بازشناسی متون فارسی با استفاده از مدل زبانی n-gram و پالایش گرامری

پریسا شیروانی[1]، مهرداد وطن‌خواه خوزانی[2] وخشایار یغمایی[3]
[1و3] دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران
[2] دانشکدهٔ مهندسی کامپیوتر، دانشگاه شفیلد هالام، شفیلد ، انگلستان

چکیده

بازشناسی متون، در سال‌های اخیر بسیار مورد توجه قرار گرفته است. ارائهٔ الگوریتم‌های بازشناسی، برگرفته از ساختار گرامری و معنایی این زبان می‌تواند روش مؤثری در پردازش‌های دیگر مربوط به خط و زبان فارسی باشد. در این مقاله با استفاده از شاخهٔ علمی پردازش زبان‌های طبیعی، یک الگوریتم سه‌مرحله‌ای به‌منظور بازشناسی متون فارسی بر مبنای بازشناسی جملات فارسی ارائه می‌شود. این روش شامل مراحل ترکیب زیرکلمات به‌منظور ساخت کلمات و سپس جملات بالقوهٔ معنی‌دار و درنهایت استفاده از دو مدل زبانی و چند قاعدهٔ گرامری، به‌منظور تشخیص جملهٔ صحیح براساس انطباق با گرامر زبان فارسی است.آزمایش‌های متعدد نشان می‌دهد که دقت روش ارائه‌شده برای مرحلهٔ ساخت کلمات و سپس جملات بالقوهٔ معنی‌دار ۹۸ درصد و برای تشخیص جملهٔ صحیح با استفاده از مدل زبانی بایگرام ۸۵ درصد و برای مدل زبانی ترایگرام ۸۸ درصد است.

واژگان کلیدی: بازشناسی متن، فارسی، مدل‌سازی زبان فارسی، پردازش زبان‌های طبیعی.

۱- مقدّمه

بازشناسی متن[1] که یکی از محوری‌ترین شاخه‌های بازشناسی الگو[2] است، با پیشرفت روزافزون رایانه، تبدیل به یکی از موضوعات مهم شد. درواقع به‌دلیل ویژگی‌هایی از قبیل انجام سریع‌تر و دقیق‌تر کارها و خستگی‌ناپذیر بودن رایانه‌ها، انجام امور به‌وسیلهٔ آنها ترجیح داده می‌شود. بنابراین می‌توان یکی از مهم‌ترین و اصلی‌ترین دلایل گسترش بی‌شمار تحقیقات در حوزهٔ بازشناسی متون را همین تمایل به انجام کارها با رایانه دانست.

تحقیقات در زمینهٔ بازشناسی متون چاپی فارسی و عربی به‌طور تقریبی از اوایل دههٔ ۱۹۸۰ آغاز شد (Amin et al., 1980 ، Badie et al., 1980).

وجود برخی ویژگی‌های خاص در نگارش این متون مانند وجود نقاط، علائم، تنوع قلم‌ها و هم‌پوشانی حروف با یکدیگر، بازشناسی متون فارسی و عربی را با پیچیدگی‌ها و دشواری‌هایی همراه کرده است.

تاکنون کارهای انجام شده در زمینهٔ بازشناسی متون بر بازشناسی حروف و کلمات متمرکز بوده‌اند (Al- Muhtaseb, 2008 ، El-Abed et al., 2008 ، Jacobs et al., 2005 ، Khosravi al., 2008 ، Sarfaraz , 2003 ، Adab, 2002 ، et al., 2008 Ebrahimi ، ابراهیمی ۱۳۸۴، رضوی و همکاران۱۳۸۴و ۱۳۸۳) (Ebrahimi et al., 2011 ، et al., Bahmani 2010). این بازشناسی‌ها به‌طور معمول با سه رویکرد کلی صورت می‌گیرند (ابراهیمی، ۱۳۸۴): در رویکرد اول جداسازی و بازشناسی دو مرحلهٔ مستقل از یکدیگر هستند. در رویکرد دوم کلمه به‌عنوان یک الگوی واحد درنظر گرفته و بازشناسی کلمه بدون جداسازی حروفش انجام می‌شود. در رویکرد سوم یا رویکرد ترکیبی اطلاعات مربوط به جداسازی حروف تشکیل‌دهندهٔ کلمه و اطلاعات

[1] Text Recognition
[2] Pattern Recognition

مربوط به شکل کلی کلمه در قالب یک سیستم ترکیبی جداسازی-بازشناسی به کار گرفته می‌شود. در این مقاله از مدل‌سازی آماری زبان[1] در حوزهٔ پردازش زبان‌های طبیعی[2] و همچنین چند قاعدهٔ گرامری به‌منظور بازشناسی جملات فارسی استفاده شده است.

