت بیست و سوم آن آزموده شد. در اینجا دقت پیش‌بینی به ۵/۷۱ درصد رسید. نتایج عالی نبودند اما برای یک چنین روش ساده‌ای به صورت باورنکردنی خوب محسوب می‌شوند. دلیل این نتیجه‌ی نسبتاً وب مشخصاً این است که بخش بزرگی از اسم‌ها در این مجله با یک حرف تعریف ظاهر می‌شوند. برای مثال "the FBI" "the و یا defendant" "the و ... یک رویکرد دیگر در حل چنین مسائلی که اتفاقاً در تشخیص واحدهای اسمی نیز مورد استفاده قرار می‌گیرد، استفاده از دسته‌بندی [۲۹] است. به جدول توجه کنید: در اینجا هدف یادگیری تابع دسته‌بندی است که بتواند مثال‌های جدید را پیش‌بینی کند. در این تابع سعی شده است که با استفاده از چند خصوصیت (جمع یا مفرد بودن) و اولین حضور، دسته‌ی مورد نظر که در اینجا حرف تعریف است، مشخص

[۲۸] Wall Street Journal

[۲۹] Classification

شود. بسیاری از کاربردهای دسته‌بندی در پردازش زبان طبیعی می‌توانند به صورت نگاشتی از یک مجموعه‌ی پیچیده به مجموعه‌ی دیگری نگریسته شوند، برای مثال:

- تجزیه[۳۰] : رشته‌ها به درخت‌ها، منظور چیزی شبیه به آنست که در شکل آورده شده است.
- ترجمه‌ی ماشینی: رشته‌ها به رشته‌ها
- تولید زبان طبیعی: مدخل پایگاه‌های داده به رشته‌ها[؟].

در ادامه‌ی بحث مخصوصا در بررسی الگوریتم‌های مختلف پیاده‌سازی تشخیص واحدهای اسمی به اصطلاح چندوزنی‌ها[۳۱] برمی‌خوریم، که به توضیح کوتاهی در این زمینه می‌پردازیم. یک چندوزنی یک زیرتوالی از n آیتم درون یک توالی است. آیتم‌ها در توالی می‌توانند واج‌ها، بخش‌ها (سیلاب‌ها)، حروف، کلمات یا جفت‌های بنیادی با توجه به کاربرد مورد نظر باشند. یک چندوزنی با اندازه‌ی یک‌وزنی[۳۲] ، اندازه۲ یک دووزنی[۳۳] ، اندازه‌ی ۳ یک سه‌وزنی [۳۴] نامیده می‌شوند. اندازه‌های ۴ و بیشتر نیز اغلب با همان اصطلاح کلی چندوزنی مشخص می‌شوند. بعضی مدل‌های زبانی که از چندوزنی ها ساخته‌شده‌اند، مدل‌های مارکوف مرتبه n-۱ هستند. یک مدل چندوزنی یک نوع از مدل احتمالی برای پیش‌بینی آیتم بعدی با استفاده از n-۱ آیتم قبلی در توالی است. مفهوم چندوزنی‌ها در ناحیه‌های مختلفی از پردازش زبان طبیعی به صورت آماری و تحلیل توالی ژنتیک استفاده می‌شوند. در این روش آماری تنها مقدار این احتمال بررسی می‌شود که مثلا اگر توالی "for ex" دیده شود، با چه احتمالی حرف بعدی "a" است؟ با چه احتمالی و الی آخر. (که طبیعتا جمع این احتمالات یک خواهد شد.)

در زیر مثال‌هایی از کلمات مرتبه‌ی سه‌وزنی و چهاروزنی را که از پیکره‌ی متنی چندوزنی شرکت گوگل استخراج‌شده‌اند به همراه تعداد تکرار آن‌ها در پیکره‌ی متنی آورده شده‌اند:

---

[۳۰] Parsing

[۳۱] n-gram

[۳۲] unigram

[۳۳] bigram

[۳۴] trigram

سه وزنی:

ceramics collectables collectibles (55)

ceramics collectables fine (130)

ceramics collected by (52)

ceramics collectible pottery (50)

ceramics collectibles cooking (45)

چهاروزنی:

serve as the incoming (92)

serve as the incubator (99)

serve as the independent (794)

serve as the index (223)

serve as the indication (72)

serve as the indicator (120)

برای دنباله‌ای از کلمات، مثلا برای جمله‌ی "The dog smelled like a skunk" ، سه وزنی‌ها می‌توانند این موارد باشند: "The dog" ، "the dog smelled" ، "dog smelled like" ، "smelled like a" ، "like a skunk" و "a skunk"

برای یک توالی از کاراکترها نیز، تعدادی از سه‌وزنی‌های ساخته شده از عبارت "good morning" موارد زیر هستند: "goo" ، "ood" ، " od" ، "d m" ، " mo" ، "mor"[۹]

عملیات تشخیص واحدهای اسمی

تشخیص واحدهای[۳۵] اسمی که به آن بازیابی واحدهای اسمی[۳۶] و نیز استخراج واحدهای[۳۷] اسمی اطلاق می‌شود، یک زیر وظیفه از استخراج اطلاعات است و به معنی پردازش مستنداتی است که در آن به دنبال مکان‌یابی عناصر اسمی در متن و دسته‌بندی آنها به رده‌های از پیش تعیین شده مانند اسامی اشخاص، سازمان‌ها (شرکت‌ها، سازمان‌های دولتی و غیره)، مکان‌ها (شهرها، کشورها، رودخانه‌ها و غیره)، عبارت‌های زمانی، کمیت‌ها، مقدارهای پولی، درصدها و غیره هستیم. تشخیص واحدهای اسمی یک وظیفه‌ی پایه‌ای است و یکی از هسته‌های پردازش زبان طبیعی محسوب می‌شود. این عملیات شامل دو وظیفه است: تشخیص اسامی مربوطه در متن و سپس دسته‌بندی این اسامی به مجموعه‌های از پیش تعیین شده. تحقیقات اصلی بر روی سیستم‌های تشخیص واحدهای اسمی بر اساس دریافت متنی مانند "Jim bought 300 shares of Acme Corp. in 2006." و تولید یک بلوک متنی به صورت زیر است:

<ENAMEX TYPE="PERSON">Jim</ENAMEX> bought <NUMEX TYPE="QUANTITY">300</NUMEX> shares of <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in <TIMEX TYPE="DATE">2006</TIMEX>.

در این مثال نشانه‌گذاری‌ها با توجه به قراردادهای تعیین شده در کنفرانس تشخیص پیام[۳۸] یا اختصارا MUC در سال ۱۹۹۰ مشخص شده‌اند که اصطلاحا به آن‌ها برچسب‌های ENAMEX گفته می‌شود. در حقیقت کنفرانس‌های MUC رویدادهایی بودند که به صورت جدی در مورد تحقیقات در این حوزه برگزار شده‌اند و نیز بستر آزمون‌هایی برای این حوزه را فراهم کرده‌اند.

برای انسان‌ها، تشخیص واحدهای اسمی ساده به نظر می‌رسد. زیرا بسیاری از واحدهای اسمی

[۳۵] Entity Identification

[۳۶] Named Entity Recognition

[۳۷] Entity Extraction

[۳۸] Message Understanding Conference

نام‌های ساده هستند. بسیاری از آن‌ها با حروف بزرگ در زبان‌هایی مانند زبان انگلیسی آغاز می‌شوند، اما برای ماشین‌ها این عمل بسیار مشکل است. ممکن است گمان کنید واحدهای اسمی به راحتی می‌توانند توسط لغت‌نامه‌هایی دسته‌بندی شوند زیرا بسیاری از اسامی، نام‌های ساده و شناخته شده هستند، اما این عقیده اشتباه است. در طول زمان اسامی جدیدی به طور متناوب ساخته می‌شوند.بنابراین، امکان ندارد که همه‌ی آن‌ها را به لغت‌نامه‌هایی اضافه کنیم. به علاوه اگر اسامی درون لغت‌نامه‌ها موجود باشند باز هم این تصمیم روی مفهوم آن‌ها مشکل است.

سیستم‌های مدرن تشخیص واحدهای اسمی برای زبان انگلیسی کارایی نزدیک به انسان تولید می‌کنند. برای مثال امتیاز F-measure بهترین سیستم ارائه شده در ششمین کنفرانس MUC به ۹۳/۳۹ درصد رسید در حالی که نماینده‌های انسانی امتیازهای ۹۷/۶۰ درصد و ۹۶/۹۵ درصد را به دست آوردند. این نتایج نشان می‌دهد که متاسفانه میزان خطای این الگوریتم‌ها درصد) بیش از دو برابر خطای انسانی درصد و ۲/۴۰ درصد) است. مهم‌ترین مسأله‌ها در تشخیص واحدهای اسمی این است که آن‌ها دارای ابهام معنایی هستند. به بیان دیگر یک اسم می‌تواند دارای مفاهیم متفاوتی با توجه به متن باشد. برای مثال چه زمانی "The White house" به معنی یک سازمان و چه زمانی به معنی یک مکان است؟ چه زمانی "June" نام یک فرد محسوب می‌شود و چه زمانی نام یک ماه؟ در عبارت "He visited Bush at White house." ، عبارت "White House" یک مکان است اما در "White house announced the list of ministry candidate." ، به معنای سازمان به کار رفته است.

تحقیقات نشان داده‌اند که سیستم‌های تشخیص واحدهای اسمی که برای یک دامنه‌ی خاص توسعه داده شده‌اند، روی دامنه‌های دیگر خوب عمل نمی‌کنند. تحقیقات اولیه بر روی این سیستم در دهه‌ی ۹۰ به طور کلی به استخراج از مقالات روزنامه و مجلات انجام شد. سپس توجه به سمت پردازش مخابره‌ها و گزارش‌های نظامی جلب شد. تقریبا از سال ۱۹۹۸ علاقه‌مندی و نیز کاربردهای جالب در زمینه‌ی استفاده از تشخیص واحدهای اسمی در زیست‌شناسی مولکولی، بیوانفورماتیک و انجمن‌های پردازش زبان طبیعی ایجاد شد. علاقه‌مندی اصلی در این زمینه نام‌گذاری ژن‌ها و تولیدهای آن‌ها بود.

ارزشیابی سیستم‌های تشخیص واحدهای اسمی برای پیشرفت علمی این عرصه بسیار مهم هستند. بیشتر ارزشیابی‌های این سیستم‌ها در طی کنفرانس‌ها یا مسابقاتی که توسط سازمان‌های دولتی برگزار می‌شوند، صورت می‌گیرد که در بسیاری از اوقات از محیط‌های دانشگاهی کمک گرفته می‌شود.

کاربردهای عملیات تشخیص واحدهای اسمی

تشخیص و استخراج واحدهای اسمی دقیق در داده‌کاوی اطلاعات از میان متون و منابع الکترونیکی بسیار مفید هستند. اگر این عملیات به درستی انجام شود، برای حل بسیاری از مشکلات در به‌روزترین زمینه‌ها در رشته‌هایی چون پاسخگویی به پرسش‌ها، سیستم‌های خلاصه سازی، بازیابی اطلاعات، استخراج اطلاعات، ترجمه‌ی ماشینی، نشانه‌گذاری ویدئویی، جستجوی معنایی وب و زیست شناسی، می‌تواند مثمر ثمر باشد[؟].

انواع واحدهای اسمی

در عبارت واحدهای اسمی کلمه‌ی «اسمی»، وظیفه تعیین شده را به تشخیص عبارات مشخص‌کننده‌ی نوع یا زمان خاص به صورتی که معنای عام نداشته باشد (مثلا اشاره به سال ۱۹۹۱ در مقایسه با اشاره به واژه‌ی عام و کلی سال) و نیز نام‌های ساده و اصطلاحات شبه طبیعی مانند گونه‌ها و مواد بیولوژیکی محدود می‌کند. یک توافق کلی برای شامل شدن اصطلاحات زمانی و بعضی از عبارات عددی مانند پول، درصد و ... جزء وظایف تشخیص واحدهای اسمی به عنوان عبارات اسمی نیز وجود دارد. البته بایستی همان طور که در بالا اشاره شد موارد کلی و عام را از این میان حذف کنیم مثلا واژه‌ی "June" در عبارت "I take my vacation in June" یک عبارت اسمی نیست چون ممکن است به ماه ژوئن در هر سالی اشاره کند. اما June 2007 یک عبارت اسمی است. البته به دلایل عملی گاهی اینگونه عبارات نیز در عملیات تشخیص واحدهای اسمی استخراج شده اند اما به طور کلی جزئی از آن محسوب نمی‌شوند[؟].

حداقل دو نوع از سلسله مراتب انواع عبارات اسمی معرفی شده اند. طبقه‌بندی‌های BBN[۳۹] که

[۳۹] BBN Categories. BBN ، نام شرکتی است که از اول نام مؤسسان آن، Bolt، Beranek و Newman گرفته شده است.

در سال ۲۰۰۲ پیشنهاد شد که برای پاسخگویی به پرسش‌ها استفاده شد و شامل ۲۹ نوع و ۶۴ زیر-نوع بود. سلسله مراتب توسعه یافته‌ی سکاین[۴۰] نیز در سال ۲۰۰۲ پیشنهاد شد و بعدها نیز بیشتر توسعه داده شد. این سلسله مراتب زیر-نوع تشکیل شده است. در بخش توضیحات بیشتری را در مورد این طبقه‌بندی‌ها خواهید یافت [؟ ؟].

پردازش زبان طبیعی و زبان‌های آسیای جنوبی

بعد از گذر سال‌های بسیاری از فعالیت‌های تحقیقاتی بر روی پردازش زبان طبیعی بر روی زبان‌های انگلیسی، اروپایی و زبان‌های آسیای شرقی، پردازش زبان طبیعی به صورت خودکار، به یک موضوع پرطرفدار برای تحقیق تبدیل شده است. اما متاسفانه همانند زبان فارسی، به زبان‌های آسیای جنوبی که شباهت‌های بسیاری به زبان ما دارند توجه کمی شده است. در این بخش به بررسی مشکلات موجود در مورد پردازش طبیعی زبان‌های آسیای جنوبی صحبت خواهیم کرد. مطالب این بخش به طور کلی به زبان‌های آسیای جنوبی پرداخته است زیرا این زبان‌ها همگی شباهت‌هایی به فارسی دارند و مشکلات بررسی شده در این بخش مشکلات زبان فارسی هم هست. این زبان‌ها از لحاظ رسم‌الخط و در خیلی از موارد دیگر از جمله دستور زبان به یکدیگر شبیه هستند. برای مثال، در مورد زبان اردو،

- اردو یک کلمه‌ی برگرفته از زبان ترکی است و معنی آن لشکر است.
- اردو یک زبان هندی-اروپایی از خانواده‌ی هندی-ایرانی است که در هند و پاکستان به این زبان صحبت می‌شود.
- رسم‌الخط زبان اردو مانند زبان‌های فارسی و عربی، از راست به چپ نگاشته می‌شود.
- این زبان، زبان رسمی کشور پاکستان است که ۱۱ میلیون گوینده دارد.
- در میان تمام زبان‌های دنیا، زبان هندی بیش‌ترین شباهت را به اردو دارد زیرا هر دوی این زبان در شهر دهلی متولد شده‌اند. البته به همان میزان که هندی بسیاری از کلمات خود را از

[۴۰] Sekine's Extended Hierarchy

زبان سانسکریت گرفته است، اردو تعداد بسیار زیادی از لغات خود را از زبان‌های فارسی و عربی گرفته است. (تعدادی کلمه نیز از زبان‌های ترکی، پرتقالی و انگلیسی وارد زبان اردو شده‌اند). البته تعداد زیادی از کلمات این زبان که در فارسی هم وجود دارند، از نظر معنی با آن متفاوتند.

- ...

یکی از ابعاد مهم دستور زبانِ زبان‌های مورد بحث در این بخش، ترتیب کلمه‌ای SOV یا فاعل- مفعول- فعل آن‌هاست که در آن‌ها در بسیاری از مواقع ضمیر فاعلی حذف می‌شود. همچنین در این زبان‌ها شباهت‌های بسیارِ دیگری به فارسی مانند طرز ساخت کلمات از ریشه‌ی فعلی وجود دارد یا کلمات مشترک زیادی بین آن‌ها موجود است که همین موارد ما را به بررسی کارهای صورت گرفته در پردازش زبان طبیعی این زبان‌ها ترغیب می‌کند.

همچنین در مورد رسم‌الخط این زبان‌ها شباهت‌های زیادی وجود دارند. برای مثال الفبای آن‌ها از زبان عربی ریشه گرفته است (الفبای زبان اردو از زبان فارسی مشتق شده است که خود الفبای فارسی از عربی مشتق شده است.) و مانند آن حروف از راست به چپ و اعداد از چپ به راست نوشته می‌شوند. در بیشتر آن‌ها نیز، مانند زبان فارسی، حروف با توجه به قرارگیری در متن تغییر می‌کنند که در مورد آن‌ها اصطلاح حساس به متن[۴۱] را به کار می‌برند. همچنین در آن‌ها اِعراب وجود دارند. با توجه به اینکه گاهی این اعراب نوشته می‌شوند و گاهی نوشته نمی‌شوند، مشکلات مشابهی با زبان فارسی در پردازش این زبان‌ها وجود دارد.

در مورد این زبان‌ها، همواره مؤسساتی سعی نموده‌اند تا استانداردهای خاصی را برای رسم‌الخط آن‌ها وضع کنند، هر چند هنوز بسیاری از این استانداردها توسط برنامه‌نویسان پیاده‌سازی نمی‌شوند[؟ ؟].

