شماره مستند: ۱۹۰/۲۵۳۷/۱/۳

جمهوری اسلامی ایران

دبیرخانه شورای عالی اطلاع رسانی

استخراج نیازمندی‌های تعیین حدود «جمله» برای پیکره متنی زبان فارسی

نسخه ۱.۰

دانشگاه علم و صنعت ایران

فروردین ۸۸

فهرست مطالب

مقدمه

۱-۱- مقدمه

از آن جایی که جمله، یک واحد متنی پایه است و بلافاصله بعد از کلمه و عبارت قرار می‌گیرد، تعیین محدوده جمله یک مسئله اساسی برای بسیاری از کاربردهای پردازش زبان طبیعی[۱]، مانند تجزیه[۲]، استخراج‌اطلاعات[۳]، ماشین‌ترجمه[۴]، خلاصه‌سازی[۵] و خلاصه‌گیری[۶]، ساخت پیکره‌های متنی[۷] زبان و برچسب‌گذاری[۸] نحوی و معنایی و ... است و علاوه بر این، برچسب‌گذاری اجزای کلام[۹] شدیداً نیازمند تعیین دقیق حدود جملات است[۱]. دقت این سیستم به طور مستقیم روی کارآیی برنامه‌های کاربردی اثر می‌گذارد. اگرچه قطعه‌بندی جمله می‌تواند از روی علائم نشانه‌گذاری مثل نقطه یا کوتیشن به دست آید، اما ابهام‌های زیادی هنوز در متون واقعی باقی می‌ماند.

جمله مهمترین واحد در بسیاری از کارهای پردازش زبان طبیعی است. برای مثال، تنظیم جملات[۱۰] در اسناد چند زبانه موازی[۱۱] نیاز دارد که اول محدوده جملات به روشی برچسب‌گذاری شود[۲].

بیشتر برچسب‌های اجزای کلام، به حذف ابهام محدوده جملات در متن ورودی نیاز دارند، معمولاً این کار با درج یک رشته کاراکتر منحصر به فرد در پایان هر جمله انجام می‌گیرد. بدین‌گونه است که ابزار[۱۲] تحلیل پردازش زبان طبیعی جملات منحصر به فرد را به راحتی می توانند تشخیص دهند.

تعیین حدود جمله در زمره مسائل پیش‌پردازش زبان فارسی قرار می‌گیرد که کاری روی آن

---

[1] Natural Language Processing (NLP)
[2] Parsing
[3] Information Extraction
[4] Machine Translation
[5] Abstraction
[6] Summarization
[7] Corpus
[8] Tagging
[9] Part-of-Speech tag (POS tag)
[10] Alignment
[11] Parallel Multi-Lingual Corpora
[12] Tools

انجام نشده است و در جاهایی که نیاز به این کار بوده، از علائم نشانه‌گذاری استفاده شده است. در این پایان‌نامه چند روش جهت شناسایی حدود جمله در متن فارسی ارائه شده و این روش‌های پیشنهادی مورد بررسی و تجزیه و تحلیل قرار داده شده است. این روشها عبارتند از بررسی ساختاری جمله و تعیین حدود جمله با استفاده از یک روش کارای تشخیص فعل در جمله، استفاده از مدل چند-تایی[۱۳] در تعیین حدود جمله، استخراج خصیصه از جملات و به کاربردن روشهای رده‌بندی، چون شبکه‌های عصبی ، برای تعیین حدود جمله.

مفاهیم اولیه مورد نیاز و مشکلات و چالشهای زبان فارسی در فصل دوم مورد بررسی قرار گرفته و در ادامه آن مسئله تعیین حدود جمله به‌طور کامل تشریح شده و مورد بررسی قرار می‌گیرد.

تعیین حدود جمله، یکی از مهم‌ترین مراحل پردازش لغوی که از جمله مراحل پیش‌پردازش در اکثر کارهای متن‌کاوی و پردازش زبان طبیعی است، می‌باشد. در این زمینه تابه‌حال در مورد زبان فارسی هیچ کاری صورت نگرفته است، اما در مورد زبان‌های دیگر، مثل انگلیسی، پرتغالی و ژاپنی و ...، کارهای زیادی با استفاده از روشهای مختلف صورت گرفته است. در این کارها مسئله تعیین حدود جمله تبدیل به مسئله رفع ابهام علائم نشانه‌گذاری شده است و با آن به صورت یک مسئله رده‌بندی پایه برخورد شده است. چندی از کارهای مهم انجام شده در زبان‌های مختلف در فصل سوم مورد بررسی قرار داده شده و در فصل چهارم یک سری روش‌های پیشنهادی جهت تعیین حدود جمله در متن فارسی ارائه و مورد بررسی قرار گرفته است. فصل پنجم به پیاده سازی و ارزیابی روش‌های ارائه شده و مقایسه‌ای بین آنها پرداخته است. در فصل آخر نتیجه‌گیری کارهای انجام شده به همراه راه‌کارهای آینده آن مورد بررسی قرار گرفته است

---

[۱۳] **مدل چند-تایی**

مفاهیم اولیه و زمینه‌ي تحقیق

۱-۲- پردازش متون فارسی

زبان فارسی دربردارنده گنجینه‌ی بزرگی از زیباترین سروده‌ها و داستانها است. زبان فارسی یکی از پربارترین زبان‌های دنیا است. کتابهایی چون مثنوی معنوی، دیوان حافظ، رباعیّات خیام و ... به زبان‌های گوناگون گیتی برگردانده شده و بارها چاپ شده‌اند. برترین ویژگی این نوشته‌ها، انسانی بودن آنها است بگونه‌ای که همه‌ی انسانها گرایشی درونی به این نوشته‌ها دارند.

متأسفانه این درخت تنومند امروزه نیاز به توجه بیشتری دارد زیرا برای دنیای نوین آماده نشده است. پیرایش و ویرایش بر روی دیگر زبان‌های دنیا خیلی پیشتر از این آغاز شده است. ساده کردن قاعده‌ها، کم کردن قاعده‌های پیچیده و استثناها در زبان روزمره (نه زبان ادبی)، یکسان کردن گفتار و نوشتار روزمره، به کارگیری تعداد کمی واژه و اصطلاح، گسترش استانداردهای آماده شده برای زبان از کارهایی است که بر روی بسیاری از زبان‌ها انجام شده است. استادان زبان انگلیسی و زبان‌شناسان، بسیاری از قاعده‌های این زبان را پیراسته‌اند و یادگیری و به کارگیری این زبان را ساده نموده‌اند. برای نمونه در نوشتار امروزی انگلیسی کمتر حرفها به هم چسبیده نوشته می‌شوند و واژه‌ها و اصطلاحهای کمی، بویژه در نوشته‌های علمی، به کار گرفته می‌شود. ویراشهای انجام شده در زبان انگلیسی بسیار بر کارهای رایانه‌ای، که بر پایه‌ی زبان انگلیسی هستند، اثر داشته است و به پیشرفت نرم افزارهای رایانه‌ای کمک نموده است. پیرایش‌هایی که در زبان انگلیسی انجام شده است، بسیاری از پیچیدگی‌های ساخت نرم افزارهایی برای این زبان را کاسته است و به نوبه‌ی خود ساخت نرم افزار رایانه‌ای گسترش استاندارد آن زبان را در پی داشته است.

در پردازش متون زبان طبیعی با زبان نوشتاری سروکار داریم. این مسئله باعث می‌شود گرچه به جهت از دست دادن اطلاعات گویشی مانند لحن گوینده، آهنگ صدا، تاکید و مکث، با مشکلات و ابهاماتی مواجه شویم، ولی در مقابل با شکل

محدودتری از زبان کار می‌کنیم. بسیاری از بی‌ترتیبی‌های زبان، متعلق به زبان گفتاری است و در زبان نوشتاری بیشتر قالب‌های دستوری رعایت می‌شوند و لذا تهیه دستور زبان پوشاننده‌ی تمام متن، ساده‌تر است.

در تلاش برای ساخت یک سیستم پردازش و درک متون فارسی با مسائل و مشکلاتی مواجه می‌شویم که بعضی در بیشتر زبان‌ها بروز کرده و برخی خاص زبان فارسی می‌باشند. همچنین برخی از این پیچیدگی‌ها به طبیعت زبان و نارسایی‌های قواعد زبان شناسی مربوط و برخی دیگر برخاسته از مشکلات ایجاد سیستم‌های هوش مصنوعی است [۳]. در بخش بعد به برخی از این مسائل اشاره می‌شود [٤].

۱-۲-۱- پیچیدگی‌های پردازش متون فارسی

با توجه به بحث اخیر می‌توان در کل اهم مشکلات فعلی پردازش متون فارسی را در چند دسته زیر خلاصه نمود[٤]:

۱) عدم وجود منابع زبانی مناسب و کافی برای زبان فارسی مانند واژگان‌های تک زبانه و چند زبانه محاسباتی، واژگان‌های معنایی و متصل به هستان شناسی (هستان شناسی‌های لغوی)، هستان شناسی جامع عمومی و تخصصی، پیکره‌های عمومی و تخصصی ساده یا برچسب خورده (با برچسب‌های اجزای کلام، کسره اضافه، نقش‌های موضوعی، مفاهیم و روابط مفهومی و غیره)، مجموعه مدون قوانین ساختواژی و دستوری پوشا، عدم وجود استانداردهای شیوه نگارش، فاصله‌گذاری و رمزگزاری حروف و علائم.

۲) مشکل تشخیص مرز کلمات (مسئله شیوه‌های نگارش متفاوت)

۳) مشکل تشخیص مرز گروه‌های اسمی (مسئله کسره اضافه نامرئی)

٤) از دست دادن اطلاعات گویشی

٥) مسئله ابهام

٦) افعال مرکب و اصطلاحات

۷) مسئله هم نگاره‌ها و تحت آن مسئله حذف مصوت‌های کوتاه (اعراب) از نوشتار
۸) معناشناسی و مشکلات تحلیل معنایی.

۱-۳- تعیین حدود جمله

۱-۳-۱- تعریف جمله

از جمله در زبان فارسی تعاریف گوناگونی ارائه شده است که برخی از آنها در زیر آورده شده است :
جمله یک یا مجموع چند واژه است که بر روی هم پیام کاملی را از گوینده به شنونده برساند.
مجموعه‌ی کلمات به هم پیوسته که پیامی را می‌رساند و دارای اجزاء و ارکان است.
به قسمتهایی که حذف آنها ساختمان را از هم می پاشد ارکان جمله گویند و اجزای جمله آنها هستند که حذفشان به ساختمان جمله آسیبی نمی‌رساند.
در زبان فارسی جمله‌ها از چند حیث دسته‌بندی می‌شوند:
اقسام جمله از لحاظ پیام‌رسانی:
۱- خبری: جمله‌ای که به صورت اِخباری یا التزامی ومثبت یا منفی درباره تحقق کاری یاحالتی سخنی گویند، مانند "هوا سرد است.".
۲- پرسشی: جمله‌ای که به وسیله آن، ظاهراً یا حقیقتاً درباره امری پرسش شود، مانند "حال شما چطور است؟".
۳- امری: که راجع تحقق کار یا حالتی درخواستی صورت می گیرد، مانند "عصبانی نشوید.". وجه امری شامل دو صورت امرو نهی، رایج در زبان عربی، است. در جمله های امری خطابی معمولاً نهاد خذف می‌شود، مانند " افسوس نخور".
٤- عاطفی: جمله‌ای که با آن عواطف و احساسات انسانی بیان می شود ،از قبیل تحسین، تمجید، تعجب، آرزو، افسوس؛ مانند "شما چقدر خوش ذوق هستید.".

اقسام جمله از لحاظ نظم:

۱- جمله مستقیم: که ارکان یا اجزای آن در جای خود قرار دارد، مانند "علی دیروز ما را به خانه خود دعوت کرد".

۲- جمله غیر مستقیم: جمله‌ای که نظم دستوری یک یا چند رکن یا جزء آن به هم خورده است که برای تنوع بخشیدن به جمله به کار می رود. از نمونه های خوب کاربرد چنین جملات درآثار جلال آل احمد به چشم می خورد، مانند "امیدوارم خسته نشده باشید؛ اگر هم هستی کمی استراحت کنید و یک نوشیدنی بنوشید تا خستگی برطرف شود و بقیه مطالب را بخوانید".

اقسام جمله از لحاظ فعل:

۱- جمله فعلی: جمله‌ای که دارای فعل تام است، مانند "ماه تابید.".

۲- جمله اسنادی: جمله که دارای فعل ربطی است، مانند " امین با هوش است.".

۳- جمله بی فعل: جمله‌ای که فعل ندارد. مانند "دریغا".

اقسام جمله دارای فعل:

۱- جمله ساده: جمله دارای یک فعل است، مانند "امین خوابید.".

۲- جمله مرکب: جمله‌ای که دارای بیش از یک فعل است، مانند "همین که به خانه رسیدم مهمان آمد".

اما با این‌حال هیچ تعریف محاسباتی دقیقی از جمله در زبان فارسی وجود ندارد تا بتوان با استناد به آن به‌طور دقیق‌تر مسئله را تعریف کرد اما به هر حال با توجه به تعاریف ارائه شده در بخش بعد به بیان مسئله پرداخته شده است.

۱-۳-۲- بیان مسئله

جملات در زبان فارسی می‌توانند به صورت ساده و مرکب، با ساختار درست و استاندارد و بدون داشتن این ساختار و با استثنائات زیاد بیاید. در جملات فارسی می-توان قسمتی از جمله را حذف کرد و استفاده ننمود. تمامی این مسائل باعث می‌شود که تعیین حدود جمله کار بسیار مشکلی در زبان فارسی باشد. اما در مسائل مختلف نیاز به دانستن حدود جمله - ابتدا و انتهای آن- داریم. در جملات ساده و استاندارد کار قابل انجام به نظر می‌رسد، اما در مورد جملات مرکب و غیر استاندارد کار ساده‌ای نیست.

هدف تعیین ابتدا و انتهای هر جمله است تا این‌کار بتواند کمک به سزایی در بالا بردن دقت عملیات متن‌کاوی و پردازش زبان طبیعی بکند.

مروري برکارهاي انجام شده

۱-۴- مقدمه

همانطور که در بخش‌های پیشین ذکر شد، تعیین حدود جمله، از مراحل پردازش لغوی که از جمله مراحل پیش‌پردازش در اکثر کارهای متن‌کاوی و پردازش زبان طبیعی است، می‌باشد. در این زمینه تابه‌حال در مورد زبان فارسی هیچ کاری صورت نگرفته است، اما در مورد زبان‌های دیگر، مثل انگلیسی، پرتغالی و ژاپنی و ...، کارهای زیادی با استفاده از روشهای مختلف صورت گرفته است. در این کارها مسئله تعیین حدود جمله تبدیل به مسئله رفع ابهام علائم نشانه‌گذاری شده است و با آن به صورت یک مسئله رده‌بندی پایه برخورد شده است. چندی از کارهای مهم انجام شده در زبان‌های مختلف در ادامه مورد بررسی قرار داده شده است.

در این فصل به بررسی سیستمهای موجود در تعیین حدود جمله در متون زبان‌های غیرفارسی می‌پردازیم. رفع ابهام محدوده جملات برای تشخیص عناصر جمله در یک پاراگراف یا مقاله کاربردی است. از آن جایی که جمله، یک واحد متنی پایه بلافاصله بعد از کلمه و عبارت است. ابهام زدایی محدوده جمله یک مسئله اساسی برای بسیاری از کاربردهای پردازش زبان طبیعی است. دقت این سیستم به طور مستقیم روی کارآیی برنامه‌های کاربردی اثر می‌گذارد. به هر حال کارهای تحقیقی گذشته در این زمینه، کارآیی خیلی بالای به دست آورده و مسئله برای جلب توجه محققان ساده به نظر می‌رسد[۱].

