تعیین مرز و نوع عبارات نحوی در متون فارسی

محمدمهدی همایون‌پور و آرمین سلیمی‌بدر
آزمایشگاه پردازش هوشمند سیگنال‌های صوتی و گفتاری، دانشکدهٔ مهندسی کامپیوتر و فناوری اطّلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

واحدسازی، از مهم‌ترین مسائل در پردازش زبان‌های طبیعی است که عبارت از فرآیند تقسیم متن به واحدهای معنادار نظیر واژه، عبارت نحوی، جمله و غیره است. واحدسازی گروه‌های نحوی یک متن، از جمله وظایف واحدسازی متن محسوب می‌شود که در بسیاری از کارهای پردازش زبان طبیعی، نظیر سامانه‌های ترجمهٔ ماشینی، استخراج اطّلاعات، پرسش و پاسخ و سامانه‌های تبدیل متن به گفتار، به‌عنوان پیش‌پردازشی مهم، می‌تواند حضور داشته باشد. واحدسازی عبارات نحوی، در هر زبان، متناسب با ویژگی‌های نوشتاری آن زبان مشکلات و پیچیدگی‌هایی دارد. زبان فارسی به‌دلیل وجود رسم‌الخط‌های مختلف، جملات بدون ترتیب، افعال مرکب، ابهامات معنایی و عدم نمایش مصوت‌ها مشکلاتی دارد. در این مقاله روشی مبتنی بر روش‌های آماری و یادگیری و اطلاعات و ویژگی‌های دستور زبانی جهت تشخیص مرز و نوع گروه‌های نحوی در متون فارسیِ فاقد علایم سجاوندی، ارائه شده است که در آن از روش‌های یادگیری ماشین بردار پشتیبان و میدان تصادفی شرطی استفاده شده است. در روش ارائه‌شده ویژگی‌های مختلف زبانی مرتبط با زبان فارسی استخراج و برای نخستین‌بار مورد بررسی و استفاده قرار گرفته‌اند. بهترین دقت به‌دست آمده توسط این سامانه، ۸۴/۰۲٪ بر اساس معیار F و ۸۷/۴۵٪، بر اساس تعداد برچسب‌های صحیح به کل در تعیین مرز، و ۷۸/۰۴٪، در تعیین مرز و نوع به صورت توأم، بوده است.

واژگان کلیدی: پردازش زبان طبیعی، تعیین مرز و نوع گروه‌های نحوی، برچسب‌گذاری مقولهٔ نحوی، ماشین بردار پشتیبان، میدان تصادفی شرطی، تبدیل متن به گفتار و ترجمهٔ ماشینی.

۱- مقدّمه

واحدسازی، از جمله مهم‌ترین و اساسی‌ترین مسائل در پردازش زبان‌های طبیعی[1] است. به‌طور کلی، واحدسازی فرآیند تقسیم متن به واژه‌ها، عبارات، جملات و یا واحدهای معنادار مشابه دیگر است و به سه نوع واحدسازی کلمات، عبارات و متون تقسیم می‌شود (کیانی و هَمکاران، ۸۷). هدف نهایی این پژوهش، واحدسازی و تعیین نوع گروه‌های نحوی[2] در متون فاقدِ علائم سجاوندی در زبان فارسی است.

تشخیص مرز و نوع عبارات نحوی[3] می‌تواند پیش‌پردازش سامانه‌های مهم دیگری در پردازش زبان‌های طبیعی باشد. از جملهٔ این سامانه‌ها می‌توان به سامانه‌های ترجمهٔ ماشینی، استخراج اطّلاعات، پرسش و پاسخ و سامانه‌های تبدیل متن به گفتار[4] اشاره کرد؛ به‌عنوان مثال، تشخیص صحیح مرز و نوع گروه‌های نحوی، می‌تواند در تعیین آهنگ و نوای مناسب در سامانه‌های تبدیل متن به گفتار، نقش به‌سزایی ایفا کند. نمونه‌ای از استفادهٔ این روّیه به‌عنوان یک پیش‌پردازش، در ترجمهٔ ماشینی از زبان پرتغالی به چینی، در (Oliveria et al., 2009) آمده است.

تعیین مرز و نوع گروه‌های نحوی، ممکن است امری ساده به‌نظر رسد؛ اما در زبان‌های مختلف، با توجّه به ویژگی‌های هر زبان، مشکلات متعددی دارد؛ به‌عنوان مثال در زبان فارسی، برخی واژه‌ها صورت املایی یکسان، اما معنای به‌طور کامل متفاوت دارند و همین امر می‌تواند منجر به اشتباه در تعیین مرز عبارات شود؛ همچنین عدم نمایش

[1] Natural Language Processing (NLP)
[2] Syntactic Phrase
[3] Phrase Chunking
[4] Text To Speech (TTS)

واکه‌های کوتاه از جمله نقش‌نمای اضافه، اتصال و یا عدم اتصال پیشوند و پسوندهای مختلف و عدم رعایت تفاوت فاصله و نیم‌فاصله در نوشتار فارسی می‌توانند امر تعیین مرز گروه‌های نحوی را با مشکل و ابهام روبه‌رو کنند. همچنین علایم سجاوندی می‌توانند در تعیین مرز عبارات نحوی مؤثر واقع شوند که استفاده از این ویژگی در متون فاقد علایم سجاوندی غیرممکن است.

در ادامهٔ این نوشته، ابتدا در بخش دوم، توضیح مختصری بر انواع گروه‌های نحوی، سپس در بخش سه، تاریخچه‌ای دربارهٔ این کار در زبان فارسی و سایر زبان‌ها، در بخش چهارم، روش‌های کلی کار شرح داده می‌شود. بخش پنجم، روش مورد استفاده و ویژگی‌های مورد استفاده را شرح می‌دهد. بخش ششم به مروری کوتاه دربارهٔ روش‌های یادگیری مورد استفاده تخصیص یافته و در بخش هفتم شرحی مختصر دربارهٔ پیاده‌سازی سامانه ارائه شده است. در بخش هشتم، نتیجهٔ آزمایش‌های صورت‌گرفته، آورده شده و در پایان نتیجه‌گیری و فعالیت‌های آتی ارائه می‌شوند.

۲- گروه‌های نحوی

به‌طور کلی، گروه‌های نحوی به پنج نوع گروه اسمی، فعلی، صفتی، قیدی و حرف اضافه‌ای طبقه‌بندی می‌شوند. هر گروه نحوی، مجموعه‌ای از کلمات است که نتوان آن را به گروه نحوی کوچک‌تر تجزیه کرد (باطنی، ۱۳۶۰). همچنین واژه‌هایی در هر زبان موجودند که مانند حروف ربط در هیچ یک از عبارات پنجگانهٔ فوق قرار نمی‌گیرند. در ادامه، به شرح مختصری دربارهٔ ساختار این گروه‌های نحوی می‌پردازیم.

گروه فعلی، مهم‌ترین عضو گزاره است و دست‌کم از یک بن فعل و شناسه درست می‌شود. به‌عبارت دیگر هر گروه فعلی، مجموعه‌ای از واژه‌هاست که در حکم فعل یک جمله بوده و یک هستهٔ فعلی دارد (باطنی، ۱۳۶۰).

فعل، از نظر ساختاری، بر سه نوع ساده، مرکب و پیشوندی است. فعل ساده، فعلی است که بن مضارع آن تنها، یک تکواژ باشد. فعل پیشوندی، فعلی است که از قرار گرفتن تکواژهایی چون "بر، در، باز، فرو، فرا، وا و ..." پیش از یک فعل ساده، ساخته می‌شوند. این تکواژها، گاهی معنای فعل را تغییر نمی‌دهند و گاه این معنی را به‌کُلی دگرگون می‌سازند. اگر به فعل ساده یا پیشوندی، یک یا چند تکواژ آزاد[1] اضافه شود، فعل مرکب ساخته می‌شود (باطنی، ۱۳۶۰).

فعل از دیدگاه زمانی، بر سه نوع کلی گذشته، حال و آینده، است. همچنین، هر یک از این افعال، زیرمجموعه‌هایی دارند. شخص فعل، توسط شناسه‌ها یا به عبارت دیگر ضمایر متصل فاعلی، مشخص می‌شود که بسته به زمان فعل می‌تواند متفاوت باشد.

گروه اسمی، مجموعه‌ای از واژه‌ها را تشکیل می‌دهد که یک هستهٔ اسمی و تعدادی وابستهٔ پیشین و یا پسین دارد (باطنی، ۱۳۶۰). وابسته‌های هستهٔ یک گروه اسمی می‌توانند صفت، اسم و ضمیر باشند و وجود آنها در یک گروه اسمی اختیاری بوده و به‌منظور شرح و توصیف هستهٔ گروه اسمی، قبل و یا بعد آن واقع می‌شوند که اگر پیش از اسم واقع شوند، به آنها وابستهٔ پیشین و در غیر این صورت، وابستهٔ پسین گفته می‌شود.

گروه قیدی، بخشی از جمله است که جمله یا جزیی از آن را مقید می‌کند، یا توضیحی به آن می‌افزاید و برخلاف اجزای اصلی جمله، می‌توان آن را حذف کرد. نشانهٔ اصلی شناخت قید، از راه معناست؛ بدین ترتیب که اگر از جمله حذف گردد، به معنای آن خللی وارد نمی‌شود.

گروه حرف اضافه‌ای، شامل گروهی از کلمات است که با یک حرف اضافه شروع می‌شوند.. نکته‌ای که بایستی بدان توجه شود، آن است که این حروف، زمانی که به‌عنوان حرف اضافه و شروع‌کنندهٔ یک گروه حرف اضافه‌ای به‌کار می‌روند، مستقل‌اند؛ درحالی‌که در زمان پیشوند بودن، جزئی از واژه هستند.

همان‌گونه که از نام گروه صفتی بر می‌آید، این گروه، مجموعه‌ای از واژه‌هاست که در کل، نقش صفت را ایفا می‌کنند. این گروه، غالباً، به‌عنوان مسند در جملات اسنادی استفاده می‌شود.

۳- کارهای مشابه قبلی

تعیین مرز و نوع عبارات نحوی، طی سال‌های اخیر، در زبان‌های مختلف، مورد بررسی قرار گرفته است. در زبان‌هایی مانند زبان انگلیسی، عربی، چینی، تایلندی، سوئدی، اسپانیایی، عبری، هندی و ... کارهای مشابهٔ بسیاری، با روش‌های متفاوت صورت گرفته‌اند که اغلب بسیار موفق بوده‌اند. شاید بتوان در زبان انگلیسی مهم‌ترین کارهای صورت‌گرفته را در این زمینه، CoNLL 2000 دانست. طی این کنفرانس، پیکره‌هایی بزرگ جهت آموزش و آزمون روش‌های ارائه شدهٔ مختلف، ایجاد شده است که بعد از این

[1] تکواژ آزاد، واژه‌ای است که معنای مستقلی دارد (باطنی، ۱۳۶۰).

اجلاس نیز، در کارهای مشابهٔ دیگری مورد استفاده قرار گرفته‌اند.

دو پژوهش بسیار موفق صورت‌گرفته در زبان انگلیسی، (Kudo and Matsumoto, 2001) و (Sha and Pereira, 2003) هستند که با استفاده از دو الگوریتم یادگیری ماشین بردار پشتیبان و میدان تصادفی شرطی، به دقّتی در حدود ۹۴٪ رسیده‌اند. استفاده از نکات و روش‌های ارائه شده در این دو کار، می‌تواند راهنمایی بسیار مفید جهت انجام پروژهٔ پیش رو باشد.