[۴۱] Context Sensitive

پردازش زبان طبیعی به کمک پیکره‌ی متنی

نیاز اصلی برای پردازش آماری زبان طبیعی بر پایه‌ی پیکره‌ی متنی یا corpora است که در تئوری کلمه‌ی Corpus (مفرد Corpora) سرواژه‌ی عبارت انگلیسی "Capable Of Representing Potentially Unlimited Selections of texts" است که در زبان فارسی به آن پیکره‌ی متنی می‌گویند. اهمیت پیکره‌ی متنی هم در زبان شناسی و هم در فناوری‌های مبتنی بر زبان دیده می‌شود. در چند سال اخیر مرکز توجه زبان‌شناسی بر پایه‌ی پیکره‌ی متنی، بر روی زبان انگیسی، زبان‌های اروپایی و آسیای شرقی بوده است. در هر حال، بر خلاف توجه بر زبان‌های یادشده و انتشار پیکره‌های متنی مختلف در مورد آن‌ها، توجه بسیار کمی به این مقوله در زبان‌های آسیای جنوبی شده است (در مورد زبان فارسی، در تابستان سال ۱۳۸۸ یک سند راهبردی در شورای عالی اطلاع رسانی برای آماده‌سازی یک پیکره‌ی متنی برای زبان فارسی نوشته شده است که امیدواریم در آینده یک پیکره‌ی متنی قابل قبول در اختیار پژوهشگران قرار گیرد). تعداد زیادی از زبان‌های مختلف در آسیای جنوبی وجود دارند ولی بر خلاف این قضیه، پیکره‌ی متنی عمومی قابل دسترسی در مورد زبان‌های آسیای جنوبی بسیار محدود هستند. همچنین در عین حال که تشکل‌های مهندسی زبان و زبان‌شناسان، مشتاق و البته نگران سریع‌تر پیش بردن تحقیقات در زبان‌های آسیای جنوبی هستند، اما با مشکل کمبود پیکره‌های متنی مناسب مواجهند. تعداد زیادی از محققان تصمیم دارند مطالعات خود را روی زبان‌های جنوب آسیا متمرکز کنند اما کمیابی منابع پیکره‌ی متنی در این زبان‌ها، خط تحقیقات را عقب می‌اندازد. این عقب‌ماندگی چند دلیل اصلی دارد. مهم‌ترین این دلایل کمبود هماهنگی و همکاری در توسعه‌ی پیکره‌ی متنی در این زبان‌هاست، که به نظر پایه‌ای‌ترین مشکل می‌باشد. همچنین هیچ سازمانی در آسیای جنوبی وجود ندارد که به صورت واقعی موازی با سازمان پیکره‌ی متنی زبان‌های اروپایی (ELRA)[۴۲] پیشروی کند. بنابراین هیچ هماهنگی برای پیشرفت پیکره‌ی متنی زبان‌های آسیای جنوبی وجود ندارد. افزون بر این هیچ استانداردی برای نگهداری و کد کردن زبان‌های مرتبط در پیکره‌های متنی مورد نظر وجود ندارد. مشکل دیگر در ساخت پیکره‌ی متنی برای این زبان‌ها در این است که دسترسی به متون الکترونیکی محدودی وجود دارد. همچنین عملاً نمی‌توانیم از اطلاعات موجود در این زبان‌ها به صورت غیر الکترونیکی نیز

[۴۲] European Language Resource Association

برای تبدیل آن‌ها با استفاده از OCR به اطلاعات الکترونیکی استفاده کنیم زیرا این فناوری در همه‌ی این زبان‌ها فعلاً بسیار نوپا است. علاوه بر این ترجیح داده می‌شود که این اطلاعات در قالب فرمت یونیکد باشند اما متأسفانه بیشتر اطلاعات موجود برای زبان‌های هندی و اروپایی با فرمت‌ها و متدولوژی‌های متفاوتی نگه‌داری می‌شوند. در صورت استفاده از فرمت یونیکد به محققان دیگرکشورها امکان بهتری برای خواندن و تحلیل‌های بیشتر را می‌دهیم. برای مثال، منابع مذکور اغلب به صورت عکس‌ها یا فرمت PDF نگهداری می‌شوند که امکان بیرون کشیدن متون واقعی به دستخط فارسی برای آن‌ها وجود ندارد. همچنین منابع PDF شامل جداول و ستون‌ها و تصاویر نیز هستند. یا مثلا برای زبان اردو این منابع از بسته‌های افزونه‌ای مانند "Urdu98" برای نمایش متون در مرورگر وب استفاده می‌کنند. (برای فارسی سیستم‌های قبل از «ویندوز ۲۰۰» نیز شرایط مشابهی داشتند. مانند سیستم عامل «پارو۹» که تغییراتی در «ویندوز ۹» ایجاد می‌کرد) علی‌رغم تمامی این مشکلات تعدادی از انیستیتوها و سازمان‌ها عملاً بر روی فراهم آوردن پیکره‌ی متنی این زبان‌ها تلاش می‌کنند. برای مثال در سال ۲۰۰۴ دانشگاه لانچستر [۴۳] پروژه‌ای را با نام پیکره‌ی متنی "EMILLE" به پایان رساند که بر روی ۱۴ زبان آسیای جنوبی کار شده بود: آسامی، بنگالی، کوجاراتی، هندی، کانادا، کشمیری، مالایالام، ماهاراتی، اوریا، پنجابی، سینهالا، تامیل، تلگو و اردو. همچنین منابع گفتاری استانداردی هم برای این زبان‌ها جمع‌آوری شده‌اند. به علاوه تمامی این منابع مبتنی بر استاندارد کدگذاری پیکره‌ی متنی[۴۴] یا CES و بر اساس زبان نشانه‌گذاری همه منظوره‌ی استانداردشده[۴۵] یا SGML نشانه‌گذاری شده‌اند. تحقیقات انجام شده بر روی زبان‌های آسیای جنوبی قابل قبول هستند اما کافی نیستند زیرا در مقایسه با زبان‌های شرقی آسیا هنوز کار زیادی روی آن‌ها انجام نشده است. در دهه‌ی گذشته توسعه‌ی سریعی روی زبان‌های آسیای شرقی وجود داشته است و پیشرفت‌های زیادی روی جمع‌آوری و ساخت پیکره‌ی متنی روی زبان‌هایی مانند چینی ماندریایی حاصل شده است.

همچنین نیاز بسیار زیادی به توسعه‌ی پیکره‌ی گفتاری در این زبان‌ها احساس می‌شود. محققان اهمیت وجود پیکره‌ی گفتاری را در توسعه‌ی زبان‌شناسی بسیار زیاد قلمداد می‌کنند. برای مثال

[۴۳] Lanchester University

[۴۴] Standard Corpus Encoding

[۴۵] Standardized Generalized Markup Language

بدون وجودپیکره‌ی گفتاری کافی نمی‌توان به توسعه‌ی منابع زبانی و در نتیجه سیستم‌های پردازش زبان طبیعی کاملا احساس می‌شود.

همچنین منابع زبانی دو زبانه نیز می‌توانند نیاز بعدی ما در زمینه‌های توسعه‌ای جدیدتر باشند. زیرا این منابع، منابع ارزشمندی برای توسعه‌ی سیستم‌های ترجمه‌ی ماشینی هستند که به طور معمول ورودی آن‌ها منابع زبانی دو زبانه است[؟].

نشانه‌گذاری ادات سخن

به منظور افزایش دادن قابلیت اطمینان و دقت هر سیستم پردازش زبان طبیعی بایستی این توانایی را داشته باشیم که اطلاعات زبان‌شناسی را به طور خودکار از یک منبع زبانی متنی ساده استخراج کنیم. وجود اطلاعات زبان‌شناسی قابل استخراج بیشتر، یک سیستم پردازش زبان طبیعی را دقیق‌تر و قابل اطمینان‌تر می‌کند. روش‌های زیادی برای نشانه‌گذاری زبان‌های انگلیسی، اروپایی و آسیای شرقی به وجود آمده‌اند. ابتدایی‌ترین روش مبتنی بر قوانین توسط کلین و سایمونز[۴۶] در سال ۱۹۶۳ و گرین و رابین[۴۷] در ۱۹۷۱ برای نشانه‌گذاری پیکره‌ی متنی براون[۴۸] توسعه داده شدند. هر دوی این نشانه‌گذاری‌ها به نرخ خطایی بهتر از سیستم‌های زمان خودشان دست‌یافتند. نشانه‌گذاری که در سال ۱۹۹۲ توسط اریک بریل معرفی شد، به درستی حدود ۹۶ درصد دست یافت. بعد از آن درستی نشانه‌گذاری تا ۹۷/۵ درصد در سال ۱۹۹۴ افزایش یافت. در چند سال اخیر تلاش‌های بیشتری روی توسعه‌ی مدل‌های آماری برای تشخیص درست نشانه‌ها به کار گرفته‌شده‌است، برای مثال استفاده از مدل مخفی مارکوف توسط ویشِدِل و مارکوس[۴۹] ، استفاده از درخت تصمیم آماری توسط ژلینک و مگرمن[۵۰]. با توجه به اطلاعات ما از میان زبان‌های آسیای جنوبی تا سال ۲۰۰۶ تنها برای زبان اردو فقط یک نشانه‌گذار ادات سخن که توسط اندرو هاردی [۵۱] توسعه داده شده

[۴۶] Klein And Simons

[۴۷] Greene And Rubin

[۴۸] Brown Corpus

[۴۹] Weischedel And Marcus

[۵۰] Jelinek And Magerman

[۵۱] Andrew Hardy

است، وجود دارد که حدود ۲۷۰ قانون نوشته شده دارد و دقتی حدود ۹۰ درصد و یک میزان ابهام بسیار بالا و ۲/۵ نشانه در هر کلمه دارد[؟].

تشخیص واحدهای اسمی

مثال‌های مختلفی از تشخیص واحدهای اسمی مانند اندیس‌گذاری خودکار کتاب‌ها، سازمان‌دهی کلی مستندات و ... وجود دارد. کنفرانس‌های فهم ماشینی از جمله MUC-6 و MUC-7 همان طور که پیش‌تر نیز اشاره شد، به طور کامل از تشخیص واحدهای اسمی به زبان انگلیسی پشتیبانی کردند. این کنفرانس‌ها توسط واحد تحقیقات کارگزار دفاعی آمریکا برگزار شدند. هدف اصلی این کنفرانس‌ها ارزیابی سیستم‌های استخراج اطلاعات بوده است. در ژاپن، تلاش واحدی چندزبانه[۵۲] یا MET ، هدفی مشابه کنفرانس‌های مزبور را برای زبان ژاپنی پیاده‌سازی کردند. اما در نهایت تلاش بسیار کمی نه تنها در زبان فارسی بلکه در مورد زبان‌های آسیای جنوبی (که در بخش در مورد آن‌ها صحبت کرده‌ایم) در این زمینه انجام شده است. در عین حال بیشتر این زبان‌ها مشکلاتی از قبیل ابهام در نقطه‌گذاری و تأکیدها، نداشتن بزرگی و کوچکی حروف، نداشتن راهنماهای املایی برای تشخیص مخفف‌ها ونیز وجود حروف در چهار حالت ابتدایی، میانی، انتهایی و جدا را دارند. همچنین موتور جستجوی خاصی برای این زبان‌ها وجود ندارد در حالی که مثلا برای زبان چینی موتور جستجویی مخصوص این زبان ساخته شده است[؟ ؟].

جمع‌بندی

در این بخش ما ابتدا به شرح کلی مسائل پردازش زبان طبیعی پرداختیم و جایگاه مسأله‌ی تشخیص واحدهای اسمی را در پردازش زبان طبیعی بررسی نمودیم. همچنین به مشکلات اصلی پیش رو در حل این گونه از مسائل پرداختیم. در این میان اصلی‌ترین تلاش‌های انجام شده در تشخیص واحدهای اسمی در زبان‌های اروپایی بیان شدند. علاوه بر این به طور خاص تحقیقات انجام شده بر روی زبان‌های آسیایی جنوبی به دلیل شباهت آن‌ها به زبان عربی بررسی شدند. البته پژوهش‌های

[۵۲] Multilingual Entity Task

انجام شده در مسأله‌ی تشخیص واحدهای اسمی در زبان عربی نیز به طور مفصل در فصل ۴ شرح داده می‌شود. در فصل آینده، به طور کامل‌تر مسأله‌ی تشخیص واحدهای اسمی را شرح داده و به بررسی راه حل‌های ممکن برای حل این مسأله خواهیم پرداخت. همچنین مثال‌های بیشتری از کاربردهای آن در عملیات‌های دیگر پردازش زبان طبیعی را در فصل بعدی خواهید خواند.

فصل ۲

بررسی کلی عملیات تشخیص واحدهای اسمی

مقدمه

در سال‌های اخیر سیستم‌های تشخیص و استخراج خودکار واحدهای اسمی به یکی از محدوده‌های پرطرفدار تحقیقاتی تبدیل شده‌اند. این سیستم‌ها را می‌توان به سه کلاس اصلی تقسیم‌بندی کرد. روش‌های مبتنی بر قوانینی که به صورت دستی مشخص شده‌اند[۱] ، سیستم‌های مبتنی بر یادگیری ماشینی[۲] و سیستم‌هایی که ترکیبی از دو روش قبل هستند.

روش‌های مبتنی بر قوانین

رهیافت‌های مبتنی بر قوانین، روی استخراج نام‌ها با استفاده از تعداد زیادی از مجموعه قوانین ساخته شده به صورت دستی تمرکز نموده اند. به طور کلی این سیستم‌ها مجموعه‌ای از الگوها

[۱] Hand-made Rule-based approaches

[۲] Machine Learning -base approaches

هستند که از مشخصه‌های گرامری (مانند ادات سخن[۳])، نحوی (مانند تقدم کلمات) و املایی (مانند بزرگی و کوچکی حروف) در کنار ترکیبی از لغت‌نامه‌ها استفاده می‌کنند. برای مثال در جمله‌ی "President Bush said Monday's talks will include discussions on security and time table for U.S. forces to leave Iraq" یک نام ساده در دنباله‌ی یک عنوان شخصی (President) آمده است، پس این نام یک نام متعلق به یک شخص است. و نیز در ادامه یک نام دیگر آمده است که با یک حرف بزرگ شروع شده (Iraq) و به دنبال یک فعل (to leave) آمده است که مشخصاً نام یک مکان است. در این نوع از رهیافت، اپلت[۴] و همکارانش، یک سیستم تشخیص اسامی بر پایه‌ی عبارات منظمی که با دقت بسیار و به صورت دستی استخراج شده‌اند را پیشنهاد داده‌اند. این عبارات منظم FASTUS نامیده می‌شوند. آن‌ها این وظیفه را به سه قسمت تقسیم کرده‌اند: تشخیص عبارات، تشخیص الگوها و سپس ترکیب مهم‌ترین یافته‌ها. همچنین ایوانسکا[۵] از منابع خاص توسعه یافته‌ای مانند فرهنگ جغرافیایی، صفحات سفید و صفحات زرد در کنار این روش استفاده کرده است. مورگان[۶] با هدفی مشابه، از یک آنالیز زبان‌شناسانه‌ی قدرتمند استفاده کرده است. این رهیافت‌ها بر پایه‌ی استفاده از قوانین کد شده به صورت دستی و مستندات تفسیر شده به صورت دستی کار می‌کنند. این نوع از مدل‌ها برای دامنه‌های خاص نتایج بهتری را حاصل کرده‌اند و توانایی تشخیص واحدهای پیچیده‌ای را دارند که روش‌ها مبتنی بر یادگیری ماشین به سختی می‌توانند آن‌ها را تشخیص دهند. به هر حال روش‌های مبتنی بر قوانین، مشکل غیر قابل حمل بودن و نیز کمبود قابلیت اطمینان را دارند. علاوه بر آن هزینه‌ی بالای ایجاد قوانین نیز به مشکلات این روش اضافه می‌شود حتی اگر تغییر کوچکی در داده‌های ما ایجاد شود، این هزینه افزایش می‌یابد. این نوع از رهیافت‌ها اغلب منحصر به زبان یا حوزه‌ی خاص می‌شوند و نمی‌توانند به طرز مناسبی با زبان‌ها و حوزه‌های خاص جدید منطبق شوند [؟].

---

[۳] Part of Speech

[۴] Appelt

[۵] Iwanska

[۶] Morgan

روش‌های مبتنی بر یادگیری ماشین

در سیستم‌های مبتنی بر یادگیری ماشین، هدف از رهیافت تشخیص واحدهای اسمی تبدیل مساله‌ی تشخیص به مساله‌ی دسته‌بندی و از یک مدل آماری دسته‌بندی برای حل این مساله استفاده می‌شود. در این روش مدل به دنبال تشخیص الگوها و یافتن رابطه‌ی آن‌ها با متن و ساختن یک مدل آماری و الگوریتم‌های یادگیری ماشین است. این سیستم‌ها نام‌ها را یافته و آن‌ها را بر اساس مدل به دست آمده با استفاده از روش‌های یادگیری ماشین به کلاس‌های از پیش تعیین شده مانند اشخاص، مکان‌ها، زمان‌ها و غیره تقسیم می‌کنند. دو نوع مدل یادگیری ماشین وجود دارند که برای تشخیص واحدهای اسمی مورد استفاده قرار می‌گیرند؛ یادگیری با ناظر از برنامه‌ای استفاده می‌کند که با استفاده از یک مجموعه از مثال‌های برچسب‌گذاری شده دسته‌بندی کردن را یاد می‌گیرد. این روند آموزش، با ناظر نامیده می‌شود زیرا افرادی که مثال‌های ذکر شده در بالا را برچسب‌گذاری کرده‌اند، تفاوت‌های درست را به برنامه آموزش داده‌اند. روش‌های با ناظر به آماده کردن داده‌های برچسب‌گذاری شده برای ساخت یک مدل آماری نیاز دارند اما به دلیل مشکل جدا افتادگی داده‌ها[۷] تا هنگامی که از تعداد زیادی داده استفاده نشود کارایی خوبی به دست نخواهد آمد. در سال‌های اخیر تعداد زیادی از روش‌های آماری بر پایه‌ی روش آموزش با ناظر ارائه شده‌اند. بایکل[۸] و همکارانش یک اسم‌یاب آموزش پذیر به نام نایمبل[۹] را مبتنی بر مدل مخفی مارکوف معرفی کرده‌اند. بورثویک[۱۰] و همکارانش منابع دانش عظیمی را با استفاده از روش بیشترین آنتروپی در تشخیص واحدهای اسمی مورد استخراج قرار داده‌اند. علامت‌گذاری نام‌های ساده‌ی ناشناخته با استفاده از درخت تصمیم توسط بِچه[۱۱] و همکارانش پیشنهاد شد. همچنین وو[۱۲] و همکارانش یک سیستم تشخیص واحدهای اسمی را مبتنی بر ماشین‌های بردار پشتیبانی نمایش داده‌اند. روش‌های آموزش بدون ناظر نوع دیگری از مدل یادگیری ماشین هستند که در آن‌ها، مدل

---

[۷] Sparseness of data

[۸] Bikel

[۹] Nymbel

[۱۰] Borthwick

[۱۱] Bechet

[۱۲] Wu

بدون ناظر، بدون هیچ بازخوردی آموزش می‌بیند. در روش بدون ناظر هدف اصلی ساختن یک نمایش از داده‌ها است. این نمایش بعدها می‌تواند برای فشرده سازی، دسته‌بندی، تصمیم‌گیری و اهداف دیگری مورد استفاده قرار گیرد. آموزش بدون ناظر یک روش پرطرفدار برای سیستم‌های تشخیص واحدهای اسمی نیست و سیستم‌هایی که از این روش استفاده می‌کنند اغلب کاملاً بدون ناظر نیستند. در این نوع از رفتارها کالینز[۱۳] و همکارانش یک مدل بدون ناظر برای دسته‌بندی با استفاده از داده‌های بدون برچسب را بررسی کرده‌اند. کوایم[۱۴] و همکارانش مدل‌های تشخیص واحدهای اسمی بدون ناظر و اثر کلی آن‌ها را پیشنهاد داده‌اند و یک لغت‌نامه‌ی واحدهای اسمی در مقیاس کوچک به همراه مجموعه‌ای از مستندات بدون برچسب را برای دسته‌بندی واحدهای اسمی استفاده نموده اند. بر خلاف رهیافت‌های مبتنی بر قوانین، روش‌های یادگیری ماشین به راحتی به حوزه‌ها و زبان‌های دیگر منتقل می‌شوند. در مجموع به طور خلاصه در مورد استفاده از روش‌های یادگیری ماشین می‌توان روش‌هایی مانند مدل مخفی مارکوف، روش بیش‌ترین آنتروپی، درخت‌های تصمیم، میدان‌های تصادفی شرطی یا CRF، ماشین بردار پشتیبان و استفاده از جداکننده‌ی naïve Bayes را برشمرد [؟ ؟].

روش‌های ترکیبی

در سیستم‌های ترکیبی، رهیافت اصلی، ترکیب کردن روش‌های مبتنی بر قوانین و روش‌های مبتنی بر یادگیری ماشین و تولید روش‌های جدیدی است که از نقاط قدرت هر کدام از این روش‌ها بهره می‌برند. در این خانواده از رهیافت‌ها، میخیف[۱۵] و همکارانش یک سیستم ترکیبی مبتنی بر استفاده از مستندات به نام سیستم LTG[۱۶] را پیشنهاد داده‌اند[؟]. سیریهاری[۱۷] و همکارانش نیز یک سیستم ترکیبی از ماشین مخفی مارکوف، بیش‌ترین آنتروپی و روش‌های گرامری که به صورت دستی استخراج شده‌اند را پیشنهاد کرده‌اند. هرچند این نوع از روش‌ها، نتایج بهتری را نسبت به بعضی

[۱۳] Collins

[۱۴] Koim

[۱۵] Mikheef

[۱۶] Language Technology Group

[۱۷] Sirihari

روش‌های دیگر حاصل می‌کنند، اما هنوز هم ضعف روش‌های مبتنی بر قوانین، که همان نیاز به تغییرات عمده هنگام تعویض دامنه است، در آن‌ها وجود دارد [؟ ؟ ؟]. در ادامه درباره ی این روش‌ها بیشتر توضیح خواهیم داد.