در حقیقت خود مسئله به سادگی که برای انجام نشان می‌دهد، نیست. ما همه می‌دانیم که جمله یک توالی از کلمات است که به یک علامت نشانه‌گذاری[14] پایان جمله، مثل .، ؟ یا !، ختم می‌شود. بیشتر جملات یک نقطه در پایان را استفاده می‌کنند. به هر حال ما باید توجه کنیم که بعضی اوقات یک نقطه پایان جمله می‌تواند بهمراه اختصار مثل «Mr.» یا نقطه اعشار مثل «12.58» در یک عدد بیاید. در در این موارد

[14] Punctuation Mark

نقطه یک قسمتی از مخفف[۱۵] یا یک عدد است. حدود جمله را با استفاده از نقطه نمی-توان تعیین کرد، زیرا نقطه می‌تواند معانی مختلفی در مکان‌های مختلف داشته باشد. از طرف دیگر نقطه پشتی یک مخفف در برخی اوقات می‌تواند پایان جمله را نیز نشان دهد. در بیشتر این موارد کلمات دنبال این نقطه یک کلمه معمولی با حروف بزرگ نوشته شده است، مثل« The President lives in Washington D.C. He likes that place **». علاوه بر این اگر کلمه یک اسم خاص یا یک قسمت از یک عبارت خاص باشد که همیشه با حروف بزرگ نوشته می‌شود، سیستم تشخیص حدود کلمه معمولاً نباید به عنوان نقطه شروع جمله بعدی آنرا به حساب آورد، بلکه باید به عنوان قسمتی از همان جمله برچسب‌گذاری شود، مثل«P.R. China». رفع ابهام یک اسم خاص از یک کلمه معمولی یک مسئله چالش‌انگیز است و مسئله ابهام حدود جمله را پیچیده‌تر می‌کند[۲۶].**

سیستم تشخیص حدود جمله[۱۶] اصلی از قوانینی که به طور دستی تولید شده به فرم عبارات منظم ساخته شده است تشکیل شده و یک لیست از اختصارات، کلمات عمومی، کلمات خاص و... به آن افزوده می‌شود.برای مثال سیستم المبیک[۱۷] [۲۷] بیش از ۱۰۰ قانون عبارت منظم را در فلکس[۱۸] برای این منظور تولید کرده است.

با این روش یک سیستم ممکن است فقط روی زبان یا پیکره‌ای که برای آن طراحی شده، خوب کار کند. با این حال ساخت و نگهداری یک سیستم بر پایه قانون[۱۹] قطعاً نیاز به کدنویسی دستی و دانش محیط کاربرد را نیاز دارد، که خیلی وقت‌گیر است. اشکال دیگر این نوع سیستم‌ها این است که بردن سیستم موجود به دامنه یا مجموعه-ای از زبان‌های دیگر، مشکل است. چنین تغییری برابر ساخت یک سیستم جدید بدون استفاده از کارهای قبلی است.

فعالیت‌های تحقیقاتی جدید در تشخیص حدود جمله روی به کارگیری تکنیک‌های

---

[15] abbreviation

[16] Sentence Boundary Detection(SBD)

[17] Olambic

[18] Flex

[19] Rule-Based System

یادگیری ماشین[۲۰] مثل درخت تصمیم‌گیری[۲۱]، شبکه عصبی[۲۲]، حداکثر آنتروپی[۲۳] و مدل مخفی مارکوف متمرکز شده است، که تشخیص حدود جمله به عنوان یک مسئله رده‌بندی[۲۴] استاندارد تلقی می‌شود. اصول کلی این سیستم‌ها عبارتند از: آموزش سیستم روی یک پیکره، با استفاده از استخراج خصیصه‌های متن محلی اطراف نشانه-گذاری جداکننده جمله و اطلاعات عمومی مانند لیستی از اختصارات و اسم خاص-ها، سپس جملات متنی واقعی با استفاده از این سیستم آموزش دیده تشخیص داده می‌شوند.

یک الگوریتم رفع‌ابهام محدوده جمله موفق باید مشخصات زیر را داشته باشد[۲۶]:

- روش باید قوی باشد و نیاز به گرامر دست‌ساز یا قوانین خاصی که بستگی زیاد به حروف بزرگ، فضاهای چندگانه بین جملات و... نداشته باشد. بنابراین روش باید به آسانی انواع متنی جدید و برخی زبان‌های جدید وفق یابد.

- روش باید به سرعت با یک مجموعه آموزش کوچک و بدون نیاز به حافظه بالاسری بزرگ آموزش ببیند.

- نتایج روش باید بسیار با دقت باشد و باید کارآیی کافی داشته باشد که پیش‌پردازش متن را به طور قابل ملاحظه‌ای کند نکند.

- روش باید توانایی تشخیص بدون ناظر در مواردی که رفع ابهام سخت است را به جای پیشنهاد دادن با استفاده از اطلاعات را داشته باشد.

چندین روش مختلف را در ادامه بحث می‌کنیم و مشکلات مختلف را در آنها بررسی می‌کنیم.

---

[20] Machine learning

[21] Decision Tree

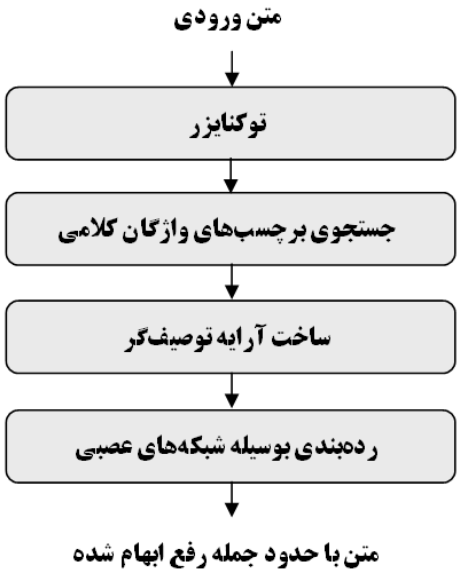
[22] Neural Network

[23] Maximum Entropy

[24] Classification

۱-۵- سیستم SATZ

در این قسمت ساختار سیستم قطعه‌بندی جمله وفقی[۲۵] SATZ [۲۶] را بررسی می‌کنیم. هدف در این سیستم نمایش متن مجاور یک علامت نشانه‌گذاری به صورت یک سری بردار احتمالی است، که این احتمال برای هر کلمه موجود در متن از احتمال اجزای کلامی اولیه به دست آمده است. بردارهای متن یا آرایه توصیف‌گر[۲۷] به عنوان ورودی یک شبکه عصبی آموزش‌دیده برای رفع ابهام محدوده جملات[۲۸] اعمال می‌شود. خروجی شبکه عصبی برای تعیین نقش علامت نشانه‌گذاری در جمله به کار می‌روند. معماری سیستم در شکل ۳-۱ نشان داده شده است.



شکل (۱-۱) معماری سیستم تعیین حدود جمله SATZ

[25] Adaptive Sentence Segmentation

[۲۶] **یک کلمه آلمانی به معنای جمله است**

[27] Descriptor Array

[28] Sentence Boundary Disambiguation

۱-۵-۲- نمایش متن

متن مجاور یک علامت نشانه‌گذاری به روشهای گوناگونی می‌تواند نمایش داده شود. ساده‌ترین و سرراست‌ترین راه برای استفاده کلمات قبل و بعد از علامت نشانه‌گذاری است، به عنوان مثال استفاده از سه کلمه هر طرف علامت نشانه‌گذاری:

At the plant. He had thought

برای هر کلمه باید احتمال پایان یا ابتدای جمله بودن را تعیین کرد. به هرحال این محاسبه برای هر کلمه در یک زبان بسیار وقت‌گیر است و حافظه زیادی را نیاز دارد و استفاده‌اش برای مراحل بعدی محتمل نیست. به عنوان یک راه چاره، می‌توان هر کلمه را با استفاده از یک برچسب اجزای کلام ساده تقریب زد. مفهوم فوق به وسیله یک سری اجزای کلام نمایش داده می‌شود:

Preposition article noun

Pronoun verb verb

نیاز به یک برچسب اجزای کلام ساده برای هر کلمه یک فرآیند چرخشی نیاز دارد، زیرا برای بیشتر برچسب‌گذارهای اجزای کلام نیاز است که از قبل محدوده جمله تعیین شده باشد، اما برای برچسب‌گذاری جملات قبل از برچسب‌گذاری اجزای کلام [۲۹] هیچ برچسب اجزای کلام اختصاص یافته‌ای برای الگوریتم تعیین حدود جمله در دسترس نیست.

برای جلوگیری از این فرآیند چرخشی و جلوگیری از نیاز برای برچسب اجزای کلام ساده برای هر کلمه متن، یک سری احتمالات اولیه‌ای که هر کلمه می‌تواند به عنوان اجزای کلام در متن قرار گیرد، نمایش داده می‌شود. بنابراین، هر کلمه در متن به وسیله یک سری احتمالات اجزای کلام و همچنین احتمال وقوع این کلمه در هر کدام از نقش‌های اجزای کلام نشان داده می‌شود، به عنوان مثال:

Preposition (1.0) article (1.0) noun (0.8)/verb (0.2)

Pronoun (1.0) verb (1.0) noun (0.1)/verb (0.9)

---

[29] Part-of-Speech Tagging

با این کار مشخص شد که «at» و «the» به احتمال یک به عنوان حرف تعریف و حرف اضافه به‌ترتیب واقع می‌شوند، «plant» به احتمال ۰.۸ به عنوان اسم و به احتمال ۰.۲ به عنوان فعل اتفاق می‌افتد.

این احتمالات بر پایه وقوع کلمات در یک پیکره برچسب‌گذاری شده به دست آمده است و وابسته به پیکره است. اطلاعات اجزای کلام اغلب برچسب‌گذارهای می‌شوند و به طور سرراست در دسترس است و نیاز به حافظه بالاسری وسیعی ندارد، به این دلیل در سیستم، تخمین متن به وسیله احتمالات اولیه اجزای کلام انتخاب شده است.

۱-۵-۳- فرهنگ لغات[۳۰]

یک جزء مهم در سیستم فرهنگ لغت است که شامل فراوانی داده‌های اجزای کلام، که احتمال آنها محاسبه شده است، می‌باشد. کلمات در فرهنگ لغت براساس یک سری برچسب‌های اجزای کلام و تکرارهای مرتبط جستجو می‌شوند. در این مرحله یک کلمه در فرهنگ لغت - اگر موجود باشد - پیدا می‌شود و احتمال اجزای کلام برگردانده می‌شود. برای کلمه انگلیسی «well» برای مثال ماژول جستجو برچسب-های زیر را بر می‌گرداند:

JJ/15 NN/18QL /68RB /634 UH/22VB /5

این مشخص می‌کند که کلمه «well» ۱۵ بار صفت، ۱۸ بار اسم، فقط ٦٨ بار کلمه توصیفی، ٦٣٤ بار قید و ٥ بار فعل رخ داده است.

[30] Dictionary

۱-۵-۴- ابداعی برای کلمات ناشناخته

اگر کلمه‌ای در فرهنگ لغت نباشد سیستم شامل یک سری ابداعات است که تلاش می‌کند معقولانه‌ترین اجزای کلام را به کلمه اختصاص دهد. یک خلاصه از این ابداعات در زیر آورده شده است:

توکِن‌های[31] ناشناخته شامل ۹-۰ عدد فرض می‌شود.

هر توکِنی که با یک نقطه پایانی، علامت سئوال یا تعجب شروع می‌شود، به عنوان پایان جمله در نظر گرفته می‌شود.

پایانی‌های ساختاری معمول تشخیص داده می‌شوند و نحو مناسب به کلمه کامل اختصاص می‌یابد.

کلمات شامل یک نشان اتصال «-» به یک مجموعه از برچسب‌ها و تعداد تکرارشان مبنی بر «کلمات ربطی ناشناخته » اختصاص می‌یابد.

کلمات شامل نقطه داخلی اختصار فرض می‌شود.

کلماتی که با حروف بزرگ نوشته شده‌اند همیشه اسم خاص نیستند حتی اگر در ابتدای جمله ظاهر نشوند. مثل «American» که اغلب به صورت وصفی به کار می‌رود، این کلمات که در فرهنگ لغات نمایش داده نشده‌اند، یک احتمال معین اسم خاص شدن (۰.۹ برای انگلیسی) به آن اختصاص می‌یابد. بعلاوه برای نمایش تعداد تکرار اجزای کلام در فرهنگ لغات، یک احتمال معین اسم خاص شدن به این کلمات اختصاص می‌یابد (۰.۵ برای انگلیسی). به عنوان آخرین دسته‌بندی، کلمات به یک سری برچسب ممکن با یک توزیع یکنواخت اختصاص می‌یابند.

این ابداعات به آسانی می‌توانند تغییر کنند و با نیازهای هر زبان خاص وفق پیدا کند، برای مثال احتمال کلمات با حروف بزرگ در انگلیسی بیشتر از آلمانی است.

---

[31] Token

۱-۵-۵- ساختار آرایه توصیف گر

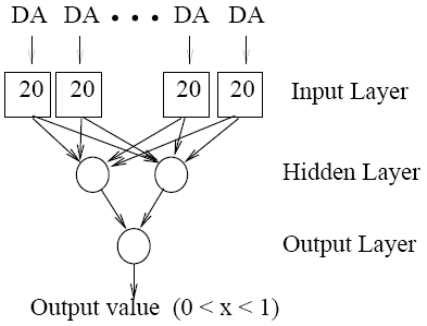
برای هر توکن در متن ورودی یک بردار از احتمالات نیاز است، بنابر توصیف عددی توکن بسازیم، این بردار، بردار توصیف‌گر شناخته می‌شود. فرهنگ لغات باید شامل ۷۰ یا ۸۰ اجزای کلام خیلی خاص باشد که ما اول نیاز داریم به دسته‌های عادی بیشتر نگاشت کنیم. برای مثال برچسب‌های فعل زمان حال، صفت مفعولی گذشته و فعل قیدی همه به دسته فعل نگاشت می‌شود. اجزای کلام به وسیله جستجوی ماژول به ۱۸دسته اصلی نگاشت شده در شکل ۳-۲ نشان داده می‌شود و تعداد تکرار هر دسته جمع می‌شوند. سپس تعداد تکرار هر دسته با تقسیم بر تعداد کل به احتمال تبدیل می‌شود. علاوه بر این ۱۸ احتمال آرایه توصیف‌گر همچنین شامل ۲ پرچم اضافی که نشان‌دهنده این است که، کلمه با یک حرف بزرگ شروع می‌شود یا نه، و آیا بعد از آن علامت نشانه‌گذاری وجود دارد یا نه، پس ۲۰ جزء محلی در هر آرایه توصیف‌گر وجود دارد.

جدول (۱-۱) اجزاء آرایه توصیف‌گر اختصاص داده شده به هر توکن

Noun	**اسم**	Verb	**فعل**
Article	**حرف تعریف**	Modifier	§§§§
Conjunction	**حرف ربط**	Pronoun	**ضمیر**
Preposition	**حرف اضافه**	Proper noun	**اسم خاص**
Number	**عدد**	Comma or semicolon	**، یا ؛**
Left parentheses	**پرانتز سمت چپ**	Right parentheses	**پرانتز سمت راست**
Non-punctuation character	**حرف غیر نشانه‌گذاری**	Possessive	**ضمیر ملکی**
Colon or dash	**: یا -**	Abbreviation	**مخفف**
Sentence ending punctuation	**علائم نشانه‌گذاری پایان جمله**	Other	**غیره**

۱-۵-۶- رده‌بندی با شبکه عصبی

آرایه توصیف‌گر، به عنوان ورودی یک شبکه عصبی پیش‌رو[۳۲] کاملاً متصل که در شکل ۳-۲ نشان داده شده است.



شکل (۱-۲) معماری شبکه عصبی (آرایه توصیف‌گر = DA)

◘ معماری شبکه عصبی

شبکه عصبی $k\times20$ **واحد ورودی را می‌پذیرد که** k **تعداد کلمات مجاور متن یک نمونه از یک علامت پایان جمله و ۲۰ تعداد المانهای موجود در آرایه توصیف‌گر است. لایه ورودی کاملاً متصل به لایه مخفی شامل** j **واحد مخفی با یک تابع فعالیت**

[32] Feed Forward Neural Network

فشرده سیگموئیدی[۳۳] است. لایه مخفی[۳۴] یک به یک واحد خروجی متصل می‌شود که نتیجه تابع را مشخص می‌کند. خروجی شبکه یک مقدار بین ۰ تا ۱ است، شدت ملاکی است که یک علامت نشانه‌گذاری که در متن واقع می‌شود، به راستی پایان جمله است.

دو آستانه[۳۵] با حساسیت قابل‌تنظیم تعریف‌شده t0 و t1 برای رفع ابهام دسته‌بندی نتایج به کار می‌رود. اگر خروجی از t0 کوچکتر باشد علامت نشانه‌گذاری محدوده جمله نیست و اگر بزرگتر یا مساوی t1 باشد محدوده جمله است. خروجی که بین دو حد آستانه قرار دارد نمی‌تواند به وسیله شبکه رفع‌ابهام شود، بنابراین آنها می‌توانند به طور خاص در پردازش‌های بعدی به کار گرفته شوند. وقتی t1 = t0 هیچ علامت نشانه‌گذاری با ابهامی باقی نمی‌ماند.