در دهه‌های اخیر با توجه به کاربردهای گستردهٔ مدل‌های زبان، تحقیقات زیادی برای مدل‌سازی زبان‌های پرکاربرد جهانی و به‌خصوص زبان انگلیسی انجام شده است. مدل‌های آماری زبان انگلیسی از سال ۱۹۸۰ در سیستم‌های واقعی به‌کار رفته‌اند (Rosenfeld, 2000). اما برای مدل‌سازی زبان فارسی هنوز تحقیقات گسترده و قابل اعتنایی صورت نگرفته است. پیش‌بینی واژگان فارسی، یکی از تحقیقاتی است که با استفاده از مدل‌سازی آماری زبان فارسی انجام شده است (Ghayoomi et al., 2004). در این تحقیق که با واردشدن حروف اولیهٔ یک واژه، سیستم واژه‌هایی را که با آن حرف آغاز می‌شود در پنجره‌ای فهرست می‌کند و به کاربر پیشنهاد می‌دهد. با واردشدن حروف بیشتر، پیشنهادها محدودتر می‌شود تا واژه مورد نظر کاربر از میان آنها یافت شود.

بررسی یک مدل آماری زبان بر اساس دسته‌های منطقی دستوری زبان فارسی برای استفاده از بازشناسی گفتار پیوسته، نوعی مدل‌سازی جدید زبان فارسی است. این مدل زبان بر اساس دسته‌های منطقی N-Gram با طول متغیر کار می‌کند و در آن به جای یافتن الگوهای آماری مربوط به دنباله‌های کلمات، روابط بین دسته‌های منطقی از کلمات، مورد بررسی قرار می‌گیرند. این تحقیق در سه مرحله خوشه‌بندی کلمات، به‌دست آوردن مدل آماری زبان با استفاده از ساختار درختی و اعمال مدل زبان به سیستم بازشناسی انجام شده است.

تقطیع و برچسب‌دهی نحوی - معنایی داده‌های نوشتاری یکی از فعالیت‌های اصلی در طراحی و ساخت هر دادگان زبانی برای استخراج مدل زبانی است (بی‌جن‌خان محمود، ۱۳۸۱). برای استخراج مدل زبان فارسی یک بستهٔ نرم‌افزاری نوشته شده، که در چارچوب فرآیند مارکوف صفر تا سه مرحله‌ای، توزیع احتمال مشروط کلمات فارسی را در چهار حالت به دست می‌دهد.

با استفاده از یک مدل آماری زبان می‌توان احتمال کلمات بعدی را پیش‌بینی کرد. همچنین مدل‌های زبانی دیگری از جمله مدل‌های توانی[3] (Rosenfeld et al., 2001) و گرامرهای مستقل از متن[4] (Corazza et al., 1993) برای مدل‌سازی زبان‌های طبیعی پیشنهاد شده‌اند؛ اما درعمل مدل‌های آماری N-gram (Rosenfeld, 2000) به‌علّت سادگی پیاده‌سازی ترجیح داده می‌شوند.

در الگوریتم پیشنهادی از مدل‌سازی زبان فارسی در جهت پالایش گرامری جملات آن استفاده شده است. برای این منظور ابتدا زیرکلمات از متن مورد پردازش استخراج شده و از ترکیب آنها کلمات و سپس جملات بالقوهٔ معنی‌دار ساخته می‌شوند. درنهایت از مدل زبانی 2-gram یا بایگرام (مدل پنهان مارکوف مرتبهٔ اول) و3-gram یا ترایگرام (مدل پنهان مارکوف مرتبهٔ دوم) و همچنین حذف برخی ترکیبات دستوری غیرممکن به‌منظور تشخیص جملهٔ صحیح از میان مجموعه‌ای از جملات بالقوه معنی‌دار استفاده شده است.

در بخش دوم به ساخت کلمات بامعنی و جملات بالقوهٔ معنی‌دار از ترکیب زیرکلمات پرداخته می‌شود. بخش سوم به تشخیص جملهٔ صحیح با استفاده از مدل‌های زبانی و همچنین حذف ترکیبات دستوری غیرممکن می‌پردازد. در بخش چهارم نتایج ارائه شده و بخش پنجم به بحث و تحلیل پیرامون آن‌ها اختصاص می‌یابد. درنهایت نتیجه‌گیری در بخش ششم ارائه خواهد شد.