بعد از دو کار ذکر شده در بالا، پروژه‌های مُشابه بسیاری در این زمینه، در سایر زبان‌ها، به‌خصوص زبان‌های متعلق به خاور دور، نظیر چینی، ژاپنی و تایلندی صورت گرفته است. در این کارها، با توجه به ساختار به‌طور کامل متفاوت این زبان‌ها نسبت به زبان‌های غربی، همچون زبان انگلیسی، ویژگی اساسی مورد استفاده، ویژگی‌های مبتنی بر ساختار و رسم‌الخط خاص این‌گونه زبان‌ها بوده است. دو کار (Tesprasit et al., 2003) و (Hansakunbuntheung et al., 2005) در زبان تایلندی، نمونه‌هایی از این کارها هستند که کار دوم، در جهت بهبود کار اوّل صورت‌گرفته است. در این دو پژوهش، الگوریتم‌های یادگیری گوناگون مانند درخت تصمیم کارت[1] و الگوریتم رایپر[2] مورد ارزیابی واقع شده‌اند.

در (Diab et al., 2004)، یک سامانه برای جداسازی واژه‌ها، برچسب‌گذاری مقولهٔ نحوی و درنهایت، تعیین مرز عبارات نحوی و نوع آنها در زبان عربی معرفی شده است. برای مشخص‌سازی مرز و نوع عبارات نحوی، از روش برچسب‌گذاری IOB استفاده شده است. روش یادگیری، روشی بانظارت، مبتنی بر ماشین بردار پشتیبان است. ویژگی‌هایی که در اینجا، برای تعیین مرز و نوع گروه‌های نحوی مورد استفاده قرار گرفته است، عبارت از واژه‌ها و برچسب مقولهٔ نحوی آنها و برچسب IOB واحدهای قبلی هستند. پیکرهٔ مورد استفاده برای آموزش سامانه از Arabic TreeBank تأمین می‌شود و دقت این سامانه، ۹۲،۰۸٪ برآورد شده است.

در (Xu and Zhao, 2006)، روشی ترکیبی، جهت تشخیص مرز گروه‌های اسمی در متون به زبان چینی ارائه شده است که از روش یادگیری ماشین بردار پشتیبان و میدان تصادفی شرطی جهت این کار بهره برده است و برای رفع ابهام و تکمیل پردازش دو روش گفته شده، از تعدادی قاعده استفاده شده است. ویژگی‌های مورد استفاده عبارت از برچسب مقولهٔ نحوی، خروجی‌های قبلی و کلمات می‌باشند. پیکرهٔ استفاده‌شده در این آزمایش‌ها، شامل پانصدهزار کلمهٔ چینی است که سیصدهزار کلمه برای آموزش استفاده شده است. در بهترین وضعیت، کارآیی سامانه ۸۹،۲۷٪ برآورد شده است.

در IJCAI 2007، رقابتی برای ارائهٔ روشی، جهت تشخیص مرز گروه‌های نحوی و برچسب‌های مقولهٔ نحوی در متون به زبان‌های جنوب آسیایی که عبارتند از سه زبان هندی، بنگالی و تلوگو[3]، برقرار شد (SPSAL-2007) (Bharathi and Mannem, 2007). از بین هشت روش ارائه‌شده ((Pattabhi et al., 2007)، (Rao et al., 2007)، (Ekbal et al., 2007)، (Ravi et al., 2007)، (Satish and Kishore, 2007)، (Sandipan, 2007)، (Himanshu, 2007) و (Avinesh et al., 2007))، روش ارائه‌شده در (Avinesh et al., 2007)، با ۷۲٫۷۱٪، بیشترین کارآیی به‌دست‌آمده به‌طور متوسط در هر سه زبان را داشته است که در آن از روش‌های یادگیری میدان تصادفی شرطی و مدل مخفی مارکف[4] برای تعیین مرز و نوع عبارات نحوی استفاده شده است.

تعیین مرز و نوع گروه‌های نحوی، در متون فارسی، تاریخچه‌ای نسبتاً کوتاه دارد. در سال‌های اخیر در زبان فارسی، تحقیقات اندکی، در این زمینه صورت گرفته است. نخستین پژوهشی که به‌طور مستقیم در این زمینه انجام شده، (Shamsfard and Mousavi, 2008) است. در این مقاله، با اعمال شصت شرط و قاعده، سعی در تعیین مرز گروه‌های نحوی در زبان فارسی شده است. دقت برآورد شده، ۷۰٪ است. کار دیگر، (کیانی و شمس‌فرد، ۱۳۸۷) است که در ادامهٔ کار قبلی و توسط همان گروه صورت گرفته و این‌بار، با استفاده از شبکه‌های عصبی، تلاشی در جهت تشخیص مرز عبارات نحوی شده است. در این کار، از برچسب‌های نوع نحوی کلمه، دو کلمهٔ قبل و بعد برای تشخیص مرز عبارات نحوی استفاده شده است. دو راه‌کار گفته شده، در (Kiani et al., 2009)، به نوعی، ترکیب شده‌اند و این‌بار از روش‌های خوشه‌بندی FCM نیز کمک گرفته شده و بهترین دقت به‌دست آمده، ۸۵٫۷٪ بوده است.

از دیگر تحقیقات در این حوزه در زبان فارسی می‌توان به پژوهش انجام شده در (شریفی آتشگاه، ۱۳۸۸) اشاره کرد که در آن الگوریتمی جهت تهیهٔ دادگان درختی

[1] Cart
[2] Ripper
[3] Telugu
[4] Hidden Markov Model (HMM)

گروه‌های نحوی فارسی در چارچوب کمینه‌گرایی و به‌صورت نیمه‌خودکار ارائه شده است. در این پژوهش برای ارزیابی کیفیت دادهٔ تهیه شده فقط به گروه اسمی بسنده شده است.

همچنین فعالیت‌هایی در زمینه‌های نزدیک به زمینهٔ تعیین مرز و نوع عبارات نحوی در متون فارسی نیز در سال‌های اخیر صورت‌گرفته است؛ نظیر تعیین محل کسرهٔ اضافه (عیسی‌پور و همکاران، ۱۳۸۶)، (بی‌جن‌خان، ۱۳۸۴)، تعیین مرز کلمات (کیانی و شمس‌فرد، ۱۳۸۷)، تعیین مرجع ضمایر (موسوی و ثانی، ۱۳۸۷)، تعیین حدود جملات بر اساس تعیین مرز افعال (مبارکه و مینایی بیدگلی، ۱۳۸۷) و تعیین نقش‌های موضوعی جملات (شمس‌فرد و صدر موسوی، ۱۳۸۵).

همان‌گونه که بیان شد، تلاش‌های محدودی در جهت تعیین مرز و نوع عبارات نحوی در متون فارسی صورت گرفته است که متمرکز بر تعیین مرز و نه نوع گروه‌های نحوی بوده است.

۴- روش‌های کلی

۴-۱- روش‌های برچسب‌گذاری

پروژهٔ مورد نظر، یک پروژهٔ برچسب‌گذاری است و هدف نهایی، در وهلهٔ نخست، مشخص‌کردن مرز گروه‌های نحوی به کمک نوعی برچسب‌گذاری است. برای تعیین مرز گروه‌های نحوی، روش‌های برچسب‌گذاری متفاوتی وجود دارد که متداول‌ترین آنها دو روش داخل- خارج و روش ابتدا - انتهاست (Kudo and Matsumoto, 2001) (کیانی و شمس‌فرد، ۱۳۸۷). در ادامه به‌اختصار به شرح این روش‌ها می‌پردازیم.

۴-۱-۱- برچسب‌گذاری به روش درون- برون

یکی از متداول‌ترین روش‌های برچسب‌گذاری گروه‌ها و عبارات نحوی در متون با زبان‌های مختلف، روش درون- برون است. این روش دو نسخهٔ[1] کلی دارد که به روش IOB و IOE شناخته می‌شوند. در روش IOB، از سه نوع برچسب سادهٔ I (حرف نخست کلمهٔ Inside، به معنای درون)، B (حرف نخست کلمهٔ Begin، به معنای شروع) و O (حرف نخست کلمهٔ Outside، به معنای برون) استفاده شده است که هر یک در پایان یک واحد[2] قرار می‌گیرد. در روش IOE، به جای برچسب B، از برچسب E برای نشان دادن پایان یک گروه نحوی، استفاده می‌شود. این روش نخستین‌بار در (Lance et al., 1995) به‌کاررفته و بعد از آن به‌صورت یک استاندارد عرفی[3] برای تعیین مرز گروه‌های نحوی در آمده است. به مثال زیر، برگرفته از (کیانی و شمس‌فرد، ۱۳۸۷) توجه کنید:

علی (B) کتاب (B) حسن (I) را (O) به (B) مدرسه (I) برد (O) . (B)

نکتهٔ قابل اشاره در این‌جا، این است که این روش، پایان عبارات نحوی را مشخص نمی‌کند و تعیین و تشخیص مرز عبارات نحوی، بر اساس برچسب B صورت می‌گیرد که نشان‌دهندهٔ شروع یک عبارت نحوی و در صورت وجود، اتمام گروه نحوی قبلی است. این روش در موارد مشابهی مورد استفاده قرار گرفته که نمونه‌هایی مهم برای آن، (Shamsfard and Mousavi, 2008)، (کیانی و شمس‌فرد، ۱۳۸۷)، (Kudo and Matsumoto, 2001) و (عیسی‌پور و همکاران، ۱۳۸۶) هستند.

۴-۱-۲- برچسب گذاری ابتدا-انتها

این روش، پنج برچسب متفاوت B، E، I، S و O دارد. برچسب B، نشان می‌دهد که واحد جاری، شروع‌کنندهٔ یک عبارت نحوی است که بیش از یک واحد دارد. برچسب E، بیان می‌کند که واحد فعلی، انتهای یک عبارت نحوی است که بیش از یک واحد دارد. برچسب بعدی، برچسب I، نشان‌دهندهٔ این موضوع است که واحد جاری، در بین یک عبارت نحوی واقع شده است که بیش از دو واحد دارد. برچسب S، نشان‌دهندهٔ این است که واحد فعلی، یک عبارت نحوی با تنها یک واحد است. درنهایت، برچسب O، معرف این است که واحد مورد نظر، در بیرون هر عبارتی قرار گرفته است (Kudo and Matsumoto, 2001).

۴-۱-۳- برچسب‌گذاری نوع نحوی

در برچسب‌گذاری نوع نحوی برای هر کلمه نوع گروه نحوی به همراه برچسب تعیین مرز آن مورد استفاده قرار می‌گیرد. به این ترتیب مشخص می‌شود که هر واژه دارای چه موقعیتی نسبت به یک گروه نحوی بوده و نوع این گروه نحوی چیست. جدول (۱) برچسب‌های نوع نحوی مورد استفاده در این مقاله را نمایش می‌دهد:

[1] Version
[2] Token
[3] De facto Standard

(جدول ۱): برچسب‌های نوع نحوی مورد استفاده

برچسب نوع نحوی	مفهوم
NP	گروه اسمی
VP	گروه فعلی
JP	گروه صفتی
AP	گروه قیدی
PP	گروه حرف اضافه‌ای

بر این اساس برای هر واژه یک برچسب تعیین مرز و نوع گروه نحوی استخراج می‌شود. این برچسب استخراج‌شده به‌صورت ترکیبی از برچسب‌های بالا و هر یک از انواع برچسب‌های تعیین مرز نظیر IOB به‌عنوان خروجی تعیین مرز و نوع گروه نحوی است؛ به‌عنوان مثال چنانچه کلمه‌ای آغازگر یک عبارت نحوی اسمی باشد، برچسب B-NP برای آن درنظرگرفته می‌شود.