برای رفع‌ابهام یک علامت نشانه‌گذاری، یک پنجره k+1 توکنی و آرایه توصیف-گرشان از متن ورودی خوانده و نگهداری می‌شود. توکن‌های k/2 ابتدا و k/2، k توکن مورد نیاز جهت لایه ورودی شبکه را فراهم می‌کند، به شرطی که توکن میانی یک علامت نشانه‌گذاری پایان جمله باشد. نتایج خروجی برچسب مناسب را با توجه به آستانه t0 و t1 مشخص می‌کند.

◘ آموزش

داده‌های آموزش[۳۶] شامل دو قسمت است که همه محدوده‌های جملات برچسب-گذاری شده است. متن اول شامل ۲۵۰ تا ۵۰۰ مورد آزمون می‌شود که هر مورد آموزش یک علامت نشانه‌گذاری مبهم است. وزنهای شبکه عصبی با قسمت اول داده-های آموزش بر اساس الگوریتم انتشار برگشتی[۳۷] آموزش داده می‌شود. قسمت دوم

---

[33] Sigmoidal Squashing Activation Function

[34] Hidden Layer

[35] Threshold

[36] Training Data

[37] Back Propagation

برای وارسی‌اعتبار[۳۸] به کار می‌رود و شامل ۱۲۵ تا ۲۵۰ مورد آزمون جدا از مجموعه آموزش است. آموزش وزن‌ها روی قسمت دوم انجام نمی‌شود، بلکه قسمت وارسی-اعتبار برای افزایش تعمیم آموزش کاربرد دارد. اینطور که وقتی کل خطای آموزش روی قسمت وارسی‌اعتبار به حداقل برسد آموزش متوقف می‌شود. سپس آزمون روی متون وابسته به متون آموزش و وارسی‌اعتبار انجام می‌گیرد.

۱-۵-۷- نتایج آزمایش‌ها

سیستم برای متن انگلیسی مجله وال استریت[۳۹] آزمون شده است. متن آموزشی از ۵۷۳ نمونه آموزش و ۲۵۸ مورد بررسی اعتبار از همین نوشته‌ها است. سپس یک متن آزمایشی جدا شامل ۲۷۲۹۴ مورد آزمون با یک حد پائین ۷۵ درصد است. فرهنگ لغات و بدین ترتیب تعداد تکرار برای محاسبه آرایه توصیف‌گر به کار می‌رود.به منظور تعیین اندازه متن لازم است جملات بدقت در متن قسمت‌بندی شود. اندازه-های متفاوت متن در جدول ۳-۲ به دست آورده شده است.

[38] Cross Validation

[39] Wall Street Journal (WSJ)

جدول (۱-۲) نتایج مقایسه اندازه متن

اندازه متن	دورهای آموزش[40]	خطای آموزش[41]	خطای متقاطع[42]	خطای آزمون[43]	خطای آزمون (بر حسب٪)
٤	۱۷۳۱	۱.۵۲	۲.۳٦	۱٤۲٤	٪۵.۲۲
٦	۲۱۸	۰.۷۵	۲.۰۱	٤۰۹	٪۱.۵۰
۸	۸۳۱	۰.۰٤۳	۱.۸۸	۸۷۷	٪۳.۲۱

خطای آموزش، حداقل مربعات خطا[44] (نصف مجموع مربعات همه خطاها) برای همه ۵۷۳ جزء در مجموعه آموزش است. خطای متقاطع مقدار مساوی برای مجموعه وارسی‌اعتبار است. این اشکال نشان می‌دهد که داده‌های آموزش قبل از توقف چگونه آموزش می‌بینند. از این داده‌ها نتیجه شد که یک متن دارای شش توکن، سه توکن قبلی و سه توکن بعدی، بهترین نتیجه را می‌دهد.

برای تعیین اندازه لایه مخفی در شبکه عصبی که بالابردن دقت خروجی را فراهم می‌کند. انواع اندازه‌های لایه مخفی به کار رفت و نتیجه در جدول۳-۳ نشان داده شده که بهترین نتیجه در مورد ۲ لایه مخفی بود.

جدول (۱-۳) نتایج مقایسه تعداد واحدهای لایه مخفی

تعداد لایه مخفی[45]	دورهای آموزش	خطای آموزش	خطای متقاطع	خطای آزمون	خطای آزمون (بر حسب٪)
۱	٦۲۳	۱.۰۵	۱.٦۱	۷۲۱	٪۲.٦٤
۲	۲۱٦	۱.۰۸	۲.۱۸	٤۰۹	٪۱.۵۰
۳	۲۳۹	۰.۳۹	۲.۲۷	٤۳۵	٪۱.۵۹
٤	۳۵۰	۰.۲۷	۱.٤۲	۱۳٤۳	٪۱.۹۲

[40] Training Epochs
[41] Training Error
[42] Cross Error
[43] Testing Error
[44] Least Mean Square Error
[45] Hidden Layer

ساختار گفته شده کلاً ۴۰۹ خطا داشت که دقت ۹۸.۵ درصد را گزارش می‌کند. این خطا در دو دسته اصلی اتفاق می‌افتد:

۱- مثبت کاذب[۴۶] که علامت نشانه‌گذاری اشتباهاً برچسب جمله خورده است.

۲- منفی کاذب[۴۷] که محدوده جملات واقعی برچسب نخورده‌اند.

جدول ۳-۴ شامل خلاصه این خطاهاست.

جدول (۱-۴) نتایج آزمون ۲۷۲۹۴ مورد (دو واحد مخفی، سه توکن قبل و بعد و $t0 = t1 = 0.5$ **)**

۲۲٤ (٥٤.٨٪) مثبت کاذب
۱۸٥ (٤٥.۲٪) منفی کاذب
٤۰۹ کل خطا از ۲۷۲۹٤ نمونه

۳۷.۶ درصد مثبت‌های کاذب در یک اختصار نام یا عنوان معمولاً به خاطر اینکه کلمات بعد از نقطه در فرهنگ لغات با برچسبی دیگر وجود دارد.

۲۲.۵ درصد منفی‌های کاذب به علت اختصار در پایان جمله.

۱۱ درصد منفی یا مثبت کاذب به علت سری کاراکترهای شامل یک نقطه و علامت کوتیشن که می‌توانند در پایان جمله هم واقع شوند.

۹.۲ درصد منفی کاذب نتیجه شده از یک اختصار بعد از علامت کوتیشن، نسبت به دو نوع قبلی.

۹.۸ درصد منفی یا مثبت کاذب که از بریده‌گویی نتیجه می‌شود که در پایان جمله اتفاق می‌افتد.

۹.۹ درصد خطاهای مختلف شامل کاراکترهای خارجی (دش، علامت ستاره)، جملات غیرگرامری، غلط املایی، جملات درون پرانتز.

دو مورد اول نشان می‌دهد که تشخیص اختصار سخت است که برای مقابله با آن سعی شده با تقسیم اختصارات به دو دسته اختصارات عنوان مثل .Mr **و** .Dr **که هرگز**

[46] False Positive

[47] False Negative

پایان جمله نمی‌آید و بقیه آنها. این دسته‌بندی جدید به طور معنی داری زمان آموزش را افزایش می‌دهد و ۱۲ خطا از ۴۰۹ خطا (۲.۹٪) را حذف می‌کند.

خروجی شبکه عصبی برای تعیین تابع علامت نشانه‌گذاری بر اساس مقادیر مرتبط به دو حد آستانه با خروجی‌هایی که اتفاق می‌افتد به کار می‌رود، برای خروجی‌های بین این دو حد آستانه تابع هنوز مبهم است که در جدول ۳-۵ با عبارت «بدون برچسب[48]» نشان داده شده است. یک آزمایش سیستماتیک با آستانه حساسیت از مقادیر اولیه ۰.۵ حرکت داده می‌شوند. بین افزایش درصد خطا با تعدیل آستانه، همچنین کاهش درصد نمونه‌های درست برچسب‌گذاری شده و افزایش درصد نمونه‌های باقیمانده با ابهام، یک تقابل وجود دارد.

جدول (۱-۵) نتایج تغییر حدود آستانه (دو واحد مخفی، سه توکن قبل و بعد و $t0 = t1 = 0.5$)

آستانه پایین	آستانه بالا	مثبت کاذب	منفی کاذب	بدون برچسب	درست	بدون برچسب(٪)	خطای آزمون (بر حسب٪)
۰.۵	۰.۵	۲۰۹	۲۰۰	۰	۰	۰.۰	٪۱.۵۰
۰.٤	۰.٦	۱۷۳	۱۷٤	۱٤٥	۸۳	٪۰.۵۰	٪۱.۲۷
۰.۳	۰.۷	۱٤۰	۱٤۸	۳۲٦	۲۰۵	٪۱.۲۰	٪۱.۰٦
۰.۲	۰.۸	۱۱۱	۱۳۳	۵٤۱	۳۷٦	٪۱.۹۸	٪۰.۸۹
۰.۱	۰.۹	۷۹	۹٤	۱۰۲۱	۷۸۵	٪۳.۷٤	٪۰.٦۳

با تغییر اندازه فرهنگ لغت به کار رفته نتایج مختلفی ثبت شده است که در جدول ۳-۶ نشان داده شده است. این نشان می‌دهد که فرهنگ لغت بزرگتر سرعت آموزش سریعتر و دقت بیشتر را سبب می‌شود، اگرچه کارآیی فرهنگ لغت کوچکتر تقریباً برابر دقت قبلی است.

[48] Not Labaled

جدول (۱-۶) نتایج تغییر اندازه فرهنگ لغات (دو واحد مخفی، سه توکن قبل و بعد و $t0 = t1 = 0.5$ **)**

تعداد کلمات فرهنگ لغت	دورهای آموزش	خطای آموزش	خطای متقاطع	خطای آزمون	خطای آزمون (بر حسب٪)
۳۰۰۰۰	۲۱۸	۰.۷۵	۲.۰۱	٤۱۱	٪۱.۵۰
۵۰۰۰	۳۷۲	۰.۵	۱.۷۵	٤۸۳	٪۱.۷۵
۳۰۰۰	۱۰۵٦	۰.۰۵	۱.۳۰	۵۵۱	٪۲.۰۰

۱-۶- سیستم Bondec

سیستم Bondec **یک سیستم تشخیص حدود جمله است**[۱]**، که سه برنامه کاربردی مستقل دارد (حداکثر آنتروپی، مدل مخفی مارکوف، برپایه قانون). مدل حداکثر آنتروپی قسمت اصلی این سیستم است. نرخ خطای کمتر از ۲٪ با فقط ۸ خصیصه دودوئی را روی مجموعه وال استریت گزارش کرده است. فعالیت‌های تحقیقاتی جدید در تشخیص حدود جمله روی به کارگیری تکنیک‌های یادگیری ماشین مثل درخت تصمیم‌گیری، شبکه عصبی، حداکثر آنتروپی و مدل مخفی مارکوف متمرکز شده است، طوری‌که تشخیص حدود جمله به عنوان یک مسئله دسته‌بندی استاندارد تلقی می‌شود.**

۱-۶-۱- مقدمه

در میان بسیاری از متدهای یادگیری ماشین برای تعیین حدود جمله مثل نایو بیز[۴۹]، درخت تصمیم‌گیری، شبکه‌های عصبی، مدل مخفی مارکوف و حداکثر آنتروپی، مدل حداکثر آنتروپی به عنوان متد اصلی برای حل این مسئله در نظر گرفته شده است. انتخاب روش در این سیستم به این علت است که اولاً مدل حداکثر آنتروپی یک پایه ریاضی محکم دارد، دوماً خصیصه های مدل حداکثر آنتروپی می‌تواند از

[49] Naïve Bayse

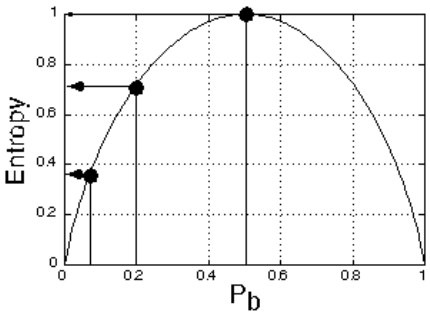
منابع ناهمگن باشد و به راحتی پیاده‌سازی می‌شود و سوماً مدل حداکثر آنتروپی نسبتاً جدید است و می‌شود از روی آن می‌توان اطلاعات جدید به دست آورد.

۱-۶-۲- آنتروپی[50]

آنتروپی یک اصطلاح فنی اصلی در زمینه تئوری اطلاعات[51] است [۲۸]. که برای تخمین مقداری که داده می‌تواند فشرده شود قبل از اینکه رو یک کانال ارتباطی ارسال شود به کار می‌رفت [۲۹]. آنتروپی H متوسط عدم قطعیت یک متغیر تصادفی منفرد x است:

$$H(p) = H(X) = \sum_{x \in X} p(x) \log_2 p(x) \qquad (۱-۱)$$

در معادله(۳-۱) $p(x)$ **تابع جمعی احتمال متغیر تصادفی** x **است و معادله فوق به ما می‌گوید که متوسط بیت‌هایی که نیاز داریم برای انتقال تمام اطلاعات موجود در** x **چقدر است. برای مثال اگر ما یک سکه را بیاندازیم و** x **تعداد شیر آمدن‌ها باشد،** x **یک متغیر تصادفی دودوئی خواهد بود. ما می‌توانیم فرض کنیم** $p(X=1) = p$ **و** $H(X)$ $p(X=0) = (1-p)$ **است:**



شکل (۱-۳) آنتروپی یک متغیر تصادفی دودوئی

[50] Entropy
[51] Information Theory

شکل ۳-۳ نشان می‌دهد که $H(x)\geq 0$، وقتی x یک مقدار ثابت دارد، چونکه هیچ اطلاعات ادغام شده‌ای در این متغییر وجود ندارد. این شکل همچنین نشان می‌دهد که $H(x)$ وقتی که $p=0.5$ است به نقطه حداکثر می‌رسد، که این بدان معنی است که x توزیع یکنواخت دارد. برای ذخیره پهنای باند کانال ارتباطی یک مدل x با آنتروپی کمتر را ترجیح می‌دهیم برای اینکه ما می‌توانیم بیت‌های کمتری را برای توصیف عدم قطعیت (اطلاعات) درون x استفاده کنیم.

به هر حال، در اینجا خواسته شده مدلی برای حداکثر آنتروپی ساخته شود. این بنظر می‌رسد که پایه قاعده کلی آنتروپی نقض شده است. دلیل اصلی برای انجام اینکار این است که باید کمترین بایاس ممکن، وقتی قطعیت نمی‌تواند از گواههای تجربی تعیین شود، حفظ شود.

۱-۶-۳- مدل حداکثر آنتروپی

اگر ما یک پاراگراف را به عنوان رشته‌ای از توکن‌ها در نظر بگیریم می‌توانیم مسئله تشخیص حدود جمله را یک فرآیند تصادفی در نظر بگیریم که حدود جمله را در پاراگراف تعیین می‌کند. چنین فرآیندی خروجی y را تولید خواهد کرد، که دامنه y همه موقعیت‌های حدود جمله ممکن در رشته است. یک متغییر تصادفی Y روی دامنه تعریف می‌شود، و y یک مقدار خاص Y است. در این فرآیند تصادفی مقدار Y روی این دامنه ممکن است به وسیله برخی از اطلاعات ضمنی x تحت تاثیر قرار بگیرد، که دامنه x همه ترکیبات متنی ممکن در رشته است. می‌توان فرض کرد که x یک مجموعه نامحدود است. شبیه y ما همچنین یک متغییر تصادفی X از این دامنه نامحدود تعریف می‌کنیم. برای حل مسئله تشخیص حدود جمله، می‌توان یک مدل آماری بسازیم، برای اینکه این فرآیند تصادفی را به طور درست شبیه‌سازی کنیم. یک چنین مدلی یک مدل احتمال شرطی است، یک رشته متنی x را می‌دهد و حدود y را تعیین می‌کند $p(y|x)$.