۲- ساخت کلمات بامعنی و جملات بالقوهٔ معنی‌دار

کلمات را در زبان فارسی می‌توان ترکیبی از زیرکلمات دانست. هر بخش از دنبالهٔ حروف که از بخش‌های قبل و بعد بتواند جدا در نظر گرفته شود، یک زیرکلمه است. زیرکلمه از یک حرف یا ترکیبی از حروف به هم پیوسته تشکیل شده است. به‌عنوان مثال کلمهٔ "مدرسه" دارای سه زیرکلمه "مد"، "ر"، "سه" است. درحالی که کلمه "شبنم" تنها شامل یک زیر کلمهٔ چهارحرفی "شبنم" است. از این رو در این تحقیق سعی بر این بوده است که کلمات و به دنبال آن جملات از ترکیب زیرکلمات ایجاد و شناسایی شوند. برای این منظور ابتدا بانک اطلاعاتی از کلمات پربسامد در زبان فارسی تهیه شد (حسنی، ۱۳۸۴).

[1] Statistical Language Models
[2] Natural Language Processing
[3] Exponential Models
[4] Context Free Grammars

با ترکیب زیرکلمات، کلمات معنی‌دار ممکن به‌دست می‌آیند. به‌عنوان مثال جمله "پدر من آمد" از زیرکلمات"پد"، "ر"، "من"، "آ"، "مد" تشکیل شده است.

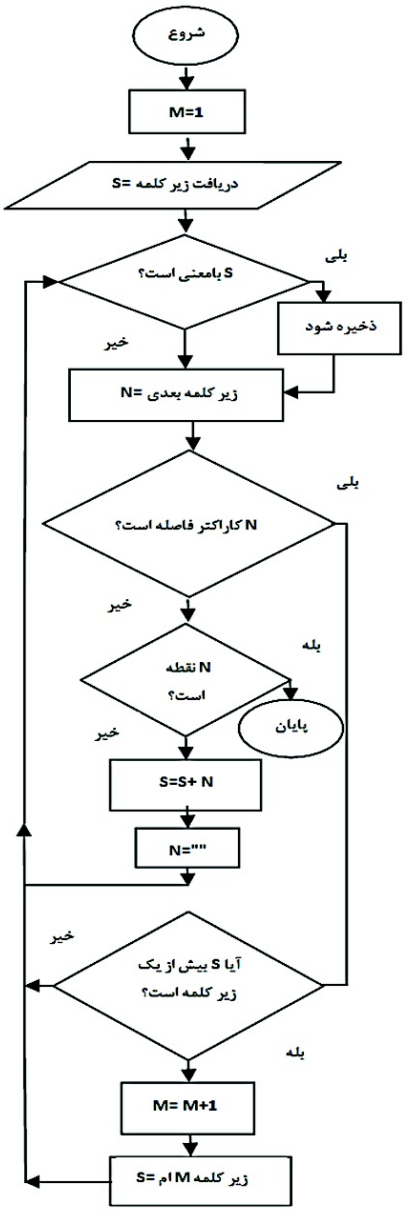
از ترکیب ترتیبی (از راست به چپ) این زیرکلمات، تنها می‌توان کلمات معنی‌دار "پدر"، "من"،" آمد" را استخراج کرد و به همین دلیل ترکیب جمله به‌راحتی قابل شناسایی است.

این در حالی است که برخی از جملات فارسی را به‌دلیل امکان ترکیب زیرکلمات در ساخت کلمات با معنی متفاوت، می‌توان به صورت‌های مختلف خواند. اگرچه تمامی این ترکیب‌ها دارای معنی نیستند (این همان اشتباهی است که در کودکان و افراد نوسواد در خواندن متن‌ها پیش می‌آید). به‌طور مثال، زیرکلمات جمله "مادربزرگ آمد"، عبارتند از"ما"، "د"، "ر"، "بز"، "ر"، "گ"، "آ"، "مد". اکنون اگر از ترکیب زیرکلمات (از راست به چپ) استفاده کنیم، چهار جمله بالقوهٔ معنی‌دار زیر قابل استخراج می‌باشند:

ما در بز رگ آمد.
ما در بزرگ آمد.
مادر بز رگ آمد.
مادربزرگ آمد.