معیارهای ارزیابی

واحد اسمی، یک شیئ اسمی مورد توجه مانند یک شخص، سازمان، یا مکان است. این عملیات شامل سه زیروظیفه با نام‌های اسامی نهادی، عبارات زمانی و عبارات عددی است. عباراتی که باید نشانه‌گذاری شوند، شناسه‌های غیرتکراری از واحدها هستند (سازمان‌ها، اشخاص، مکان‌ها) یا ENAMEX، زمان‌ها (تاریخ‌ها، زمان‌ها) یا TIMEX، و کمیت‌ها (مقادیر پولی، درصدها) یا NUMEX. این وظیفه متشکل از شناختن همه‌ی نمونه‌های این سه نوع از عبارات در هر متن از مجموعه‌ی آزمون و دسته‌بندی عبارات به آن‌ها است. از آن جایی که سیستم یا روش باید یک خروجی منحصر به فرد و بدون ابهام برای هر یک از رشته‌های مربوطه در متن بیابد، ارزیابی نمی‌تواند بر اساس دیدی از معماری سیستم خط لوله که در آن تشخیص واحدهای اسمی باید به طور کامل مانند یک پیش پردازش جمله‌ای و تحلیل کلامی کنترل شود، پایه‌گذاری‌گردد. در این عملیات، سیستم بایستی تشخیص دهد یک رشته چه چیزی را نمایش می‌دهد، نه اینکه فقط ظاهرصوری آن چیست. گاهی اوقات جواب درست ظاهر صوری است، حتی می‌توان گفت در اکثر مواقع همین‌طور است، اما نه همواره و می‌توان از طریق تطبیق الگوی محلی به این هدف دست یافت. در موارد دیگر جواب صحیح، ظاهر صوری نیست مانند یک کلمه که با حرف بزرگ شروع شده است و نشانگر یک مکان، شخص یا سازمان است، و پاسخ بایستی با استفاده از تکنیک‌هایی که اطلاعات را از یک متن بزرگ‌تر یا فهرست‌های ارجاع بیرون می‌کشند، به دست آید. یک مدل امتیازدهی که برای ارزشیMUC و عملیات واحدهای چندزبانی[۱۸] مورد استفاده قرار می‌گیرد،دو معیار دقت[۱۹] یا اختصاراً P و بازخوانی[۲۰] یا اختصاراً R است. این معیارها در حقیقت از تعاریف موجود در

[۱۸] Multilingual Entities

[۱۹] Precesion

[۲۰] Recall

حوزه‌ی بازیابی اطلاعات گرفته شده‌اند. و تعریف آن‌ها به صورت زیر است:

$$P = \frac{\text{تعداد پاسخ‌های صحیح}}{\text{تعداد پاسخ‌ها}} \quad (\text{۱})$$

$$R = \frac{\text{تعداد پاسخ‌های صحیح}}{\text{تعداد کلیدهای صحیح}} \quad (\text{۲})$$

از ترکیب این دو معیار از کارایی، یک معیار دیگر حاصل می‌شود که در حقیقت متوسط میانگین همساز[۲۱] این دو معیار است و به آن معیاF یا F-measure می‌گویند:

$$F = \frac{\text{RP}}{\text{۱/۲(R+P)}} \quad (\text{۳})$$

همچنین در این فرمول‌ها، اصطلاح پاسخ[۲۲] در فرمول‌های بالا، اشاره به جواب‌های تحویل داده شده توسط یابنده‌ی اسامی می‌کند و اصطلاح کلید[۲۳] یا فایل کلید[۲۴] به معنای یک فایل نشان‌گذاری شده که حاوی پاسخ‌های صحیح می‌باشد، به کار رفته است.

در MUC-7 ، یک پاسخ صحیح از یک یابنده‌ی نام، پاسخی است که هم برچسب آن و هم هر دوی محدوده‌های آن درست باشند.

سه نوع از برچسب‌ها وجود دارند که هر کدام از آن‌ها از یک صفت[۲۵] را برای مشخص کردن یک واحد خاص استفاده می‌کنند. انواع برچسب‌ها و واحدهایی که به آن‌ها اشاره می‌کنند در زیر آورده شده‌اند:

- هویت یا Entity یا :ENAMEX اشخاص، سازمان‌ها و مکان‌ها.
- عبارت زمانی یا Time expression یا :TIMEX تاریخ، زمان.
- عبارت عددی یا Numeric expression یا :NUMEX زمان، درصد.

---

[۲۱] Harmonic Mean

[۲۲] Response

[۲۳] Key

[۲۴] Key file

[۲۵] Attribute

یک پاسخ نیمه-صحیح، پاسخی است که برچسب (هم نوع و هم صفت) درست است اما فقط یکی از محدوده‌ها صحیح می‌باشد. به گونه‌ای دیگر،یک پاسخ نمیه صحیح است اگر فقط نوع برچسب (و نه صفت آن) و هر دوی محدوده‌های صحیح هستند.

مهم‌ترین طبقه‌بندی‌های واحدهای اسمی

طبقه‌بندی BBN

این طبقه‌بندی توسط شرکت BBN Technologies برای پاسخ‌گویی به پرسش‌ها معرفی شد. شرکت BBN که نام آن برگرفته از نام سه مؤسس آن یعنی بولت، برانک و نیومن[۲۶] است، یک شرکت با فناوری پیشرفته است که سرویس‌های توسعه و تحقیقات را ارایه می‌دهد و بیشتر از هر چیز دیگر به خاطر توسعه‌ی سوئیچینگ بسته‌ای برای شبکه‌های آرپانت و اینترنت شناخته شد.

در مارس سال ۲۰۰، این شرکت یک طبقه‌بندی از واحدهای اسمی با نام «دستورالعمل‌های نشانه‌گذاری برای انواع پاسخ‌ها» را منتشر کرد. شرکت BBN طبقه‌بندی‌ها و دستورالعمل‌های نشانه‌گذاری را برای وظیفه‌ی پاسخگویی به پرسش‌ها ارائه نمود. ۲۹ نوع از طبقه‌بندی برای پاسخ‌ها در جدولی ارائه شد که در زیر برای مثال چند تا از آن‌ها را خواهیم دید. در ستون توضیحات جدول، توضیحات کلی از یک موضوع، بعضی از مثال‌ها و رشته SGML که در نشانه‌گذاری مورد استفاده قرار گرفته است، موجود می‌باشد. برای هر نوعی که شامل زیر-نوع می‌باشد، ستون زیر-نوع مثال‌هایی از هر یک از زیر-نوع‌های این نوع خاص را نشان می‌دهد و مهم‌ترین واحد برای آن نوع و زیر-نوع، درون <> نشان داده شده‌اند همچنین رشته SGML برای زیر-نوع نیز داده شده است. اغلب مثال‌ها از داده‌های درخت‌بانک[۲۷] از وال استریت ژورنال آورده شده است.

این ۲۹ مقوله شامل پنج نوع EDT، یعنی شخص، سازمان، مکان، GPE ، واحدهای تسهیلاتی به همراه بعضی از اصلاحات جزئی به آن دستورالعمل‌ها است. به عنوان مثال، تحت رهنمودهای EDT، موزه‌ها جزئی از واحدهای تسهیلاتی هستند در حالی که ممکن است آن‌ها جزئی از سازمان‌ها

[۲۶] Bolt, Beranek and Newman

[۲۷] Treebank

تلقی شوند.

همچنین طبقه‌بندی‌های ارائه شده، شامل طبقه‌بندی‌های MUC، یعنی پول، درصد، زمان و تاریخ، مجددا با برخی اصلاحات می‌باشند. برای مثال، در طبقه‌بندی ارائه شده، دسته‌ی تاریخ، شامل مدت، تاریخ (نسبی و مطلق) و سن است که هر کدام به عنوان یک زیر-نوع در این نشانه‌گذاری معرفی شده‌اند.

دسته‌های دیگری نیز در طبقه‌بندی BBN وجود دارند که بر اساس ادبیات پاسخ به پرسش‌ها اضافه شده‌اند و نیز تعدادی از آن‌ها موجود هستند که بر اساس آزمایش‌های خود شرکت BBN بر روی داده‌های پاسخ‌گویی به پرسش‌ها مشخص شده‌اند. در جدول زیر، برای آشنایی با این نوع از نشانه‌گذاری، چند نمونه از رده‌های این نوع از طبقه‌بندی در جدول آمده است [؟].

طبقه‌بندی توسعه‌داده‌شده‌ی سکاین

طبقه‌بندی توسعه داده شده‌ی واحدهای اسمی[۲۸] برای تامین نیاز روز افزون برای محدوده وسیع‌تری از واحدهای اسمی توسعه یافته است. این طبقه‌بندی از اولین مجموعه‌ی واحدهای اسمی، مشتق شده است که در سال ۱۹۹۶ توسط گریشمن[۲۹] و همکارانش در MUC تعریف شد. همچنین در توسعه‌ی این طبقه‌بندی به طبقه‌بندی واحدهای اسمی که توسط سکاین و همکارانش در سال ۲۰۰۰ تحت عنوان طبقه‌بندی IREX معرفی شد، توجه شده است. در طول زمان نسخه‌های تازه‌تری از طبقه‌بندی سکاین معرفی شده‌اند. برای مثال در سال ۲۰۰۲ یک طبقه‌بندی شامل ۱۵۰ نوع توسط سکاین معرفی شد. در ژوئن سال ۲۰۰، نسخه‌ی ۶.۱.۰ این طبقه‌بندی معرفی شد و در ۲۰۰۷ نسخه‌ی ۷ آن توسعه داده شد. آنچه که در این گزارش بررسی شده است، نسخه‌ی سال ۲۰۰۳ این طبقه‌بندی است. چون مستندات این طبقه‌بندی به زبان ژاپنی نوشته شده است و سپس به انگلیسی ترجمه گشته، ترجمه‌ی مناسبی از نسخه‌ی ۷ به زبان انگلیسی موجود نبود تا در اینجا بررسی شود. به هر حال چون هدف از آوردن این طبقه‌بندی‌ها در این گزارش، آشنایی اولیه با طبقه‌بندی‌های پیچیده‌تر واحدهای اسمی است، بررسی این نسخه نیز خالی از لطف نیست.

[۲۸] Extended name entity hierarchy

[۲۹] Grishman

جدول طبقه‌بندی BBN

نوع	زیرنوع‌ها	خلاصه‌ی توضیحات
اسم شخص	زیر-نوعی ندارد	نام‌های ساده از مردم، شامل مواردی چون شخصیت‌های داستانی، نام کوچک، نام خانوادگی، نام فرد یا خانواده، لقب منحصر به فرد، نشانگر نسل‌ها مانند جونیور و IV هستند. برای مثال، Ray Garrett Jr. , Henry IV SGML: <ENAMEX TYPE = ”PERSON”> Michael Henderson </ENAMEX>
توصیف‌گر شخص	زیر-نوعی ندارد	هر کلمه‌ای سر از اسم مشترک با اشاره به فرد یا گروهی از مردم است. و نیز هر عنوان شغلی. به عنوان مثال در President Ahmadinezhad ما یک توصیف‌گر (President) و یک نام (Ahmadinezhad) داریم. عنوان‌های تجلیلی مثل Mr.، Mrs، Sir، و غیره. مثال‌ها: his top <aides> die-hard <fans> <analysts> said State court <Judge> Richard Curry The <owner> , who begs anonymity <Chief Executive> Sir Christopher Hogg <Chairman> Jay B. Langner SGML: <ENAMEX TYPE = ”PER-DESC”> </ENAMEX>

ادامه‌ی جدول در صفحه‌ی ۳۲

ادامه‌ی جدول

نوع	زیرنوع‌ها	خلاصه‌ی توضیحات
NORP	ملیت : فرم صفتی از یک تغییر دهنده‌ی GPE مکان، (یک آمریکایی، ملوان چینی، سازمان‌های اروپایی، و غیره). the <Brazilian> development <European> euphoria twice as many Nobel science prizes as the <Japanese> the <Dutch> publishing group a fellow <Texan> a conservative <Californian> <Irish-American> (note that this is marked as one entity, not two) مذهب : هر گونه صفتی از دین، بدون در نظر گرفتن چهار چوب متن. Mr. Sohmer , who is <Muslim> a devout <Catholic> The <Hindu> newspaper سیاسی: Some <Democrats> a <Communist> official France's <Socialist> government بقیه: <African-American> activist <Arab> uprising <Francophone> < Mafioso>	این نوع از روی تعریف زیر-نوع‌های آن یعنی ملیت، مذهب، سیاسی و بقیه (Other) تعریف می‌شود. تمایز بین NORP و انواع دیگر، در ریخت شناسی آن‌ها بدون در نظر گرفتن چهار چوب متنی است که در آن آمده‌اند. American و Americans ملیت هستند، در حالی که America و US از نوع GPE هستند. فرم‌های صفتی از نام‌های GPE و نام مکان‌ها (مانند American). فرم صفتی از مذاهب نام برده، میراث فرهنگی و وابستگی‌های سیاسی (نظیر حزب دموکرات، چینی آمریکایی، و یهودی). سر واژه‌هایی که اشاره به اشخاصی دارد که به نهادی (اغلب GPE، محل سکونت یا سازمان) وابسته هستند. به عنوان مثال، «سه شخص دموکرات»، «مسلمانان» و غیره.
پول	زیر-نوعی ندارد.	هر مقدار پولی که شامل همه گروه‌های مالی بشود. واحد پولی باید صریح باشد.برای مثال: ¥ ۵، یک میلیون دلار، ۱۷،۰۰۰ £، ۱۰.۲۰ $. فقط باید ارزش‌ها نشانه‌گذاری شوند نه رجوع‌های عمومی. به عنوان مثال، در «پول در سرمایه‌گذاری...» است، هیچ نشانه‌گذاری روی «پول» نباید وجود داشته باشد. در عباراتی از قبیل نرخ £ در هر واحد، واحد باید در مقدار وجود داشته باشد. به عنوان مثال، در۳ دلار در هر سهم»، تنها دلار» جز مقدار است.

ادامه‌ی جدول

کاربردهای این طبقه‌بندی، شامل سیستم‌های پرسش و پاسخ که به تحلیل متون عمومی مانند مقالات روزنامه‌ها می‌پردازند، و همچنین استخراج اطلاعات، ترجمه‌ی ماشینی، تلخیص و بازیابی اطلاعات و سیستم‌های دیگر مرتبط با سیستم‌های پردازش زبان طبیعی می‌باشد[؟].

به عنوان مثال، سیستم‌های پرسش و پاسخ اطلاعاتی را ارائه می‌دهد که شخصی به دنبال دانستن آن است یا می‌خواهد آن را از مقالات استخراج کند. این اطلاعات را می‌توان به تعداد ثابتی از کلاس‌ها به همراه سلسله مراتب آن‌ها طبقه‌بندی کرد. در تحلیل و توسعه‌ی مدل گسترش یافته‌ی تشخیص واحدهای اسمی فرض شده است که مثلاً در یک سیستم پرسش و پاسخ و یا یک سیستم بازیابی اطلاعات، آنچه که شخص می‌خواهد بداند اساسا در یک شکل از عبارت اسمی با نام‌های خاص و یا ارزش عددی نهفته شده است. به عبارت دیگر، آن شخص به دنبال یک مفهوم یا کلاس کلی نمی‌باشد، بلکه دنبال نام و یا مفهوم چیزی است که می‌توان اشاره‌ی فیزیکی به آن نمود.

سلسله مراتب گسترش یافته‌ی تشخیص واحدهای اسمی به سه دسته اصلی تقسیم شده است؛ نام، زمان و عبارات عددی (این سه کلاس همان کلاس‌هایی هستن که در سلسله مراتب تعیین شده در MUC و پروژه IREX وجود داشتند). بر اساس مشاهدات صورت گرفته توسط سکاین و همکارانش، می‌دانیم که هر سوالی در مورد یک موضوع خاص، اغلب در یکی از این طبقه‌بندی‌ها می‌گنجد. با داشتن این سه کلاس در بالای سلسله مراتب سیستم توسعه داده شده‌ی تشخیص واحدهای اسمی؛ سیستم پرسش و پاسخ و سیستم استخراج اطلاعات، با توجه به مفاهیم و واژه‌هایی که بطور کلی به عنوان دانش رایج در مقالات روزنامه‌های معمولی و دائره المعارف ها وجود دارند، ساخته شده‌اند.

طراحی طبقه‌بندی توسعه داده شده‌ی واحدهای اسمی

کلاس‌های تعریف شده بر اساس این معیار انتخاب شده‌اند که کلمات و عبارات اسمی که بیشتر تکرار شده‌اند باید با توجه به معنی و نحوه استفاده از آن طبقه بندی و طبقه‌بندی شوند. در عمل، روش‌های زیر به منظور توسعه سلسله مراتب استفاده می‌شوند:

عبارات اسمی نامزد شده از مقالات روزنامه‌های انگلیسی زبان استخراج شده‌اند. به ویژه، واژه‌ها با حروف بزرگ (اسامی ساده) و عبارت‌های عددی (تعدادی از چیزها) از هزاران

متن استخراج شده‌اند، سپس برچسب کلاس و مفهوم‌های مناسبی برای آن اسامی ساده و عبارات عددی تخصیص داده شده است.

اصطلاحنامه های موجود و هستی شناسی موجود در صفحه آغازین وردنت[۳۰] ، برای یافتن اطلاعاتی که منطبق بر طبقه‌بندی توسعه یافته هستند، مورد ارجاع و استفاده قرار گرفته بودند. همچنین، تحقیقات مربوط به این حوزه در طبقه بندی نهایی بی تأثیر نبوده‌اند.

در نهایت تجدید نظری بر روی این طبقه‌بندی از روی مقالات روزنامه‌ای که به صورت دستی برچسب‌گذاری شده بودند، صورت گرفت.

آن‌ها ۳۰۰۰ سؤال را در یک سیستم پرسش و پاسخ برای ایجاد و آزمایش‌های خود مورد استفاده قرار داده‌اند. سپس نوع پاسخ‌های دریافتی با توجه به طبقه‌بندی پیشنهاد شده برچسب‌گذاری شده و کلاس جدید در صورت نیاز افزوده شده‌اند.

واژه‌های ایندکس گذاری شده (حدود ۱۱۰۰۰) در دائرة المعارف انسایکلوپدیا[۳۱] با توجه به طبقه‌بندی گسترش یافته‌ی واحدهای اسمی، طبقه بندی شدند، و بر این اساس کلاس‌های جدید در صورت لزوم اضافه شد.