نمونه‌ای از این برچسب‌گذاری در کنار برچسب‌گذاری مرز IOB در زیر آمده است:

علی (B-NP) کتاب (B-NP) حسن (I-NP) را (O) به (B-PP) مدرسه (I-PP) برد (B-VP) . (O)

در این جمله گروه‌های نحوی بر اساس این برچسب‌گذاری به‌صورت زیر قابل تفسیر خواهد بود:

علی: گروه اسمی

کتاب حسن: گروه اسمی

به مدرسه: گروه حرف اضافه‌ای

برد: گروه فعلی

۴-۲- راه‌کارها برای تعیین مرز و نوع عبارات نحوی

برای تشخیص مرز گروه‌های نحوی و به‌طور کلی پردازش متن، روش‌هایی مورد استفاده قرار گرفته است که در این بخش به شرح و توضیح مختصر آنها می‌پردازیم. بر اساس (کیانی و شمس‌فرد، ۱۳۸۷) و (Kiani et al., 2009)، روش‌های مورد استفاده جهت تشخیص مرز عبارات نحوی، به چهار روش کلی روش مبتنی بر قواعد، روش مبتنی بر فرهنگ واژه‌ها، روش آماری و روش یادگیری ماشین تقسیم می‌شود. در بخش بعد، چگونگی عمل‌کرد و معماری سامانه‌های مشابه را به‌اختصار شرح می‌دهیم.

۴-۲-۱- روش مبتنی بر قواعد

سنتی‌ترین روشی که برای جداسازی و تشخیص مرز گروه‌های نحوی، به‌خصوص تشخیص مرز و نوع گروه‌های اسمی، در متون به زبان‌های مختلف استفاده شده است، روش مبتنی بر قواعد[1] است. این روش مبتنی بر اطلاعات و دانش تخصصی زبان‌شناسی است و بیشتر توسط زبان‌شناسان ارائه شده است. استفاده از درخت‌های اشتقاق[2] و داشتن مجموعه‌ای از قواعد برای شناسایی مرز عبارات نحوی، از جمله راه‌کارهای مورد استفاده در این روش است.

این روش، معایبی دارد. از جمله آن که به‌سختی قابل بهبود است؛ به‌خصوص برای زبان‌هایی چون زبان فارسی که پیچیدگی و استثناهای بسیاری دارد. دقّت این روش‌ها اندک است. همچنین این روش مبتنی بر دانش فرد خبره است و قواعد به‌صورت خودکار استخراج نمی‌شوند. در (Shamsfard and Mousavi, 2008)، نمونه‌ای از روش مبتنی بر قواعد جهت تعیین مرز و نوع گروه‌های نحوی متون فارسی آورده شده است.

۴-۲-۲- روش مبتنی بر فرهنگ‌واژه

در این‌گونه روش‌ها، مرزبندی عبارات، با تطابق عناصر جمله با مدخل‌های یک فرهنگ‌واژه صورت می‌پذیرد. میزان موفقیت این روش‌ها به پوشش فرهنگ‌واژه بستگی دارد و در رویارویی با یک کلمهٔ جدید، شکست می‌خورند (کیانی و شمس‌فرد، ۱۳۸۷). در این روش‌ها، برای کاهش میزان لغاتی که توسط فرهنگ‌واژه باید پوشش داده شوند، از روش‌های ریشه‌یابی واژه‌ها استفاده می‌شود؛ بدین ترتیب، واژه‌ها به ریشه‌ها تقلیل می‌یابند. در (Sanders and Taylor, 1995)یک روش مبتنی بر فرهنگ‌واژه برای زبان چینی آورده شده است.

۴-۲-۳- روش آماری

این روش‌ها، به‌کمک اطلاعات آماری و مدل زبانی، به قطعه‌بندی عبارات نحوی می‌پردازند. این اطلاعات آماری، از منابع زبانی، مانند پیکره‌های پردازش‌شده، اسناد وب، خروجی موتورهای جست‌وجو و ... به‌دست می‌آیند. این اطلاعات آماری، می‌توانند شامل تعیین عبارات پررخداد زبان، تعیین فراوانی و احتمال وقوع عبارات مختلف در متون گوناگون باشد (کیانی و شمس‌فرد، ۱۳۸۷).

[1] Rule Based
[2] Parse Tree

این روش‌ها، مزایا و معایبی دارند. نخستین مزیت این روش‌ها، این است که به دانش زبان‌شناسی زیادی نیاز ندارند. میزان موفقیت آنها، تا حدّ زیادی، به مرجع آمارگیری مورد استفاده بستگی دارد. مزیت دیگری که این روش‌ها از آن برخوردارند، قابل حمل‌بودن آنهاست. این روش‌ها به‌راحتی از زبانی به زبان دیگر، با ویژگی‌های مشترک، قابل تعمیم هستند (کیانی و شمس‌فرد، ۱۳۸۷).

۴-۲-۴- روش‌های یادگیری

این روش‌ها، مبتنی بر یادگیری ماشین، با یکی از الگوریتم‌های یادگیری ماشین می‌باشند. طی این روش‌ها، ماشین، خود اطلاعاتش را از منبع ورودی و نمونه‌های آموزشی به‌دست می‌آورد و آنها را یاد می‌گیرد. این اطلاعات، می‌توانند اطلاعات آماری مورد نیاز سامانه، مدل زبانی و یا قواعد نحوی و معنایی مورد استفاده در مرزبندی عبارات باشند (کیانی و شمس‌فرد، ۱۳۸۷).

عدم وجود پیکره‌های مناسب، با برچسب‌های مورد نیاز، در بسیاری از زبان‌ها از جمله زبان فارسی، دلیل ضعف این روش‌هاست. مشخص است که هر چه نمونه‌های آموزشی، بیش‌تر باشد، دقت این روش‌ها نیز بیش‌تر خواهد بود.

۵- راه‌کار پیشنهادی

در این بخش، بر اساس مطالعهٔ کارهای مشابه قبلی، روشی برای تعیین مرز و نوع گروه‌های نحوی ارائه می‌شود. این روش بر اساس روش‌های یادگیری بوده و در آن از دو الگوریتم ماشین بردار پشتیبان[1] ((Vapnik, 1995) و (Vapnik, 1998)) و میدان تصادفی شرطی[2] (Lafferty et al., 2001)، استفاده شده است. همچنین، برای افزایش کارآیی، تعداد محدودی قاعده قطعی نیز به نتیجهٔ الگوریتم‌های یادگیری اعمال می‌شود. روش برچسب‌گذاری درون-برون از نوع IOB و IOE و نیز روش ابتدا- انتها (IOEBS) مورد استفاده و ارزیابی واقع می‌شوند. ابتدا به ویژگی‌های مورد استفاده در این کار می‌پردازیم و سپس ویژگی‌های پیکرهٔ آموزشی و قواعد قطعی را شرح می‌دهیم.

۵-۱- ویژگی‌های مورد استفاده

انتخاب ویژگی‌های مناسب، در هر کار گروه‌بندی[3]، مانند هدف پیش رو، اهمیت بسیاری دارد. انتخاب صحیح ویژگی‌ها، موجب افزایش کارآیی سامانه می‌شود. در اینجا، به بررسی ویژگی‌ها و عوامل مناسب، جهت تعیین مرز و نوع گروه‌های نحوی می‌پردازیم و هر یک از این ویژگی‌ها را به‌اختصار توضیح می‌دهیم. لازم به‌ذکر است که با توجه به وابستگی موقعیت یک واژه در یک گروه نحوی به واژگان قبل و بعد از آن، برای هر واژه ویژگی‌های کلمات با شعاع دو در اطراف آن نیز به ویژگی‌های آن اضافه می‌شود.

۵-۱-۱- برچسب مقولهٔ نحوی

برچسب‌های مقولهٔ نحوی، از جمله مهم‌ترین ویژگی‌های استفاده شده در تعیین مرز و نوع گروه‌های نحوی است. این ویژگی در اکثر کارهای مشابه قبلی، گاه حتی به‌عنوان تنها ویژگی، مورد استفاده قرار گرفته است. این برچسب‌ها، تعیین کنندهٔ نوع یک واژه است. برچسب مقولهٔ نحوی انواع بسیار و زیرمجموعه‌های فراوانی دارد. در این مقاله، برخی از دسته‌های مربوط به مقوله‌های نحوی را بر اساس شباهت، با هم ادغام کرده‌ایم؛ زیرا در این پژوهش، پیکرهٔ بزرگی در اختیار نداریم و پوشش‌دادن همهٔ حالات مقوله‌های نحوی با چنین پیکره‌ای مشکل‌آفرین است؛ هم‌چنین برخی از برچسب‌های مقولهٔ نحوی تفاوت‌هایی را شامل می‌شوند که در تعیین مرز عبارات نحوی نقش چندانی ندارند. به‌عنوان مثال، تعداد بسیاری از برچسب‌های مقولهٔ نحوی به تمییز افعال از نظر زمان و شخص تخصیص یافته که در این پژوهش این دو، نقش چندانی ندارند و آنچه مؤثر است فعل‌بودن یک کلمه، و نه زمان و شخص آن است. در این کار، از چهارده برچسب نوع نحوی استفاده شده است. جدول (۲)، فهرست برچسب‌های مقولهٔ نحوی مورد استفاده را نمایش می‌دهد. برچسب مقولهٔ نحوی خود واژه و دو واژهٔ بعدی و دو واژهٔ قبلی به‌عنوان ویژگی‌های واژه در نظر گرفته می‌شوند.

۵-۱-۲- خروجی‌های سامانه برای کلمات قبلی

این ویژگی نیز، در بسیاری از کارهای مشابه قبلی استفاده شده است. استفاده از این ویژگی، به معنای استفاده از خروجی کلمات قبلی سامانه جهت تشخیص موقعیت و نوع واژهٔ فعلی است و نوعی بازخورد[4] برای سامانه مورد نظر محسوب می‌شود. در این کار، از برچسب خروجی دو واژهٔ قبل، به‌عنوان ویژگی استفاده شده است. اطلاع از موقعیت کلمات پیشین نسبت به یک گروه نحوی و نوع گروه آنها

[1] Support Vector Machine (SVM)
[2] Conditional Random Field (CRF)
[3] Classification
[4] Feed Back

می‌تواند در تعیین نوع گروه نحوی واژهٔ فعلی اثر مثبت داشته باشد. مقدار خروجی برچسب دو واژه قبل به‌عنوان مقدار این ویژگی در نظر گرفته می‌شود که بسته به نوع برچسب‌گذاری مورد استفاده در تعلیم سامانه متفاوت است.