روندنمای قسمت ساخت کلمات بامعنی از ترکیب زیرکلمات در شکل (۱) نشان داده شده است. همان‌طور که در این شکل مشاهده می‌کنید، ابتدا زیرکلمهٔ اول دریافت می‌شود. سپس اگر این زیرکلمه به تنهایی معنی‌دار باشد، ذخیره و زیرکلمهٔ بعدی از ورودی دریافت می‌شود. همچنین اگر زیرکلمهٔ دریافتی اول بامعنی نباشد، زیرکلمهٔ بعدی از ورودی دریافت می‌شود. اگر زیرکلمهٔ دریافت‌شده، نقطه باشد به معنای رسیدن به انتهای جمله و پایان زیر کلمات جمله است. اگر نویسهٔ فاصله باشد به معنای پایان زیرکلمات مربوط به یک کلمه است؛ زیرا بین زیرکلمات در یک کلمهٔ فارسی حداکثر نیم‌فاصله وجود دارد و فاصلهٔ کامل بین کلمات ظاهر می‌شود. از این‌رو زیرکلمهٔ دریافتی دوم درصورتی‌که نقطه یا نویسهٔ فاصله نباشد، در کنار زیرکلمه اول قرار می‌گیرد و معنی‌داربودن ترکیب ساخته شده بررسی می‌شود. در این روندنما M به‌عنوان یک اندیس برای انتخاب زیرکلمات در نظر گرفته شده است. به‌عنوان مثال دیگر زیرکلمات"پز"، "شک"،"با"،"عجله"،"به"،"د"، "ا"، "ر"، "و"، "خا"، "نه"،"ر"، "فت" را درنظر بگیرید. کلمات "پزشک، با، عجله، به، داروخانه، دارو خانه، دار و خانه، رفت" از زیرکلمات استخراج می‌شوند. سپس این کلمات در کنار هم قرار می‌گیرند و سه جملهٔ بالقوهٔ معنی‌دار زیر را می‌سازند:

پزشک با عجله به داروخانه رفت.
پزشک با عجله به دارو خانه رفت.
پزشک با عجله به دار و خانه رفت.



(شکل-۱): روندنمای ساخت کلمات بامعنی

هرچند تعداد این جملات بالقوه معنی‌دار در بسیاری از موارد محدود است؛ اما برای شناسایی درست جملات

ضروری است که با استفاده از قواعد دستوری و معنایی فارسی، جمله یا جملات معنی‌دار از این ترکیبات بالقوه استخراج شوند.

۳- تشخیص جملهٔ صحیح

به‌منظور تشخیص جملهٔ صحیح از میان مجموعه جملات بالقوهٔ معنی‌دار ساخته شده، از دو مدل زبانی بایگرام و ترایگرام و همچنین حذف ترکیبات دستوری غیرممکن استفاده کردیم.

۳-۱- اعمال مدل‌های زبانی

در این مرحله از دو مدل زبانی بایگرام و ترایگرام (Srihar et al., 2007) که برای پالایش گرامری زبان فارسی تغییر یافته‌اند، استفاده می‌کنیم. دادگان آموزشی مورد استفاده در این تحقیق، متون فارسی استاندارد موجود در نشریات و مجلات هستند. درحقیقت، این داده‌ها از طیف وسیعی از متون فارسی در زمینه‌ها و موضوعات مختلف جمع آوری شده‌اند. برای تهیهٔ داده‌های آموزشی با حجم زیاد تنها راه عملی استفاده از متون موجود در شبکه اینترنت است. این آموزش در مدل زبانی بایگرام به‌صورت دوتایی و در مدل زبانی ترایگرام به‌صورت سه‌تایی انجام می‌شود.

به این ترتیب که در مدل زبانی اول (بایگرام) متون آموزشی به‌صورت دو کلمه دو کلمه، بررسی می‌شوند و تعداد تکرار برای هر دو نوع متوالی کلمه در متون آموزش داده شده بر حسب درصد، در یک جدول به نام "فراوانی توالی دوتایی مقولات واژگانی"ذخیره می‌شود؛ درحالی‌که در مدل زبانی دوم (ترایگرام) متون آموزشی به‌صورت سه‌کلمه سه‌کلمه، مورد بررسی قرار می‌گیرند. در این حالت تعداد تکرار برای هر سه نوع متوالی کلمه در جدول "فراوانی توالی سه تایی مقولات واژگانی" نشان داده می‌شود.