پرسش‌های آزمونی

برای ما راحت‌تر است که به منظور تعیین سلسله مراتب مورد نظر، ازسیستمی شبیه سیستم پرسش و پاسخ استفاده کنیم. در ادامه به توضیحاتی راجع به این نوع استفاده از سیستم پرسش و پاسخ می‌پردازیم. همچنین، در عین حال که این توضیحات برای هدف برچسب‌گذاری به کار می‌رود، اصطلاحی مانند "Tag X" را استفاده می‌کنیم، نه "X is Y class of word" . در این بخش معیار قضاوت توضیح داده شده است، یعنی کدام نوع از کلمات در کدام کلاس قرار می‌گیرند. پرسش‌های انتخاب شده در این گزارش به طور خلاصه و بر اساس معیارهای اساسی توضیح داده شده است. در عین حال، ممکن است گاهی ابهامی وجود داشته باشد و ممکن است نیاز به اتخاذ

[۳۰] WordNet

[۳۱] Encyclopedia

بعضی از تصمیم‌های خودسرانه باشد. به منظور حل این مشکل، برای بسیاری از تعاریف نمونه‌های خاصی آورده شده است.

نام‌ها

اساسا، هنگامی که ما یک سوال آزمون مانند «نام X را به من بگو» (که در آن X نام کلاس مورد نظر است) را ایجاد می‌کنیم، یک کلمه از یک پاسخ رضایت بخش به این سوال می‌تواند به عنوان یک نمونه از کلاس واحد اسمی آن تصور شود.

- PERSON the is "What person: a of Name
- COMPANY the is "What company: a of Name
- SEA the is "What sea: the of Name
- FOOD the is "What food: the of Name
- RULE the is "What rule: a of Name
- INCIDENT the is "What incident: an of Name
- SHOW the is "What show: a of Name
- INISECT the is "What insect: an of Name

به این ترتیب، به عنوان معیار برای برچسب زدن روشن است که «سوسک»، به عنوان مثال، می‌تواند به عنوان یک نام از حشرات برچسب زده شود، اما «انگل» نمی‌تواند. در مورد «سوسک»، مناسب‌تر آن است که در مورد نام پرسیده شود. اما در مورد «انگل»، مناسب‌تر است راجع به نوع سؤال شود تا نام.

زمانی که کلمات به نام شخص، سازمان و یا محل اشاره می‌کنند، به راحتی می‌توانیم برچسب نام ساده را به آن‌ها نسبت دهیم اما در مورد نام محصولات این اطمینان وجود ندارد. به عنوان مثال، پرسش‌ها آزمونی برای نام‌های مواد غذایی می‌تواند اسم عام محسوب شود نه نام ساده. مانند

«اسفناج» و یا «پیتزا». البته، غذاهایی که متعلق به هر دو مجموعه‌ها باشند نیز وجود دارد. به عنوان مثال، بسته به چهار چوب متن مورد نظر، «نیهون شو» (یک غذای ژاپنی) بهتر است برای نام خود برچسب‌گذاری شود تا برای نوعش. در چنین شرایطی، به نمونه‌های ذکر شده در زیر تعریف کلاس‌ها مراجعه می‌شود تا به تصمیم مناسب اتخاذ گردد.

عبارات زمانی

برای عبارات زمانی، اینکه یک عبارت آیا می‌تواند به عنوان عبارت زمانی برچسب‌گذاری شود یا خیر، باید در یکی از پرسش‌ها آزمونی زیر صدق کند:

- was (TIME) "When Time:
- was (PERIOD) long "How Period:

با این تعریف، بسیاری از اصطلاحات زمانی از تعریف عبارت زمانی خارج می‌شوند. به عبارت دیگر، عباراتی از قبیل «امروز»، «دو روز پیش»، «سال گذشته»، «در آن زمان» یا «همان روز» به زمان دقیق یک واقعه مشخص اشاره نمی‌کند و در این تعریف نمی‌گنجند. به هر حال، در هنگامی هم که این قانون به طور دقیق و با سخت‌گیری اعمال شود، عبارت ماه می» نیز نمی‌تواند دقیقاً سال وقوع را تعیین کند. به همین ترتیب، در عبارت «روز بیست و دوم»، ماه وقوع روشن نیست. اما بر خلاف عبارت «امروز»، «بیست و دوم ماه می» به یک تاریخ خاص از یک ماه محدود می‌شود، و بنابراین باید نشانه‌گذاری شود و آن طور که در بالا توضیح داده شد، نباید در مورد آن سخت‌گیری شود.

عبارات عددی

به همین ترتیب برای عبارات عددی، سوال آزمون‌هایی از قبیل « X آن را به من بگو» ایجاد می‌شود. اگر پاسخ داده شده رضایت‌بخش بود، عبارت مربوطه یک عبارت عددی است.

- MONEY: "Tell me its amount of MONEY".
- RANK: "Tell me it's RANK".

- PHYSICAL EXTENT: “Tell me it’s PHYSICAL EXTENT (length)”.
- TEMPERATURE: “Tell me it’s TEMPERATURE”.
- N-ORGANIZATION: “Tell me its number of ORGANIZATION”.
- N-COUNTRY: “Tell me its number of COUNTRY”.
- N-PRODUCT: “Tell me its number of PRODUCT”.

این کلاس به عنوان عبارات عددی تعریف شده است، با این حال، اگر پاسخ به پرسش‌های بالا رضایت‌بخش باشد، عبارت عددی لزوماً شامل مقادیر عددی نمی‌باشد. به عبارت دیگر، «بالا» و یا «قهرمان» نیز، در محدوده‌ی عبارات عددی گنجانده می‌شوند. با این توضیحات نمونه‌هایی از تعاریف طبقه‌بندی سکاین در زیر آورده شده است [؟]: 0. TOP

Type: TOP

Child categories: NAME TIME_TOP NUMEX

Definition: Root of all the extended named entity categories.

1. NAME

Type: Intermediate node

Parent Category: TOP

Child categories: NAME-OTHER PERSON ORGANIZATION LOCATION FACILITY PRODUCT EVENT NATURAL-OBJECT TITLE UNIT VOCATION DESEASE GOD ID-NUMBER COLOR

Definition: Names of things, whether it is concrete or abstract.

Problematic Points: nothing

1-0. NAME-OTHER

Type: Leaf node

Parent category: NAME

Prototypical words: Name

Definition: Names that do not belong to any of the lower class (i.e., name of pet).

Problematic Points: nothing

Examples:

Examples	Judge
Rover, Whiskers	OK

1-1. PERSON

Type: Leaf node

Parent Category: NAME

Definition: Name of a person including legendary or fictional characters. Nick-names are also included.

Problematic Points:

Fictional NE

Initials and alphabets such as “X Hospital”

Nicknames

Name of a person that is succeeded (i.e., one’s father’s name)

Commonly known “name of a person and his/her rank”

Expressions that is not included in “era”

Ranks with numerical expressions and ordinal numbers

Examples:

Examples	Judge
George W. Bush, Bush, George, Hillary Clinton, Hillary, Edgar Allan Poe, J.Lo, Jennifer, Jen, Michael Chriton, Ichiro, Hamlet, Ophelia, Elizabeth II, Tom Cruise, Rocket (Clemens), Harry Potter	OK
Mr. Clinton (Mr.=TITLE OTHER, Clinton=NAME)	NG

روش‌های پیاده‌سازی تشخیص واحدهای اسمی

در این فصل به بررسی روش‌های مختلف برای پیاده‌سازی تشخیص واحدهای اسمی می‌پردازیم. در ادامه به بررسی روش‌های موجود و توسعه داده شده مانند استفاده از قواعد انجمنی برای تشخیص واحدهای اسمی، استفاده از ماشین بردار پشتیبان، مدل مخفی مارکوف، روش‌های مبتنی بر وب در تشخیص واحدهای اسمی، الگوریتم‌های تشخیص نقل قول در مقالات، ارتباط میان اسامی و رفع ابهام از اسامی مشترک میان چند فرد خواهیم پرداخت.

قواعد انجمنی

در دهه‌ی اخیر قواعد انجمنی و استخراج قواعد انجمنی توجه بیشتری درحوزه‌های بانک اطلاعات و داده‌کاوی را به خود جلب کردند. رفته رفته قواعد انجمنی دارای کاربردهای زیادی در دامنه‌های مختلف شدند. یک قاعده‌ی انجمنی سنتی یک رابطه به فرم $X \rightarrow Y$ است که در آن X و Y مجموعه‌هایی ازآیتم‌های میان مجموعه داده مورد مطالعه هستند. فاکتور پشتیبانی[۳۲] به معنی نسبتی از تعدادی از آیتم‌ها می‌باشد که شامل آیتم‌های X با هم است، نسبت به کل آیتم‌ها. فاکتور اطمینان[۳۳] نسبت آیتم‌های موجود در X و Y است تقسیم بر تعداد آیتم‌ها در X . یا در حقیقت:

$$conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} \qquad ($$

برای کسب اطلاعات بیشتر در مورد قواعد انجمنی می‌توانید به [؟] مراجعه کنید. در ادامه برای بیشتر آشنا شدن با چگونگی استفاده از قواعد انجمنی برای تشخیص واحدهای اسمی الگوریتمی را ذکر می‌کنیم که در سال ۲۰۰۳ توسط بودی و برسان[۳۴] ارائه شد. طبیعتا می‌توان با اعمال تغییراتی آن را بهبود بخشید[؟].

کاویدن قواعد انجمنی شامل یافتن قواعدی است که بیش‌ترین پشتیبانی و بیش‌ترین اطمینان را دارند. در استخراج اسم، مجموعه داده‌ها، شامل توالی از اصطلاحاتی با خصیصه‌ها و کلاس‌های

[۳۲] Support

[۳۳] Confidence

[۳۴] Indra Budi And Stephane Bressan

اسمی (مانند مکان‌ها، اسامی اشخاص، اسامی سازمان‌ها، تاریخ‌ها، مقادیر پولی، زمان‌ها و درصد) هستند. یک مثال از خصیصه‌ها می‌تواند بزرگی یا کوچکی حروف باشند. X و Y در بالا می‌توانند یکی از این سه حالتی باشند که در ادامه آمده است:

- ترم‌ها: $\langle t_2 \rangle \Rightarrow nc_2$ که به آن «قاعده‌ی لغت‌نامه» می‌گوییم.
- دنباله‌ای از ترم‌ها: $\langle t_1, t_2 \rangle \Rightarrow nc_2$ که به آن «قاعده‌ی بایگرم» می‌گوییم.
- خصیصه‌ها: $\langle t_1, f_2 \rangle \Rightarrow nc_2$ که به آن «قاعده‌ی خصیصه» می‌گوییم.

که در آن‌ها $\langle t_1, t_2 \rangle$ یک دنباله از ترم‌هاست که در آن f_2 یک خصیصه از t_2 و nc_2 نام کلاس t_2 است. برای روشن شدن مطلب مثالی در زیر آمده است: جمله‌ی Prof. Hasibuan conducted a lecture on information retrieval را در نظر بگیرید. در یک پیکره‌ی متنی آموزشی که در آن کلاس‌های مورد نظر داده شده‌اند، تفسیر پیکره‌ی متنی مشخص می‌کند که‌ترم Hasibuan از کلاس «نام شخص» است. ما یک قاعده‌ی لغت‌نامه‌ای به صورت زیر می‌نویسیم:

$$\langle Hasibuan \rangle \Rightarrow person - named(Hasibuan)$$

که در آن پشتیبانی و اطمینان بستگی به تعداد ظاهر شدن این کلمه در متن و جاهایی که این کلمه دارای برچسب Hasibuan می‌شود، دارد. ما یک قاعده‌ی بایگرم نیز به صورت زیر تولید می‌کنیم:

$$\langle prof., Hasibuan \rangle \Rightarrow person - named(Hasibuan)$$

که در آن پشتیبانی و اطمینان بستگی به تعداد ظاهر شدن $Prof.Hasibuan$ در متن و جاهایی که این عبارت دارای برچسب Hasibuan می‌شود، دارد. همچنین یک قاعده‌ی خصیصه نیز تولید می‌کنیم:

$$\langle prof., capitalised - word(X) \rangle \Rightarrow person - named(X)$$

که در آن پشتیبانی و اطمینان بستگی به تعداد ظاهر شدن Prof. X در متن و جاهایی که این عبارت دارای برچسب X می‌شود، دارد. بعد از تولید این قواعد، آن‌هایی که میزان پشتیبانی و اطمینانشان

از مقدار معینی بالاتر هستند، در پردازش تشخیص واحدهای اسمی شرکت می‌کنند. برای هر جفت از ترم‌ها در متن، الگوریتم تشخیص واحدهای اسمی مبتنی بر قواعد انجمنی، در زیر آمده است:

For every pair of terms $\langle t_1, t_2 \rangle$

Find the set of R rules $X \Rightarrow nc$

Such that X matches $\langle t_1, t_2 \rangle$

And support and conference are above that threshold

If R is not empty

Then

Choose in R the rule with highest conference $X \Rightarrow nc$

Assign nc as the class name of t_2

Else

Assign "Not-a-name" as the class name of t_2

Endfor

که در آن قواعد کم‌ترین پشتیبانی و بیش‌ترین اطمینان مشخص می‌شوند. در جاهایی که تساوی صورت بگیرد، یکی از کلاس‌ها به صورت تصادفی انتخاب می‌شوند. در صورتی که قاعده‌ای یافت نشود (که احتمال آن کم است)، کلاس Not-a-name به آن تعلق می‌گیرد. برای ارزیابی الگوریتم ارائه شده در بالا از یک پیکره‌ی متنی استاندارد تشخیص واحدهای اسمی که در MUC-7 معرفی شده، استفاده شده است. این پیکره‌ی متنی شامل مقالات خبری به زبان انگلیسی است که در آن ترم‌ها به ۷ کلاس تقسیم بندی شده‌اند. ۲۰۰ مقاله برای آموزش و ۱۰۰ مقاله برای آزمون استفاده شده است. این الگوریتم در این آزمایش به زبان C++ نوشته شده است. آزمایش‌ها روی مقالات یک روزنامه‌ی اندونزیایی به نام کامپاس[۳۵] نیز انجام شده است که مقالات ۲۰۱ تا ۱۵۳۲ ترمی را از آن روزنامه انتخاب کرده‌اند و به صورت دستی با ۳ کلاس برچسب‌گذاری نموده‌اند. تأیید این روش با استفاده از دو معیار دقت و بازخوانی که در فصل سوم تعریف شده‌اند سنجیده شده‌اند. پاسخ‌های به دست آمده در جدول آورده شده است [؟].

[۳۵] Kompas

جدول نتایج آزمایش‌ها با استفاده از الگوریتم ارائه شده برای قواعد انجمنی

قاعده	MUC-7		Kompas	
	بازخوانی	دقت	بازخوانی	دقت
Bigram	۳۴/۳۷	۹۳/۲۱	۲۸/۴۵	۸۶/۵۲
Feature	۴۴/۸۴	۶۷/۷۵	۵۱/۵۸	۷۷/۶۵
Bigram + Dict	۶۰/۴۴	۸۹/۵۹	۴۸/۴۲	۸۲/۶۱
Feature	۶۶/۳۴	۸۳/۴۳	۶۲/۸۰	۸۱/۳۵

استفاده از ماشین بردار پشتیبان

هدف تشخیص واحدهای اسمی این است که مسأله‌ی شناسایی[۳۶] را به مسأله‌ی دسته‌بندی یا کلاس‌بندی[۳۷] تبدیل کند و یک مدل آماری کلاس‌بندی را برای حل آن به‌کارگیرد. در رهیافت استفاده از ماشین بردار پشتیبان برای بهتر کردن این کلاس‌بندی استفاده می‌شود. البته ماشین بردار پشتیبان در کلاس‌بندی کردن چندتایی ضعف دارد (در روش‌های نرمال هر واحد اسمی متعلق به یک کلاس ثابت با توجه به خصیصه‌هایش است.) که این ضعف را می‌توان به روش‌های گوناگون مانند آن چه که منصوری[۳۸] و همکارانش پیشنهاد داده‌اند (استفاده از یک ماشین بردار پشتیبان فازی[۳۹])، برطرف نمود.

ماشین بردار پشتیبان چیست؟

ماشین بردار پشتیبان یکی از روش‌های یادگیری ماشین با ناظر معروف برای کلاس‌بندی دوتایی در طیف وسیعی از مجموعه داده‌ها است. ماشین بردار پشتیبان بهترین پاسخ را در مواقعی که مجموعه داده‌ها پراکنده هستند و یا هنگامی که مجموعه‌ی آموزشی کوچک است تولید می‌کند و با استفاده از توسعه‌های مختلف از این الگوریتم می‌تواند برای مسائل با چند کلاس نیز به کار رود. برای حل یک مسأله‌ی کلاس‌بندی با استفاده از یک الگوریتم با ناظر، همواره ماشین با استفاده از یک

[۳۶] Identification

[۳۷] Classification

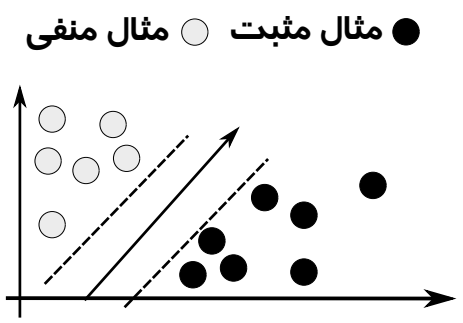
[۳۸] Alireza Mansouri

[۳۹] Fuzzy Support Vector Machine

مجموعه‌ی آموزشی که در آن کلاس‌های هر مورد مشخص می‌باشد، آموزش می‌بیند. هدف در اینجا به دست آمدن یک دقت بالا بعد از اعمال الگوریتم روی داده‌های آزمون است. برای هر ماشین بردار پشتیبان، دو مجموعه داده وجود دارد که در آن ماشین با استفاده از مجموعه داده‌ی آموزشی این داده‌ها را بر اساس خصیصه‌هایشان کلاس‌بندی می‌کند. هر داده‌ی آموزشی، با یک برچسب مثبت یا منفی برچسب گذاری شده است:

$$(x_1, y_1)...(x_n, y_n) \quad where \quad x_i \in R^n, y_i \in \{+1, -1\} \qquad (1$$

که در آن N تعداد مثال‌های مشتق شده از مجموعه‌ی آموزشی است. (شکل ۱ را ببینید) در فرم پایه‌ای



شکل ۱ کلاس‌بندی خطی ماشین بردار پشتیبان

یک ماشین بردار پشتیبان می‌آموزد که چگونه یک ابر صفحه را برای جداسازی مثال‌های مثبت و منفی با بیش‌ترین حاشیه بیابد. بیش‌ترین حاشیه می‌تواند به صورت زیر بیان شود:

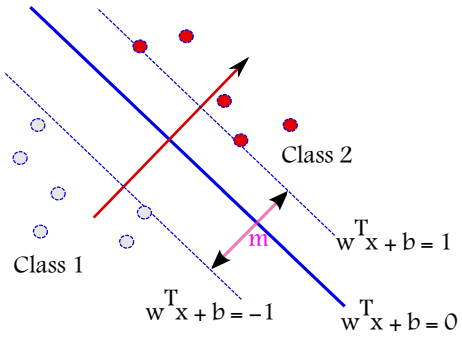
$$(w.x) + b = 0, \quad (w \in R^n, b \in R) \qquad (2$$

ابر صفحه داده‌های آموزشی را به دو قسمت مثبت و منفی تقسیم می‌کند، به گونه‌ای که:

$$y_i(w.x_i) \geq 1 \qquad (3$$

به هر حال تعداد زیادی از این ابر صفحه‌ها می‌توانند وجود داشته باشند، اما ماشین بردار پشتیبان سعی در یافتن ابر صفحه‌ی دارد که بیش‌ترین حاشیه را تا نزدیک‌ترین مثال‌ها به خود داشته باشد. (شکل ۱ را ببینید) حاشیه‌ی M و خط‌ها می‌توانند به صورت زیر بیان‌شوند:

$$w.x + b = +1, M = 2/\|w\| \qquad (4$$



شکل بهینه‌سازی ابر صفحه در کلاس‌بندی خطی ماشین بردار پشتیبان

بیشینه کردن این حاشیه معادل کمینه کردن مقدار $\|w\|$ است. یعنی می‌توان مسأله‌ی بهینه‌سازی زیر را حل نمود:

$$\frac{۱}{۲}\|w\| \quad subjectto: y_i|(w, x_i) + b| \geq ۱ \qquad (۲$$

ماشین بردار پشتیبان خطی می‌تواند برای هر مجموعه داده یک برچسب کلاس را بیابد[۲۹]:

$$C(x_i) = sign(w.x_i + b) \qquad (۲$$

$$C(X) = \begin{cases} +۱ & \text{if} \quad w.x + b > ۰, w \in R^n, b \in R \\ -۱ & \text{Otherwise} \end{cases} \qquad (۲$$

استفاده از ماشین بردار پشتیبان در تشخیص واحدهای اسمی

بررسی صورت گرفته در مورد ماشین‌های بردار پشتیبانی، نشان داد که نتایج قابل قبولی با استفاده از این روش گرفته شده است، اما نزدیک‌ترین نتیجه به نتیجه‌ی مورد نیاز ما در زبان فارسی، یک بررسی صورت گرفته بر روی زبان هندی بود که در آن به بازخوانی حدود ۸۸ درصد و دقتی در حدود ۸۱ درصد دست‌یافته‌اند. در مورد زبان هندی مسائل زیر قابل توجه هستند:

- بر خلاف زبان‌های انگلیسی و اروپایی، در زبان هندی نیز مانند فارسی، حروف بزرگ و کوچک نداریم.
- بسیاری از اسامی اشخاص در زبان هندی در معانی غیر اسمی و به صورت صفت در جمله‌های

این زبان به کار می‌روند.