(جدول ۲): برچسب‌های مقولهٔ نحوی مورد استفاده

برچسب مقولهٔ نحوی	مورد استفاده	مقدار ویژگی در بردار ویژگی
V	فعل	۰
N	اسم	۱
Ad	قید	۲
Intj	شبه‌جمله	۳
Pr	حرف اضافه	۴
Po	نشانهٔ مفعول	۵
Co	حرف عطف	۶
Det	حرف تعریف	۷
Nu	عدد	۸
Ps	لقب	۹
A	صفت	۱۰
Q	استفهام	۱۱
C	حرف ربط	۱۲
Exp	عبارت	۱۳

۵-۱-۳- ویژگی‌های آماری

استفاده از مدل زبانی آماری[1] چندتایی[2]، می‌تواند تأثیر مثبتی بر کارآیی سامانهٔ تعیین مرز و نوع گروه‌های نحوی داشته باشد. این ویژگی، در کارهای برچسب‌گذاری نظیر برچسب‌گذاری نوع نحوی و نیز پژوهش‌های مشابه قبلی نظیر (Murphy, 2003)، در تعیین مرز و نوع عبارات نحوی، مورد استفاده قرار گرفته است. ما در این مقاله، از ویژگی‌های دوتایی[3] و نیز اندیس کلمه، در پنجره‌ای به شعاع دو، استفاده کرده‌ایم. ویژگی‌های بیان‌شده از این جهت می‌توانند در تعیین مرز و نوع گروه‌های نحوی مؤثر واقع شوند که بیان‌گر احتمال واقع‌شدن گروهی از کلمات در کنار هم می‌باشند و بنابراین می‌توانند احتمال تشکیل یک گروه نحوی توسط گروهی از کلمات را بیان نمایند. مقدار به‌دست آمده توسط مدل ساخته‌شده برای هر کلمه مقدار این ویژگی است.

۵-۱-۴- ویژگی‌های ساختاری

ویژگی‌های ساختاری، ویژگی‌هایی هستند که بر اساس ساختار کلمات در زبان فارسی، مورد استفاده قرار گرفته‌اند. کلیهٔ این ویژگی‌ها از نظر مقداری باینری (دودویی) و مقدار آنها صفر یا یک است که مقدار یک به معنای برقرار بودن ویژگی است. این ویژگی‌ها عبارتند از:

۱- نشانه‌های افعال ماضی و مضارع و مستقبل، مانند پسوند "می". فعل‌بودن یک واژه احتمال قرارگرفتن آن به‌عنوان هستهٔ گروه فعلی را افزایش می‌دهد.

۲- نشانه‌های جمع (پسوندهای "ها"، "ان" و "ات")، که نشانهٔ اسم‌بودن یک واژه هستند. این نشانه می‌تواند نشانه‌ای از اسم‌بودن یک واژه است و احتمال تعلق آن را به گروه فعلی کاهش می‌دهد.

۳- نشانهٔ صفت عالی (پسوند "ترین")؛ صفت عالی از وابسته‌های پیشین گروه‌های اسمی است که بعد از آن هستهٔ گروه اسمی قرار می‌گیرد. با توجه به توضیح مذکور، صفات عالی می‌توانند نشانه‌ای جهت تعیین مرز گروه‌های اسمی باشد؛ زیرا کلمه واقع‌شده بعد از این صفت به‌حتم داخل یک گروه اسمی بوده و خود صفت عالی نیز شروع‌کننده و یا گاهی عضوی از گروه اسمی است.

۴- نشانهٔ صفت تفضیلی (پسوند "تر")؛ صفات تفضیلی در زبان فارسی، نشانه‌ای مشخص دارند (پسوند "تر") که به‌سادگی قابل تشخیص است. صفت تفضیلی به‌طور عمومی بعد از هستهٔ یک گروه اسمی قرار می‌گیرد؛ بنابراین می‌تواند نشانه‌ای جهت تشخیص مرز عبارات اسمی باشد؛ زیرا واژهٔ قبل از آن قطعاً متعلق به گروه اسمی بوده و هستهٔ آن است و خود کلمه نیز متعلق به گروه اسمی است.

۵- نشانهٔ نکره؛ در زبان فارسی، نشانهٔ معرفه‌بودن (مانند حرف "The" در زبان انگلیسی) وجود ندارد؛ اما نشانهٔ نکره‌بودن، موجود است. این نشانه، به‌صورت یک حرف "ی"، در انتهای اسم مورد نظر واقع می‌شود. البته لازم به ذکر است که نشانه‌های دیگری نیز وجود دارد؛ اما در اینجا تنها این نشانه در نظر گرفته شده است. این نشانه می‌تواند ایجاد ابهام نیز کند؛ زیرا می‌تواند نشانهٔ صفات نسبی باشد. با این وجود، این ابهام، ایجاد اختلال نخواهد کرد؛ زیرا در هر حالت نشانه‌ای برای اسم می‌باشد. اسم‌بودن یک واژه احتمال قرارگیری آن در گروه فعلی را کاهش می‌دهد.

۶- نشانهٔ اعداد ترتیبی نوع اول، مانند "اولین" به‌عنوان یک وابستهٔ پیشین برای هستهٔ گروه اسمی، واقع می‌شود و نکتهٔ قابل توجه آن است که بین ویژگی‌های صفت

[1] Statistical Language Model
[2] N-Gram
[3] Bigram

عالی، نشانهٔ نکره و اعداد ترتیبی، تنها یکی می‌تواند در عبارت حاضر شود. واژهٔ بعد از این کلمات حتماً متعلق به گروه نحوی اسمی و هسته آن است و این کلمات نیز شروع‌کنندهٔ گروه نحوی هستند.

۷- ضمایر اشاره؛ ضمایر اشاره به‌طور عمومی به جای مرجع خویش در جمله قرار می‌گیرند و به‌طور عمومی خود به تنهایی تشکیل یک گروه اسمی را می‌دهند. این ضمایر می‌توانند در تعیین مرز و نوع گروه‌های نحوی به‌خصوص گروه‌های اسمی تأثیر مثبت داشته باشند. این ضمایر عبارت‌اند از: "این"، "آن"، "همین"، "همان"، "آنها"، "اینها"، "همین‌ها" و "همان‌ها". به‌طور عمومی این کلمات یا به‌تنهایی تشکیل یک گروه نحوی را می‌دهند و یا شروع‌کنندهٔ گروه نحوی اسمی بوده و واژهٔ بعد از آنها متعلق به گروه اسمی است.

۸- ضمایر تأکیدی؛ ضمیر تأکیدی "خود"، می‌تواند شناسه‌ای برای شناسایی مرز برخی عبارات، به‌ویژه گروه‌های اسمی باشد. این ضمیر به‌طور عمومی برای تأکید به همراه ضمایر متصل، جانشین اسم شده و خود یک گروه اسمی را تشکیل می‌دهد.

۹- کلمات استفهام مانند "چرا" و "کجا"؛ می‌توانند نشانه‌ای از پایان و شروع گروه نحوی باشند.

۱۰- حروف شرط؛ مانند "اگر"، می‌توانند نشانه از پایان و شروع دو گروه نحوی باشند.

۱۱- نقش‌نمای مفعول؛ یعنی کلمه "را"، که نشانهٔ بسیار مناسبی جهت تعیین مرز پایانی یک گروه اسمی است و به‌سادگی قابل تشخیص است.

۱۲- حروف اضافه؛ حروف اضافه از پر کاربردترین حروف در زبان فارسی هستند. این حروف به‌طور عمومی نشانهٔ شروع یک گروه حرف اضافه‌ای هستند و در مواردی نادر، در داخل یک گروه (اغلب گروه حرف اضافه‌ای) واقع می‌شوند.

۱۳- حروف عطف؛ مانند "و" و "با". این حروف از آن جهت مؤثرند که نوع گروه نحوی دو واژه قبل و بعد آنها یکسان است.

۱۴- صفات مبهم؛ مانند "هر" و "همه"، که نشانه‌ای برای گروه اسمی می‌باشند.

۱۵- افعال اسنادی؛ مانند "است"، "بود" و "گشت". فعل بودن یک واژه احتمال قرار گرفتن آن به‌عنوان هستهٔ گروه فعلی را افزایش می‌دهد.

این ویژگی‌ها برای خود کلمه، دو کلمهٔ قبل و بعد مورد بررسی قرار می‌گیرند و درصورت وجود این نشانه‌ها عدد یک و در غیر این صورت، مقدار صفر در بردار ویژگی درج می‌شود.

استفاده از این ویژگی‌ها در کنار ویژگی برچسب مقولهٔ نحوی ممکن است در نگاه نخست حشو به‌نظر رسد؛ اما لازم به ذکر است که این ویژگی‌ها از یک سو می‌توانند مکملی برای ویژگی برچسب مقولهٔ نحوی باشند و از سوی دیگر نیاز به این ویژگی را برای سامانه کاهش می‌دهند. استخراج ویژگی برچسب مقولهٔ نحوی یک پیش‌پردازش توأم خطاست و استفاده از ویژگی‌های بیان‌شده در فوق، از یک سو مکمل این ویژگی بوده و از سوی دیگر نیاز به این پیش‌پردازش را حذف می‌کند. در پایان لازم به‌ذکر است که کلیهٔ ویژگی‌های ساختاری بیان‌شده برای نخستین‌بار در این پژوهش استخراج و مورد استفاده واقع شده‌اند.

۵-۲- پیکرهٔ آموزشی

پیکرهٔ مورد استفاده برای آموزش و آزمون الگوریتم یادگیری، پیکره‌ای کوچک با حدود شش‌هزار واژهٔ مجزاست که می‌توانند تکراری نیز باشند؛ هم‌چنین تعداد جملات این پیکره، ۵۸۹ جمله و فرهنگ لغات به‌کار رفته برابر ۲۹۱۳ واژه است. این پیکره شامل ۲۵۴۰ گروه نحوی است که ۹۷۴ گروه اسمی، ۸۳۰ گروه فعلی، ۵۲۰ گروه حرف اضافه‌ای، ۱۵۱ گروه قیدی و ۶۵ گروه صفتی را در بر می‌گیرد. پیکرهٔ مورد استفاده در این پژوهش توسط یک شخص خبره برچسب‌گذاری شده است. در این پیکره، پس از آمدن هر جمله، واژه‌های آن، به‌صورت مجزا فهرست شده‌اند و بعد از هر واژه، برچسب مقولهٔ نحوی و بعد از آن، برچسب تعیین مرز و نوع واژه‌ها به روش IOEBS، آورده شده است. نمونه‌ای از یک جملهٔ این پیکره در زیر ارائه شده است.

رنگ سبز را دوست دارم

رنگ	N	B-NP
سبز	N	E-NP
را	Po	O
دوست	N	B-VP
دارم	V	E-VP

۵-۳- قواعد قطعی

برخی واژه‌ها، برچسب ثابت و مشخصی دارند. این واژه‌ها، به‌طور عمومی پرتکرار بوده و از سوی دیگر، تنوع اندکی

دارند. می‌توانیم تعدادی قاعدهٔ مشخص و ساده را که به‌صورت قطعی، موقعیت واژه را نسبت به یک گروه نحوی مشخص می‌نماید، وضع نماییم و با اعمال این قواعد، بر خروجی الگوریتم‌های یادگیری، کارآیی را بهبود بخشیم. واژه‌های نقش‌نمای مفعول (را)، حروف ربط، حروف استفهام برچسب O دارند و بدین ترتیب، می‌توانند برچسب واژهٔ قبل و بعد خود را نیز مشخص نمایند. همچنین، حروف اضافه، به غیر از واژه‌ی "از"، به‌طور عمومی شروع‌کنندهٔ گروه حرف اضافه‌ای می‌باشند. این قواعد برای افزایش کارآیی بر خروجی الگوریتم‌های جداکننده مبتنی بر یادگیری اعمال می‌شوند. بنابراین قواعد زیر قابل استفاده است:

۱- اگر واژهٔ فعلی حرف اضافه است آنگاه واژه شروع‌کننده گروه نحوی حرف اضافه‌ای بوده و واژه بعد از آن متعلق به گروه حرف اضافه‌ای است.