به‌عنوان نمونه جدول "فراوانی توالی دوتایی مقولات واژگانی" برای مدل زبانی بایگرام در شکل (۲) نشان داده شده است. همچنین روندنمای اعمال مدل زبانی بایگرام در شکل (۳) آورده شده است. همان‌طور که در این شکل مشاهده می‌کنید کلمات موجود در جمله به‌صورت دوتایی مورد بررسی قرار می‌گیرند به این ترتیب که ابتدا دو کلمهٔ اول جمله درنظر گرفته شده و نقش دستوری آنها مشخص می‌شود؛ سپس عدد مربوط به ترکیب دستوری از جدول "فراوانی توالی دوتایی مقولات واژگانی" خوانده می‌شود. این روند برای کلمات بعدی نیز تکرار می‌شود تا زمانی که به نویسهٔ نقطه برسیم. درنهایت درصدهای خوانده شده از جدول "فراوانی توالی دوتایی مقولات واژگانی" مربوط به ترکیبات گرامری دوتایی واژگان جمله با هم جمع می‌شوند و درصد حاصل از اعمال مدل زبانی به جمله بالقوه معنی‌دار مورد بررسی، بدست می‌آید؛ سپس از بین تمام جملات بالقوهٔ معنی‌دار که به هر کدام درصدی براساس اعمال مدل زبانی داده شده است، جمله با بیشترین مقدار درصد به‌عنوان جملهٔ صحیح تشخیص داده خواهد شد.

به‌عنوان مثال جمله "من بادام دارم" را در نظر بگیرید. ابتدا دو کلمه اول این جمله یعنی "من" و "بادام" بررسی می‌شوند و نقش دستوری هر یک مشخص می‌شود. "من" به عنوان ضمیر و "بادام" به عنوان اسم در نظر گرفته می‌شوند یعنی دارای ترکیب "ضمیر+ اسم" هستند. سپس عدد مربوط به این ترکیب دستوری از جدول "فراوانی توالی دوتایی مقولات واژگانی" خوانده می‌شود. درواقع درصدهای موجود در جداول بایگرام و ترایگرام، اطلاعات آماری هستند که از متون آموزش داده شده، استخراج شده‌اند. در مرحله بعد، نقش دستوری کلمهٔ بعدی (کلمه "دارم" در مثال مذکور) مشخص می‌شود و درصد مربوط به ترکیب گرامری این کلمه با کلمهٔ قبل از آن با استفاده از جدول "فراوانی توالی دوتایی مقولات واژگانی"تعیین می‌گردد. در اینجا "دارم" به‌عنوان فعل و کلمه قبل از آن یعنی "بادام" به‌عنوان اسم، ترکیب دستوری "اسم+فعل" را تشکیل می‌دهند. درنهایت درصدهای مربوط به ترکیب‌های گرامری دوتایی کلمات جمله با هم جمع و عدد حاصل از اعمال مدل زبانی بایگرام به جمله مشخص می‌شود. در مثال بیان‌شده از جمع دو درصد مربوط به ترکیب‌های گرامری "ضمیر+ اسم" و "اسم + فعل" ، درصد مربوط به مدل زبانی بایگرام جمله تعیین می‌شود.

در مدل زبانی ترایگرام، ابتدا سه کلمهٔ اوّل جمله درنظر گرفته شده و نقش دستوری‌شان مشخص می‌شود؛ سپس کلمهٔ اوّل حذف شده و سه کلمهٔ بعدی بررسی می‌شوند. به‌عنوان مثال جملهٔ "پدر به خانه آمد" را در نظر بگیرید. در ابتدا نقش دستوری سه کلمه اول یعنی کلمات "پدر" ، " به" ، "خانه" به صورت "اسم + حرف اضافه + اسم" تعیین می‌شود. سپس کلمهٔ اول یعنی پدر حذف شده و نقش دستوری سه کلمهٔ بعدی "به"، "خانه"، "آمد" به‌صورت "حرف اضافه + اسم + فعل" مشخص می‌شود. تا اینجا دو ترکیب دستوری سه تایی "اسم + حرف اضافه +

اسم" و "حرف اضافه + اسم+فعل" حاصل شده است که با استفاده از جدول "فراوانی توالی سه تایی مقولات واژگانی"درصدهای مربوط به هر ترکیب مشخص شده و این مقادیر با هم جمع می‌شوند. این مجموع، درصد مربوط به اعمال مدل زبانی ترایگرام به جمله مذکور را می‌دهد.