- زبان هندی یک زبان تقریباً ترتیب-آزاد است.
- پیکره‌ی متنی برای زبان هندی ضعیف است.
- زبان هندی دارای تاریخچه و منبع بسیار قدیمی می‌باشد[؟].

در روش استفاده از ماشین بردار پشتیبان، ابتدا داده‌های موجود به دو قسمت داده‌های آموزشی و داده‌های آزمون تقسیم می‌شوند. در قدم بعدی، مجموعه خصیصه‌ها انتخاب می‌شوند. در انتخاب مجموعه‌ی خصیصه، اغلب خصیصه‌هایی انتخاب می‌شوند که مختص آن زبان خاص هستند، مانند خصیصه‌های پیشنهاد شده توسط هیرائو و همکارانش[۴۰]:

- اطلاعات لغوی وابسته به فرهنگ لغت.
- حروف پسوند و پیشوند تا
- اطلاعات واحد اسمی که در گذشته به دست آمده‌اند.
- کلاس‌های واحد اسمی ممکن.

سپس در مرحله‌ی بعدی ممکن است با تغییرات دیگری در الگوریتم مزبور و با استفاده از پس‌پردازش‌ها نتیجه را بهبود ببخشیم. در حالت عادی دقت حاصل شده از روش هیرائو در بهترین حالت به بیش از ۶۵ درصد رسیده است.

اما در بعضی موارد نیز خصیصه‌های انتخاب شده، مستقل از زبان هستند، مانند موارد زیر:

- کلمه‌ی اول جمله بودن
- شامل رقم بودن
- وجود دو رقم داخل کلمه

---

[۴۰] Hirao

- حاوی رقم و ویرگول بودن
- حاوی علامت درصد بودن
- تعداد کلمات قبل و بعد از کلمه در جمله
- و.. [؟ ؟].

تحقیق نامبرده شده بر زبان هندی، از روش انتخاب خصیصه‌های مستقل از زبان استفاده کرده است و تنیجه بسیار خوب بوده است [؟].

در روش استفاده از ماشین بردار پشتیبان، بعد از انتخاب خصیصه‌ها، کلمات را در یک فضای n بعدی که n تعداد خصیصه‌ها است، قرار می‌دهند و سپس فضا را تقسیم‌بندی می‌کنند. البته لزومی ندارد که تقسیم‌کننده‌ی فضا لزوماً یک ابر صفحه باشد و با استفاده از ماشین‌های هسته‌ای[۴۱] به جای ماشین بردار پشتیبان می‌توان از یک تقسیم‌کننده‌ی غیرخطی استفاده نمود. (ماشین بردار پشتیبان نوع خاصی از ماشین‌های هسته‌ای محسوب می‌شود.)

مدل مخفی مارکوف

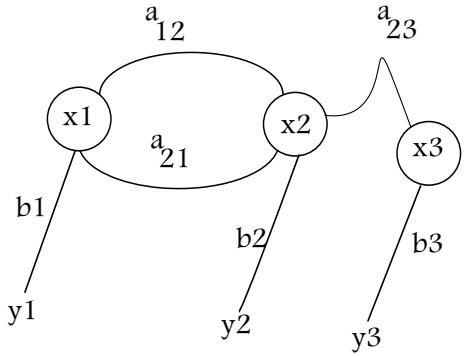
انتخاب بهترین تصمیم مستلزم دانستن نتایج ابعاد مختلف مسئله و مهمتر از همه عواقب بعدی آن است . تصمیم‌گیرنده باید بتواند بین دستاوردهای کوتاه‌مدت و درازمدت (که ممکن است گاهی در تضاد با یکدیگر باشند) انتخاب‌کند یعنی این که نگاه کند که با توجه به اتفاقات گذشته و پارامترهای موجود کدام تصمیم برای آینده بهترین (بهینه) است البته همان طور که می‌دانید بهینه بودن به این معنی نیست که این تصمیم از همه نظر بهترین است بلکه باید در نگاه کلان‌نگر و آینده‌نگر و با توجه به همه اجزاء سیستم بهترین باشد. این موضوع دست مایه اصلی مدل مخفی مارکوف است و در کاربردهای بسیار زیادی مورد استفاده قرار گیرد .مدل مخفی مارکوف در اواخر دهه ۱۹۶۰ میلادی معرفی گردید و درحال حاضر به سرعت در حال گسترش دامنه کاربردها می‌باشد. دو دلیل مهم برای این مساله وجود دارد؛ اول این که این مدل از لحاظ ساختار ریاضی

[۴۱] Kernel Machine

بسیار قدرتمند است و به همین دلیل مبانی نظری بسیاری از کاربردها را شکل داده است . دوم اینکه مدل مخفی مارکوف اگر به صورت مناسبی ایجاد شود می‌تواند برای کاربردهای بسیاری مورد استفاده قرارگیرد.

تعریف

در مدل مارکوف معمولی، هر حالت متناظر با یک رویداد قابل مشاهده است اما در مدل مخفی مارکوف مشاهدات توابع احتمالاتی از حالت‌ها هستند. در این صورت مدل حاصل یک مدل تصادفی با یک فرآیند تصادفی مخفی است و تنها توسط مجموعه‌ای از فرآیندهای تصادفی که دنباله مشاهدات را تولید می‌کنند قابل مشاهده است. اطلاق کلمه مخفی به موضوع مورد بحث ما به این دلیل است که درباره مسائلی صحبت می‌کنیم که طریقه انجام آن‌ها از دید ما پنهان است و البته ماهیت آماری دارد. یعنی اینکه نه تنها نمی‌دانیم نتیجه چه خواهد بود، بلکه نوع اتفاق و احتمال آن اتفاق نیز باید از پارامترهایی که در دسترس است، نتیجه‌گیری شود. احتمالات خروجی را نشان b احتمالات انتقال و a ، نشان‌دهنده حالات قابل مشاهده است y نشان‌دهنده حالت‌های مخفی و x همان طور که در شکل می‌بینید می‌دهد.

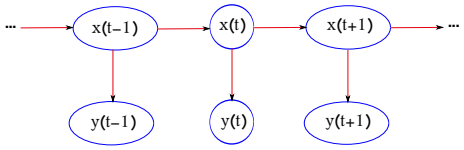


شکل حالات مخفی و قابل مشاهده در مدل مخفی مارکوف

معماری مدل مخفی مارکوف

در شکل زیر معماری مدل مخفی مارکوف نشان داده شده است . هر شکل بیضی بیانگر یک مقدار متغیر تصادفی است. x_t مقدار متغیر تصادفی است که مقدار تغییر پذیرش در واحد زمان مخفی

است و y_t مقدار متغیر تصادفی است که مقدار پذیرش در زمان t قابل مشاهده است (شکل را ببینید). مدل مخفی مارکوف با پارمترهای زیر تعریف می‌شود:



شکل معماری مدل مخفی مارکوف

- N : تعداد حالت‌ها (متناهی)
- M : تعداد خروجی‌های ممکن
- A : ماتریس گذار حالت
- B : توزیع‌های احتمالاتی برای خروجی‌های ممکن در هر حالت
- Π : توزیع احتمالاتی حالات اولیه.
- برای تعریف مدل به صورت کامل باید پارامترهای B ،A ،N ،M و Π[؟].

مثال

برای درک بهتر مدل مخفی مارکوف به عنوان مثال فرض کنید یک دوست دارید که دور از شما زندگی می‌کند و شما با او در مورد اینکه هر روز چه کاری انجام می‌دهد از طریق تلفن صحبت می‌کنید. دوست شما تنها به سه کار علاقه‌مند است: پیاده‌روی ، خرید و نظافت آپارتمان خود. انتخاب او کاملا با وضعیت هوایی هر روز در ارتباط است. شما هیچ اطلاعی از آب و هوای محلی که دوست شما در آن زندگی می‌کند ندارید، اما بر حسب آنچه که او هر روز از کارهای خود تعریف می‌کند شما سعی می‌کنید که آب و هوای محل زندگی دوستتان را حدس بزنید.
شما قبول دارید که هوا مانند یک حلقه‌ی مارکوف گسسته عمل می‌کند. دو وضعیت ممکن است وجود داشته باشد: هوا بارانی باشد یا آفتابی باشد. اما شما نمی‌توانید آن‌ها را مستقیما مشاهده کنید

زیرا آن‌ها از شما مخفی هستند. هر روز این شانس وجود دارد که دوست شما یکی از عملیات‌های walk، shop و یا clean را با توجه به وضعیت هوا انجام دهد. دوست شما در مورد فعالیتی که انجام می‌دهد به شما توضیحاتی می‌دهد که به آن‌ها مشاهدات می‌گوییم. این گونه سیستم‌ها را مدل مخفی مارکوف می‌گویند. شما وضعیت کلی هوا و اینکه دوستتان تمایل دارد چه کاری را انجام دهد می‌دانید. به بیان دیگر پارامترهای مدل HMM مشخص است. می‌توان مدل HMM مورد نظر را به صورت نمادین زیر بیان نمود:

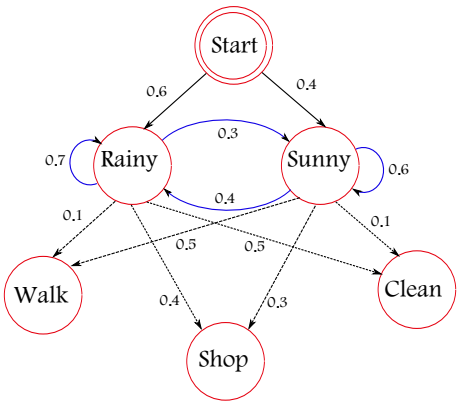
```
'Sunny') ('Rainy'، = states
'clean') 'shop'، ('walk'، = observation
{۰. : 'Sunny' ،۰. : {'Rainy' = start_probability
} = transaction_probability
،{۰ : 'Sunny' ،۰. : {'Rainy' : 'Rainy'
،{۰ : 'Sunny' ،۰. : {'Rainy' 'Sunny':
{
} = probability emission_
،{۰ : 'clean' ،۰. : 'shop' ،۰. : {'walk' : 'Rainy'
،{۰ : 'clean' ،۰. : 'shop' ،۰. : {'walk' : 'Sunny'
{
```

در شکل نمودار مدل مخفی مارکوف نیز درباره‌ی این مسأله نشان داده شده است[؟]:

استفاده از مدل مخفی مارکوف در تشخیص واحدهای اسمی

مدل مخفی مارکوف مانند روش‌های دیگر یادگیری ماشین در تشخیص واحدهای اسمی به کار می‌رود. یکی از کاربردهای اصلی مدل مخفی مارکوف تشخیص محتمل‌ترین دنباله از حالت‌های مخفی (با استفاده از الگوریتم ویتربای[۴۲]) است.

[۴۲] Viterbi Algorithm



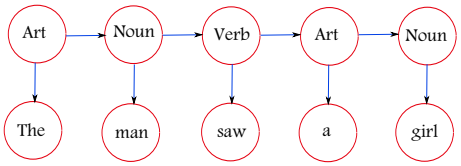
شکل نمودار مدل مخفی مارکوف

مشابه با روش استفاده از مدل مخفی مارکوف در تشخیص الگو[۴۳]، در تشخیص واحدهای اسمی از آن استفاده می‌شود. همانطور که در بخش‌های بعدی خواهید دید ما از ایده‌ی برچسب‌گذاری توالی‌ها (میدان‌های تصادفی شرطی) در پیاده‌سازی سامانه‌ی خود استفاده خواهیم کرد. از نظر مفهومی، و نه از لحاظ ساختار داخلی، مدل‌های میدان‌های تصادفی شرطی و مدل مخفی مارکوف سعی دارند با محاسبه‌ی پارامترهای داخلی یک توالی (برای مسأله‌ی ما جملات) را برچسب‌گذاری نمایند.

همان طور که می‌دانیم جمله در یک زبان مانند انگلیسی دارای کلماتی است که هر کدام نقشی را بازی می‌کنند (و این نقش‌ها می‌توانند به عنوان ادات سخن شناخته‌شده و برچسب‌گذاری شوند.) با در نظر گرفتن برچسب‌ها به عنوان حالت‌های مدل مخفی مارکوف، می‌توانیم این موقعیت را با استفاده از پردازش تصادفی[۴۴] مدل کنیم. برای مثال برای جمله‌ی "he man saw a girl" ، ما برچسب‌ها را مانند شکل قرار می‌دهیم: در واقع ما وقتی مدل مخفی مارکوف را مدل می‌کنیم، این برچسب‌ها را با استفاده از الگوریتم ویتربای تخمین می‌زنیم و مشکل کلیدی این است که چگونه مدل مخفی را با استفاده از داده‌های آموزشی بسازیم. در استفاده از این روش و روش‌های مشابه فرض می‌شود داده‌های آموزشی کافی هستند و ما به تعدادی زیادی از گذارهای بین برچسب‌های ادات سخن دست یافته‌ایم تا مدل (O,Q,A,B,Π) را بسازیم. که در آن x و y به ترتیب کلمات

[۴۳] Pattern Recognition

[۴۴] Stochastic Processing



شکل ساخت حالات از روی برچسب‌های ادات سخن

جدول نتایج حاصل از به‌کارگیری روش مدل مخفی مارکوف توسط آگئیشی و میورا

نرخ درستی	تعداد خطهای آموزشی
۶۱/۷۹	۱۰۰۰
۸۳/۱۲	۵۰۰۰
۸۷/۱۱	۱۰۰۰۰
۸۹/۱۹	۲۰۰۰۰
۹۰/۰۲	۳۰۰۰۰
۹۰/۴۶	۴۰۰۰۰
۹۰/۶۱	۴۵۰۰۰
۹۰/۶۷	۵۰۰۰۰
۹۰/۶۹	۵۲۰۰۰
۹۰/۸۱	۵۵۰۸۲

و برچسب‌ها هستند. استفاده از مدل مخفی مارکوف در بسیاری از موارد به نتایج خوب و قابل قبولی منجر شده است. مثلا برای آگئیشی و میورا[۴۵] بر حسب تعداد داده‌های ورودی نتایج زیر به دست آمدند (جدول) که با استفاده از روشی به نام هموارکردن مقادیر دقت تا ۹۳/۳۷ درصد بهتر شدند[؟]. مدل مخفی مارکوف نیز برای تشخیص واحدهای اسمی زبان هندی و بنگالی توسط اکبال و باندیوپازیای[۴۶] به کار برده شده است. جالب است که در اینجا نیز از ویژگی‌های مستقل از زبان استفاده شده است و نتیجه نیز بسیار خوب بوده است (برای زبان بنگالی بازخوانی در بهترین حالت به ۹۰/۰۳ و برای زبان هندی به ۸۲/۵ رسید)[؟].

[۴۵] Ryohei Ageishi and Takao Miura

[۴۶] Asif Ekbal and Sivaji Bandyopadhyay

روش‌های مبتنی بر وب

در میان روش‌های گوناگونی که برای تشخیص واحدهای اسمی وجوددارد، می‌توان به روش‌های مبتنی بر وب هم اشاره نمود که البته روش‌های مستقلی نیستند و در واقع، ترکیبی از استفاده از منابع اینترنتی با روش‌هایی که در قسمت‌های قبل به آن‌ها اشاره شد هستند. شاید نام بهتر برای این بخش، «روش‌های مبتنی بر منابع وب» باشد. برای روشن شدن منظور خود، در ادامه دو مثال را ذکر می‌کنیم که یکی از آن‌ها درباره‌ی استفاده از منابع اینترنتی در ترسیم روابط بین انجمن‌ها و دیگری استفاده از یک منبع اینترنتی برای رفع ابهام از اسمی مشترک میان چند فرد به کار برده شده است.

بیان یک مثال از کاربرد منابع اینترنتی در تشخیص واحدهای اسمی: ترسیم روابط ارتباط انجمن‌ها

کشف این که چه کسانی با چه کسانی روی چه پروژه‌ای و با کدام مشتری کار کرده‌اند، یک وظیفه‌ی کلیدی در مدیریت دانش محسوب می‌شود. همچنین بیشتر سازمان‌ها مدل‌هایی از ساختارهای سازمانی را نگه‌داری می‌کنند اما این مدل‌ها نمی‌توانند دقت کافی داشته‌باشند. در این مقاله‌ای که توسط ژو[۴۷] و همکارانش در سال ۲۰۰۹ ارائه شد، یک روش کاویدن متن با نام CORDER معرفی شده است که در آن ابتدا واحدهای اسمی از انواع خاصی از صفحات اینترنتی تشخیص داده می‌شوند و سپس ارتباط میان یک واحد اسمی خاص با واحدهای اسمی دیگر با آن اتفاق افتاده‌اند، کشف می‌شود. این روش بر روی وب‌سایت دانشگاه نویسندگان آن انجام شده است. در اینجا ابتدا با استفاده از CORDER واحدهای اسمی مربوطه را در ۴ نوع سازمان‌ها، اشخاص، پروژه‌ها و محل‌های پژوهشی از صفحات وب‌سایت استخراج نموده و بر اساس تکرار همزمان با افراد دپارتمان خود رتبه‌بندی کرده‌اند و برای هر شخص ۲۰ رابطه را مشخص کرده‌اند. روش CORDER از مکانیزم ساده‌ای برای تشخیص اسامی که با یکدیگر در متن ظاهر شده‌اند، استفاده می‌کند و با اعمال یک معیار شباهت اسامی را دسته‌بندی می‌کند[؟].