مانند دیگر متدهای یادگیری ماشین، یک فایل آموزش نیاز داریم. کار ساخت یک مدل است که مقدار حداکثر درست‌نمایی[۵۲] را با فایل آموزش دارد. در مرحله اول فرآیند تصادفی را شبیه‌سازی می‌شود. یک تعداد زیادی از نمونه‌ها $(y1،x1)$، $(y2, x2)$، **...** $(y_N، x_N)$ **از مجموعه آموزش استخراج شده‌اند. یک توزیع تجربی الحاقی روی** x **و** y **از این نمونه‌ها تعریف شده است:**

$$\tilde{p}(x,y) = \frac{1}{N} \times number\ of\ \ (x,y) \qquad (۲-۱)$$

گذشته از این فایل آموزش را نمونه‌برداری کرده، همچنین می‌توانیم برخی از خصیصه‌ها را از نمونه آموزش که کاملاً مفید است برای مسئله دسته‌بندی استفاده کرد. مثلاً، مشاهده شده است که که اگر کلمه بعد از یک نقطه با حروف بزرگ نوشته شود با یک احتمال زیاد حدود جمله است. می‌توان یک تابع دودوئی f **را برای این خصیصه‌ها مطرح کنیم:**

$$f(x,y) = \begin{cases} 1\ if\ x\ is\ a\ \ capitalized\ word\ following \\ \ \ a\ period\ y, the\ period\ \ is\ a\ boundary. \\ 0\ otherwise \end{cases} \qquad (۳-۱)$$

ممکن است خصیصه‌های زیادی را برای مسئله تعیین حدود جمله معرفی کنیم، هریک از آنها یک محدودیتی روی مدل قرار می‌دهد. امید ریاضی خصیصه‌ها از نمونه آموزش باید شبیه امید ریاضی این خصیصه‌ها در مدل باشد.

$$\sum_{x,y} \tilde{p}(x)p(y|x)f(x,y) = \sum_{x,y} \tilde{p}(x,y)f(x,y) \qquad (٤-۱)$$

که $\tilde{p}(x)$ **محدودیت توزیع تجربی** x **در نمونه آموزش است.**

به هر حال هنوز مدلهای شرطی زیادی وجود دارد که می‌تواند محدودیت‌های مطابق معادله بالا را ارضاء کند. برای پیدا کردن بهترین برای مجموعه آموزش باید

[52] Maximum likelihood

مدلی با حداکثر آنتروپی به دست آوریم که این به این معنی است کـه یـک توزیـع نرمال بیشتر روی خصیصه‌ها ناشناخته نیاز داریم. از اینرو مدل بهتر p* **برای تشـخیص حدود جمله باید شرایط آنتروپی روی** p(x|y) **را حداکثر کند.**

$$H(p) = -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x) \qquad (۵-۱)$$

$$p^* = \arg\max_{p} H(p) \qquad (۶-۱)$$

و با توجه به محدودیت‌های همزمان زیر:

$p(y|x) \geq 0$.For all x ، y.

$\sum_y p(y|x) = 1$. This and the previous condition guarantee that p(y|x) is a conditional probability distribution.

$$\sum_{x,y} \tilde{p}(x)p(y|x)f_i(x,y) = \sum_{x,y} \tilde{p}(x,y)f_i(x,y)$$
$$i \in \{1,2...n\}$$

برای همه خصیصه‌های انتخاب شده (محدودیت‌ها).

این یک مسئله بهینه‌سازی اجباری عمومی است و ما می‌توانیم متد ضرایب لاگرانـژ[۵۳] را برای حل آن استفاده می‌کنیم. نتیجه نهایی این مسئله یک مدل خطی نمایی است.

$$p^*(y|x) = Z(x)\exp\left(\sum_i \lambda_i f_i(x,y)\right) \qquad (۷-۱)$$

که Z(x) **فاکتور نرمال‌سازی است که داده شده به وسیله:**

$$Z(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right) \qquad (۸-۱)$$

[53] Lagrange multiplier

که $\lambda_i \; i \in \{1,2...n\}$ ضرایب لاگرانژ مربوط به محدودیت‌های fi است و آن همچنین معیار وزن (اهمیت) خصیصه fi است. ثابت شده است که مدل از معادله (۲-۷) درست‌نمایی توزیع تجربی الحاقی $\tilde{p}(x,y)$ را حداکثر می‌کند:

$$L_{\tilde{p}}(\Lambda) = \sum_{x,y} \tilde{p}(x,y) \log p_{\Lambda}(y \mid x) \qquad (۹-۱)$$

که Λ بردار وزن $\{\lambda_1, \lambda_2 .. \lambda_n\}$ است.

۱-۶-۴- سنجش چرخشی تعمیم‌یافته[54]

برای معادله (۳-۷) هیچ مدل تحلیلی، برای به دست آوردن مقدار λ_i در این توزیع خطی لگاریتمی وجود ندارد. بنابراین سنجش چرخشی تعمیم‌یافته را به عنوان روش عددی برای به دست آوردن بردار Λ^* است. این فرآیند چرخشی به توزیع p* همگرا خواهد شد [۲۸].

۱-۶-۵- معماری سیستم

- **مجموعه آموزش و آزمون**

فایل آزمون و آموزش از فایل وال استریت حاشیه‌نویسی شده توسط دکتر پالمر[55] استفاده شده است، برای این پروژه سه فایل train.dat، test.dat ، heldout.dat ساخته شده است. از این فایل داده‌های خام train.dat به منظور آموزش مدل مخفی مارکوف و حداکثر آنتروپی استفاده شده است. ۲۱۰۲۶ جمله در مجموعه وجود

[54] Generalized Iterative Scaling (GIS)

[55] David Palmer

دارد. در میان این جمله‌ها ۲۰۰۲۸ در حدود ۹۵.۲۵٪ آنها توسط نقطه مرزیابی شده-اند، ۳.۴۷٪ از جملات یعنی ۷۲۷ جمله با یک علامت نقل‌قول پایان می‌پذیرد و ۱۴۶ جمله یعنی ۰.۶۹٪ با علامت سئوال پایان می‌پذیرد. مجموعه heldout، ۹۷۲۱ جمله دارد که برای وارسی‌اعتبار و تنظیم کارآیی استفاده شده است، در صورتیکه مجموعه آزمون حدود ۹۷۵۸ جمله دارد که فقط برای اندازه‌گیری کارآیی نهایی در دسترس است. توزیع علائم نشانه‌گذاری در دو فایل شبیه به فایل آموزش است. این ضمانت می‌کند که مدل آموزش مجموعه آزمون را تحت تاثیر خود قرار نمی‌دهد. در این سه مجموعه ۲۳۶ جمله هیچ علامت نشانه‌گذاری محدوده‌ای ندارند. معمولاً این جملات به فرم یک آدرس یا یک عنوان است.

- **ابزار محک[۵۶]**

برای مقایسه کارآیی روش یادگیری ماشین گفته شده با سیستم برپایه قانون یک ابزار ساخته شده و آنرا به عنوان یک معیار در این سیستم تلقی شده است. این برنامه کاربردی حدود جملات را فقط بر اساس قوانین گرامری به دست آورد. مطابق با توزیع واقعی در مجموعه آزمون سه معیار در این برنامه کاربردی تشخیص حدود کلمه ساده پیاده‌سازی شده است:

اگر کلمه بعد از علامت سئوال با حروف بزرگ نوشته شود این علامت سئوال حدود جمله است.

اگر توکن قبل از دابل‌کوتیشن یک نقطه یا یک علامت سئوال است، پس گیومه محدوده جمله است.

اگر کلمه بعد از نقطه با حروف بزرگ نوشته شده و کلمه قبل از این نقطه اینطور نیست، این نقطه محدوده جمله است.

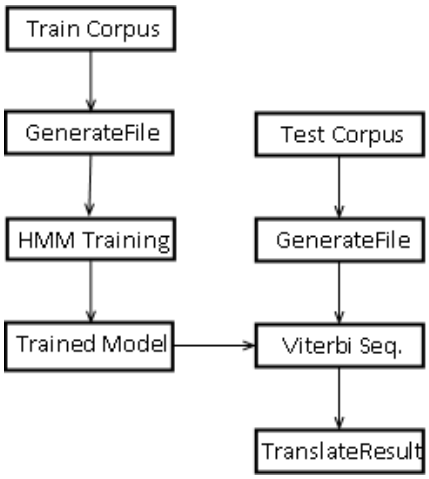
---

[56] Benchmark Tool

با این قوانین نحوی ساده سیستم موارد آزمون را با یک صحت پیشگویی ۹۹.۵۶٪ و فراخوانی ۷۶.۹۵٪ انجام می‌دهد. مقدار متوسط کارآیی - معیار F - در حدود ۸۶.۸۱٪ است.

- **مدل مخفی مارکوف**

در این سیستم یک مدل مخفی مارکوف برای تشخیص حدود جمله ساخته می‌شود. می‌خواهیم این مدل را با حداکثر آنتروپی مقایسه کنیم که قسمت اصلی معماری این مدل است. به طور مستقیم بسته edu.stanford.nlp.ie.hmm **استفاده شده است.**



شکل (۱-۴) فلوچارت برای اعمال مدل مخفی مارکوف در Bondec

شکل ۳-۴ چگونگی ارتباط موجود بسته مدل مخفی مارکوف را با مسئله تشخیص حدود جمله نشان می‌دهد. سیستم پیش‌پردازش[۵۷] شامل کلاس Generatefile **است که تابع جملات موجود در فایل آموزش/آزمون (شکل ۳-۵) را به فرمت سند مور نیاز برای بسته[۵۸] مدل مخفی مارکوف (شکل ۳-۶) تبدیل می‌کند.**

[57] Preprocessing
[58] Package

<s> BEVERLY HILLS hardly seems the place for the little guy to get an even break, but some small-business owners recently struck a blow for equality in this ghetto of glitz. </s><s> The arena for this victory was the Beverly Hills Chamber of Commerce. </s>

شکل (۱-۵) فرمت جمله نسخه اصلی در فایل‌های آزمون/ آموزش

BEVERLY HILLS hardly seems the place for the little guy to get an even break, but some small business owners recently struck a blow for equality in this ghetto of glitz <boundary>. </boundary>The arena for this victory was the Beverly Hills Chamber of Commerce <boundary>. </boundary>For ENDOFDOC

شکل (۱-۶) یک قسمت سند برای بسته مدل مخفی مارکوف

بنابراین بسته می‌تواند یک مدل مخفی مارکوف را آموزش دهد. این مدل بعداً برای پیش‌بینی محدوده جملات در فایل آزمون فرمت‌بندی شده مورد استفاده قرار خواهد گرفت. خروجی این مدل به وسیله کلاس Translate Result **به منظور ارزیابی، ترجمه می‌شود. طوریکه در شکل ۳-۶ نشان داده شده است از برچسب «<boundary>» برای نشان دادن فیلد هدفی که به مدل مخفی مارکوف آموزش داده می‌شود استفاده می‌شود.**

چون فقط یک هدف در تشخیص حدود جمله وجود دارد و این متن هدف فقط یک توکن دارد، می‌توان یک مدل مخفی مارکوف را با فرمت ساده از پیش تعیین کرد و اجازه داد که بسته مدل مخفی مارکوف به طور موثری پارامترهای این مدل را استنتاج کند.

◘ حداکثر آنتروپی

یک مدل حداکثر آنتروپی برای تشخیص حدودجمله پیاده‌سازی شده است، که وابسته به دامنه خاص نیست به جز در برخی فرضیات پایه در مورد زبان انگلیسی مثلاً علائم نشانه‌گذاری زیر می‌تواند به عنوان نقطه (وقفه کامل) جمله به کار رود: «؟، !،.». برخی فرضیات دیگر برای آماده کردن خصیصه‌هایی که قابلیت تشخیص بهتری ارائه می‌دهد نیز فراهم شده است. به هر حال مطابق چارچوبه کاری حداکثر آنتروپی، آنها برای مدل کردن و تغییر دادن و انتخاب و وزن‌دهی اتوماتیک ساده هستند.

چنانچه قبلاً اشاره شده است، در مدلسازی حداکثر آنتروپی مقادیر خصیصه‌ها به وسیله برخی متن و دسته‌بندی‌های ورودی تعیین می‌شود. در تئوری کل سند یا مجموعه داده می‌تواند متن باشد. عملاً محدودیت‌های پیچیدگی محاسباتی و پراکندگی داده‌های متون به توکن‌های همسایه محدود شده‌اند. در این پیاده‌سازی، متن شامل یک کلمه قبلی و دو کلمه بعدی محدوده جمله مفروض است، که آنرا کاندید می‌نامیم.

کلمات بلافاصله قبل و بعد از کاندید به ترتیب چپ و راست نامیده می‌شوند و کلمه دوم سمت راست کاندید دنباله[۵۹] نامیده می‌شود. متن در صورت نیاز با خصیصه‌های جدید قابل بسط است.

از خصیصه‌های نحوی و لغوی استفاده شده است، که پایداری بیشتری دارد و مانند مدل مدل چند-تایی تنکی[۶۰] داده‌ها را اجازه نمی‌دهد.

همچنین خصیصه‌های لغوی با مقادیر کم با برخی کلمات خروجی منابع خاص مانند لیست ۸۸۷۹۹ نام خانوادگی معمولی، ۵۴۹۴ نام کوچک معمولی از سرشماری سال ۹۰ آمریکا و ۲۰ عنوان افتخاری معمول اضافه شده است. به طور اتوماتیک اختصارات از

---

[59] Tail

[60] sparseness

مجموعه آزمون و آموزش از روش سند وسط‌چین شده [۳۰] استخراج شده است، تمام کمک‌های بالا مفید هستند اما اساسی نیستند.

از ساخت تعداد کمی خصیصه‌های دستی و مدل حداکثر آنتروپی بهینه با استفاده از الگوریتم سنجش چرخشی تعمیم‌یافته کار شروع شده است. سپس مدل روی مجموعه داده heldout **ارزیابی شده و قوانین بیشتری برای رسیدگی به موارد دسته-بندی نادرست اضافه شده است. تعداد کمی از خصیصه‌های معلوم برنامه‌کاربردی معمولاً خیلی سریع همگرا می‌شود. از آنجایی که وابستگی‌های داخلی در یک مسئله مدلسازی حداکثر آنتروپی انتظار نمی‌رود افزودن خصیصه‌های جدید بسیار ساده است و سیستم سریعاً به معیار** F **(بر اساس تشخیص حدود جمله) می‌رسد، بهتر از ۹۸٪ با ۱۳ خصیصه.**

در زیر لیست ۸ خصیصه انتخاب شده به وسیله الگوریتم یادگیری قیاسی[۶۱] [۳۱] آورده شده است و بر اساس انتخاب الگوریتم مرتب شده است:

1.Left is a lowercased word ,sentence boundary (SB);

2.Right is a lowercased word ,NSB ;

3.Right is '.', '?', ',', «,) , }, -, NSB;

4.Left is an honorific ,Candidate is '.', NSB;

5.Candidate is « ,an odd quote ,NSB ;

6.Left is an initi al, Candidate is '.', and not a sentence boundary (NSB);

7.Left is ',', Candidate is «,NSB ;

8.Candidate is '.', Right is ', Tail is 's', NSB;

[61] Induction Learning Algorithm

۱-۶-۶- کارآیی سیستم

سه متد به کار رفته استاندارد درستی[۶۲]، فراخوانی[۶۳]، معیار F و نرخ خطا[۶۴] کارآیی را ارزیابی کرده‌اند که در جدول ۳-۷ نمایش داده شده است. درستی، به عنوان جمع کل محدوده جملات استخراج شده درست روی کل تعداد محدوده‌های جمله استخراج شده به وسیله سیستم تعریف می‌شود. فراخوانی، به عنوان تعداد کل محدوده جمله استخراج شده درست روی تعداد کل جمله موجود در مجموعه تعریف می‌شود. معیار F، به عنوان اعداد متقابل درستی و فراخوانی (دو برابر مجموع معکوس دقت و معکوس فراخوانی)، که یک نشان‌دهنده تک رقمی خوب سیستم است، تعریف می‌شود. نرخ خطا نیز یک معیار عمومی است که به عنوان مجموع منفی‌های کاذب و مثبت‌های کاذب روی همه حدود جملات ممکن تعریف می‌شود. چیزی که قابل توجه است این است که نرخ خطا ممکن نیست عیناً مثل تعریف محدوده جمله تعریف شود و برخی اوقات بین نویسندگان مختلف متفاوت است. در جدول زیر تعریف نرخ خطا برای سیستم برپایه قوانین و سیستم حداکثر آنتروپی یکسان است شامل همه موارد «.، ؟، ! » و در مجموعه آزمون نقاط جزئی درون اعداد برحسب درصد، مقادیر پولی و یا اعداد حقیقی توسط یک توکنایزر یکسان جایگزین شده‌اند. بسته مدل مخفی مارکوف استفاده شده توکنایزر خودش را دارد که همه دابل کوتیشن‌ها را حذف می‌کند و بنابراین نرخ خطای تعریف شده کمی فرق می‌کند.