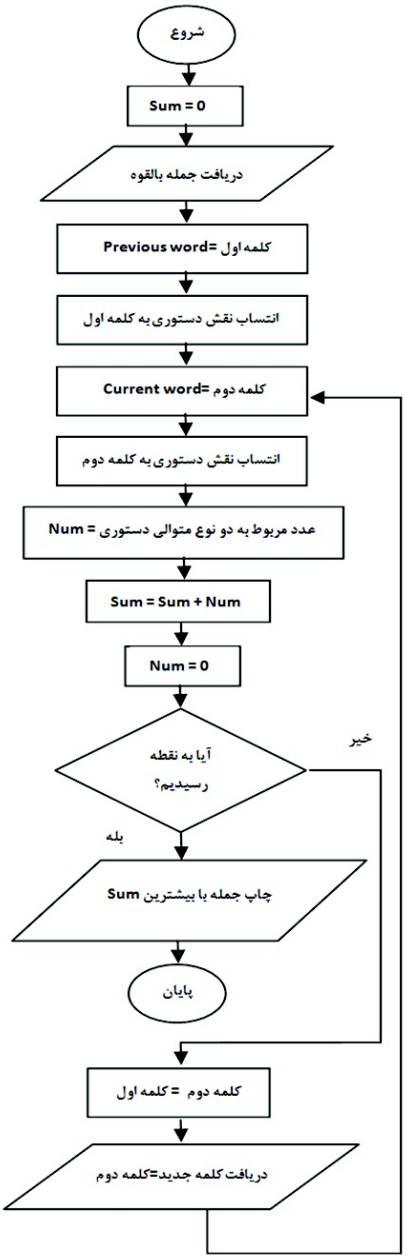
۳-۲- حذف ترکیبات دستوری غیرممکن

به‌منظور بهبود عملکرد روش پیشنهادی و افزایش دقت علاوه‌بر استفاده از مدل‌های زبانی بیان شده، از حذف برخی از ترکیبات غیرممکن دستوری نیز به‌منظور پالایش جملات استفاده کردیم. به این صورت که ابتدا مدل زبانی به جملات بالقوه معنی‌دار اعمال می‌شود و درصد متناسب با هر جمله با توجه به نقش دستوری کلمات آن، با استفاده از جدول گرامر رایج مشخص می‌شود.

(شکل-۲): فراوانی توالی دوتایی مقولات واژگانی

سپس جملاتی که دارای ترکیبات غیرممکن هستند، حذف می‌شوند و درصدهای مربوط به جملات باقی‌مانده با هم مقایسه شده و جمله با بیشترین درصد به‌عنوان جملهٔ صحیح، تشخیص داده می‌شود. ترکیبات غیرممکن دستوری به‌کار برده شده در الگوریتم پیشنهادی، به‌منظور پالایش جملات عبارتند از:

ترکیب غیرممکن اول: در گرامر فارسی دو فعل پشت سر هم قرار نمی‌گیرند. بنابراین جملات دارای ترکیب "فعل+ فعل" در مدل زبانی بایگرام و جملات دارای یکی از ترکیبات "اسم +فعل + فعل"، "صفت + فعل + فعل"، "فعل + فعل + فعل"، "ضمیر + فعل + فعل"،"قید + فعل + فعل"، "حرف + فعل + فعل"، "فعل + فعل + اسم"،"فعل + فعل + ضمیر"، "فعل + فعل + حرف"، "فعل + فعل + قید"، "فعل + فعل + صفت" در مدل زبانی ترایگرام از نظر گرامر زبان فارسی صحیح نبوده و حذف می‌شوند. به‌عنوان مثال جمله " در جریان زندگی او خود را به سرنوشت سپرد" را در نظر بگیرید که از آن دو جملهٔ بالقوهٔ معنی‌دار زیر استخراج می‌شود:



(شکل-۳): روندنمای تشخیص جملهٔ صحیح با استفاده از مدل زبانی بایگرام

در جریان زندگی او خود را به سرنوشت سپرد.

در جریان زندگی او خود را به سر نوشت سپرد.

در جملهٔ دوم به‌علت این که دو فعل "نوشت" و "سپرد" پشت سر هم قرار گرفته‌اند، یعنی جملهٔ دارای ترکیب "فعل + فعل" در مدل زبانی بایگرام و ترکیب "اسم + فعل + فعل" در مدل زبانی ترایگرام است از نظر قواعد گرامری زبان فارسی صحیح نبوده و حذف می‌شود. از این رو جملهٔ اول از میان دو جملهٔ بالقوهٔ معنی‌دار مذکور، به‌عنوان جملهٔ صحیح تشخیص داده می‌شود.

ترکیب غیرممکن دوم: حروف اضافه قبل از فعل قرار نمی‌گیرند؛ یعنی ترکیب "حرف اضافه + فعل" در جملات فارسی وجود ندارد. از این رو جملات دارای ترکیب "حرف اضافه + فعل"در مدل زبانی بایگرام و جملات دارای ترکیبات"اسم + حرف اضافه + فعل"، "صفت + حرف اضافه + فعل"، "فعل + حرف اضافه + فعل"، "ضمیر + حرف اضافه + فعل"،"قید + حرف اضافه + فعل"، "حرف اضافه + حرف اضافه + فعل"،"حرف ربط + حرف اضافه + فعل"،"حرف اضافه + فعل + اسم"، "حرف اضافه + فعل + صفت"، "حرف اضافه + فعل + ضمیر"، "حرف اضافه + فعل + قید"، "حرف اضافه + فعل + حرف اضافه"، "حرف اضافه + فعل + حرف ربط" در مدل زبانی ترایگرام صحیح نبوده و حذف می‌شوند.