۲- اگر واژهٔ فعلی نقش نمای مفعولی ("را") است آنگاه خود کلمه دارای برچسب O بوده و کلمهٔ قبلی متعلق به گروه اسمی است.

۳- اگر واژهٔ فعلی حرف استفهام یا حرف شرط باشد، آنگاه خود واژه دارای برچسب O بوده و کلمهٔ قبلی و بعدی به‌ترتیب کلمه پایانی و آغازین دو گروه نحوی مجزا می‌باشند.

۶- مروری بر تئوری الگوریتم‌های یادگیری

در این بخش به مروری کوتاه دربارهٔ تئوری الگوریتم‌های یادگیری مورد استفاده می‌پردازیم. این الگوریتم‌های یادگیری عبارتند از ماشین بردار پشتیبان و میدان تصادفی شرطی که در ادامه به شرح آنها می‌پردازیم.

۶-۱- ماشین بردار پشتیبان

الگوریتم یادگیری ماشین بردار پشتیبان، در سال ۱۹۹۲ توسط وپنیک[1] بر پایهٔ تئوری یادگیری آماری[2] ارائه شده است (Vapnik 1995). این الگوریتم دسته‌بندی، جزو شاخهٔ روش‌های هسته‌ای[3] است. شهرت این الگوریتم به‌دلیل موفقیت آن در تشخیص حروف دست‌نویس است که با شبکه‌های عصبیِ به‌دقّت تنظیم‌شده برابری می‌کند؛ یعنی ۱/۱٪ خطا. الگوریتم ماشین بردار پشتیبان، در پردازش زبان طبیعی، دسته‌بندی متن، تحلیل ساختار وابستگی نحوی و نیز تعیین مرز و نوع گروه‌های نحوی مورد استفاده قرار گرفته و نتایج خوبی را موجب شده است (Kudo and Matsumoto, 2001).

ایدهٔ اساسی الگوریتم ماشین بردار پشتیبان، جداسازی داده‌ها با استفاده از یافتن یک ابرصفحه[4] است؛ بدین معنا که با فرض این‌که دسته‌های دادهٔ مورد ارزیابی، به‌صورت خطی جداپذیر باشند، ابر صفحه‌هایی را با حداکثر حاشیه به‌دست می‌آورد که دسته‌ها را جدا کنند. در مسائلی که داده‌ها به‌صورت خطی جداپذیر نباشند، داده‌ها به فضایی با ابعاد بیش‌تر نگاشت می‌شوند، تا بتوان آنها را در این فضای جدید به‌صورت خطی جدا کرد. در الگوریتم ماشین بردار پشتیبان، هدف یافتن بهترین ابرصفحه‌ای است که دو دسته را از یکدیگر جدا می‌سازد.

فرض کنیم هر داده‌ای به یک گروه از دو گروه مثبت و یا منفی تعلق داشته باشد. داده‌های آموزشی را به‌صورت زوج مرتب‌های به فرم (x_i,y_i) به‌طوری که x_i بردار ویژگی برای نمونهٔ آموزشی iام و y_i برچسب تعیین گروه این نمونه است (که مقدار ۱ و یا ۱- را داراست) درنظر می‌گیریم. با فرض داشتن l دادهٔ آموزشی و n ویژگی برای بردار ویژگی می‌توانیم به‌صورت خلاصه داشته باشیم:

$$(x_i,y_i) \ni (1 \le i \le l \ , x_i \in R^n , y_i \in \{+1,-1\}) \quad (۱)$$

در روش ماشین بردار پشتیبان پایه، سعی می‌کنیم نمونه‌های مثبت و منفی را به‌کمک یک ابرصفحه از هم جدا کنیم که به‌صورت زیر نمایش می‌یابد:

$$(w.x) + b = 0 \ , (w \in R^n , b \in R) \quad (۲)$$

در الگوریتم یادگیری ماشین بردار پشتیبان، ابرصفحه بهینه، ابرصفحه‌ای است که نمونه‌ها را به دو دسته تقسیم کند؛ به قسمی که حاشیه را بیشینه کند شکل (۱). به‌طور دقیق می‌توانیم معادلهٔ دو خط تعیین کننده حاشیه را به‌صورت رابطهٔ زیر بنویسیم که در آن x بردار ویژگی، w بردار ضریب و b مقدار بایاس است:

$$w.x + b = \pm 1 \quad (۳)$$

و نیز می‌توانیم مقدار حاشیه را از رابطهٔ ۳-۴ به دست آوریم:

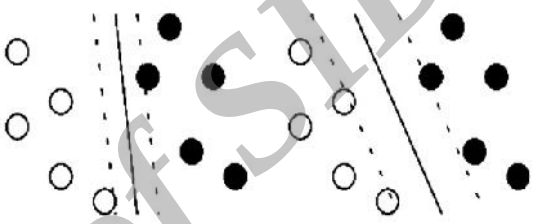
[1] Vapnik
[2] Statistical Learning Theory
[3] Kernel Methods
[4] Hyperplane

$$M= \frac{2}{||w||} \quad (۴)$$

برای بیشینه‌کردن حاشیه، با توجه به رابطهٔ ۴، بایستی مقدار ||w|| را کمینه کنیم؛ به‌عبارت دیگر، این مسأله معادل با حل مسألهٔ بهینه‌سازی زیر است:

$$\text{Minimize} : L(w) = \frac{1}{2}||w||^2 \quad (۵)$$
$$\text{Subject to} : y_i[(w.x_i) + b] \geq 1 , (i = 1, \dots, l)$$

نمونه‌های آموزشی‌ای را که به دو خط نقطه‌چین چسبیده‌اند بردار پشتیبان[1] می‌نامیم. تنها بردارهای پشتیبان در نمونهٔ آموزشی اهمیت دارند؛ بدین معنا که اگر همهٔ داده‌های آموزشی به‌جز بردارهای پشتیبان استخراج‌شده را حذف کنیم، باز هم می‌توانیم تابع تصمیم مناسب را به‌دست آوریم.



الف) حاشیهٔ بزرگ ب) حاشیهٔ کوچک

(شکل ۱)- دو ابرصفحهٔ ممکن برای جداسازی داده‌های آموزشی به دو رده

علاوه‌بر آنچه ذکر شد، الگوریتم ماشین بردار پشتیبان، توانایی رده‌بندی غیر خطی نیز دارد. این عمل با تعریف مسألهٔ بهینه‌سازی مورد نظر در یک فرم دوتایی، به‌کمک ضرب داخلی بردارهای ویژگی و استفاده از حقه هسته صورت می‌گیرد. به‌کمک در نظر گرفتن ضرب داخلی هر دو بردار ویژگی x_i و x_j، به صورت دوتایی تحت یک تابع کرنل (هسته) خاص $K(x_i,x_j)$، ماشین بردار پشتیبان (SVM)، می‌تواند فرضیه‌های غیر خطی را نیز پوشش دهد. ضرب داخلی، با توجه به اینکه نشانه‌ای از زاویهٔ بین دو بردار دارد، می‌تواند معیاری از شباهت دو بردار باشد. توابع هستهٔ معروف موجود عبارتند از خطی[2]، چندجمله‌ای[3]، RBF[4] و سیگموید[5].

[1] Support Vector
[2] Linear
[3] Polynomial
[4] Radial Basis Function
[5] Sigmoid

۶-۲- میدان تصادفی شرطی

الگوریتم میدان تصادفی شرطی، نخستین‌بار در (Lafferty et al., 2001) در سال ۲۰۰۱، ارائه شده است. این الگوریتم مبتنی بر احتمال شرطی و نظریهٔ گراف و برنامه‌سازی پویا[6] است و توانایی تشخیص ترتیبی از نمونه‌ها را دارد. اگر شبکه‌های عصبی قادر به تشخیص یک حرف در یک لحظه، در تشخیص دستخط، باشند، میدان تصادفی شرطی، قادر به تشخیص دنباله‌ای از حروف دستنویس در آن واحد است. در اینجا به‌اختصار، به توضیح این الگوریتم با فرض آگاهی از مفاهیم برنامه‌سازی پویا و زنجیره‌های مارکف می‌پردازیم.

فرض کنیم z دادهٔ خام، مانند یک متن، و x، مفاهیم و یا الگوها، نظیر دسته‌های یک متن باشند. برای هر z، فرض می‌کنیم x، مشروط بر z، یک زنجیر مارکف غیر مستقیم می‌باشد و داریم (Truyen and Phung 2008):

$$P(x|z)= \frac{1}{Z(z)} \prod_{t\in[1,T-1]} \psi_t(x_t,x_{t+1},z) \quad (۶)$$

که داریم :

$$Z(z)= \sum_x \prod_{t\in[1,T-1]} \psi_t(x_t,x_{t+1},z) \quad (۷)$$
$$\psi_t(x_t,x_{t+1},z) = \exp\left(\sum_{k\in[1,K]} w_k f_k(x_t,x_{t+1},z)\right) \quad (۸)$$

فرض کنیم :

$$F_k(x,z) = \sum_{t\in[1,T-1]} f_k(x_t,x_{t+1},z) \quad (۹)$$

در این صورت رابطه‌ی ۶، به صورت زیر ساده خواهد شد:

$$P(x|z) = \frac{1}{Z(z)} \exp\left(\sum_{t\in[1,T-1]} \sum_{k\in[1,K]} w_k f_k(x_t,x_{t-1},z)\right) \quad (۱۰)$$
$$= \frac{1}{Z(z)} \exp\left(\sum_{k\in[1,K]} w_k F_k(x_t,x_{t-1},z)\right)$$

در اینجا دو مسألهٔ اساسی به‌وجود می‌آید، آموزش[7] و رمزگشایی[8]. در آموزش، با مجموعه‌ای از داده‌های مشهود و نمونه، سعی در تخمین پارامتر w می‌شود. در رمزگشایی، الگوی بهینه‌ی $x^* = (x_1^*,\dots,x_T^*)$ برای یک مجموعه دادهٔ خام z به‌دست می‌آید.

[6] Dynamic Programming
[7] Learning
[8] Decoding

ابتدا به مبحث آموزش می‌پردازیم. فرض کنیم مجموعه دادهٔ D به‌صورت زیر موجود باشد :

$$D=(\tilde{x}^1,z^1);(\tilde{x}^2,z^2);\cdots;(\tilde{x}^D,z^D) \quad (11)$$

در میدان تصادفی شرطی، از مفهوم گرادیان برای آموزش و بهینه‌سازی استفاده می‌شود. بر اساس دو رابطهٔ زیر عمل آموزش صورت می‌گیرد:

$$\tilde{\mathbb{E}}[F_k] = \frac{1}{D}\sum_{d\subset[1,D]}\sum_{t\subset[1,T\ \ 1]} f_k(\tilde{x}_t,\tilde{x}_{t+1},z^d) \quad (12)$$

$$\begin{aligned}\mathbb{E}[f_k] &= \frac{1}{D}\sum_{d\in[1,D]}\sum_{x} P(x|z^d)\sum_{t\in[1,T-1]} f_k(x_t,x_{t+1},z^d) \\ &= \frac{1}{D}\sum_{d\in[1,D]}\sum_{x} P(x|z^d)\sum_{t\in[1,T-1]} f_k(x_t,x_{t+1},z^d) \\ &= \frac{1}{D}\sum_{d\in[1,D]}\sum_{t\in[1,T-1]}\sum_{x_t,x_{t+1}} P(x_t,x_{t+1}|z^d) f_k(x_t,x_{t+1},z^d)\end{aligned} \quad (13)$$

در زمان آزمون، زمانی‌که پارامتر تخمین زده شد، دادهٔ خام z داده شده و نیاز است تا x^* را به دست بیاوریم. یافتن این مقدار در بخش ۶-۲-۱، شرح داده شده است.