[۴۷] Jian-Han Zhu

ارتباط میان اسامی و رفع ابهام از اسامی مشترک میان چند فرد

با توجه به بررسی‌های صورت گرفته در بخش‌های قبل، مشخص شد که برای زبانی مانند زبان انگلیسی مسأله‌ی تشخیص و نشانه‌گذاری واحدهای اسمی تا حد قابل قبولی انجام‌شده‌است اما هدف یا چالش بعدی، تشخیص درست نمونه‌ی اصلی است که آن واحد اسمی به آن اشاره دارد. در مقاله‌ای که در سال ۲۰۰۹ توسط والتینگر و مِلِر[۴۸] ارائه شد، از میان روش‌های مختلف برای برخورد با این چالش، از یک روش مبتنی بر وب استفاده‌شده‌است. در این مقاله برای تشخیص واحدهای اسمی از یک نشانه‌گذار مبتنی بر مدل مخفی مارکوف استفاده‌شده‌است و تمرکز بر روی رفع ابهام بوده است و روش جدیدی برای نشانه‌گذاری ارائه نشده است. هدف اصلی مقاله یافتن راهی برای رفع ابهام میان اسامی، مخصوصا اسامی اشخاص است. برای درک بهتر این ابهام، مثال زیر را در نظر بگیرید: اگر کسی در مورد بوکس، اوکراین، قهرمان جهان و برادر کوچک‌تر صحبت کند می‌توان اطلاعات را به ولادیمیر کیلشکو[۴۹] نسبت داد. اگر عبارت برادر کوچک‌تر با اسم کلیشکو جایگزین شود، ۵۰ درصد شانس درست حدس زدن داریم زیرا دو برادر بوکسور با نام‌های ولادیمیر و ویتالی[۵۰] با این شرایط وجوددارند. در این رهیافت آن‌ها از ویکی‌پدیا[۵۱] به عنوان یک منبع کمک گرفته‌اند. با پوییدن "wikipedia-categury:people" یک پایگاه داده با ۱۸۳۵۵۴ مقاله‌ی مختلف در رابطه با افراد ساخته شد. در قدم بعدی همه‌ی ابرپیوندها را در این صفحات که به افراد دیگر متصل بودند، استخراج کرده‌اند. در قدم سوم، قطعات باقی‌مانده را به صورت یک گراف به نام گراف مردم (اشخاص) تشکیل دادند. این گراف یک گراف وزن‌دار است که در آن وزن‌ها به شیوه‌ی خاصی اختصاص‌داده شده‌اند؛ برای مثال، اسامی که در چند خط اول آمده‌اند، وزن بیشتری گرفته‌اند. حال با بررسی هر اسم در مقاله و با توجه به گراف مردم، اسامی درست تشخیص داده شده‌اند[؟ ؟].

---

[۴۸] Ulli Waltinger and Alexander Mehler

[۴۹] Vladimir Klitschko

[۵۰] Vitali

[۵۱] ویکی Http://www.wikipedia.org

تشخیص نقل قول در مقالات با استفاده از تشخیص واحدهای اسمی

شاخص‌های نقل قول به طور فزاینده‌ای نه تنها به عنوان ابزار هدایت برای محققان، بلکه به عنوان یک شاخص اساسی برای محاسبه‌ی کارایی و نفوذ یک تحقیق بر تحقیق‌های دیگر به کار می‌آیند. این به این معناست که میزان اطمینان از ابزاری که برای استخراج نقل قول‌ها استفاده می‌شود، بسیار مورد نیازتر است. به این منظور، تکنیک‌هایی برای استخراج نقل‌قول‌ها با دقت بالا معرفی‌شده‌اند که برای نظم‌بندی و شکل‌های مختلف مستندات به‌کار می‌آیند. با استفاده از ترکیب استخراج نقل‌قول، تجزیه‌ی منابع و تشخیص واحدهای اسمی می‌توان یک روش استخراج نقل قول با دقت بالا ساخت. برای روشن‌تر شدن مطلب، به مثال‌های زیر توجه کنید (در آخرین قسمت منظور از منبع این گزارش نیست و برای مثال آورده شده): Textual-Syntactic

par- their to according verbs ۳۰۰۰ over of classification a provides **Levin** ... alternations in ticipation

-———————————————————————

Parenthetical – Textual

(Miller WordNet are classification verb English to approaches current Two et al.، ۱۹۹) and Levin classes، **Levin** (۱۹۹ **)**.

-———————————————————————

Prosaic

properties... syntactic their of analysis an on based verbs groups **Levin**

-———————————————————————

Pronominal

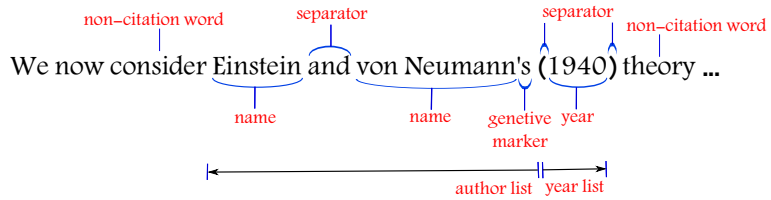
verb a of behavior syntactic the that assumption the reflects approach **Her** meaning. its by part large in determined is

-———————————————————————

Numbered

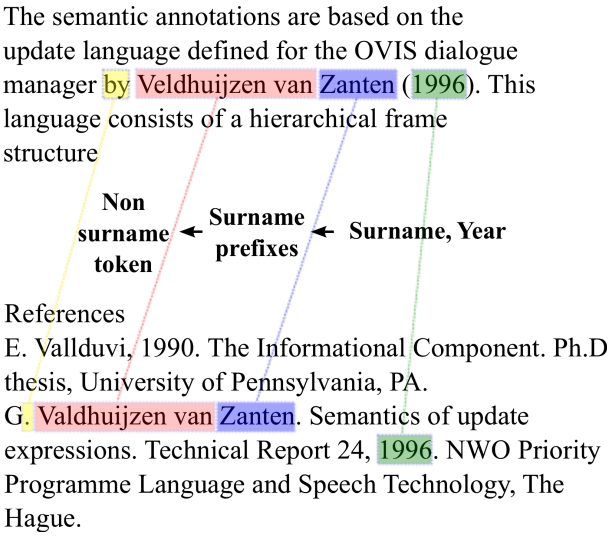
groups verb among behavior diathesis of patterns are There در هر یک از

مثال‌های بالا به نوعی یک نقل‌قول آمده است و به راحتی دیده می‌شود که تشخیص یک نقل‌قول همیشه آسان نیست. امّا مشاهده می‌شود که حتی در مثال‌های بالا می‌توان با کمک تشخیص واحدهای اسمی، نقل قول‌ها و ارتباط میان‌ها را تشخیص داد. مثلا در شکل این مطلب به خوبی درک می‌شود.



شکل استخراج اطلاعات نقل‌قول

در شکل یک نمونه از نقل قول دیده می‌شود. همان‌طور که مشخص است با استفاده از تشخیص نام‌های اشخاص، تاریخ، نام کتاب‌ها و یا کنفرانس‌ها یا سازمان‌های خاص می‌توان از روی الگوهای از پیش تعیین‌شده یک نقل قول را استخراج نمود[؟].



شکل تشخیص نقل قول با استفاده از تشخیص واحدهای اسمی

جمع‌بندی

در این فصل به بررسی دقیق‌تر مسأله‌ی تشخیص واحدهای اسمی پرداختیم و انواع طبقه‌بندی‌های واحدهای اسمی را معرفی نمودیم. عمده‌ی تمرکز ما در این بخش بر روی چگونگی برخورد با مسأله‌ی تشخیص واحدهای اسمی بود. به این معنا که روش‌های فائق آمدن بر این مسأله که عمده‌ی آن‌ها روش‌های یادگیری ماشین بودند بررسی شدند. در این فصل سعی کردیم به اختصار روش‌های همده‌ی یادگیری ماشین را در این مسأله‌ی خاص بررسی نماییم. ما به معرفی قواعد انجمنی و چگونگی استفاده از آن، ماشین بردار پشتیبان و مدل مخفی مارکوف پرداختیم. در قسمت بعد، به طور مفصل‌تر به میدان‌های تصادفی شرطی که یکی دیگر از روش‌های یادگیری ماشین است و اساس پیاده‌سازی سامانه‌ی معرفی شده در این پایان‌نامه بوده است می‌پردازیم.

فصل ۳

آشنایی با مدل‌های مبتنی بر میدان‌های تصادفی شرطی

مقدمه

در موتور تشخیص اسامی خاص که در این پایان‌نامه به شرح پیاده‌سازی آن به اختصار پرداخته‌ایم، از مدل میدان‌های تصادفی شرطی استفاده شده است. در این فصل این مدل را مختصرا شرح خواهیم داد.

مدل میدان‌های تصادفی شرطی یک روش مدل‌سازی آماری است که اغلب در شناسایی الگوها مورد استفاده قرار می‌گیرد. به طور دقیق‌تر، این روش نوعی از مدل گرافیکی احتمالاتی غیرمستقیم تبعیضی[۱] است. این نوع از مدل‌سازی برای کدگذاری روابط شناخته شده بین مشاهدات و ساختن تفسیرهای سازگار است. این روش اغلب برای برچسب‌گذاری و تجزیه‌ی داده‌های ترتیبی مانند متون زبان طبیعی، گفتار موجود برای پردازش و توالی‌های زیستی و نیز بینایی ماشین مورد استفاده قرار می‌گیرد. و به طور خاص‌تر، میدان‌های تصادفی شرطی، کاربردهایی در تجزیه‌ی سطحی، تشخیص واحدهای اسمی، استخراج برچسب‌های ادات سخن و نیز یافتن ژن‌ها در مقایسه با روش‌های مبتنی

[۱] discriminative undirected probabilistic graphical model

بر مدل مخفی مارکوف مورد استفاده قرار می‌گیرند. میدان‌های تصادفی شرطی در مقوله‌ی بینایی ماشین، غالبا برای تشخیص اشیا و قطعه قطعه کردن تصاویر مورد استفاده قرار می‌گیرند.

معیار شباهت و رگرسیون لاجیستیک

مبانی معیار بیشینه‌ی شباهت

برای توضیح دادن در مورد میدان‌های تصادفی شرطی، نیاز است که به مفهوم معیار شباهت یا likelihood بپردازیم. یک خانواده از توزیع‌های احتمالاتی که با مجموعه‌ای از پارامترهای θ مشخص شده‌اند را در نظر بگیرید. فرض کنید که ما یک نمونه‌ی تصادفی از یک عضو ثابت اما ناشناخته از این خانواده را بیرون کشیده‌ایم. نمونه‌ی تصادفی یک مجموعه‌ی آموزش از n نمونه $x_۱$ تا x_n است. فرض می‌کنیم این نمونه‌ها، مستقل از هم نیز می‌باشند، بنابراین احتمال وقوع این مجموعه، حاصل‌ضرب احتمالات تک تک نمونه‌ها است:

$$f(x_۱, ..., x_n; \theta) = \prod_j f_\theta(x_j; \theta). \qquad (۵$$

ما همیشه در مورد چنین مسأله‌ای به صورت یک توزیع ثابت θ نمونه‌های در حال تغییر x_j فکر می‌کنیم. در هر حال ما می‌توانیم به جای در نظر گرفتن یک توزیع با پارامترهای معلوم و داده‌های آموزشی تصادفی، مسأله را به صورت داده های معلوم و پارامترهای متفاوت فرض کنیم. این نکته، ایده‌ی نهفته در مفهوم معیار شباهت است:

$$L(\theta; x_۱, ..., x_n) = f(x_۱, ..., x_n; \theta). \qquad (۵$$

اصل بیشینه‌ی شباهت یا maximum likelihood یا مختصرا ML به صورت زیر تعریف می‌شود:

$$\hat{\theta} = argmax_\theta L(\theta; x_۱, ..., x_n). \qquad (۵$$

به مقدار $\hat{\theta}$ تخمین‌گر بیشینه‌ی شباهت یا MLE گفته می‌شود. در این فرمول هر x_j یک بردار از مقادیر و θ یک بردار از پارامترها با مقادیر حقیقی است. برای مثال در یک توزیع گاوسی،

$\theta = \langle \mu, \sigma^2 \rangle$. پس در حقیقت تخمین‌گر بیشترین شباهت، همیشه مجموعه‌ای از پارامترها را اختیار می‌کند که احتمال به وقوع پیوستن نمونه‌های x_1 تا x_n را بیشینه می‌کند. توجه کنید که در تعریف معیار شباهت، مسأله‌ی یافتن احتمال وقوع نمونه‌های x_1 تا x_n در یک توزیع با پارامترهای مشخص θ را به یافتن بهترین پارامترهای θ در صورت داشتن نمونه‌های x_1 تا x_n تبدیل کردیم. در اینجا در حقیقت نمونه‌های ذکر شده، داده‌های آموزشی ما هستند.

معیار بیشینه شباهت برای توزیع‌های برنولی

در اینجا برای بهتر فهمیده شدن مطلبی که در بخش قبل عرضه شد، یک توزیع برنولی را در نظر گرفته و معیار بیشینه‌ی شباهت را برای آن محاسبه می‌کنیم.
پارامتر یک توزیع برنولی را در نظر بگیرید. یک متغیر تصادفی با این توزیع فرموله کردن پرتاب یک سکه است. مقدار متغیر تصادفی، برابر یک با احتمال θ و صفر یا احتمال $1 - \theta$ است. فرض کنید X یک متغیر تصادفی برنولی باشد. داریم:

$$P(X = x) = \begin{cases} \theta & \text{if } x=1 \\ 1-\theta & \text{if } x=0 \end{cases} \quad (1$$

یا به طور معادل

$$P(X = x) = \theta^x (1 - \theta)^{1-x} \quad (2$$

فرض کنید n مشاهده‌ی x_1 تا x_n در دست هستند. معیار بیشینه شباهت را به صورت زیر می‌نویسیم:

$$L(\theta; x_1, ..., x_n) = f(x_1, ..., x_n; \theta) = \theta^h (1 - \theta)^{n-h} \quad (3$$

که در آن $h = \sum_i x_i$ در حقیقت در این مثال فرض شده است که مشاهدات x_1 تا x_n را داریم و می‌خواهیم محاسبه نماییم با چه پارامتر θای احتمال به وقوع پیوستن این مشاهدات، بیشینه است. جواب این سوال به طور شهودی، مشخص است. ما در اینجا به صورت ریاضی پارامتر مورد نظر را

با مشتق‌گیری محاسبه می‌نماییم:

$$\frac{\partial}{\partial_p}\theta^h(۱-\theta)^{n-h} = h\theta^{h-۱}(۱-\theta)^{n-h} + \theta^h(n-h)(۱-\theta)^{n-h-۱}(-۱)$$
$$= \theta^{h-۱}(۱-\theta)^{n-h-۱}[h(۱-\theta) - (n-h)\theta] \quad ($$

که وقتی این مشتق صفر می‌شود که θ برابر یکی از مقادیر صفر، یک یا $\theta = \frac{h}{n}$ باشد که بیشینه زمانی خواهد بود که $\theta = \frac{h}{n}$. پس برای معیار بیشینه‌ی شباهت خواهیم داشت:

$$\hat{\theta}_{MLE} = \frac{h}{n} \quad ($$

معیار شباهت شرطی

معیار شباهت شرطی یا Conditional likelihood در حقیقت یافتن بهترین θ برای بیشینه کردن یک احتمال شرطی مانند $y|x$ است. در نتیجه برای x های مختلف، توزیع احتمالاتی y خواهیم داشت. چون θ ها یکی هستند اما مقادیر آن‌ها متفاوتند. به طور نمادین یک معیار شباهت شرطی به صورت $L(\theta; y|x) = f(y|x;\theta)$ نمایش داده می‌شود. در این صورت ما همیشه به عنوان فرض مقادیر x و θ را داریم و با آن احتمال y را حساب می‌کنیم.

تابع likelihood lrlog لگاریتم تابع معیار شباهت است. به دلیل اینکه تابع معیار شباهت یک تابع یکنوای اکیدا صعودی است، پس بیشینه کردن تابع لگاریتم آن، بیشینه کردن خود آن است.

رگرسیون لاجیستیک

همان طور که می‌دانید، در علم آمار، برای نشان دادن ارتباط یا همبستگی میان دو یا چند متغیر از مفهومی به نام رگرسیون استفاده می‌کنند. در ساده‌ترین حالت برای نشان دادن یک همبستگی خطی (در صورت وجود) دو بردار با مقادیر حقیقی $x_۱$ تا x_n و $y_۱$ تا y_n ، از فرمول زیر استفاده می‌شود:

$$y = a + bx \quad ($$

که در آن:

$$\frac{N\sum XY - (\sum X)(\sum y)}{N\sum X^{۲} - (\sum X)^{۲}} \quad ($$

و

$$\frac{\sum Y - b(\sum X)}{N} \quad ($$

برای شرایط مختلف، فرمول‌های مختلفی برای محاسبه‌ی رگرسیون توسط ریاضیدانان ارائه شده است که یکی از آن‌ها رگرسیون لاجیستیک است. اگر y یک خروجی دودویی و x یک بردار با مقادیر حقیقی است، مدل شرطی

$$p = p(y|x;\alpha,\beta) = \frac{۱}{۱ + e^{\alpha+\sum_{j=۱}^{d}\beta_j x_j}}. \quad ($$

رگرسیون لاجیستیک نامیده می‌شود. ما از j برای اندیس گذاری روی مقادیر خصیصه‌ی $x_۱$ تا x_d از یک نمونه‌ی تکی از بُعد d است و از i برای اندیس‌گذاری روی نمونه‌های ۱ تا n استفاده می‌شود.

یادگیری شیب‌دار تصادفی

شیب رگرسیون لاجیستیک

در این قسمت صحبت در مورد یک نوع خاص از رگرسیون لاجیستیک را پی می‌گیریم. وقتی یک نمونه‌ی آموزشی تکی را که شامل مقادیر x و y است داشته باشیم، لگاریتم معیار شباهت شرطی وقتی y=۱ برابر مقدار $lnL(\beta;x,y) = lnp$ است. و وقتی y=۰ برابر $lnL(\beta;x,y) = ln(۱-p)$ است. هدف از آموزش بیشینه کردن مقدار لگاریتم معیار شباهت شرطی است. بنابراین بیایید مقدار مشتق جزئی آن را نسبت به هر β_j محاسبه کنیم. برای ساده کردن بخش آتی، برای هر نمونه‌ی x فرض می‌کنیم $\alpha = \beta.$ و $x. = ۱$. برای وقتی که $y = ۱$ داریم:

$$\frac{\partial}{\partial\beta_j} ln\ p = \frac{۱}{p}\frac{\partial}{\partial\beta_j} p \quad ($$

رای وقتی که $y = ۰$ داریم:

$$\frac{\partial}{\partial \beta_j} ln\ (۱ - p) = \frac{۱}{۱ - p}(-\frac{\partial}{\partial \beta_j} p) \quad (۶۲$$

فرض کنید $e = e^{-\sum_j \beta_j x_j}$ که در آن سیگما روی $j = ۰$ تا $j = d$ عمل می‌کند. بنابراین $p = ۱/(۱ + e)$ و $۱ - p = (۱ + e - ۱)/(۱ + e) = e/(۱ + e)$. در این صورت خواهیم داشت:

$$\begin{aligned} \frac{\partial}{\partial \beta_j} &= (-۱)(۱ + e)^{-۲} \frac{\partial}{\partial \beta_j} e \\ &= (-۱)(۱ + e)^{-۲}(e) \frac{\partial}{\partial \beta_j}[-\sum_j \beta_j x_j] \\ &= (-۱)(۱ + e)^{-۲}(e)(-x_j) \\ &= \frac{۱}{۱ + e} \frac{e}{۱ + e} x_j \\ &= p(۱ - p) x_j. \end{aligned} \quad (۶۳$$

بنابراین $(\partial/\partial \beta_j) ln\ p = (۱ - p) x_j$ و $(\partial/\partial \beta_j) ln(۱ - p) = -p x_j$. با داشتن نمونه‌های $\langle x_۱, y_۱ \rangle$ تا $\langle x_n, y_n \rangle$ ، مجموع مشتقات جزئی لگاریتم معیار شباهت نسبت به β_j برابر

$$\sum_{i: y_i = ۱} (۱ - p_i) x_{ij} + \sum_{i: y_i = ۱} -p_i x_{ij} = \sum_i (y_i - p_i) x_{ij} \quad (۶۴$$

که در آن x_{ij} مقدار jامین خصیصه از iامین نمونه‌ی آموزشی است. قرار دادن مشتقات جزئی برابر صفر، فرمول زیر را حاصل می‌کند:

$$\sum_i y_i x_{ij} = \sum_i p_i x_{ij} \quad (۶۵$$

ما برای هر پارامتر β_j یک معادله به این صورت خواهیم داشت.