به‌عنوان مثال از جمله "اهداف جدیدی برای زندگی خود برگزیدم" دو جملهٔ بالقوهٔ معنی‌دار زیر استخراج می‌شود:

اهداف جدیدی برای زندگی خود برگزیدم.

اهداف جدیدی برای زندگی خود بر گزیدم.

جملهٔ دوم به‌علّت این که حرف اضافه "بر" قبل از فعل "گزیدم" قرار گرفته است یعنی دارای ترکیب "حرف اضافه + فعل" در مدل زبانی بایگرام و ترکیب "ضمیر + حرف اضافه + فعل" در مدل زبانی ترایگرام است صحیح نبوده و حذف می‌شود. از این رو جملهٔ بالقوهٔ معنی‌دار اول به‌عنوان جملهٔ صحیح در نظر گرفته می‌شود.

ترکیب غیرممکن سوم: در زبان فارسی، نشانهٔ مفعول (را) بعد از اسم یا صفت قرار می‌گیرد. از این رو ترکیبات نحوی "قید + را"، "حرف اضافه + را"، "حرف ربط + را" و همچنین"را + را"در گرامر فارسی وجود ندارند و نادرست هستند. از این رو ترکیبات بالا از نظر مدل زبانی بایگرام و به‌طور مشابه ترکیب سه‌تایی ترکیبات دوتایی بالا از نظر مدل زبانی ترایگرام صحیح نبوده و حذف می‌شوند.

البته از نظر صرفی کلمهٔ مشترک بین صفت و قید و همچنین بین ضمیر و حرف ربط وجود دارد که در برخی از موقعیت‌ها پس از آنها در جملهٔ نشانهٔ مفعول نیز می‌تواند قرار گیرد. به‌عنوان مثال کلمهٔ "خوب" بین صفت و قید و کلمهٔ "که" بین ضمیر و حرف ربط مشترک است.

۴- نتایج تجربی

در این قسمت نتایج حاصل ازاعمال الگوریتم مربوط به بخش چهارم (یعنی تشخیص جملهٔ صحیح) بر روی تعدادی از جملات که از یک متن بیست صفحه‌ای استخراج شده‌اند، ارائه می‌شود. این نتایج در جداول(۱ و ۲) نشان داده شده‌اند. در این جداول تعداد چهارصد و بیست جمله مورد بررسی قرار گرفته‌اند.که از این تعداد، چهل و یک جمله به دو صورت، بالقوهٔ معنی‌دار ویک جمله به سه صورت، بالقوهٔ معنی دار هستند. همان‌طور که انتظار می‌رود در متون فارسی تعداد جملاتی که دارای حالات بالقوهٔ معنی‌دار هستند کم می‌باشد. در جدول (۲) نتایج حاصل از اعمال مدل‌های زبانی بایگرام و ترایگرام به جملات انتخابی آورده شده است. همان‌طور که در جدول مشخص است، تعداد جملات اشتباه تشخیص داده شده توسط مدل‌های زبانی اعمال‌شده بسیار کم است. مدل ترایگرام عملکرد بهتری نسبت به مدل زبانی بایگرام داشته است؛ زیرا تعداد جملات اشتباه تشخیص داده شده توسط مدل ترایگرام کمتر از مدل بایگرام است. با توجه به اعمال مدل‌های زبانی بر جملات مختلف و جداول (۱) و (۲) کارایی مدل زبانی بایگرام ۸۵ درصد و کارایی مدل زبانی ترایگرام ۸۸ درصد محاسبه شده است. به‌عنوان مثال نمونه‌ای از این جملات در زیر بررسی شده است:

جملهٔ ورودی: با مداد نرم سرعت نوشتن خود را زیاد کنید.

از جملهٔ فوق، چهار جملهٔ بالقوهٔ معنی‌دار زیر را می‌توان استخراج کرد:

با مداد نرم سرعت نوشتن خود را زیاد کنید.

بامداد نرم سرعت نوشتن خودرا زیاد کنید.

با مداد نرم سرعت نوشتن خود راز یاد کنید.

بامداد نرم سرعت نوشتن خود راز یاد کنید.