۷- پیاده‌سازی

در پیاده‌سازی نرم‌افزار تعیین مرز و نوع عبارات نحوی در متون فارسی، از زبان برنامه‌نویسی C++ استفاده شده است. این برنامه برای کار در سیستم‌عامل ویندوز هفت توسعه یافته است. برای استفاده از الگوریتم‌های یادگیری میدان تصادفی شرطی و ماشین بردار پشتیبان و نیز برای تولید مدل آماری زبانی، از ابزارهای آماده استفاده شده است. برای الگوریتم ماشین بردار پشتیبان از LibSVM و برای الگوریتم میدان تصادفی شرطی از ابزار PocketCRF استفاده شده است. برای تولید مدل آماری زبانی نیز از ابزار CMU-Cambridge استفاده شده است.

۸- نتایج آزمایش‌ها

ابتدا ضرایب مورد نیاز الگوریتم ماشین بردار پشتیبان، بر اساس آزمایش، به‌دست آمده که برای ضریب C، مقدار ۱۰، و برای گاما، مقدار ۰/۰۴ به‌دست آمده است. در این کار، کرنل RBF برای این الگوریتم، مبنا قرار گرفته و البته، سایر کرنل‌ها نیز مورد بررسی قرار گرفته‌اند. در ذیل، به گزارش آزمایش‌های صورت‌گرفته می‌پردازیم.

لازم به‌ذکر است که در کلیهٔ ارزیابی‌های صورت‌گرفته از ارزیابی متقاطع ۱۰تا شده‌ای[1] استفاده شده است؛ که در آن کلیهٔ پیکرهٔ آموزشی به ده قسمت به‌طور تقریبی برابر تجزیه شده و نه قسمت برای آموزش و قسمت باقیمانده جهت آزمون مورد استفاده قرار می‌گیرد و این عمل ۱۰ بار تکرار شده و درنهایت میانگین خطا به‌عنوان خطای مورد ارزیابی درنظر گرفته می‌شود.

برای دسته‌بندی با برچسب‌های مختلف تعداد دسته‌های مختلفی استفاده شده است. در دسته‌بندی برای تعیین مرز عبارات، تعداد رده‌ها با استفاده از هر یک از برچسب‌گذاری‌های IOB و IOE برابر ۳ (رده B، I و O بر هر کلمه) و با استفاده از برچسب‌گذاری IOEBS برابر ۵ (رده B، I، O، E و S برای هر کلمه) است. در مسألهٔ تعیین مرز و نوع عبارات تعداد رده‌ها بیشتر خواهد بود چنانچه برای هر گروه، دارای ۲ یا ۴ ردهٔ تعیین مرز (بسته به انتخاب نوع برچسب‌گذاری تعیین مرز) هستیم (به‌عنوان مثال برای گروه اسمی با انتخاب برچسب‌گذاری IOB دارای برچسب‌های NP-B، NP-I) و یک رده برای برچسب O در نظر گرفته می‌شود. بنابراین با توجه به داشتن ۵ گروه نحوی مختلف، با استفاده از برچسب‌زنی IOB و IOE برای تعیین مرز دارای ۱۱ رده و با استفاده از برچسب‌زنی IOEBS دارای ۲۱ رده هستیم. برای بررسی کارآیی سامانه، نیازمند معیاری متناسب با کارکرد آن سامانه هستیم. سه معیار مهمی که برای بررسی کارآیی چنین سامانه‌هایی، در کارهای مشابه، مورد استفاده قرار گرفته است، معیار دقت[2]، یادآوری[3] و امتیاز F، موسوم به Fβ=1. معیار دقت، تعداد عبارت‌های صحیح تشخیص داده شده توسط سامانه و معیار یادآوری، درصد برچسب‌های تعیین شده در پیکره را که توسط برنامهٔ مورد نظر، تشخیص داده شده است، نشان امتیاز F، به‌صورت زیر محاسبه می‌شود، می‌دهند (Rijsbergen 79):

$$F=\frac{2\times percision\times recall}{percision+recall} \quad (38)$$

[1] 10fold-cross-validation
[2] Precision
[3] Recall

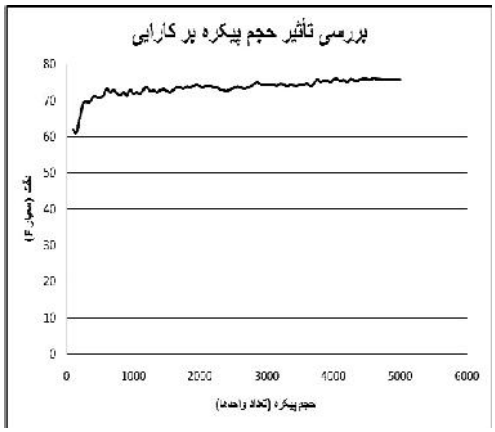
این معیارها در بسیاری از پژوهش‌های پیشین نظیر (Kudo and Matsumoto, 2001) و (Diab et al., 2004) مورد استفاده قرار گرفته‌اند.

۸-۱- بررسی تأثیر ویژگی‌ها و روش‌های برچسب‌گذاری

در بخش ۵-۱، ویژگی‌هایی معرفی شده‌اند. ممکن است برخی از این ویژگی‌ها، مفید واقع نشوند. طی آزمایش‌های بسیاری، تأثیر هر ویژگی مورد بررسی قرار گرفت. این آزمایش‌ها، بدین ترتیب صورت گرفت که هر ویژگی به ترکیب ویژگی‌های قبلی افزوده شد و در صورت افزایش کارآیی، به‌عنوان ویژگی مفید درنظر گرفته شد و در غیر این‌صورت کنار گذاشته شد. اگر یک ویژگی، موجب کاهش کارآیی شد، در مرحلهٔ بعدی، ویژگی بعدی را هم به ترکیب بدون ویژگی مضر قبلی و هم به ترکیب دارای ویژگی مضر قبلی می‌افزاییم؛ زیرا ممکن است تأثیر یک ویژگی، در کنار سایر ویژگی‌ها مشاهده شود. با اعمال این آزمایش‌ها، مشاهده شد که ویژگی‌های منتخب مفید بوده و تنها، ویژگی‌های آماری، برای کار فعلی مناسب نمی‌باشند. شرح کامل این آزمایش‌ها در (سلیمی بدر، ۱۳۸۹) آورده شده است. همچنین در طی این آزمایش‌ها، روش‌های برچسب‌گذاری IOB، IOE و IOEBS، برای تعیین مرز و تعیین مرز و نوع گروه‌های نحوی، مورد ارزیابی قرار گرفته‌اند که بر اساس این آزمایش‌ها، روش برچسب‌گذاری IOE برای تعیین مرز و روش برچسب‌گذاری IOB برای تعیین مرز و نوع گروه‌های نحوی کارآیی مناسب‌تری را به همراه داشته‌اند.

۸-۲- بررسی تأثیر حجم پیکره بر کارآیی

حجم پیکرهٔ آموزشی، می‌تواند تأثیر به‌سزایی در کارآیی سامانهٔ مبتنی بر الگوریتم یادگیر با نظارت داشته باشد. در حال حاضر، پیکره‌ای نسبتاً کوچک، برای آموزش سامانه تعیین مرز و نوع در اختیار است. بررسی تأثیر حجم و بزرگی پیکره بر کارآیی سامانه نشان می‌دهد که با بزرگ‌ترشدن حجم پیکره، آیا کارآیی سامانه افزایش خواهد یافت و یا خیر؟ شکل (۲)، تأثیر اندازهٔ پیکره را بر روی سامانهٔ تعیین مرز با کمک الگوریتم ماشین بردار پشتیبان نشان می‌دهد. به کمک این نمودار، می‌توان تأثیر حجم پیکره را بر روی الگوریتم میدان تصادفی شرط و تعیین مرز و نوع گروه‌های نحوی، به‌دست آورد. در این آزمایش، پیکره‌ای ثابت، با تعداد واحدهای هزار را به‌عنوان پیکرهٔ آزمون در نظر می‌گیریم و پیکرهٔ آموزشی را از پنجاه واحد تا پنج‌هزار واحد، با گام پنجاه واحد، افزایش داده و دقت سامانهٔ تعیین مرز با الگوریتم ماشین بردار پشتیبان، با روش برچسب‌گذاری IOE را بر حسب معیار F به‌دست آورده‌ایم. با توجه به شکل (۱)، متوجه می‌شویم که این نمودار، صعودی است و با وجود افت‌وخیزهایی، رو به افزایش است. می‌توانیم نتیجه بگیریم، با افزایش حجم پیکره، کارآیی سامانه، افزایش خواهد یافت؛ اما آهنگ افزایش آن از حدی به بعد کاهش یافته و به سمت ثبات می‌رود.



(شکل ۲)- بررسی تأثیر حجم پیکره بر کارآیی

۸-۳- مقایسهٔ کارآیی دو الگوریتم میدان تصادفی شرطی و ماشین بردار پشتیبان

جدول (۳)، نمایندهٔ کارآیی سامانه، با استفاده از هر یک از دو الگوریتم یادگیری پیشنهادی برای تعیین مرز گروه‌های نحوی و جدول (۴)، بیان‌گر کارآیی سامانه برای تعیین مرز و نوع گروه‌های نحوی به‌صورت توأم است. در این آزمایش‌ها، از ترکیب بهترین ویژگی‌ها استفاده شده است. همچنین، برای ارزیابی سامانه، از روش ارزیابی متقاطع چندباره[1] ده‌تایی، استفاده شده است.

مشاهده می‌شود که کارآیی سامانه در تعیین مرز، با استفاده از الگوریتم میدان تصادفی شرطی بیشترین کارآیی را داشته؛ اما، در تعیین مرز و نوع به‌صورت توأم، با استفاده از الگوریتم ماشین بردار پشتیبان، به بیش‌ترین کارآیی خود رسیده است.

[1] k-fold-cross validation

۸-۴- بررسی کارآیی هسته‌های الگوریتم ماشین بردار پشتیبان

در این بخش، هسته‌های دیگر الگوریتم ماشین بردار پشتیبان را مورد بررسی قرار می‌دهیم. لازم به ذکر است که این هسته‌ها نیز پارامترهایی دارند که تعیین مقدار آنها، می‌تواند در کارآیی سامانه تأثیرگذار باشد. در اینجا، تنها با مقادیر پیش‌فرض، این بررسی‌ها را انجام می‌دهیم. جدول (۵)، کارآیی سامانه تعیین مرز در هر یک از این هسته‌ها را نشان می‌دهد. آزمایش‌های تعیین مرز، با روش برچسب‌گذاری IOE تعیین مرز و نوع، با روش برچسب‌گذاری IOB صورت گرفته‌اند.

مشاهده می‌شود که هسته‌های دیگر، کارآیی کمتری نسبت به هسته‌ی RBF داشته‌اند. ممکن است با بهینه‌کردن ضرایب مربوط به این هسته‌ها، کارآیی بهبود یابد که این کار به کارهای آینده محول می‌شود.