جدول (۱-۷) مقایسه کارآیی سه متد

---

[62] precision
[63] Recall
[64] Error Rate

Method	Precision	Recall	F1	Error Rate
RuleBased	99.56%	76.95%	86.81%	16.25%
HMM	91.43%	94.46%	92.92%	10.00%
MaxEnt	99.16%	97.62%	98.38%	1.99%

در جدول فوق می‌توانیم ببینم که مدل حداکثر آنتروپی به بهترین نحو کار می‌کند و این تعجب‌انگیز نیست زیرا:

سیستم بر پایه قانون می‌تواند موارد صریح را با دقت بالا مشخص کند. به هر حال برای بررسی موارد غیرمعمول بسیار کمتر، تعداد قوانین افزایش می‌یابد و متعادل کردن قوانین مخالف مشکل است، بنابراین ما به تعداد زیادی از قوانین اهمیت نمی‌دهیم و بنابراین فراخوانی افت می‌کند.

مدل مخفی مارکوف اصلاً یک مدل bigram **تقویت شده، با فرضیات مستقل قوی بین انتقال[۶۵] و انتشار[۶۶] است، حتی اگر اهداف بیشتر مراحل داخلی، مراحل پس‌زمینه و مدلسازی بهتر متن تاحدی کمک کنند، مدل مخفی مارکوف نمی‌تواند براحتی برخی خصیصه‌های پیچیده موثر زیاد را در تشخیص حدود متن مدل کند. خصیصه‌های لغوی ساده را با استفاده از تجزیه خصیصه مدل می‌کند.**

حداکثر آنتروپی برای یکپارچه کردن خصیصه‌های پیچیده مختلف برای منابع دانش ناهمگن یک چارچوب کاری خوب است. خصیصه‌های لغوی، نحوی و bigram**ها می‌توانند به طور طبیعی به عنوان خصیصه در همان مدل، مدل شوند. برای مثال رشته «** his compromise bill. " A committee staffer **» خیلی مبهم است، نمی‌توان از متن محلی بالا متوجه شد که نقطه یا دابل‌کوتیشن باید محدوده جمله باشد. به هر حال حداکثر آنتروپی تعادل دابل‌کوتیشن را به عنوان یک خصیصه اضافه می‌کند، که خیلی موثراست اما نمی‌تواند فقط از متن محلی به‌دست آید.**

---

[65] Transition

[66] Emission

مدلهای حداکثر آنتروپی فقط نمی‌توانند خصیصه‌های پیچیده مختلف را براحتی ثبت کنند، بلکه خصیصه‌های دارای اشتراک را خیلی‌خوب به کار می‌برند. در این موارد، یک خصیصه پیچیده از ترکیبات منطقی خصیصه‌های ساده در مدل به جای یک تعدادی از خصیصه‌های ساده اضافه شود. خصیصه‌های ساده می‌توانند همزمان با خصیصه‌های پیچیده، برای جبران اطلاعاتی که از خصیصه پیچیده برنمی‌آید، وجود داشته باشند. و این نوع از مدلسازی نه تنها دقت پیش‌بینی بهتری را تولید می‌کند بلکه به الگوریتم انتخاب خصیصه مثل یادگیری قیاسی کمک می‌کند. یادگیری قیاسی یک الگوریتم حریصانه است و فقط خصیصه‌های یک به یک را جمع می‌کند. بنابراین یک تعداد خصیصه ساده که فقط باهم کار می‌کنند ممکن است به وسیله این الگوریتم از دست برود، زیرا یکی از آنها به تنهایی اطلاعات با ارزش بیشتری برای شرح داده-های آموزش ارائه می‌دهد.

مسئله بالقوه دیگر نیاز به هموارسازی[۶۷] است. برخی خصیصه‌های کمیاب اما مفید، به طور قابل اعتمادی در یک مجموعه آموزش خاص تخمین زده می‌شوند. در چنین مواردی، سعی شده تا موارد شبیه درون یک خصیصه منفرد که موارد بیشتری در مجموعه آموزش دارد دسته‌بندی شود (مثلاً به جای داشتن یک خصیصه جدا برای مدل کردن یک نقطه که غیرمحتمل است حدود جمله باشد اگر سمت راست آن '}' باشد، خصیصه به انضمام علائم نشانه‌گذاری دیگر طوریکه پیشتر آورده شده مدل می-شود.).

سرانجام، ساخت مدل می‌تواند از لحاظ محاسباتی بسیار هزینه‌بر باشد. با استفاده از سنجش چرخشی تعمیم‌یافته آموزش یک مدل حداکثر آنتروپی با یک مجموعه ۱۷ خصیصه‌ای روی مجموعه آموزش گفته شده در حدود ۱۰ دقیقه روی یک واحد پردازنده مرکزی تک ۲۰۰۰ سان‌بلید[۶۸] طول می‌کشد. حدود ۹۰ دقیقه برای الگوریتم

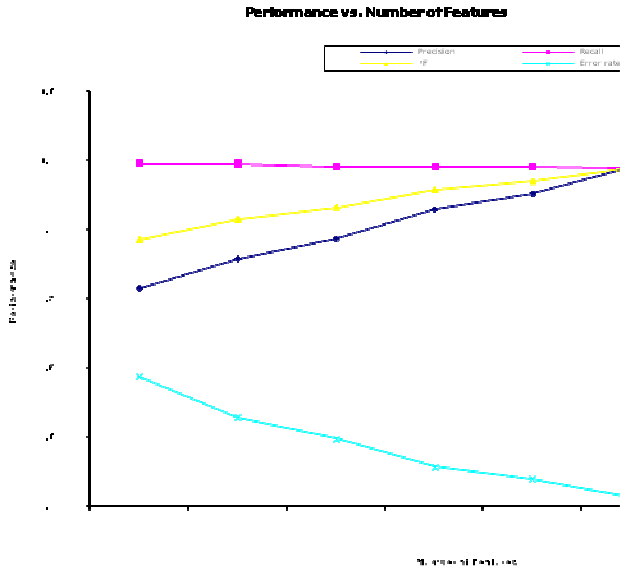
---

[67] smoothing

[68] Sun Blade

یادگیری قیاسی، برای آموزش یک مدل حداکثر آنتروپی و انتخاب ۸ خصیصه بهتر از ۱۷ خصیصه، طول می‌کشد.

تشخیص حدود جمله یک مسئله خوب تعریف شده است و با یک تعداد خیلی محدود از خصیصه‌ها حداکثر آنتروپی به یک کارآیی قابل قبول می‌رسد. شکل ۳-۷ و جدول ۳-۸ بهبود کارآیی با رشد تعداد خصیصه‌های انتخاب شده به وسیله الگوریتم آموزش قیاسی را نشان می‌دهد، ما می‌توانیم ببینیم که مدل حداکثر آنتروپی به سرعت با ۶ خصیصه به کارآیی خوبی می‌رسد.



شکل (۱-۷) کارآیی حداکثر آنتروپی با تعداد خصیصه‌ها روی داده‌های Held out

جدول (۱-۸) کارآیی حداکثر آنتروپی با تعداد خصیصه‌ها روی داده‌های Held out

Number of features	1	2	3	4	5	6	7	8
Precision	63.1%	71.7%	77.4%	86.0%	90.4%	98.2%	98.8%	99.4%
Recall	99.2%	99.0%	98.0%	98.0%	98.0%	97.8%	97.8%	97.8%
F1	77.2%	83.2%	86.5%	91.6%	94.1%	98.0%	98.3%	98.6%
Error rate	37.4%	25.5%	19.5%	11.4%	7.9%	2.6%	2.2%	1.8%

نه تنها مدلهای حداکثر آنتروپی می‌توانند خصیصه‌های پیچیده مختلف را براحتی ثبت کنند بلکه سنجش چرخشی تعمیم یافته با خصیصه‌های دارای اشتراک خیلی-

خوب کار می‌کند، این یک مزیت است، آنچه بسیاری از الگوریتم‌های دیگر مثل نایوبیز و مدل مخفی مارکوف ندارند.
آموزش مدلهای حداکثر آنتروپی می‌تواند از لحاظ محاسباتی گران باشد، بعلت اینکه متدهای عددی چرخشی - سنجش چرخشی تعمیم یافته - ممکن است همگرایی کندی داشته باشند. بنابراین مدلسازی و انتخاب خصیصه‌های بالقوه علاوه بر خوبی، مهم هم است. برای مسئله تشخیص حدود جمله نشان داده شده که مدلهای حداکثر آنتروپی با استفاده از خصیصه‌های کم خوب مدل شده به کارآیی قابل قبولی می‌رسند.
خصیصه‌های اطلاعات اختصار توسط این الگوریتم انتخاب نشده است که به این معنی است که با وجود خصیصه‌های موجود به کارکرد مدل کمکی نمی‌کند. بنظر می‌رسد که ترکیب شدن با دانش، برای مسائلی چون کلمه نوشته شده با حروف بزرگ اسم خاص یا کلمه عادی است، مفید باشد.

۱-۷- نتیجه‌گیری

با توجه به مطالعات انجام شده روی زبان‌های غیرفارسی چون انگلیسی، تحقیقات انجام شده در زمینه تعیین حدود جمله به رفع ابهام علائم نشانه‌گذاری خلاصه شده است و با استفاده از روشهای مختلفی که خلاصه‌ای از آن در بالا آورده شد به تشخیص حدود جمله پرداخته‌اند.
در زبان فارسی نیز فقط در پروژه شیراز [۸] بصورت بسیار جزئی به این موضوع پرداخته شده است. با توجه به اینکه هیچ کار جامعی روی این مسئله مهم انجام نشده است و این مسئله از جمله مسائل پیش‌پردازش مهم در کارهای پردازش زبان فارسی است اقدام به بررسی روشهای مختلف و ممکن برای اینکار پرداختیم، که این روشها در فصل‌های آینده مورد بررسی قرار گرفته است.

روش‌هاي پيشنهادي

۱-۸- مقدمه

جهت تعیین حدود جمله در متون زبان فارسی می‌توان روشهایی که در فصل قبل جهت تعیین حدود جمله با استفاده از رفع ابهام علائم نشانه‌گذاری در زبان انگلیسی معرفی شد را به کار برد، اما این روش در زبان فارسی دقت مطلوب را ارائه نمی‌کند، به خاطر اینکه همه متون علائم نشانه‌گذاری دقیقی ندارند، همچنین استاندارد دقیقی برای نشانه‌گذاری وجود نداشته و از همه مهم‌تر اینکه بسیاری از جملات فارسی به علائم نشانه‌گذاری ختم نمی‌شوند.

به خاطر اینکه تعیین حدود جمله از مراحل پیش‌پردازش کارهای پردازش زبان طبیعی و متن‌کاوی است، دقت آن تاثییر زیادی روی دقت مراحل بعدی دارد، بنابراین نیاز به روشی داریم که با توجه به ساختار زبان فارسی ایجاد شده باشد تا بتواند دقت مطلوب مورد نظر را ارائه دهد.

در این فصل به بررسی روش‌های پیشنهادی جهت تشخیص حدود جملات در پیکره-های متنی زبان فارسی می‌پردازیم.

۱-۹- تعیین حدود جمله با استفاده از تشخیص فعل

فارسی یک زبان هندواروپایی است که اصولاً در ایران، تاجیکستان و بخشی از افغانستان با آن صحبت می‌کنند. الفبای فارسی حاوی ۳۲ حرف است و از راست به چپ نوشته می‌شود. برخی دیگر زبان‌ها مثل عربی، کردی و اردو از خط فارسی استفاده می‌کنند اما خصوصیات خاص خود را دارند. فارسی نیز خصوصیات خاص خود را دارد مثل اینکه فارسی در نوشتار از مصوت‌ها استفاده نمی‌کند - مگر در برخی موارد خاص- و حالت چندریختی در نوشتار دارد[۳۲].

ریخت‌شناسی[۶۹] فارسی یک سیستم الحاقی حاوی تعداد زیادی پسوند و تعداد کمی پیشوند است. سیستم صرف فعل در فارسی کاملاً منظم است و به وسیله ترکیب پیشوند، ریشه، صرف فعل و افعال کمکی ساخته می‌شود. فارسی زبانی است که جملات آن دارای ساختار مرتبی با ترتیب فاعل - مفعول - فعل[۷۰] است هر چند استثناءهای زیادی دارد. فعل به واسطه زمان و وجه مشخص می‌شود و از لحاظ شخص و تعداد با فاعل مطابقت می‌کند.

با توجه به مشخصات ارائه شده زبان فارسی بر آن شدیم تا به عنوان اولین روش انتخابی برای تعیین حدود جمله از مشخصات ساختاری زبان، یعنی ساختار مرتب فاعل - مفعول - فعل جملات فارسی، استفاده کنیم و با تشخیص فعل در جمله کران جمله را مشخص کنیم. هرچند استثنائات زیادی وجود دارد اما باز هم تعداد زیادی از جملات فارسی از این قاعده پیروی می‌کنند. در پیکره ایجاد شده جهت اعتبارسنجی روش بررسی شد در حدود بیش از ۸۵ درصد از جملات از این ساختار پیروی می‌کنند، بنابراین می‌توان از این روش استفاده نمود.

۱-۹-۱- تشریح کامل روش

پس با توجه به مطالب ارائه شده برای این رهیافت نیاز به روشی برای تشخیص فعل در جمله است تا بتوان آنرا به عنوان پایان جمله و شروع جمله بعد محسوب کرد. برای تشخیص فعل از ساختار فعل در زبان فارسی استفاده شده است که به طور کامل در زیر تشریح شده است:

- **روش تشخیص فعل**

---

[69] Morphology

[70] Subject-Object-Verb(SOV)

ساختار نوشتاری زبان فارسی کاملاً مبهم است و مشکلات خاصی را در تجزیه اتوماتیک متن دارد. تشخیص افعال در متن فارسی نیز به علت ابهامات خاص خود پیچیده است[۳۳].

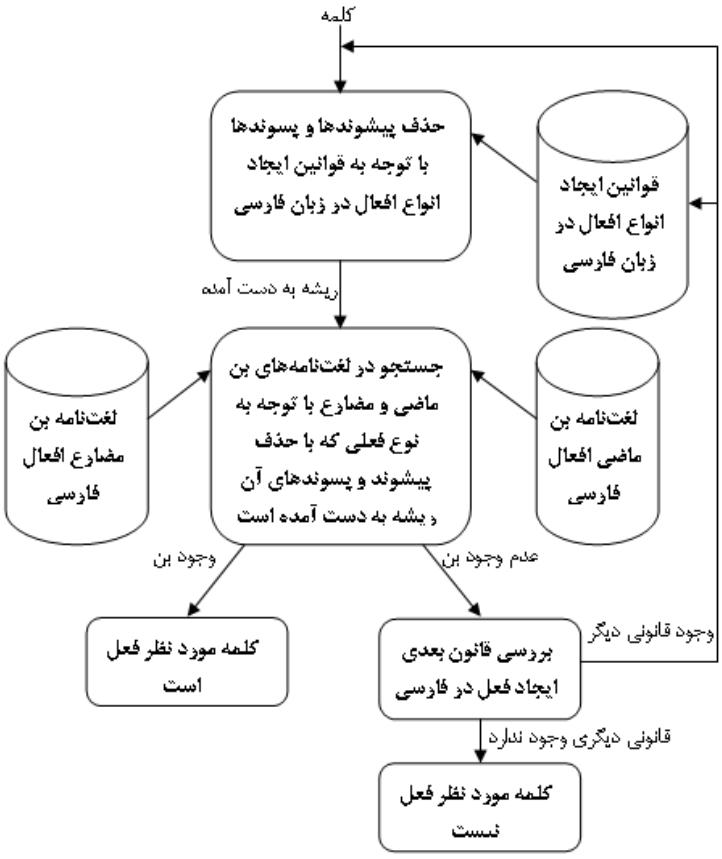
به خاطر در دسترس نبودن لغت نامه تصمیم گرفته شده که بر پایه برخی قواعد افعال و نوع آنها شناسایی انجام شود. در زبان فارسی شناسایی رده یک واژه (اسم، فعل، ...) به سادگی امکانپذیر نمی‌باشد. افعال در زبان فارسی رده بسیار بزرگی هستند که به کمک آنها بسیاری از اسم‌ها و صفت‌ها و ... ساخته می‌شوند، بنابراین با شناسایی این رده بزرگ، می‌توان بسیاری از واژه‌های فارسی را ریشه‌یابی نمود و با استفاده از روش پیشهادی می‌توان حدود جملات را تعیین نمود.