شیب‌دهی صعودی، یک نمونه در هر لحظه

همان طور که در بالا مشاهده شد، برای محاسبه‌ی بهترین مقدار هر β_j نیاز است که فرمول $\sum_i y_i x_{ij} = \sum_i p_i x_{ij}$ برقرار باشد. اما طبق تعریف دیدیم که خود p_i به تمام مقادیر β_j وابسته است. یکی از روش‌های معمول برای محاسبه‌ی بهترین مقدار برای β_j استفاده از روش شیب‌دهی است. به این صورت که، ابتدا برای تمام مقادیر β_j یک مقدار پیش‌فرض در نظر گرفته می‌شود. سپس با مشاهده‌ی هر نمونه، تنها یکی از β_j ها طبق فرمول ارائه شده، اصلاح می‌شوند. و این عملیات با دیدن هر نمونه‌ی آموزشی تکرار می‌شود. اثبات می‌شود که این تکرار به بهترین مقدار β_j همگرا می‌شود.[؟]

مدل‌های خطی لاجیستیک

در این بخش به معرفی مدل‌های خطی لاجیستیک یا اصطلاحاً Log-linear models می‌پردازیم.

فرض کنید x یک نمونه باشد، و y یک برچسب ممکن برای آن. یک مدل خطی لاجیستیک فرض می‌کند که

$$p(y|x;w) = \frac{e^{\sum_j w_j F_j(x,y)}}{Z(x,w)} \quad (1)$$

که در آن Z تابع افراز[۲] نامیده می‌شود و برابر است با:

$$Z(x,w) = \sum_{y'} e^{\sum_j w_j F_j(x,y')} \quad (2)$$

بنابراین، با داشتن ورودی x، برچسب پیش‌بینی شده توسط مدل

$$\hat{y} = argmax_y p(y|x;w) = argmax_y \sum_j w_j F_j(x,y) \quad (3)$$

خواهد بود. هر کدام از عبارات $F_j(x,y)$ یک تابع خصیصه نامیده می‌شود.

از نظر ریاضیاتی، مدل‌های خطی لاجیستیک خیلی ساده هستند: برای هر تابع خصیصه، یک و فقط یک ضریب حقیقی مرتبط وجود دارد. دستکاری‌های زیادی برای فرمول وجود دارد.

[۲] partition function

میدان‌های تصادفی شرطی

مدل میدان‌های تصادفی شرطی در حقیقت نوع خاصی از مدل‌های خطی لاجیستیک است. منظور ما از میدان‌های تصادفی شرطی در این پایان‌نامه، به طور کلی میدان‌های تصادفی شرطی با حلقه‌ی خطی[۳] است.

یک کاربرد عمومی از میدان‌های تصادفی شرطی

برای شروع، پرداختن به نمونه‌ای از عملیات یادگیری با استفاده از میدان‌های تصادفی شرطی مفید به نظر می‌رسد. هدف ما تخصیص برچسب‌های ادات سخن به کلمات یک جمله به عنوان ورودی است. یک تعداد مشخص از برچسب‌های ادات سخن وجود دارند: اسم، فعل، قید، حرف‌اضافه و ... هر جمله یک نمونه‌ی جداگانه‌ی یادگیری یا آزمون محسوب می‌شود. ما یک جمله را با استفاده از تابع‌های خصیصه بر روی کلمات آن نمایش می‌دهیم. توابع خصیصه می‌توانند بسیار متفاوت از هم باشند:

- بعضی از توابع خصیصه می‌توانند مبتنی بر مکان باشند. برای مثال، شروع یا انتهای یک جمله، یا جمع تمام مکان‌های موجود در جمله.
- بعضی می‌توانند مربوط به تنها یک کلمه باشند. برای مثال پیشوندها یا پسوندها.
- بعضی از خصیصه‌ها می‌توانند به یک یا چند کلمه از کلمات قبل و یا بعد وابسته باشند.

دقیق‌ترین برنامه‌هایی که به عملیات برچسب‌گذاری ادات سخن می‌پردازند، در حال حاضر بیش از ۱۰۰ هزار تابع خصیصه دارند. یک محدودیت مهم برای میدان‌های تصادفی شرطی این است که یک تابع می‌تواند حداکثر به دو تا از برچسب‌های همسایه‌ی خود وابسته باشد.

---

[۳] Linear-chain CRF

تعریف میدان‌های تصادفی شرطی

برچسب‌گذاری ادات سخن یک نمونه از عملیات پیش‌بینی ساخت‌یافته[۴] است. طبق تعریف هدف این نوع از عملیات، پیش‌بینی یک دنباله‌ی پیچیده از برچسب‌ها برای یک دنباله‌ی ورودیِ پیچیده است. به سه دلیل انجام این عملیات، مشکل است. نخست، اطلاعات زیادی هنگام استفاده از یک رده‌بند کلمه به کلمه در هنگام یادگیری گم می‌شوند. تاثیر کلمات همسایه باید مورد توجه قرار گیرند. دوم، جملات مختلف، طول‌های مختلفی دارند؛ بنابراین واضح نیست که چگونه باید تمام جملات با استفاده از بردارهایی با طول یکسان نمایش داده شوند. سوم، تعداد مجموعه‌ی همه‌ی توالی‌های ممکن از برچسب‌ها، یک مقدار نمایی است.

یک میدان تصادفی شرطی، راهی است برای اعمال نمودن یک مدل خطی لاجیستیک به این نوع از عملیات است. در ادامه، استفاده از یک خط بر روی یک حرف، مانند $\bar{x}$ نشان‌گر یک توالی از برچسب یا داده‌ها است. به طور خاص، فرض کنید $\bar{x}$ به معنی یک توالی از n کلمه و $\bar{y}$ توالی n برچسب متناسب با آن کلمات است.

مدل خطی لاجیستیک معرفی شده در فرمول را در نظر بگیرید. فرض کنید که هر تابع خصیصه‌ی F_j واقعا یک مجموع روی جمله است، برای ۱ $i = 1$ تا $i = n$ که n طول $\bar{x}$ است داریم:

$$F_j(\bar{x}, \bar{y}) = \sum_i f_j(y_{i-1}, y_i, \bar{x}, i). \qquad (\text{۶۵})$$

این نمایش به این معنی است که خصیصه تنها بر اساس برچسب جاری، برچسب قبلی، کل جمله و مکان کلمه در جمله تعریف می‌شود. این محدودیت در حقیقت از تعریف میدان‌های تصادفی شرطی به عنوان حالت خاصی از مدل‌های خطی لاجیستیک حاصل می‌شود.

[۴] structured prediction task

استنتاج و یادگیری در میدان‌های تصادفی شرطی

آموزش یک مدل CRF به معنای پیدا کردن بردار وزن‌های w است به گونه‌ای که بهترین پیش‌بینی ممکن را برای هر نمونه‌ی آموزشی $\bar{x}$ ارائه دهد:

$$\bar{y}^* = argmax_{\bar{y}} p(\bar{y}|\bar{x}; w) \quad (1$$

در هر حال، قبل از اینکه بخواهیم در مورد فرآیند آموزش سخن بگوییم بایستی، باید متوجه این باشیم که دو مشکل اساسی در مرحله‌ی استنتاج برای ما وجود دارد: نخست، چگونه می‌توانیم معادله‌ی را برای هر $\bar{x}$ و هر مجموعه‌ای از وزن‌های w به صورت کارا محاسبه نماییم. این محاسبه تا هنگامی که تعداد دنباله‌های مختلف برای برچسب‌های $\bar{y}$ نمایی است.

دوم، با داشتن هر $\bar{x}$ و $\bar{y}$ ما باید مقدار زیر را ارزیابی کنیم:

$$p(\bar{y}|\bar{x}; w) = \frac{1}{Z(\bar{x}, w)} e^{\sum_j w_j F_j(\bar{x}, \bar{y})} \quad (2$$

مشکل در اینجا مخرج کسر است که باز روی تمام دنباله‌های $\bar{y}$ عمل می‌کند:

$$Z(\bar{x}, w) = \sum_{\bar{y}'} e^{\sum_j w_j F_j(\bar{x}, \bar{y}')} \quad (3$$

برای هر دوی این موارد به روش‌های ابتکاری و میان‌بر نیاز داریم که بدون پردازش لحظه‌ای روی همه‌ی در $\bar{y}$ ها، همه‌ی آن‌ها به گونه‌ای کارا پردازش شوند.این حقیقت که در میدان‌های تصادفی شرطی، هر تابع خصیصه تنها به دو برچسب که کنار هم قرار دارند وابسته است، به ما برای پیدا کردن چنین راه حلی کمک خواهد کرد.

برای آشنایی با شیوه‌ی استنتاج در میدان‌های تصادفی شرطی به روش کارا، می‌توانید به [؟] یا [؟] یا [؟] مراجعه نمایید.

در ادامه اشاره‌ای به چگونگی آموزش میدان‌های تصادفی شرطی خواهیم داشت.

وقتی که مجموعه‌ای از نمونه‌های آموزشی را در اختیار داریم، فرض می‌کنیم که هدف ما پیدا کردن پارامترهای w_j ای است که احتمال شرطی نمونه‌های آموزشی بیشینه شود. برای این منظور همان طور که در بالا شرح داده شد، می‌توانیم از روش شیب‌دار استفاده نماییم. پس نیاز است که مشتق جزئی معیار شباهت شرطی را برای یک نمونه‌ی آموزشی برای هر w_j محاسبه نماییم. به خاطر

بیاورید که بیشینه کردن p همان بیشینه کردن $ln\ p$ است:

$$\frac{\partial}{\partial_{w_j}} ln\ p(y|x;w) = F_j(x,y) - \frac{\partial}{\partial_{w_j}} logZ(x,w)$$

$$= F_j(x,y) - \frac{۱}{Z(x,w)} \sum_{y'} \frac{\partial}{\partial_{w_j}} e^{\sum_{j'} w_{j'} F_{j'}(x,y')}$$

$$= F_j(x,y) - \frac{۱}{Z(x,w)} \sum_{y'} e^{\sum_{j'} w_{j'} F_{j'}(x,y')} F_j(x,y') \quad (۶۷$$

$$= F_j(x,y) - \sum_{y'} F_j(x,y') \frac{e^{\sum_{j'} w_{j'} F_{j'}(x,y')}}{\sum_{y''} e^{\sum_{j''} w_{j''} F_{j''}(x,y'')}}$$

$$= F_j(x,y) - \sum_{y'} F_j(x,y') p(y'|x;w)$$

$$= F_j(x,y) - E_{y' \sim p(y'|x;w)}[F_j(x,y')].$$

به بیان دیگر مشتقات جزئی نسبت به وزن iام مقدار تابع خصیصه‌ی iام برای برچسب درست y است منهای مقدار متوسط تابع خصیصه برای همه‌ی برچسب‌های ممکن y'. توجه کنید که این مشتق‌گیری مقدار حقیقی را برای توابع خصیصه مجاز می‌سازد، نه فقط مقادیر صفر و یک را.

شیب‌دهی معیار شباهت شرطی وقتی که تمامی مجموعه‌ی آموزشی T در دست باشد، مجموع شیب‌دهی‌ها برای هر نمونه‌ی آموزشی است. بیشینه‌ی مطلق کل این شیب‌دهی‌ها برابر صفر است، پس داریم:

$$\sum_{\langle x,y \rangle \in T} F_j(x,y) = \sum_{\langle x,. \rangle \in T} E_{y \sim p(y'|x;w)}[F_j(x,y)] \quad (۶۷$$

این معادله برای همه‌ی نمونه‌های آموزشی درست است نه برای تک تک نمونه‌ها.

سمت چپ معادله‌ی بالا مقدار مجموع تابع خصیصه‌ی j روی همه‌ی مجموعه‌ی آموزشی است. سمت راست، مقدار مجموع تابع خصیصه‌ی j که توسط مدل پیش‌بینی شده است.

در نهایت در هنگام بیشینه کردن معیار شباهت شرطی با روش شیب‌دهی برخط، دستکاری وزن w_j به صورت زیر خواهد بود:

$$w_j := w_j + \alpha(F_j(x,y) - E_{y' \sim p(y'|x;w)}[F_j(x,y')]) \quad (۶۷$$

جمع‌بندی

در این بخش مدل میدان‌های تصادفی شرطی به عنوان یک مسأله‌ی کلی با کارکرد برچسب‌گذاری روی توالی‌ها بررسی شد. این بخش بدون شک یکی از مهمترین بخش‌های این پایان‌نامه است زیرا اساس سامانه‌ی تشخیص واحدهای اسمی ما را تشکیل می‌دهد. بعد از مطرح نمودن این بحث تئوری، در فصل آینده به مسأله‌ی تشخیص واحدهای اسمی به طور خاص در زبان عربی خواهیم پرداخت. در بخش **؟؟** به مطالب بیان شده در این فصل بازگشته و روش پیشنهادی خود را بر روی مدل میدان‌های تصادفی شرطی بنا می‌نهیم.

فصل ۴

تشخیص واحدهای اسمی برای زبان عربی

مقدمه

همان‌طور که می‌دانید عملیات تشخیص واحدهای اسمی امکان تشخیص اسامی خاص و نیز عبارات عددی یا زمانی را در یک متن بدون دامنه‌ی مشخص می‌دهد. سامانه‌های تشخیص واحدهای اسمی به عنوان یکی از نیازهای جانبی خیلی مهم برای عملیات‌های گوناگون پردازش زبان طبیعی شناخته شده است. متاسفانه تلاش‌های اصلی برای ساخت سامانه‌های تشخیص واحدهای اسمی قابل اطمینان برای زبان عربی در چارچوب‌های تجاری عرضه شده‌اند و رهیافت آن‌ها و نیز دقت کارایی آن‌ها شناخته شده نیستند. در میان تلاش‌های انجام شده، تعدادی از سیستم‌ها وجود دارند که کماکان محصولات متن-بسته محسوب می‌شوند اما لااقل مستندات کافی در مورد چگونگی عملکرد آن‌ها و نیز معیارهای ارزیابی‌شان وجود دارد. یکی از این موارد، برنامه‌ی ANERsys است که توسط یاسین بن‌جیبا و همکارانش پیاده‌سازی شده است. در ادامه سعی داریم، این سامانه را به عنوان سامانه‌ای که مخصوص متون عربی بر پایه‌ی وزنی-n و بیشترین آنتروپی ساخته شده‌است مورد بررسی قرار دهیم. همچنین روش اکتشافی مربوط به زبان و نیز پیکره‌ی متنی که برای بهتر شدن این سیستم استفاده شده است نیز توضیح داده می‌شود. آن‌ها پیکره‌ی متنی و فرهنگ[1] خودشان

[1] Gazetteers

را برای آزمودن، آموزش، ارزیابی و بهتر نمودن سامانه‌ی خود، ساخته‌اند. در ساخت این پیکره‌ی متنی آن‌ها تلاش زیادی کرده‌اند تا مطمئن شوند که چارچوب آن‌ها در هنگام ارزیابی، منطبق بر ارزیابی‌های صورت گرفته در کنفرانس CONLL 2002 است. آن‌ها همچنین آزمایش‌ها زیادی را انجام دادند و نتایج اولیه نشان دادند که این رهیافت خوبی برای برخورد با مساله‌ی تشخیص واحدهای اسمی در زبان عربی است.

مقدمه

آن‌ها در مقاله‌ی خود[؟] تحقیق گسترده‌ای را روی ابزارها و منابع کلی (مانند پیکره‌های متنی، فرهنگ‌ها و نشانه‌گذارهای ادات سخن) برای پردازش زبان طبیعی در زبان عربی انجام داده‌اند. نتایج به این منجر شد که در مقایسه با زبان‌های دیگر، زبان عربی منابع لغوی کافی، مخصوصا منابع آزاد برای تحقیق، ندارد. تعدادی از مهمترین منابع که هر زبانی با آن نیاز دارد، سیستم‌های تشخیص واحدهای اسمی هستند که به شناسایی اسامی خاص در یک متن عام کمک می‌کنند. مطالعه‌ی روزنامه‌های انگلیسی و فرانسوی ثابت کرد که این واحدها حدود ۱۰ درصد از مقالات را دربرمی‌گیرند. همان‌طور که قبلا نیز اشاره شد، در ششمین کنفرانس MUC عملیات تشخیص واحدهای اسمی در سه زیر-عملیات مشخص شد: ENAMEX (برای اسامی خاص یا خاص)، TIMEX (برای عبارات زمانی) و NUMEX (برای عبارات عددی). اولین زیر عملیات، عملیاتی است که توجه ما در این پایان نامه به آن معطوف شده است. ENAMEX به این صورت تعریف شده است: استخراج اسامی خاص و دسته‌بندی هر کدام از آن‌ها به (i) سازمان (برای مثال شرکت‌ها، سازمان‌های دولتی و یا سازمان‌های دیگر)؛ (ii) اماکن (نام مکان‌های مشخص شده به صورت سیاسی یا جغرافیایی) یا (iii) شخص (نام‌های شخصی یا خانوادگی) برای زبان عربی یا فارسی، پیکره‌های متنی کافی برای این منظور وجود ندارند. برای مثال در کنفرانس CONLL 2002، تنها پیکره‌های متنی برای زبان‌های چینی، انگلیسی، فرانسوی، ژاپنی، پرتقالی و اسپانیایی موجود بودند. این دلیلی بود برای اینکه آن‌ها سعی کنند به تولید پیکره‌ی متنی خودشان دست بزنند و هدف آن‌ها در تولید این پیکره‌ی متنی این بوده است که بتوانند آن را با محققان دیگر تشخیص واحدهای اسمی در زبان عربی به اشتراک بگذارند. البته قابل ذکر است که شرکت‌های دیگر

سامانه‌های تشخیص واحدهای اسمی دلخواه خود را برای اهداف تجاری ساخته‌اند. برای مثال برنامه‌ی سراج توسط شرکت صخر[۲]، نرم‌افزار ClearTags توسط شرکت ClearForest[۳]، نرم افزار NetOwlExtractor توسط شرکت NetOwl[۴] و InxightSmartDiscoveryEnti-tyExtractor توسط شرکت Inxight[۵] . متاسفانه با توجه به اینکه مستندات کافی در مورد این نرم‌افزارها وجود ندارد، مطالعه‌ای مقایسه‌ای میان آن‌ها امکان‌پذیر نیست. دو تکنیک اصلی برای ساخت سامانه‌های تشخیص واحدهای اسمی در زبان عربی وجود دارد. آن‌ها به ترتیب براساس استفاده از مجموعه‌ای از کلمات کلیدی و افعال خاص به عنوان رهاکننده‌ها[۶]، و یک مجموعه از قوانین برای استخراج اسامی خاص، و دومی استفاده از یک تحلیل لغوی با دقت بالا.