۸-۵- ترکیب الگوریتم‌ها

ترکیب دو الگوریتم، به‌صورت متوالی ممکن است موجب بهبود کارآیی سامانه شود. در اینجا، این دو الگوریتم را به‌صورت متوالی، با هم و با خود ترکیب و تأثیر این ترکیب را بر روی کارآیی سامانه مشاهده می‌کنیم. جدول (۶)، نتایج تأثیر این ترکیبات را به نمایش می‌گذارد. در اینجا لازم است، بیان شود که برای ساخت پیکرهٔ آموزشی، پیکرهٔ اصلی را به سه قسمت تقسیم کرده، هر دو قسمت به‌عنوان پیکرهٔ آموزشی و قسمت سوم، به‌عنوان پیکرهٔ آزمون درنظر گرفته شد و خروجی این آزمون‌ها، با هم ترکیب شده و پیکرهٔ آموزشی را تشکیل می‌دهد. در اینجا، تعیین مرز مورد بررسی قرار گرفته است و از بهترین ترکیب ویژگی‌ها و روش برچسب‌گذاری IOE استفاده شده است.

(جدول ۳): کارآیی سامانه تعیین مرز، بر اساس معیارهای دقت، یادآوری و F

معیار F	یادآوری	دقت	ویژگی‌های مبتنی بر POS	الگوریتم یادگیری
۷۶٫۶۸	۷۴٫۸۷	۷۸٫۹۱	بله	SVM
۸۳٫۹۷	۸۲٫۱۷	۸۵٫۸۶	بله	CRF
۷۰٫۲۳	۶۵٫۷۹	۷۵٫۹۵	خیر	SVM
۷۶٫۸۷	۷۳٫۹۰	۷۹٫۹۱	خیر	CRF

(جدول ۴): کارآیی سامانه تعیین مرز و نوع، بر اساس معیار دقت

الگوریتم یادگیری	کارآیی (دقت)
SVM	۷۸٫۰۴
CRF	۷۶٫۶۰

(جدول ۵): مقایسهٔ کارآیی هسته‌های مختلف الگوریتم ماشین بردار پشتیبان

کارآیی در تعیین مرز و نوع با معیار تعداد صحیح به کل	کارآیی در تعیین مرز با معیار تعداد صحیح به کل	کارآیی در تعیین مرز با معیار F	کارآیی در تعیین مرز با معیار یادآوری	کارآیی در تعیین مرز با معیار دقت	هسته
۷۰٫۱۵	۷۴٫۸۴	۷۱٫۵۵	۶۸٫۸۹	۷۴٫۹۴	خطی
۷۲٫۰۹	۷۳٫۰۸	۷۱٫۴۲	۷۱٫۵۹	۷۱٫۵۶	چند جمله‌ای
۷۸٫۰۴	۷۸٫۷۸	۷۶٫۶۸	۷۴٫۸۷	۷۸٫۹۱	RBF

همان‌طور که در جدول (۶)، مشاهده می‌شود، ترکیب متوالی این الگوریتم‌ها، افزایش کارآیی را در پی نداشته است. الگوریتم‌ها را می‌توان به نوعی دیگر نیز با یکدیگر ترکیب کرد، از طریق موازی. این کار زمانی میسّر است که بیش از دو الگوریتم یادگیری موجود باشد و از طریق رأی‌گیری وزن‌دار، بین برچسب‌های پیشنهادی از سوی الگوریتم‌های مختلف، یکی را انتخاب می‌کنیم. در اینجا، با توجه به در اختیار داشتن دو الگوریتم، این کار، معنای مناسبی ندارد؛ لذا این کار انجام نشده است.

(جدول ۶): نتایج حاصل از ترکیب متوالی الگوریتم‌های یادگیری

کارآیی بر اساس نسبت برچسب صحیح به کل	کارآیی بر اساس معیار F	کارآیی بر اساس یادآوری	کارآیی بر اساس دقت	الگوریتم دوم	الگوریتم اول
۷۸٫۵۲	۷۷٫۶۴	۷۱٫۱۰	۸۵٫۵۰	SVM	CRF
۸۵٫۸۵	۸۳٫۷۸	۷۹٫۲۹	۸۸٫۸۱	CRF	CRF
۷۹٫۱۱	۷۰٫۷۹	۶۱٫۶۲	۷۹٫۷۷	CRF	SVM
۷۶٫۳۱	۷۵٫۰۴	۶۸٫۶۱	۸۲٫۸۱	SVM	SVM

۸-۶- بررسی تأثیر قواعد قطعی

واحد مبتنی بر قاعده، در پایان کار، با اعمال پاره‌ای قواعد حتمی و ساده که پیش از این در بخش ۵-۳ ارائه شدند، اصلاحاتی بر روی خروجی سامانه انجام می‌دهد. جدول (۷)، تأثیر اعمال این واحد را بر کارآیی سامانه نشان می‌دهد. این آزمایش‌ها برای تعیین مرز، با الگوریتم میدان تصادفی شرطی و با در نظر گرفتن همهٔ ویژگی‌ها، صورت گرفته است. این آزمایش‌ها نشان می‌دهد که این قواعد ساده، می‌توانند کارآیی سامانه را اندکی بهبود بخشند.

(جدول ۷): تأثیر اعمال واحد مبتنی بر قاعده بر کارآیی سامانه

کارآیی بر اساس معیار F	کارآیی بر اساس معیار یادآوری	کارآیی بر اساس معیار دقت	نوع برچسب‌گذاری	استفاده از واحد مبتنی بر قاعده
۸۲٫۷۴	۸۱٫۷۶	۸۳٫۷۴	IOB	بله
۸۳٫۳۵	۸۱٫۶۵	۸۵٫۱۲	IOB	خیر
۸۴٫۰۲	۸۲٫۱۷	۸۵٫۹۵	IOE	بله
۸۳٫۹۷	۸۲٫۱۷	۸۵٫۸۶	IOE	خیر

۸-۷- بررسی سایر الگوریتم‌های یادگیری

با کمک نرم‌افزار وکا[1]، الگوریتم‌های یادگیری دیگر را برای تعیین مرز مورد بررسی قرار می‌دهیم. این پژوهش، برای کارهای آینده است. به کمک این آزمایش‌ها، انواع الگوریتم‌های یادگیری را مورد بررسی قرار داده‌ایم. معیار بررسی کارآیی، در این آزمایش‌ها، معیار دقت است. در اینجا، مبنا را تشخیص مرز قرار داده‌ایم و روش برچسب‌گذاری، IOE است.

مشاهد می‌شود که در این آزمایش‌ها، درخت‌های تصمیم و نیز برخی روش‌ها استخراج‌گر قاعده، کارآیی به‌نسبه خوبی در بر داشته‌اند. در این میان، درخت کارت[2] و الگوریتم رایپر[3]، نسبت به سایر الگوریتم‌ها، بهتر عمل کرده‌اند. کارآیی دو الگوریتم ماشین بردار پشتیبان و میدان تصادفی شرط، در شرایط یکسان، به ترتیب برابر ۷۸٫۷۸ و ۸۷٫۲۴ درصد بوده است. در این آزمایش‌ها نیز از ارزیابی متقاطع چندباره[4] ده‌تایی، استفاده شده است. نتیجهٔ این آزمایش‌ها، در جدول (۸)، آورده شده است.

[1] www.cs.waikato.ac.nz/ml/weka/
[2] Cart
[3] Ripper
[4] k-fold-cross validation

۹- نتیجه‌گیری و فعالیت‌های آتی

با توجه به آنچه در این نوشته آمد، می‌توان نتیجه گرفت، دو الگوریتم میدان تصادفی شرطی و ماشین بردار پشتیبان، برای تعیین مرز و نوع گروه‌های نحوی، در متون فارسی می‌توانند مؤثر واقع شوند. لازم به‌ذکر است که با بررسی تعداد دیگری الگوریتم یادگیری، به کمک نرم‌افزار وکا، مشاهده می‌شود که الگوریتم میدان تصادفی شرط، برای تعیین مرز، و الگوریتم ماشین بردار پشتیبان، برای تعیین مرز و نوع، در بهترین وضعیت خود، بیشترین کارآیی را داشته‌اند. بعد از این الگوریتم، درخت تصمیم کارت و روش رایپر، بیشترین دقت را به همراه دارند. بر اساس آزمایش‌های صورت‌گرفته، روش برچسب‌گذاری می‌تواند تأثیر نسبتاً محسوسی بر کارآیی سامانه داشته باشد و روش برچسب‌گذاری IOE، برای تعیین مرز و روش برچسب‌گذاری IOB، برای تعیین مرز و نوع گروه‌های نحوی، بهترین کارآیی را به همراه داشته‌اند. براساس آزمایش‌های انجام‌شده، ویژگی‌های مدل زبانی، تأثیر مثبتی بر کارآیی سامانه نداشته‌اند. همچنین، بر مبنای آزمایش‌ها، عدم استفاده از ویژگی برچسب مقولهٔ نحوی، دقّت کمتری نسبت به حالت استفاده از این ویژگی را فراهم می‌نماید؛ با این وجود این کارآیی، کارآیی به‌نسبه مناسبی است. افزودن تأثیر واحد مبتنی بر قاعده که تعداد اندکی قاعده قطعی را بررسی می‌کند، کارآیی سامانه را اندکی بهبود بخشیده است؛ لذا، استفاده از چنین قواعد ساده و اندکی مفید بوده است و این واحد، توانسته خروجی الگوریتم یادگیری را اصلاح کند؛ بنابراین، از اقداماتی که می‌توان در راستای این پروژه انجام داد، اضافه‌کردن تعداد بیشتری قاعده و توسعهٔ واحد مبتنی بر قاعده است. همچنین، با توجه به بخش ۷-۲، افزایش حجم پیکره، می‌تواند کارآیی سامانه را افزایش دهد و با توسعهٔ این پیکره، می‌توان امید به افزایش کارآیی سامانه تعیین مرز و نوع گروه‌های نحوی داشت؛ بنابراین، گسترش و توسعهٔ پیکرهٔ آموزشی، می‌تواند منجر به افزایش کارآیی سامانه شود.

از فعالیت‌هایی که می‌توانند در راستای تحقیق فعلی، در جهت تکمیل و توسعهٔ آن و یا بهره‌گیری از آن، صورت گیرند، می‌توانیم به موارد ذیل اشاره کنیم :

۱- بررسی سایر الگوریتم‌های یادگیری ماشین؛ به‌عنوان مثال بررسی دقیق‌تر الگوریتم درخت

یادگیری کارت[1] و روش یادگیری رایپر[2] و یا آنتروپی بیشینه.

۲- ترکیب موازی چند روش یادگیری از طریق رأی‌گیری وزن‌دار است.

۳- پیاده‌سازی نرم‌افزاری جهت تعیین هسته و وابسته‌های گروه‌های نحوی.

۴- تعیین محل کسرهٔ اضافه.

۵- بهره‌گیری در سامانه‌های ترجمهٔ ماشینی و تبدیل متن به گفتار، به‌عنوان پیش‌پردازش.

درنهایت، بر اساس آزمایش‌های مختلف صورت‌گرفته، بیشترین کارآیی برآورد شده، به‌کمک این نرم‌افزار، در تعیین مرز ۸۴٫۰۲٪ بر اساس معیار F و ۸۷٫۴۵٪، براساس تعداد برچسب‌های صحیح به کل است که با استفاده از الگوریتم میدان تصادفی شرط و استفاده از واحد مبتنی بر قاعده، با روش برچسب‌گذاری IOE و استفاده از همهٔ ویژگی‌های منتخب، به جز ویژگی‌های حاصل از مدل زبانی، به دست آمده است. در تعیین مرز و نوع نیز، با برچسب‌گذاری IOB و با شرایطی مشابه شرایط مذکور برای تعیین مرز، با الگوریتم ماشین بردار پشتیبان، به دقتی برابر ۷۸٫۰۴٪ رسیده‌ایم.