در این روش هر کلمه مورد بررسی را در صورت وجود پیشوند و پسوندهای فعلی آن را حذف می‌کنیم. برای این‌کار از یک روش ریشه‌یابی شبیه به پورتر[۳۴]، با این تفاوت که هم پیشوندها و هم پسوندها را حذف می‌کند، استفاده شده است. بعد از به دست آوردن ریشه با توجه به پسوند و پیشوندهای حذف شده با توجه به ساختار افعال فارسی که در جدول ۴-۱ آمده است، ریشه را در فرهنگ لغات بن‌های ماضی و مضارع جستجو می‌کنیم در صورت وجود داشتن ریشه به دست آمده، کلمه فعل تشخیص داده شده و نوع آن با توجه به پسوند و پیشوندهای حذف شده تعیین می‌گردد. روش کار در شکل ۴-۱ نشان داده شده است.

فرهنگ لغات بن‌های ماضی و مضارع مورد نیاز جهت استفاده در این روش موجود نبودند و برای انجام این‌کار نیاز به ایجاد چنین فرهنگ لغاتی وجود داشت که با استفاده از [۳۵] تهیه شده است.

جدول (۱-۹) ساختار افعال فارسی

	نوع فعل	ساختار	شناسه
ماضی (گذشته)	ماضی ساده	بن ماضی + شناسه	م ی – یم ید ند
	ماضی استمراری	می + بن ماضی + شناسه	م ی – یم ید ند
	صفت مفعولی	بن ماضی + ه	–
	ماضی نقلی	صفت مفعولی + شناسه	ام ای است ایم اید اند
	ماضی نقلی استمراری	می + صفت مفعولی + شناسه	ام ای است ایم اید اند
	ماضی بعید	صفت مفعولی + شناسه	بودم بودی بود بودیم بودید بودند
	ماضی التزامی	صفت مفعولی + شناسه	باشم باشی باشد باشیم باشید باشند
مضارع (حال)	مضارع ساده	بن مضارع + شناسه	م ی د یم ید ند
	مضارع اخباری	می + بن مضارع + شناسه	م ی د یم ید ند
	مضارع التزامی	ب + بن مضارع + شناسه	م ی د یم ید ند
مستقبل (آینده)	مستقبل	شناسه + بن ماضی	خواهم خواهی خواهد خواهیم خواهید خواهند



شکل (۱-۸) روش تشخیص فعل

اما در این روش با چند مشکل برخورد می‌کنیم که دقت سیستم را پایین می‌آورد، که در زیر در مورد آنها بحث می‌کنیم :

- ابهام هم‌نویسه‌ها: در تشخیص فعل در یک جمله در برخی موارد با مشکل ابهام هم-نویسه‌ها برخورد می‌کنیم، مثل کلمه « مَردُم »، که فعل از ریشه مردن است و « مُردَم » که یک اسم جمع است، اینها از لحاظ نوشتاری کاملاً شبیه‌اند و معنا و تلفظ متفاوتی دارند که با استفاده از این روش ساختاری که معرفی شده قابل حل نیست. چون عمل تشخیص فعل در این سیستم روی کلمات یک متن درحال انجام است، می‌توان از مدلهای مدل چند-تایی جهت رفع این ابهام و بالا بردن دقت سیستم بهره برد.

پس ما یک سری قوانین را برای برخورد با این مشکل استفاده می‌کنیم. با استفاده از یک ساختار bi-gram **که از روی یک پیکره که برچسب‌های اجزای کلام خورده، آموزش دیده است در صورتی که کلمه فعل تشخیص داده شود احتمال فعل بودن**

آن با استفاده از فرمول ۳-۱ از bi-gram **محاسبه می‌شود و یک حد آستانه برابر احتمال ۰.۵ برای پذیرش فعل بودن کلمه مورد نظر در نظر گرفته شده است.**

$$P(s) \cong \prod_{t=1}^{n} P(W_t|W_{t-1}) \cong \prod_{t=1}^{n} P(C_t|C_{t-1}) \prod_{t=1}^{n} P(W_t|C_t) \qquad (۱۰-۱)$$

تغییرات ساختاری: در اتصال برخی پیشوندها به بن‌های فعلی جهت ساخت فعل برخی تغییرات در آنها صورت می‌گیرد. در اتصال پیشوند «ب» یا «ن» به بن‌هایی که با حرف «ا» آغاز می‌شوند، باعث تغییر ساختار نسبت به قانون کلی می‌شود، در برخی موارد «ا» به «ی» تبدیل می‌شود(ب+افکن=بافکن در صورتی که به بیفکن تبدیل می‌شود)، یا اینکه «ا» به «یا» تبدیل می‌شود (ب+ا(بن مضارع آمدن)=با در صورتی که به بیا تبدیل می‌شود). این مشکل را نیز می‌توان با افزودن چند قانون ساده به سیستم که در صورت وجود «ی» / «یا» در ابتدای بن و نیافتن آن در فرهنگ لغت آنها را با «ا» جایگزین کرده و مجدداً جستجو نماید حل کرد.

فعل سبک[۷۱]: نوعی از افعال هستند که از ترکیب یک بخش غیرفعلی، شامل اسم، صفت، حروف اضافه و ..، و یک بخش فعلی تشکیل شده است. در شناسایی آن افعالی که بخش غیر فعلی به بخش فعلی بعد از خودش نمی‌چسبد روش مشکلی ندارد، به خاطر اینکه ترکیب شامل دو توکن مجزا است که یکی از آنها فعل می‌باشد، اما در مورد حروف اضافه، چون فرا، فرو، باز، بر و ...، به عنوان بخش غیرفعلی چون به فعل می‌چسبند و به صورت یک توکن واحد در می‌آید، مشکل وجود دارد. این مشکل را نیز با حذف این حروف اضافه و تشخیص مجدد کلمه در صورت فعل نبودن حل نمودیم.

[71] Light Verb

۱-۱۰- تعیین حدود جمله با استفاده از مدل چند-تایی

مدل چند-تایی ها قسمت اصلی تکنولوژی تشخیص گفتار فعلی هستند. واقعاً همه محصولات تشخیص گفتار تجاری از برخی انواع مدل چند-تایی ها استفاده می‌کنند. یک مدل چند-تایی مسئله تخمین به وسیله مدلسازی زبان را از لحاظ ابعادی به یک منبع مارکوف مرتبه n-1 کاهش می‌دهد :

(۱۱-۱)

$$P(w_t|h_t) \approx P(w_t|w_{t-n+1},\ldots,w_{t-1})$$

مقدار n پایداری تخمین، مثل واریانس، را جایگزین تناسب آن، مثل بایاس، می‌کند. یک trigram (n=3) یک انتخاب معمولی با مجموعه آموزشی بزرگ (میلیونها کلمه) است درحالیکه یک bigram (n=2) اغلب برای میزان کمتری از آن استفاده می‌شود. احتمالات trigram و bigram به دست آمده، حتی با مجموعه‌های خیلی بزرگ، هنوز یک مسئله تخمین پراکنده است[۳۶].

(۱۲-۱)

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\ldots P(w_n|w_1^{n-1}) = \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

از ویژگی های قابل انتظار مدلهای مدل چند-تایی این است که دقت (کارایی) مدل با افزایش مقدار N افزایش می یابد. علیرغم این واقعیت، عملا در بیشتر کاربردها از مدلهای bigram یا حداکثر trigram استفاده می‌شود؛ زیرا مدلهای مرتبه بالاتر از ۳ برای آموزشِ مناسب، احتیاج به پیکره بسیار بزرگی دارند و در غیر این صورت نمی-توان تخمین‌های مناسبی برای احتمالات به دست آورد. علاوه بر این، مدلهای مرتبه بالاتر از ۳ بسیار بزرگ هستند و ذخیره و استفاده از آنها احتیاج به حافظه زیادی دارد. مرتبه حافظه مورد نیاز یک مدل مدل چند-تایی برابر است با تعداد ترکیبهای

مختلف N کلمه ای که دارای حد بالای VN است (V اندازه لغت‌نامه مورد استفاده است) .

ویژگی دیگر مدلهای مدل چند-تایی وابستگی زیاد آنها به پیکره (به خصوص نوع و اندازه) آموزشی است. یکی از روشهای معمول برای مشاهده عملکرد کیفی یک مدلمدل چند-تایی تولید رشته‌های تصادفی کلمات با استفاده از مدل است. به این ترتیب که اولین کلمه بر اساس احتمالunigram انتخاب می‌شود و سپس دومین کلمه با مدلbigram و پس از آن کلمات بعدی می‌توانند با مدلbigram یا trigram تولید شوند.

نظر به اینکه احتمالات در یک مدل آماری همچون مدل چند-تایی از یک مجموعه آموزشی استخراج می‌شوند، این مجموعه را باید با دقت انتخاب کرد. اگر مجموعه آموزشی تنها مربوط به یک زمینه (موضوع) خاص باشد، احتمالات نتیجه شده نمی-توانند برای جملات جدید به شکلی مناسب تعمیم یابند. از طرفی اگر مجموعه نوشته-جات آموزشی بسیار عمومی و کلی باشد، ممکن است احتمالات نتیجه شده برای کاربرد خاص مناسب نباشند. برای آموزش و سپس محاسبه کارایی یک مدل، همانند سایر مسائل یادگیری محاسباتی، پیکره موجود را به دو مجموعه مجزای آموزشی و آزمایشی تقسیم می‌شود؛ مدل بر اساس مجموعه آموزشی ساخته شده و سپس کارایی آن با استفاده از معیاری به نام سرگشتگی[۷۲]، که معیاری مشابه با آنتروپی است، روی مجموعه آزمایشی محاسبه می‌شود. البته در برخی موارد به بیش از یک مجموعه آزمایشی احتیاج داریم.

۱-۱۰-۲- تشریح کامل روش

در این روش پیشنهادی از مدلهای مدل چند-تایی برای شناسایی حدود جمله استفاده می‌شود. یعنی نیاز به پیکره‌ای که حدود جملات در آن برچسب خورده‌اند،

[72] Perplexity

وجود دارد. برخلاف روش قبل که یک روش ساختاری بود این روش کاملاً آماری است و بر مبنای مفهوم مدل چند-تایی کار می‌کند. در این روش یک مدل مدل چند-تایی با استفاده از پیکره برچسب خورده ایجاد شده آموزش داده می‌شود و سپس احتمال اینکه بعد از یک سری از کلمات برچسب ابتدا یا انتهای جمله باشد مشخص می‌گردد. اگر احتمال به دست آمده از آستانه در نظر گرفته شده بیشتر بود، برچسب مورد نظر قرار داده شده و آن محل به عنوان ابتدا یا انتهای جمله برچسب-گذاری می‌شود.

۱-۱۱- استخراج بردار ویژگی و استفاده از رده‌بندها

یکی از روشهای دیگری که برای تعیین حدود جمله در پیکره‌های متنی زبان فارسی ارائه شده است، استخراج بردار ویژگی‌های از متن و انجام تصمیم‌گیری راجع به اینکه ابتدا و انتهای جملات در متن کجا می‌باشد. برای اینکار نیاز به یک پیکره متنی زبان فارسی داریم که برچسب حدود جملات در آن وجود داشته باشد تا هم بتوان رده‌بند را آموزش داد و هم آن را آزمون کرد. این روش به صورت زیر عمل می-کند:

هر متن از این پیکره مورد بررسی قرار می‌گیرد، به این صورت که، یک پنجره n تایی به همراه یک برچسب، وجود یا عدم وجود جمله جدید و محل آن در میان این n کلمه موجود در پنجره، در نظر گرفته می‌شود و این پنجره یک کلمه به یک کلمه جلو می‌رود و با استفاده از متون برچسب خورده بردارهای موجود در متن را به همراه ویژگی‌ها استخراج می‌کند و می‌توان برای آموزش رده‌بند و آزمون آن مورد استفاده قرار گیرد.

برای هر کدام از کلمات پنجره بسیاری از موارد چون فعل بودن، حرف بودن، عدد بودن، پرانتز، علائم نشانه‌گذاری و غیرنشانه‌گذاری و ... مورد بررسی قرار می‌گیرد. ما

در این مدل از دو رده‌بند شبکه‌های عصبی استفاده کرده‌ایم که انها را مورد بررسی قرار می‌دهیم.

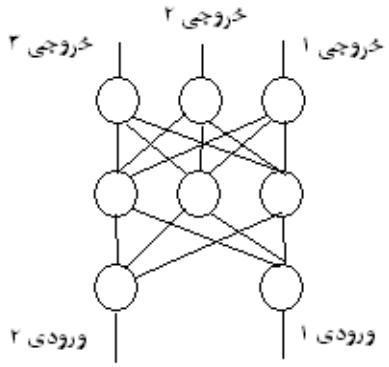
۱-۱۱-۱- شبکه عصبی مصنوعی[۷۳]

□ معرفی شبکه عصبی مصنوعی

شبکه عصبی مصنوعی از آن ایده که توانایی هوش مغز انسان را به طور ریاضی‌وار مدل کنند، نشأت گرفته است. به علت موازی‌کاری بسیار بالا، شبکه عصبی تبدیل به یک رده‌بند باارزش شده است و همچنین تاثیر زیادی بر تئوری و کاربرد شناسایی‌الگو داشته است. این نوع رده‌بند از نوع رده‌بندهای ناپایدار می‌باشد؛ یعنی تغییر کوچک در مجموعه آموزشی منجر به تغییر زیادی در رده‌بند دارد (چه در ساختار و چه در پارامترهایش). یک رده‌بند مبتنی بر شبکه‌ی عصبی می‌تواند یک مجموعه‌داده آموزشی را با هر دقت دلخواه یاد بگیرد. منتها زمان بسیار زیادی ممکن است نیاز داشته باشد. شبکه عصبی ساختارهای متنوعی دارد از جمله MLP، RBF، PNN، Probabilistic **و .... در اینجا به بررسی یک نوع متداول شبکه عصبی به نام** MLP**، که مورد استفاده قرار گرفته است، می‌پردازیم.**

MLP **یکی از شبکه‌های عصبی اساسی و اولیه می‌باشد. الگوریتم آموزش آن اینگونه است که ابتدا ورودی‌ها داده می‌شود و خروجی متناظر به دست می‌آید و سپس بنا به میزان خطای ایجاد شده، تغییرات وزن را طوری انتخاب می‌کنیم تا در خلاف جهت مشتق خطا، حرکت کنیم. آنگاه خطا را به لایه‌های عقب نیز بر می‌گردانیم و خطای نرون‌های لایه‌های قبل را نیز محاسبه می‌کنیم. شکل ۱، یک** MLP **سه‌لایه را که داده‌های دوبعدی به عنوان ورودی می‌پذیرد، نشان می‌دهد.**

[73] Artificial Neural Network (ANN)



شکل (۱-۹) ساختار یک شبکه عصبی MLP، با سه لایه

در این شکل لایه خروجی سه نرون، لایه میانی (یا مخفی) سه نرون و لایه ورودی دو نرون دارد. سه نرون لایه خروجی در اینجا می‌تواند مثلا برای تشخیص سه رده c1، c2 و c3 باشند، یعنی هر نرون برای یک رده. اگر یک داده(x1،x2) را به عنوان ورودی به این شبکه بدهیم منجر به یک خروجی (c1،c2،c3) می‌شود؛ که ci یک است اگر آن ورودی متعلق به رده‌ی ci باشد و در غیر اینصورت صفر می‌باشد. برای مطالعه‌ی بیشتر شبکه‌ی عصبی و انواع مدل‌های آن به کتاب‌های شبکه عصبی مراجعه کنید.

◘ روش استفاده

با توجه به مطالب ارائه شده در ابتدای این بخش، ما در طی روش ارائه شده متن را به بردارهایی از پنجره‌های متوالی تبدیل می‌کنیم که برچسب حدود جمله را نیز دارد. با استفاده از یک MLP به روشهای گوناگون که در ابعاد پنجره، تعداد لایه‌های شبکه، لایه ورودی، خروجی، لایه‌های مخفی و کدگذاری ورودی و خروجی متفاوتند، یک شبکه ایجاد شده و مورد بررسی قرار گرفت. نتایج و روشهای مختلف بررسی شده در فصل بعد مورد بررسی قرار گرفته است.