با توجه به سامانه‌های مستقل از زبان تشخیص واحدهای اسمی، کارهای زیادی با این روش انجام شده‌اند: در عملیات مشترک کنفرانس CONLL 2002 و CONLL 2003 برای آزمایش پیکره‌ی متنی زبان‌های انگلیسی، اسپانیایی و آلمانی، بیشتر آن‌ها از روشی مبتنی بر رهیافت بیشترین آنتروپی استفاده کرده‌اند، در حالی که بعضی دیگر ترجیح داده‌اند که تلاش‌های لغوی و وابسته به متن را با یکدیگر ترکیب کنند. علاوه بر این، در [؟] نتایج بسیار خوبی با استفاده از مدل وز-n سطح کاراکتری به دست آمده‌اند و در [؟] یک مقایسه بین روش مدل مخفی مارکوف (مقدار F-measure معادل ۳۱/۸) و بیشترین آنتروپی انجام شده است (البته افزودن امکانات بیشتر و یک مجموعه از نام‌های اول به صورت یک منبع خارجی باعث بهبود نتیجه تا به ترتیب ۸۲/۲۴ و ۸۵/۶۱ می‌شود). در نهایت، در کنفرانس NAACL/HLT 2004 یک سامانه‌ی تشخیص واحدهای اسمی بر پایه‌ی بیشترین آنتروپی برای انگلیسی، چینی و عربی،مقدار F-measure ی برابر ۶۸/۵ برای زبان عربی و ۶۸/۶ برای زبان چینی حاصل نمود. پیکره متنی عربی که در این روش‌ها مورد استفاده قرار گرفت متشکل از بیش از ۱۶۶ هزار حرف بوده است و از کنسرسیوم داده‌های زبانی[۷] گرفته شده است. این داده‌ها در حال حاضر برای دیگران قابل دسترس نیستند. به علاوه، یک روش جداسازی متن برای زبان عربی مورد استفاده قرار گرفته است که جدا بودن داده‌ها را کم نماید. زیرا زبان

---

[۲] برای اطلاعات بیشتر به این آدرس مراجعه کنید: http://siraj.sakhr.com/

[۳] برای اطلاعات بیشتر به این آدرس مراجعه کنید: http://www.clearforest.com/index.asp

[۴] برای اطلاعات بیشتر به این آدرس مراجعه کنید: http://www.netowl.com/products/extractor.html

[۵] برای اطلاعات بیشتر به این آدرس مراجعه کنید: http://www.inxight.com/products/smartdiscovery/ee/index.php

[۶] trigger

[۷] Language Data Consortium

عربی زبانی است که کلمات آن بسیار با یکدیگر پیوند می‌خورند[۸] . بنابراین، با توجه به سامانه‌هایی که در بالا راجع به آن‌ها صحبت شد، بن‌جیبا و همکارانش به این نتیجه رسیدند که روش بیشترین آنتروپی، بهترین و کاراترین روش ممکن برای عملیات تشخیص واحدهای اسمی است (البته این اظهار نظر آن‌ها در مقاله مذکور، بسیار نادقیق است. همان طور که در ادامه‌ی این بخش خواهیم دید، آن‌ها خلاف این موضوع را در مقاله‌های بعدی خود ثابت کردند).

تشخیص واحدهای اسمی در زبان عربی

سامانه‌های مستقل از زبانِ شرکت کننده در کنفرانس CoNLL که پیشتر ذکر شدند، از یک روش کلی مبتنی بر ویژگی‌های عمومی همه‌ی زبان‌ها استفاده می‌کردند. وقتی با زبان عربی کار می‌کنیم، بعضی از ویژگی‌های مهم باید مورد توجه قرار گیرند:

- یک کاراکتر ممکن است با توجه به مکان قرارگیری‌اش در کلمه، تا سه حالت متفاوت نمایشی داشته باشد: اول،چسبان و آخر.
- زبان عربی حساس به حروف کوچک و بزرگ نیست. این خصوصیت یکی از بزرگترین موانع در عملیات تشخیص بهتر واحدهای اسمی است، زیرا یکی از مهمترین خصوصیات برای انجام این عملیات در زبان‌های دیگر می‌باشد.
- این زبان دارای صداهای کشیده‌ی بلند و کوتاه می‌باشد، اما صداهای کشیده‌ی کوتاه دیگر در روزنامه‌ها نوشته نمی‌شوند و این موضوع، متون را بسیار مبهم‌تر می‌کند.
- و در نهایت، این زبان دارای لغت‌شناسی بسیار پیچیده‌ای است.

آخرین خصوصیت ذکر شده، از دیدگاه تشخیص واحدهای اسمی، از همه‌ی موارد دیگر مهمتر هستند. زبان عربی، چسبندگی بسیار زیادی دارد،زیرا فرم اصلی یک کلمه به صورت زیر است:
پیشوند(ها) + ریشه + پسوند(ها)
تعداد پیشوندها و پسوندها ممکن است هیچ یا بیشتر باشند. پسوندها به ریشه متصل می‌شوند تا

[۸] اصطلاحا این زبان یک زبان Inflected یا fusional است.

اصطلاح مورد نیاز را تولید کنند. یک مثال می‌تواند این باشد: کلمه‌ی عربی «منزل» در انگلیسی به معنی House است و «المنزل» به معنی The house. این مثال نشان می‌دهد که یک کلمه‌ی عربی می‌تواند به دو کلمه‌ی انگلیسی ترجمه شود. یک مثال پیچیده‌تر، می‌تواند کلمه‌ی «وَسَيَكتُبونَها» باشد که به معنی «و آن‌ها آن را خواهند نوشت» یا "and they will write it" می‌باشد. اگر ما این کلمه را در فرم اصلی که در بالا ذکر کردیم، نمایش دهیم، خواهیم داشت:

و + س + ی + کتب + ون + ها

از دید مبتنی بر تشخیص واحدهای اسمی، این ویژگی عجیب زبان عربی، یک مانع بسیار بزرگ خواهد بود چون باعث جداافتادگی داده‌ها خواهد شد.

در سیستم تشخیص واحدهای اسمی که در [؟] توضیح داده شده است،یک مجموعه از قوانین و کلمات کلیدی، به منظور استخراج نام‌های خاص، استفاده شده‌اند (مشکل جداافتادگی داده‌ها در مقاله‌ی مزبور بررسی نشده است). در [؟] مولفان روی این مشکل تمرکز نموده‌اند و یک الگوریتم جداسازی متن را مورد استفاده قرار داده‌اند. (در [؟] معرفی شده است.) این الگوریتم بر اساس مدل زبانی چند-وزنی ساخته شده است و احتمالات سه-وزنی واکی[۹] ، را محاسبه می‌نماید. برای انجام چنین عملیاتی، آن‌ها از یک پیکره‌ی متنی جداشده توسط دست، استفاده کرده‌اند. نتیجه این‌طور گزارش داده شده است که دقتی برابر ۹۷ درصد حاصل گشته. مهم است که روی این موضوع تایید نماییم که چنین الگوریتمی به راحتی قابل پیاده‌سازی نیست زیرا به یک پیکره‌ی متنی بزرگ که با دست جدا شده است، برای آموزش نیاز دارد.

در برنامه‌ی ANERsys، سعی شده است که با مشکل جداافتادگی داده‌ها نیز برخورد شود. به جای اعمال یک عملیات جداسازی متن، آن‌ها از یک روش اکتشافی که فقط به پیشوندها توجه می‌کند استفاده نموده‌اند.

[۹] Morpheme trigram probabilities

رهیافت بیشترین آنتروپی

تکنیک بیشترین آنتروپی، نه تنها در تشخیص واحدهای اسمی، بلکه در بسیاری از عملیات‌های دیگر پردازش زبان طبیعی نیز موفق عمل کرده است [؟ ؟ ؟]. بگذارید این رهیافت را با مثال ساده‌ای معرفی کنم. جمله‌ی زیر را که از روزنامه‌ی انگلیسی الجزیره انتخاب شده است در نظر بگیرید:

"Sudan's Darfur region remains the most pressing humanitarian problem in the world, the Food and Agriculture Organisation says."

ما نیاز داریم که کلمه‌ی Darfur را به عنوان یکی از این کلاس‌ها طبقه بندی کنیم: Pers: نام خاص یک شخص، Loc: نام خاص یک مکان، Org: نام خاص یک سازمان و lrO: خارج بودن از اسامی خاص.

اگر فرض کنیم که هیچ اطلاعاتی در مورد کلمه نداریم، آنگاه بهترین توزیع احتمالاتی این است که احتمال تمام چهار کلاس بالا را برابر هم در نظر بگیریم. بنابراین، می‌توانیم توزیع را به صورت زیر بنویسیم:

$$p(O) = p(Pers) = p(Loc) = p(Org) = ۰/۲۵ \qquad (۱$$

به بیانی دیگر، این توزیع، توزیعی است که آنتروپی را در بیشترین حالت قرار می‌دهد (در این قسمت منظور ما از «بهترین توزیع احتمالاتی» توزیعی است که معیار فاصله‌ی Kullback-Leibler از توزیع واقعی را به کمترین مقدار خود برساند)

حال فرض کنید ما موفق می‌شویم مقداری داده‌ی آماری را از یک پیکره‌ی متنی بیرون بکشیم و متوجه شویم که ۹۰ درصد کلماتی که با حرف بزرگ آغاز می‌شوند و اولین کلمه‌ی یک جمله نیستند، نام‌های خاص هستند. بنابراین، توزیع احتمالاتی جدید به صورت زیر خواهد بود:

$$p(O) = ۰/۱ \quad and \quad p(Pers) = p(Loc) = p(Org) = ۰/۳ \qquad (۲$$

این مثال به طور خلاصه نشان می‌دهد که یک رده‌بند بیشترین آنتروپی چگونه عمل می‌کند. هرگاه ما نیاز داشتیم که اطلاعات تکمیلی را اضافه نماییم، این مدل بهترین توزیعی را که آنتروپی را به

بیشترین مقدار می‌رساند، محاسبه می‌کند. ایده‌ی پشت این رهیافت، این است که بهترین توزیع زمانی حاصل می‌شود که ما از هیچ اطلاعات اضافی دیگری استفاده نمی‌کنیم، جز آنچه که در فاز آموزش حاصل شده است، و هیچ اطلاعاتی در مورد بعضی از کلاس‌ها موجود نباشد، باقی توده‌ی احتمالاتی[۱۰] به صورت یکنواخت بین آن‌ها توزیع شده‌اند.

در این مثال، ما ساخت و محاسبه‌ی توزیع احتمالاتی را به راحتی مدیریت کردیم، زیرا تعداد کلاس‌ها در این‌جا کم بود و ما اطلاعات آماری کم و ساده‌ای را راجع به اسامی خاص می‌دانستیم. متاسفانه، این فرضیات هرگز برای مثال‌های واقعی درست نیست. و ما همیشه تعداد بیشتری کلاس و برد بیشتری از اطلاعات متنی داریم. بنابراین، یک محاسبه‌ی دستی از توزیع آماری ممکن نیست. در نتیجه، یک مدل رده‌بند قابل اطمینان مورد نیاز است. اثبات شده است که یک مدل نمایی یک رهیافت مناسب برای مسئله‌هایی است که از منابع اطلاعاتی گوناگون استفاده می‌کنند. همان‌طور که در زیر نشان داده شده است:

$$Z(x) = \sum_{c'} exp(\sum_{i} \lambda_i . f_i(x, c')) \qquad ($$

که در آن c یک کلاس است، x یک اطلاعات متنی یا context information است و $f_i(x, c)$ i، امین خصیصه یا feature است. خصیصه‌های ما، توابعی دودویی هستند که بیانگر چگونگی ارتباط کلاس‌های متفاوت با یک یا چند اطلاعات متنی هستند. برای مثال:

$$f_i(x, c) = ۱ \quad if\ word(x) = "Darfur"\ and\ c = B\text{-}LOC, ۰\ otherwise \qquad ($$

برای هر خصیصه، یک وزن با نام λ_i موجود است در حالی که هر خصیصه‌ای مربوط به کلاسی است و بنابراین ممکن است یک تاثیر کوچک یا بزرگ در تصمیم رده‌بند برای یک کلاس یا دیگری داشته باشد. وزن‌ها با استفاده از الگوریتم مقیاسگیری تکراری کلی [۱۱] یا GIS تخمین زده شده است که ما را از همگرا بودن به یک وزن درست بعد از تعدادی تکرار مطمئن می‌کند.
از یک نگاه کلی، ساخت یک رده‌بند بیشترین آنتروپی، نیازمند گام‌های زیر است:

- با استفاده از مشاهدات و آزمایش‌ها فهرستی از ویژگی‌هایی در رابطه با متن که در آن واحدهای

[۱۰] Probability mass

[۱۱] General Iterative Scaling

اسمی بیشتر ظاهر می‌شوند. این کار به طور کلی خیلی ساده نیست زیرا اثبات شده است که تعدادی از این ویژگی‌ها، آنقدرها هم مفید نیستند و نیاز است که جایگزین شوند. بنابراین ممکن است که ما به این گام مدام برگردیم و تا فهرست مورد نظر را اصلاح نماییم.

- برای تخمین مقدار وزن‌های λ_i از الگوریتم GIS استفاده نمایم.
- ساخت یک رده‌بند که به صورت پایه‌ای، برای هر کلمه، احتمالاتی که باید به هر کدام از کلاس‌ها تخصیص داده می‌شود را محاسبه نماید: $p(\text{B-PERS}|w_i)$ ، $p(\text{I-PERS}|w_i)$ و غیره. استفاده از فرمول بیشترین آنتروپی و سپس تخصیص کلاس با بیشترین احتمال به آن کلمه.

مجموعه‌ی خصیصه‌هایی که برای پیاده‌سازی ANERsys به طور مفصل توضیح داده خواهد شد.

منابع توسعه داده شده

همان طور که در بالا نیز اشاره شد، یک پیکره‌ی متنی مناسب و آزاد برای زبان عربی مخصوص تشخیص واحدهای اسمی، تا قبل از تلاش بن‌جیبا و همکارانش وجود نداشت. بنابراین آن‌ها تصمیم گرفتند تا پیکره‌ی متنی خود را بسازند. به علاوه آن‌ها یک فرهنگ نیز برای آزمودن تاثیر یک منبع اطلاعاتی بیرونی بر روی سامانه ساخته‌اند. حاصل تمام این تلاش‌ها برای استفاده‌ی محققان دیگر آزاد گذاشته شده است.

اهمیت بررسی این پیکره‌ی متنی این است که آن‌ها قبل از تولید آن، به خوبی تمام استانداردهای لازم را رعایت کرده‌اند به طوری که نتیجه، قابل مقایسه با کنفرانس معتبر CoNLL است. به همین دلیل ما نیز در این پایان‌نامه، ابتدا پیکره‌ی متنی خود را که پیکره‌ی متنی فوق‌العاده بزرگ و قابل اطمینانی است، با استفاده از یک نرم‌افزار جانبی که خود آن را تهیه کرده‌ایم به این فرمت استاندارد در آورده و سپس آن را به الگوریتم خود برای ارزیابی داده‌ایم.

ANERcorp، دو پیکره‌ی متنی برای آموزش و آزمودن

همان طور که در کنفرانس CoNLL 2002 گزارش داده شد، پیکره‌ی متنی نشانه‌گذاری شده بایستی حاوی کلماتی از یک متن به همراه نوع مرتبط با آن باشد. همان کلاس‌هایی که در کنفرانس MUC-6 (سازمان، مکان و شخص) مشخص شده بودند، در این پیکره نیز وجود داشتند. «متفرقه» یا "Miscellaneous" یک کلاس تکی است که برای واحدهای اسمی اضافه شده است به این معنا که متعلق به هیچ کلاس دیگری نمی‌باشد. بنابراین، هر کلمه در متن باید به صورت یکی از موارد زیر برچسب خورده باشد:

- B-PERS : شروع نام یک شخص
- I-PERS : ادامه یا انتهای نام یک شخص
- B-LOC : شروع نام یک مکان
- I-LOC : ادامه یا انتهای نام یک مکان
- B-ORG : شروع نام یک سازمان
- I-ORG : ادامه یا انتهای نام یک سازمان
- B-MISC : شروع نام یک واحد اسمی که جز موارد بالا نیست.
- I-MISC : ادامه یا انتهای نام یک واحد اسمی که جز موارد بالا نیست.
- I-PERS : واحد اسمی نسیت.

در کنفرانس CoNLL همچنین تصمیم گرفته شد که یک فرمت واحد برای داده‌های آموزشی همه‌ی زبان‌ها مورد استفاده قرار گیرد. در این فرمتی فایلی با دو ستون وجود دارد: ستون اول برای کلمات و دومی برای برچسب‌ها. شکل روش استفاده شده در CoNLL برای زبان انگلیسی و روش استفاده شده در پیاده‌سازی‌های ما را به همراه هم نشان می‌دهد. با توجه به کنفرانس CoNLL 2002 ، ما سه پیکره‌ی متنی برای زبان عربی نساخته‌ایم (یکی برای آموزش، دیگری برای آزمون اول که برای بهینه

فرانکفورت B-LOC	with O
، O	Del B-PER
اعلن O	Bosque I-PER
اتحاد B-ORG	in O
صناعة I-ORG	the O
السيارات I-ORG	final O
فی O	years O
المانيا B-LOC	of O
امس O	the O
الاول O	seventies O
ان O	in O
	Real B-ORG
	Madrid I-ORG
	. O

شکل نحوه‌ی برچسب‌گذاری متون انگلیسی و عربی در کنفرانس CoNLL

کردن پارامترها استفاده می‌شود و دیگری برای آزمون پایانی) اما دو پیکره‌ی متنی آماده شده است که یکی برای آموزش و دیگری برای آزمودن است. قبل از آن، ما یک الگوریتم نرمال‌سازی متنی را برای جلوگیری از تاثیر جداافتادگی داده‌ها به متن اعمال کرده‌ایم. برای مثال، به خاطر ویژگی‌های زبانی، اگر نرمال‌سازی وجود نداشته باشد، ممکن است کلمه‌ای مانند «ایران» را به دو صورت پیدا کنیم. متاسفانه، نرمال‌سازی یک متن عربی با روش منحصر به فردی کار نمی‌کند، با نگاه کردن به کنفرانس TREC 2001[۱۲] و 2002 ، استخراج داده[۱۳]های دو طرفه‌ی عربی/انگلیسی، اغلب با جایگذاری تعدادی کاراکتر با معادل آن‌ها صورت می‌پذیرد. این عملیات ساده، تاثیر خوبی روی عملیات استخراج داده‌ها دارد، ولی برای تشخیص واحدهای اسمی، روش مناسبی نیست، زیر باعث از بین رفتن بعضی داده‌های باارزش می‌شود که در عملیات تشخیص واحدهای اسمی مهم هستند. بنابراین، برای بومی‌کردن تعریف نرمال‌سازی برای کار ما، و همچنین در ANERcorp[۱۴]، ما تنها، فرم‌های متفاوت را کم کرده‌ایم. مثلا برای کاراکتر «ا» فقط یک فرم را انتخاب کرده‌ایم.

پیکره‌ی متنی ANERcorp متشکل از ۳۱۶ مقاله است. در این پیکره‌ی متنی سعی شده است که مقالات بیشتر از یک موضوع و یک روزنامه انتخاب نشوند و تمام موضوعات را در بر بگیرند.

[۱۲] نام سرآیند The Tenth Text REtrieval Conference

[۱۳] Information Retreival

[۱۴] برای دریافت این پیکره‌ی متنی به آدرس http://www.dsic.upv.es/ ybenajiba مراجعه نمایید.