(جدول ۸): بررسی کارآیی برخی الگوریتم‌های یادگیری دیگر با نرم‌افزار وکا

نوع الگوریتم یادگیری	الگوریتم یادگیری	دقت محاسبه‌شده (بر حسب درصد)
درخت تصمیم	J48	۸۳٫۸۵
درخت تصمیم	Random Forest	۸۱٫۶۴
درخت تصمیم	Random Tree	۷۱٫۵۸
درخت تصمیم	J48graft	۸۴٫۳۴
درخت تصمیم	SimpleCart	۸۴٫۵۸
درخت تصمیم	BFTree	۸۳٫۱۷
درخت تصمیم	Id3	۷۵٫۳۱
درخت تصمیم	DecisionStump	۵۵٫۰۰
درخت تصمیم	FT	۷۸٫۴۲
درخت تصمیم	LADTree	۷۷٫۲۲
درخت تصمیم	RFPTree	۸۲٫۴۸
قاعده	JRIP (RIPPER)	۸۴٫۲۰
قاعده	PART	۸۱٫۷۰
قاعده	Ridor	۸۱٫۱۹
قاعده	ConjuctiveRule	۵۵٫۰۰
قاعده	ZeroR	۴۹٫۲۶
قاعده	OneR	۶۶٫۰۱
قاعده	DecisionTable	۸۱٫۶۴
بیز	NaiveBayesSimple	۷۸٫۴۰
بیز	NaiveBayesUpdateable	۷۸٫۴۰
بیز	NaiveBayes	۷۸٫۴۰
بیز	BayesNet	۷۸٫۵۲
بیز	HNB	۸۱٫۹۶
بیز	WAODE	۸۳٫۵۳
بیز	AODE	۸۰٫۴۳
بیز	AODEsr	۸۱٫۵۸

[1] Cart
[2] Ripper

منابع

باطنی، محمدرضا. توصیف ساختمان دستوری زبان فارسی. تهران: امیرکبیر، ۱۳۶۰.

بی‌جن‌خان، محمود؛ امکان‌سنجی برای تجزیه و تحلیل کسرهٔ اضافه زبان فارسی با روش انطباق الگو، پژوهشگاه فرهنگ، هنر و ارتباطات- پژوهشکدهٔ ارتباطات-گروه زبان فارسی و فناوری اطلاعات، ۱۳۸۴.

سلیمی بدر، آرمین. طراحی و پیاده‌سازی نرم‌افزار تعیین مرز و نوع گروه‌های نحوی در متون فارسی. پایان‌نامهٔ کارشناسی مهندسی کامپیوتر (گرایش نرم‌افزار)، دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، شهریور ۱۳۸۹.

شریفی- آتشگاه، مسعود. تولید نیمه‌خودکار درخت بانک گروه‌های نحوی در متون فارسی. پایان‌نامه دکتری، دانشکدهٔ ادبیات و علوم انسانی، دانشگاه تهران، ۱۳۸۸.

شمس‌فرد، مهرنوش؛ صدر موسوی، مریم؛ استخراج نقش‌های موضوعی در جملات فارسی، پانزدهمین کنفرانس مهندسی برق ایران، تهران، ۱۳۸۵.

عیسی‌پور، شهریار؛ همایون‌پور، محمد مهدی؛ بی‌جن‌خان، محمود؛ "شناسایی محل کسرهٔ اضافه در زبان فارسی با استفاده از گرامر مستقل از متن احتمالاتی"، سیزدهمین کنفرانس انجمن کامپیوتر ایران، دانشگاه صنعتی شریف، جزیرهٔ کیش، ایران، ۱۳۸۶.

کیانی، سهیلا؛ شمس‌فرد، مهرنوش؛ "تعیین مرز کلمات و عبارات در متون نوشتاری فارسی"، چهاردهمین کنفرانس ملی انجمن کامپیوتر ایران، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، تهران، ایران، ۲۰-۲۱ اسفند ۱۳۸۷.

موسوی، نفیسه سادات؛ ثانی، غلامرضا، به‌کارگیری دسته‌بندی‌کننده و رتبه‌بندی‌کننده آنتروپی بیشینه در فرآیند تعیین مرجع ضمایر زبان فارسی، چهاردهمین کنفرانس ملی انجمن کامپیوتر ایران، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، تهران، ایران، ۲۰-۲۱ اسفند ۱۳۸۷.

ایران‌پور مبارکه، مجید؛ مینایی بیدگلی، بهروز؛ تعیین حدود جملات در پیکره‌های متنی زبان فارسی با استفاده از یک روش کارای ارائه شده برای تشخیص فعل، چهاردهمین کنفرانس ملی انجمن کامپیوتر ایران، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، تهران، ایران، ۲۰-۲۱ اسفند ۱۳۸۷.

Avinesh, P., Karthik, G., "Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Bharathi, A., Mannem, P.R., "Introduction to the Shallow Parsing Contest for South Asian Languages", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Chandra Pammi , S., Prahallad, K., "POS Tagging and Chunking using Decision Forests", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Delip R., Yarowsky, D., "Part of Speech Tagging and Shallow Parsing of Indian Languages", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Diab, M., Hacioglu, K., Jurafsky D., "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks", HLT-NAACL, 2004.

Ekbal, A., Mandal, S., Bandyopadhyay S., 2007. POS Tagging Using HMM and Rule-based Chunking, In Proceedings of the IJCAI-2007, Hyderabad, India.

Hansakunbuntheung, C., Thangthai A., Wutiwi-watchai C., Siricharoenchai R., "Learning Methods and Features for Corpus-based Phrase Break Prediction on Thai", 9th Euoropean Conference on Speech Communication, 2005.

Himanshu, A., "POS tagging and Chunking for Indian Languages", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Kiani, S., Akhavan, T., Shamsfard, M., Developing A Persian Chunker Using a Hybrid Approach, IMCSIT'09-Computational Linguistics and Applications (CLA'09), Mragowa, Poland, ,October 2009.

Kudo, T., Matsumoto, Y., "Chunking with Support Vector Machines", NAAL, 2001.

Lafferty, J., McCallum, A. and Pereira, F., “Conditional Random Fields: Probabilistic models for segmentation and labeling sequence data”, In Proc. ICML-01, pages 282-289, 2001.

Lance, A. Ramshaw and Mitchel P. Marcus, “Text Chunking Using Transformation-based Learning”, In Proc. of the 3rd ACL Workshop on Very Large Corpora, Cambridge, 1995.

Murphy Richard, C., "Phrase Detection and Associative Memory Neural Network", Proceedings of the International Joint Conference, 4, 2003, 2599 – 2603.

Oliveria, F., Wong F., Dong M., “Systematic Noun Phrase Chunking by Parsing Constraint Synchronous Grammar in application to Portuguese Chinese Machine Translation”, In Proc. Of the 6th

Xu F., Zong C., Zhao J., "A Hybrid Approach to Chinese Base Noun Phrase Chunking", SIGHAN Workshop On Chinese Language Processing, 2006.

محمد مهدی همایون‌پور در سال ۱۳۳۹ در شهر شیراز متولد شد. تحصیلات تا مقطع دیپلم را در شهر شیراز سپری و دیپلم متوسطه خود را در سال ۱۳۵۸ دریافت کرد. وی تحصیلات خود در مقطع کارشناسی را در رشتهٔ مهندسی برق در دانشگاه صنعتی امیرکبیر (سال ۱۳۶۶)، کارشناسی ارشد را در رشتهٔ برق (مخابرات)، از دانشگاه خواجه نصیرالدین طوسی (سال ۱۳۶۹)، کارشناسی ارشد دوم خود را در زمینهٔ فونتیک (۱۳۷۴) در دانشگاه سوربون جدید در فرانسه و هم‌زمان دورهٔ دکترای خود را در دانشگاه پاریس ۱۱ در زمینهٔ مهندسی برق (۱۳۷۴) به‌پایان رسانید.

نامبرده از سال ۱۳۷۴ در سمت عضو هیأت علمی دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیرکبیر به تدریس و تحقیق مشغول است. ایشان علاوه‌بر تدریس، راهنمایی پروژه‌های کارشناسی، کارشناسی ارشد و دکتری در زمینه‌های مهندسی کامپیوتر و فناوری اطلاعات و نیز هدایت تعداد زیادی پروژه‌های صنعتی و ملّی را برعهده داشته است. از موضوعات مورد علاقهٔ ایشان می‌توان به پردازش سیگنال، پردازش گفتار و پردازش متن اشاره کرد.

نشانی رایانامهٔ ایشان عبارت است از:
homayoun@aut.ac.ir

آرمین سلیمی بدر در سال ۱۳۶۷ در تهران متولد شد. تحصیلات در مقطع کارشناسی را در رشتهٔ مهندسی کامپیوتر گرایش نرم‌افزار (۱۳۸۹) و مقطع کارشناسی ارشد در رشتهٔ مهندسی کامپیوتر گرایش هوش مصنوعی (۱۳۹۱) در دانشگاه صنعتی امیرکبیر به پایان رسانیده است. وی هم‌اکنون دانشجوی مقطع دکتری در رشتهٔ مهندسی کامپیوتر در دانشگاه صنعتی امیرکبیر است. از موضوعات مورد علاقهٔ وی می‌توان به شبیه‌سازی مغز، شبکه‌های عصبی مصنوعی، سامانه‌های

International Conference on Information Technology and Applications (ICITA 2009), Hanoi, Vietnam, 2009, 292-295.

Pattabhi R.K. Rao, T., Vijay Sunder Ram R., Vijayakrishna R., Sobha L., "A Text Chunker and Hybrid POS Tagger for Indian Languages", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Ravi Sastry, G.M., Sourish Chaudhuri, P. Nagender Reddy, "An HMM based Part-Of-Speech tagger and statistical chunker for 3 Indian Languages", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Sanders, E., Taylor. P., "Using Statistical Models to predict phrase boundaries for speech synthesis", 4th European Conference on Speech Communication and Technology, Spain, 1995.

Sandipan, D. "Part-of-Speech Tagging and Chunking with Maximum Entropy Model", In: Proceedings of the IJCAI-2007, Hyderabad, India, 2007.

Sha, F. and Pereira F., "Shallow Parsing with Conditional Random Fields", Proceedings of HLT-NAACL, 2003.

Shamsfard, M., Sadr Mousavi Maryam, "Thematic Role Extraction Using Shallow Parsing", International Journal of Computational Intelligence 4, 2, 2008, 126 - 132.

Tesprasit, V., Charoenpornsawat P., Sornlertlamvanich V., "Learning Phrase Break Detection in Thai Text-To-Speech", Proceeding of 8th European Conference on Speech Communication and Technology (EuroSpeech 2003): Geneva Switzerland.

Truyen, T., Phung D., "A Tutorial on the Maths behind Conditional Random Fields for Sequential Labeling", http://www.computing.edu.au/~trantt2/pubs/crf_intro.pdf, 2008.

Van Rijsbergen, C.J., "Information Retrieval", Butter-sworth, 1979.

Vladimir Vapnik, Statistical Learning Theory, Wiley_Inte-rscience, 1998.

Vladimir Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, New York, NY, 1995.

خبرهٔ فازی، الگوریتم‌های تکاملی، کنتـرل بهینـه و پـردازش زبان طبیعی اشاره کرد.

نشانی رایانامهٔ ایشان عبارت است از:
armin.salimibadr@aut.ac.ir