نتایج تجربی و ارزیابی

۱-۱۲- مقدمه

در این فصل نتایج تجربی به دست آمده از روشهای پیشنهادی فصل قبل پرداخته شده است. برای انجام این کار به یک پیکره متنی که حدود جمله در آن برچسب-گذاری شده باشد، نیاز است. به علت عدم وجود چنین پیکره‌ای، نسبت به ایجاد آن اقدام شده است. برای تولید این پیکره از بخشی از پیکره بی‌جن‌خان [۳۷] استفاده شده است، و با استفاده از چند متخصص زبان و ادبیات فارسی پیکره‌ای حاوی ۱۰۲۳۲ جمله برچسب خورده برای آموزش و آزمون سیستم‌های پیشنهادی آماده شده و مورد استفاده قرار گرفته است.

در این پیکره از برچسب <SB> **جهت نشان دادن ابتدای جمله و از برچسب** </SB> **برای نشان دادن انتهای جمله استفاده شده است. بخشی از یکی از متون برچسب خورده در شکل ۵-۱ نشان داده شده است.**

<SB>محیط دانشگاه و نسل جوان آن ، نوگرا و آرزوخواه است .</SB><SB>
در این محیط ، شعارها و خواسته‌های جدید ، نوبه‌نو عرضه می‌شوند
</SB><SB> و مشتریان خود را می‌یابند .</SB><SB> این نکته ، مزیت
دانشگاه است ،</SB><SB> و نه منقصت آن .</SB><SB> اما
مدیریتهای علمی در دانشگاهها ، بایستی توجه کنند </SB><SB> که نباید
در مسیر پرپیچ‌وخم شعارهای زودگذر قرار گیرند</SB><SB> و برای حفظ
اعتبار و یا دیگر مسایل ، در جوزدگی و غوغاسالاری عمل کنند</SB><SB>
و محکوم نسیمهای زودگذر و یا طوفانهای سهمگین باشند .</SB><SB>

شکل (۱-۱۰) نمونه متونی که برچسب حدود جمله خورده‌اند

به علت عدم وجود کار انجام شده‌ای در این زمینه در زبان فارسی و قابل مقایسه نبودن کارهای انجام شده در زبان‌های دیگر، مقایسه فقط بین روشهای ارائه شده صورت گرفته است، که در بخش‌های بعدی آورده شده است.

۱-۱۳- تعیین حدود جمله با استفاده از تشخیص فعل

در این روش با توجه به مطالب گفته شده در فصل قبل یک روش جهت تشخیص فعل در متون فارسی ارائه شده بود و با توجه به ساختار فاعل- مفعول - فعل زبان فارسی پایان هر فعل انتهای جمله قبل و انتهای جمله بعدی را نشان می‌دهد.
روش ارائه شده جهت تشخیص فعل در متون فارسی از یک سری قوانین ایجاد افعال فارسی از بن‌های ماضی و مضارع استفاده می‌کند و جهت برطرف کردن اشکالاتی چون ابهام هم‌نویسه‌ها، افعال سبک و تغییرات ساختاری یک سری قانون دیگر هم اضافه شده است تا دقت سیستم را افزایش دهد. نتایج بررسی تشخیص فعل روی پیکره بی‌جن‌خان صورت گرفته است که برچسب اجزای کلام را داراست و افعال در آن مشخص هستند. نتایج روش به کار رفته در تشخیص فعل روی متون این مجموعه در ٤ دسته آورده شده است. نتایج تجربی نشان می‌دهد که تشخیص فعل با به کار بردن روشهای معرفی شده جهت رفع ابهامات موجود در این روش، بسیار کارا بوده است. دو نوع خطای مثبت و منفی معرفی شده است که خطای مثبت معرف افعالی است که فعل تشخیص داده نشده است و خطای منفی معرف کلمات غیر فعلی است که فعل تشخیص داده شده است. مشاهده می‌شود که در این روش خطای منفی در ابتدا نیز بسیار پایین‌تر از خطای مثبت است و در تشخیص کلمات غیر فعل بهتر عمل می‌کند اما روشهای رفع ابهام بطور قابل توجهی این خطاها را کم کرده است طوری که در روش کلی دقت به دست آمده برابر ۹۸.۹٪ خواهد بود.

جدول (۱-۱۰) نتایج به دست آمده از روش تشخیص فعل

(برحسب٪) قوانین	تشخیص فعل	خطای مثبت	تشخیص فعل نبودن	خطای منفی	خطای کل	دقت کل
ساخت فعل	۸۸.۱	۱۱.۹	۹۷.٤	۲.٦	۳.٦	۹٦.٤
ساخت فعل + رفع ابهام تغییرات ساختاری	۸۹.۸	۱۰.۲	۹۷.٦	۲.٤	۳.۲	۹٦.۸
ساخت فعل + رفع ابهام تغییرات ساختاری + رفع ابهام افعال سبک	۹۱.٦	۸.٤	۹۸.۳	۱.۷	۲.۳	۹۷.۷
ساخت فعل + رفع ابهام تغییرات ساختاری + رفع ابهام افعال سبک + رفع ابهام هم-نویسه‌ها	۹٥.۸	٤.۲	۹۹.۳	۰.۷	۱.۱	۹۸.۹

با توجه به این روش کارای ساده جهت تشخیص فعل از این روش جهت شناسایی پایان جملات بهره برده شد و سیستم ارائه شده در حدود ۷۹.۵٪ دقت را در تشخیص حدود جملات در متن فارسی نشان می‌دهد که با توجه به سیستم ساده استفاده شده کارآیی بسیار خوبی محسوب می‌شود.

اما این روش مشکلاتی نیز دارد، از جمله این که جملاتی که با استفاده از این روش قابل شناسایی است، صرفاً جملات ساده می‌باشند و این روش توانایی شناسایی صحیح جملات مرکب را ندارد. در این روش فقط جملاتی تشخیص داده می‌شوند که به فعل ختم شوند و در آنها فعل در آخر جمله قرار گرفته شده باشد، که با توجه به استثنائات زیادی که در ساختار جملات زبان فارسی وجود دارد، این موارد نیز در متون فارسی، خصوصاً داستانها و نوشته‌های عامیانه به چشم می‌خورد.

از مزایای این روش نیز می‌توان به سرعت بالا، عدم نیاز به آموزش به دلیل استفاده از یک روش ساختاری و عدم وابستگی به مراحل بعد از این مرحله در کارهای پردازش زبان طبیعی و متن‌کاوی نام برد.

۱-۱۴- تعیین حدود جمله با استفاده از مدل چند-تایی

برخلاف روش قبل که یک روش ساختاری بود این روش کاملاً آماری است و بر مبنای مفهوم مدل چند-تایی کار می‌کند. در این روش یک مدل مدل چند-تایی با استفاده از پیکره برچسب خورده ایجاد شده آموزش داده می‌شود و سپس احتمال اینکه بعد از یک سری از کلمات برچسب ابتدا یا انتهای جمله باشد مشخص می‌گردد. اگر احتمال به دست آمده از آستانه در نظر گرفته شده بیشتر بود، برچسب مورد نظر قرار داده شده و آن محل به عنوان ابتدا یا انتهای جمله برچسب‌گذاری می‌شود.

در این روش مدلهای ۱-تایی[۷۴]، ۲-تایی[۷۵] و ۳-تایی مورد بررسی قرار گرفته است، دقت مطلوبی به دست نیامد، زیرا اولاً نیاز به استفاده از مدلهای ۳-تایی[۷۶] و بالاتر برای دقت بالا نیاز است و ثانیاً این مدلها برای آموزش نیاز به دامنه وسیع و همه منظوره‌ای از متون را در پیکره مورد بررسی نیاز دارند و ثالثاً پیکره باید بسیار بزرگ باشد[۳۶]، تا بتوان یک مدل کارا با استفاده از آن تهیه کرد و متاسفانه پیکره ایجاد شده وسعت مورد نیاز را جهت جوابگویی برای چنین مدلی را دارا نیست.

مدل ۱-تایی نیز حتی در صورت ایجاد مدلی کامل، به خاطر اینکه روی یک کلمه تصمیم می‌گیرد، توانایی حل این مسئله را ندارد. مدل ۲-تایی نیز دقت مطلوب را نخواهد داشت و مدل ۳-تایی و حتی مدلهای مرتبه بالاتر مناسب‌تر به نظر می‌رسند، اما با این پیکره محدود ایجاد شده توانایی ساخت مدلی کارا وجود ندارد.

---

[۷۴] unigram

[۷۵] bigram

[۷۶] trigram

۱-۱۵- تعیین حدود جمله با استفاده از شبکه عصبی

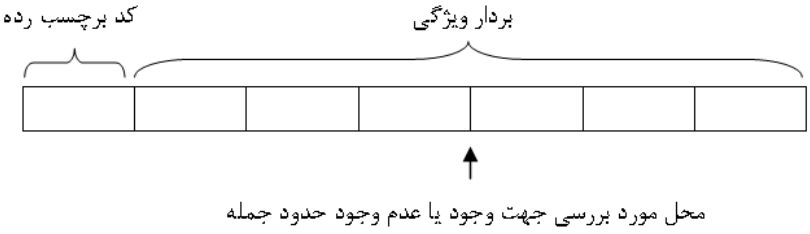
طبق مطالب ارائه شده در فصل قبل در این روش نیاز به استخراج برداری از ویژگیها از متن داریم و سپس این بردار به همراه برچسب حدود جمله جهت آموزش به یک شبکه عصبی اعمال می‌شود و پس از آموزش شبکه از آن جهت برچسب‌گذاری حدود جملات استفاده می‌شد.

بردار ویژگی را از روی خواص کلمات محصور در پنجره‌ای n خانه‌ای که روی متن یک کلمه یک کلمه حرکت می‌کند ساخته می‌شود. اندازه پنجره لغزنده در ابعاد مختلف ۴، ۵، ۶، ۷ و ۸ عنصری مورد بررسی قرار گرفت.

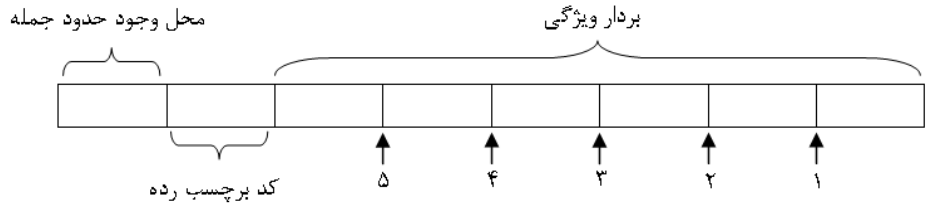
دو نوع خروجی، برای شبکه عصبی در نظر گرفته شد، یک حالت اینکه شبکه عصبی حالت تک خروجی در نظر گرفته شد- فقط برچسب رده به عنوان خروجی درنظر گرفته می‌شود- که وجود و یا عدم وجود محدوده جمله، مانند ابتدای جمله <SB> یا انتهای جمله </SB> یا انتهای جمله قبل و ابتدای جمله بعد <SB></SB> یا عدم وجود محدوده جمله(جدول ۵-۲)، را در محل خاصی از پنجره n عنصری مثلاً وسط آن، بین عنصر n/2 و n/2+1، مشخص می‌کند(شکل۵-۲). نوع دیگر خروجی شبکه عصبی می‌تواند دو مقدار باشد، یکی محل قرار گرفتن محدوده جمله که یک عدد بین ۱ تا n-۱ خواهد بود، طوری‌که مکانهای بین عنصر اول و دوم را عدد ۱، عنصر دوم و سوم را عدد ۲ و ... نشان خواهد داد و خروجی دیگر همانند خروجی حالت تک خروجی می‌باشد(شکل۵-۳).

جدول (۱-۱۱) برچسب‌های رده و کدهای مربوط به آن

برچسب رده	کد برچسب رده
حدود جمله نیست	۰
<SB>	۱
</SB>	۲
<SB></SB>	۳



شکل (۱-۱۱) بردار ویژگی ۶ عنصری به همراه برچسب رده (خروجی نوع اول)



شکل (۱-۱۲) بردار ویژگی ۶ عنصری به همراه برچسب رده و محل وجود حدود جمله (خروجی نوع دوم)

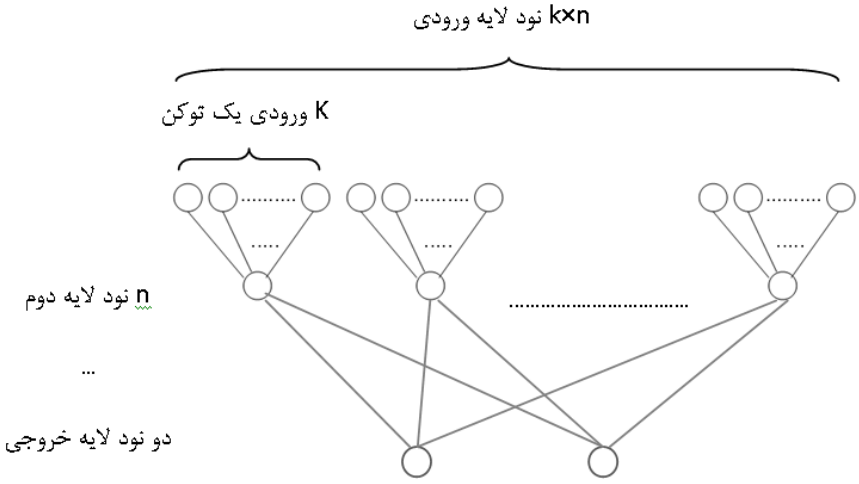
در مورد ویژگی‌هایی که از هر کلمه مورد نظر مورد بررسی قرار می‌گیرد نیز دو روش مورد بررسی قرار گرفته شده است. در مدل اول برای هر کلمه موجود در هر خانه پنجره ۴ ویژگی علامت نشانه‌گذاری، حرف، فعل بودن یا نبودن مورد بررسی قرار گرفت و در حالت بعد تعداد ویژگی‌ها بسیار بیشتر شده و مثلاً به جای علائم نشانه‌گذاری خود علائم "."، "؟"، "!"، ":"، "؛"و "،" به عنوان خصیصه در نظر گرفته شد. برای حروف نیز خود آنها مثل "و"، "چون"، "که"، "تا" و ... جایگزین شده و مورد بررسی قرار گرفت.

با توجه به مطالب ارائه شده و استخراج ویژگی‌ها با استفاده از یک پنجره لغزان روی متن و تبدیل مسئله به یک مسئله رده‌بندی کلاسیک، حال با استفاده از یک رده‌بند و آموزش آن می‌توان عمل تعیین حدود جمله را انجام داد. در این مورد رده‌بند شبکه عصبی MLP در نظر گرفته شده است که در بخش بعد نتایج به دست آمده تشریح شده است.

۱-۱۵-۲- ساختار شبکه عصبی

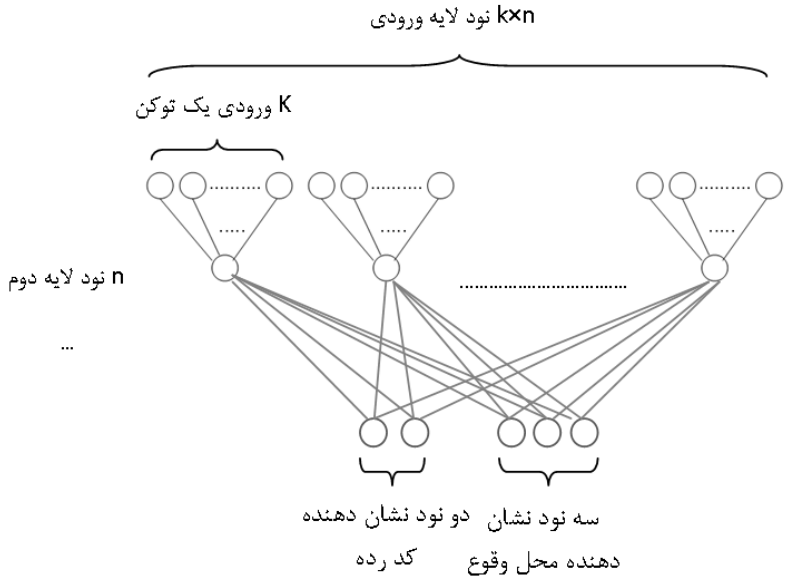
با توجه به اینکه یکی از k ویژگی‌ها مطرح شده می‌تواند برای هر توکن موجود در پنجره مورد بررسی، وجود داشته باشد، k×n گره که n تعداد توکن‌های موجود در پنجره است به عنوان گرههای لایه ورودی در نظر گرفته شده است. سپس در لایه بعد k گره مربوط به ویژگی‌های هر توکن به یک گره در لایه میانی متصل شده - می‌توان لایه‌های مخفی بیشتری داشت - و سپس همه گرههای لایه مخفی آخر به گرههای لایه خروجی متصل می‌شود.

گرههای لایه آخر با توجه به نوع خروجی تعیین می‌شود اگر خروجی از نوع اول نیاز به دو گره خروجی برای ایجاد کد خروجی به صورت دودوئی است. شکل ۵-۴ نشان دهنده شبکه عصبی با خروجی از نوع اول است.



شکل (۱-۱۳) ساختار شبکه عصبی (خروجی نوع اول)

خروجی نوع دوم علاوه بر نیاز به ۲ گره برای نشان دادن رده، نیاز به تعداد گرهی دارد که محل وقوع آن را نشان دهد. شکل ۵-۵ نشان دهنده شبکه عصبی با خروجی از نوع دوم با n برابر ۶ است.



شکل (۱-۱۴) ساختار شبکه عصبی (خروجی نوع دوم n=6)

برای آموزش شبکه عصبی بردارهای ویژگی به همراه برچسب‌های مورد نیاز برای آن از پیکره ایجاد شده استخراج شد که در حدود ٤٥٠٣٢٠ بردار به دست آمد. تعداد ۳۷۰۳۲۰ بردار جهت آموزش به طور تصادفی انتخاب شد و شبکه با استفاده از تکنیک انتشار به عقب و حداقل مربعات خطا آموزش داده شد و از مابقی بردارهای باقیمانده جهت آزمون و ارزیابی استفاده شد. نتایج به دست آمده برای حالات مختلف در جداول ٥-٣ تا ٥-٦ آمده است.

شبکه‌های عصبی با لایه‌های مخفی بیشتر نیز مورد بررسی قرار گرفت که این افزایش لایه‌های میانی منجر به کاهش دقت گردید. بهترین ساختار و نتایج بهینه ارائه شده با ساختار دارای یک لایه مخفی که در اشکال ٥-٥ و ٥-٦ نشان داده شده است به دست آمده است.

در جداول ارائه شده <SB> **نشان دهنده دقت تشخیص رده ابتدای جملات،** </SB> **نشان دهنده دقت تشخیص رده انتهای جملات و** <SB></SB> **نشان دهنده دقت تشخیص رده انتهای یک جمله و ابتدای جمله بعد می‌باشد.**

جدول (۱-۱۲) نتایج با استفاده از خروجی نوع اول و ۴ خصیصه با یک لایه مخفی

اندازه پنجره (n)	<SB>	</SB>	<SB></SB>	دقت کل	خطای کل
٤	٪٤٧	٪٤٨	٪٩٢	٪٨٥.٣	٪١٤.٧
٥	٪٥٠	٪٥٠	٪٩٤	٪٨٧.٥	٪١٢.٥
٦	٪٥٠	٪٥١	٪٩٥	٪٨٨.٣	٪١١.٧
٧	٪٥٠	٪٥١	٪٩٥	٪٨٨.٣	٪١١.٧
٨	٪٥٠	٪٥١	٪٩٥	٪٨٨.٣	٪١١.٧

جدول (۱-۱۳) نتایج با استفاده از خروجی نوع اول و ۲۳ خصیصه با یک لایه مخفی

اندازه پنجره (n)	<SB>	</SB>	<SB></SB>	دقت کل	خطای کل
٤	٪٥١	٪٥٣	٪٩٥	٪٨٨.٥	٪١١.٥
٥	٪٥٥	٪٥٥	٪٩٧	٪٩٠.٧	٪٩.٣
٦	٪٥٥	٪٥٧	٪٩٨	٪٩١.٧	٪٨.٣
٧	٪٥٥	٪٥٧	٪٩٨	٪٩١.٧	٪٨.٣
٨	٪٥٥	٪٥٧	٪٩٨	٪٩١.٧	٪٨.٣

جدول (۱-۱۴) نتایج با استفاده از خروجی نوع دوم و ۴ خصیصه با یک لایه مخفی

اندازه پنجره (n)	<SB>	</SB>	<SB></SB>	دقت کل	خطای کل
٤	٪٤٠	٪٤٠	٪٩٠	٪٨٢.٥	٪١٧.٥
٥	٪٤٠	٪٤١	٪٩١	٪٨٣.٥	٪١٦.٥
٦	٪٤٠	٪٤٠	٪٩٠	٪٨٢.٥	٪١٧.٥
٧	٪٣٧	٪٣٨	٪٨٨	٪٨٠.٥	٪١٩.٥
٨	٪٣٣	٪٣٥	٪٨٨	٪٧٩.٩	٪٢١.١

جدول (۱-۱۵) نتایج با استفاده از خروجی نوع دوم و ۲۳ خصیصه با یک لایه مخفی

اندازه پنجره (n)	<SB>	</SB>	<SB></SB>	دقت کل	خطای کل
٤	٪٤٥	٪٤٥	٪٩٢	٪٨٥	٪١٥
٥	٪٤٥	٪٤٦	٪٩٣.٥	٪٨٧	٪١٣
٦	٪٤٤	٪٤٥	٪٩١	٪٨٤	٪١٦
٧	٪٤٢	٪٤٢	٪٩٠.٥	٪٨٣.٢	٪١٧.٨
٨	٪٣٩	٪٤١	٪٩٠	٪٨٢.٥	٪١٧.٥

با توجه به نتایج به دست آمده می‌توان دریافت که استفاده از ۲۳ خصیصه که هر

کدام از حروف و علائم نشانه‌گذاری به طور مجزا خصیصه در نظر گرفته می‌شوند نسبت به اینکه همه حروف به عنوان خصیصه در نظر گرفته شود بهتر عمل می‌کند اما زمان بیشتری صرف استخراج خصیصه‌ها و آموزش شبکه می‌شود.

همچنین خروجی نوع اول که فقط رده - وجود حدود جمله در محل خاصی از پنجره و نوع آن- را بررسی می‌کند با بزرگتر شدن پنجره دقت بیشتری را ارائه می-هد، اما از پنجره به ابعاد ٦ به بعد تقریباً ثابت می‌ماند. اما در مورد خروجی نوع دوم دقت پایین‌تر از نوع اول است، چون تعیین محل وقوع نیز باید تعیین گردد و با بزرگ شدن اندازه پنجره کاهش چشمگیری در دقت دیده می‌شود که به خاطر این است که در موارد بیشتری، بیش از یک محدوده جمله در بردار وجود دارد و این تصمیم‌گیری را مشکل می‌کند.

جمع‌بندي و پيشنهادها

۱-۱۶- مقدمه

با توجه به مطالب ارائه شده مختلف و بررسی چند روش مختلف می‌توان فهمید که با مسئله تعیین حدود جمله در پیکره‌های متنی زبان فارسی می‌توان به صورت یک مسئله رده‌بندی برخورد کرد. استخراج خصیصه‌های مناسب تاثیر به سزایی در نتیجه خواهد داشت. ایجاد پیکره‌های مناسب و استاندارد جهت به دست آوردن نتایج جامع مورد نیاز است. فرهنگ لغات مختلف محاسباتی استاندارد در این کار و کارهای مرتبط مورد نیاز است و باعث صرفه‌جویی بسیار زیاد در وقت و هزینه‌ها خواهد شد.

کارهایی که در زبان انگلیسی صورت گرفته مثل ایجاد یک دستور زبان جامع و از بین بردن استثناءهای زیاد کمک بسیاری به استاندارد شدن زبان و مناسب بودن آن برای کارهای متن‌کاوی و پردازش زبان طبیعی خواهد کرد.

۱-۱۷- جمع‌بندی

با توجه به مطالب ارائه شده مختلف و بررسی چند روش مختلف می‌توان فهمید که با مسئله تعیین حدود جمله در پیکره‌های متنی زبان فارسی می‌توان به صورت یک مسئله رده‌بندی برخورد کرد. استخراج خصیصه‌های مناسب تاثیر به سزایی در نتیجه خواهد داشت. برچسب اجزای کلام می‌تواند از جمله بهترین خصیصه‌ها برای انتخاب در این کار باشند اما برچسب‌گذاری نیاز به دانستن حدود جملات دارد و این یک حالت چرخشی را ایجاد می‌کند.

استخراج خصیصه‌های ساختاری و آماری و استفاده از رده‌بند یک روش مناسب جهت تعیین حدود جمله است.

۱-۱۸- پیشنهادها

این مسئله در زبان فارسی یک مسئله باز و کار نشده است و می‌توان نتایج بسیار خوب و دقیقی از آن به دست آورد. استخراج خصیصه‌های مختلف، استفاده از رده‌بندهای گوناگون و همچنین ترکیب رده‌بندها می‌تواند در این مسئله مورد بررسی قرار گیرد. انجام عمل تعیین حدود جمله به همراه برچسب‌گذاری اجزای کلام می‌تواند نتایج را در هر دو مورد بهبود بخشد. همچنین ایجاد پیکره‌ها و فرهنگ لغات استاندارد نیز کمک شایانی به افزایش دقت به دست آمده خواهد کرد.

مـر ا جـع

مراجع

[1] Haoyi Wang, Yang Huang, "Bondec - A Sentence Boundary Detector", PhD Thesis of Berkeley Engineering Informatics,2001 .

[2] Mikheev, A., "Tagging Sentence Boundaries",NACL '2000Seattle , pp264 27,2000 .

[۳] مهرنوش شمس‌فرد، ″ درک متن فارسی″، پایان‌نامه‌ی کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۴.

[۴] مهرنوش شمس‌فرد، ″پردازش متون فارسی: دستاوردهای گذشته، چالشهای پیش رو″، دومین کارگاه پژوهشی زبان فارسی و رایانه، ص ۱۷۲-۱۸۹، تهران، ۱۳۸۵.

[5] Cole R. A, J.Mariani, H.Uszkoreit, G.Varile, A.Zaenen, V.Zue, A.Zampoili, "Survey of the State of the Art in Human Language Technology", Cambridge University Press (Eds)(1997)

[۶] داداش میری، ″تشخیص انتهای کلمات و ایجاد فاصله میان کلمات″، پایان‌نامه‌ی کارشناسی، دانشگاه علم و صنعت ایران، ۱۳۸۰.

[۷] بی‌جن‌خان، ″تشخیص کسره‌ی اضافه″، طرح تحقیقاتی، پژوهشگاه فرهنگ هنر و ارتباطات، تهران، ۱۳۸۴.

[8] Megerdoomian Karin, Remi Zajac, "Processing Persian Text: Tokenization in the Shiraz Project", NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-322 ,2000

[۹] رضانیا، ″ پی‌ریزی طرح کلی واژگان و طراحی و پیاده‌سازی پردازشگر ساختواژی برای زبان فارسی″، پایان‌نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۶.

[۱۰] اسلامی، شریفی، علیزاده، زندی، ″واژگان زایای فارسی″، مجموعه سخنرانی‌های اولین کارگاه پژوهشی زبان فارسی و رایانه، تهران، ۱۳۸۳.

[۱۱] پورحسن، ″ تحلیل‌گر ساختواژی زبان فارسی″، پایان‌نامه‌ی کارشناسی، دانشگاه شهید بهشتی، تهران، ایران، ۱۳۸۵.

[۱۲] قاسمی‌زاده، رحیمی، سالاریان، ترکمنی، سمیاری، کوچاری، نم‌نبات، براری، ″روشی نوین برای صرف واژه-های فارسی″، یازدهمین کنفرانس انجمن کامپیوتر ایران، تهران، ۱۳۸۴.

[13] Arabsorkhi,Shamsfard ," Unsupervised Discovery of Persian Morphemes",11 th Conference of the European Chapter of the Association for Computational Linguistics (EACL),Italy ,2006

[۱۴] نوربالا، ″طراحی یک واژگان جامع برای پردازشگر زبان فارسی″، پایان‌نامه‌ی کارشناسی ارشد، دانشگاه علم و صنعت ایران، تهران، ۱۳۸۰.

[۱۵] رستم‌پور، ″واژگان محاسباتی زبان فارسی″، پایان‌نامه‌ی کارشناسی، دانشگاه شهید بهشتی، تهران، ۱۳۸۵.

[۱۶] بی‌جن‌خان، ″پیکره متنی زبان فارسی″، مجموعه سخنرانی‌های اولی‌ن کارگاه پژوهشی زبان فارسی و رای‌انه، تهران، ۱۳۸۳.

[17] Oroumchian F., Darrudi E., Hejazi M.R., "Assessment of a Modern Farsi Corpus", Proceeding of the 2nd Workshop on Information Technology & Its Disciplines (WITID), Iran,2004

[18] Sheykh Esmaili K., Abolhassani H., Neshati M., Behrangi E., Rostami A. Mohammadi Nasiri M., "Mahak: A Test Collection for Evaluation of Farsi Information Retrieval

System",Proceedings of 5th ACS/IEEE International Conference on Computer System and Applications (AICCSA-07) ,Amman ,Jordan ,May 2007

[19] Sheykh Esmaili K., Rostami A., "List of Persian StopWords", Technical Report No. 2006-03,Semantic Web Research Laboratory ,Sharif University of Technology , Tehran,Iran ,June 2006

[20] Mazdak N., Hassel M., "FarsiSum-a Persian Text Summarizer", Master Thesis, Department of Linguistics Stockholm University,2004

[۲۱] اشراق، سارابی، "سیستم پرسش و پاسخ به زبان فارسی"، پایان‌نامه‌ی کارشناسی، دانشگاه شهید بهشتی، تهران، ۱۳۸۵.

[۲۲] قیومی، "پیش‌بینی رایانه‌ای کلمه"، مجموعه سخنرانی‌های اولین کارگاه پژوهشی زبان فارسی و رایانه، تهران، ۱۳۸۳.

[۲۳] محروقی، " طراحی و پیاده‌سازی سیستمی برای ویرایش ادبی جملات ساده زبان فارسی"، پایان‌نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، ۱۳۷۵.

[۲۴] زاهدی، "طراحی و پیاده سازی یک برنامه هوشمند برای اعراب گذاری در متون فارسی"، پایان‌نامه کارشناسی ارشد، دانشگاه تهران، ۱۳۷۷.

[۲۵] باقری، "استنباط موضوعات مشترک از جملات مرتبط به هم"، پایان‌نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، ۱۳۷۲.

[26] Palmer D.et al ., "Adaptive Sentence Boundary Disambiguation",Proceeding of the 4th ACL Conference on Applied Natural Language Processing,Stuttgart ,pp78-83,13 -15 October 1994.

[27] Aderdenn et al.," Regular Expression Rules in Alembic System",1995

[28] Manning C.D., Schutze H., "Foundations of Statistical Natural Language Processing", MIT Press,Cambridge ,London ,2002 .

[29] Shannon C.E., "A Mathematical Theory of Communication", Bell System Technical Journal,pp 379-423,1948

[30] Mikheev A., "Document Centered Approach to Text Normalization",2000 .

[31] Berger A., "A Brief Maxent Tutorial", http://www-2.cs.cmu.edu /~aberger /maxent.html,1996 .

[32] Rayner J.C., Ratnaparkhi A., "A Maximum Entropy to Identifying Sentence Boundaries",Proceeding of the ANLP 97, Washington D.C.,1997.

[۳۳] یوسفان، صالحی، مینایی بیدگلی، "دشواری‌های ریشه‌یابی فارسی و روشی برای ریشه‌یابی فعل‌های ساده فارسی"، دومین کارگاه پژوهشی زبان فارسی و رایانه، ص. ۱۷۲-۱۸۹، تهران، ۱۳۸۵.

[34] Porter M.F.," An Algorithm for Suffix Stripping",Program , vol.14,no3,pp 130-137, 1980

[۳۵] طباطبایی، " فعل بسیط فارسی و واژه‌سازی"، نشر دانشگاهی ایران، تهران، ۱۳۷۴.

[36] Rosenfeld R., "Maximum Entropy Approach to Adaptive Statistical Language Modeling",Computer Speech and Language ,vol .10,pp 187-228,Longer Version Published as " Adaptive Statistical Language Modeling: A Maximum Entropy Approach", Phd Thesis, Computer Science Department, Carnegie Mellon University, TR CMU-CS-94-138,April 1994

[۳۷] بی‌جن‌خان، "نقش پیکره در نوشتن دستور زبان: مقدمه‌ای بر یک نرم افزار"، مجله زبانشناسی ایران، ش. ۱۹/۲، ۱۳۸۳.