ابتدا نقش دستوری کلمات هر جمله مشخص می‌شود و درصد متناسب با ترکیب‌های گرامری دوتایی آن با استفاده از جدول "فراوانی توالی مقولات واژگانی"تعیین می‌شود. در مرحلهٔ بعد به‌دلیل این که هیچ‌کدام از جملات بالا دارای

ترکیبات دستوری غیرممکن نیستند، از این رو هیچ یک از آنها حذف نمی‌شوند؛ سپس، درصدهای مربوط به این چهار جمله با یکدیگر مقایسه می‌شوند و مدل زبانی بایگرام جملات اول و سوم را صحیح تشخیص می‌دهد. درحالی‌که مدل زبانی ترایگرام تنها جملهٔ اول را صحیح تشخیص می‌دهد.

البته غالب جملات موجود در متون استاندارد تنها به یک صورت خوانده می‌شوند و دارای چندین جملهٔ بالقوهٔ معنی‌دار نیستند. به‌عنوان مثال دیگر، جملهٔ زیر را در نظر بگیرید:

جملهٔ ورودی: شکارچیان شورا را برای فراری دادن شیر از جنگل تشکیل دادند.

از جمله فوق، چهار جمله بالقوه معنی‌دار زیر استخراج می‌شوند:

شکارچیان شورا را برای فراری دادن شیر از جنگل تشکیل دادند.

شکارچیان شو را را برای فراری دادن شیر از جنگل تشکیل دادند.

شکارچیان شورا را برای فراری دادن شیراز جنگل تشکیل دادند.

شکارچیان شو را را برای فراری دادن شیراز جنگل تشکیل دادند.

جملات دوم و چهارم به‌علّت ترکیب دستوری غیرممکن "را + را"حذف می‌شوند؛ سپس درصدهای مربوط جملات باقی‌مانده یعنی جملات اول و سوم با یکدیگر مقایسه می‌شوند و هر دو مدل زبانی، جمله اول را به‌عنوان جملهٔ صحیح تشخیص می‌دهند.

مثال سوم را به‌صورت زیر در نظر بگیرید:

جملهٔ ورودی: دود کارخانه هوا را آلوده می‌کند.

دو جملهٔ بالقوهٔ معنی‌دار زیر را می‌توان از آن استخراج کرد:

دود کارخانه هوا را آلوده می‌کند.

دود کار خانه هوا را آلوده می‌کند.

بعد از تعیین نقش دستوری کلمات در جملات بالقوهٔ معنی‌دار فوق، عدد مربوط به هر یک از این جملات مشخص می‌شود؛ سپس از نظر ترکیبات دستوری غیرممکن مورد بررسی قرار می‌گیرند و به‌دلیل این که هیچ‌کدام دارای ترکیبات دستوری غیرممکن نیستند، حذف نمی‌شوند. مدل زبانی بایگرام و همچنین مدل زبانی ترایگرام هر دو جمله را به‌عنوان جملات صحیح تشخیص می‌دهند. برای رفع مشکلاتی از این قبیل بایستی از قواعد معنایی زبان فارسی استفاده کرد؛ زیرا از نظر قواعد دستوری (استفاده از جداول "فراوانی توالی مقولات واژگانی" و حذف ترکیبات غیرممکن)، هر دو جملهٔ صحیح تشخیص داده می‌شوند.

۵- بحث و تحلیل

با توجه به نتایج حاصل از جدول (۲) مشاهده می‌شود اشتباهات، بیشتر در مواردی است که جمله کلمه‌ای داشته که شامل یک حرف و یک اسم معنی‌دار باشد؛ مانند: "باهوش، بامزه، برنامه و ...". به این دلیل که با توجه به جدول گرامر رایج بایگرام (ترایگرام) احتمال ترکیب دوتایی (سه تایی) "حرف اضافه + اسم" در جملات فارسی زیاد است. بنابراین این کلمات به‌صورت ترکیب "حرف اضافه + اسم" صحیح تشخیص داده می‌شوند. به‌عنوان مثال کلمهٔ "برچسب" به‌صورت دو کلمهٔ مجزا "بر" به‌عنوان حرف اضافه و "چسب" به‌عنوان اسم تشخیص داده می‌شود. همچنین حضور کلماتی که از دو اسم معنی‌دار تشکیل شده باشند مانند: "کارخانه، آشپزخانه، داروخانه و ..." احتمال تشخیص بیش از یک جمله به‌عنوان جملهٔ صحیح را افزایش می‌دهند. بیش‌تر در این‌گونه موارد، دو جمله به‌عنوان جملات صحیح تشخیص داده می‌شوند.

نتیجهٔ دیگری که از جداول (۱ و ۲) می‌توان گرفت این است که نتایج حاصل از مدل ترایگرام بهتر از نتایج حاصل از مدل بایگرام است. درواقع به‌دلیل افزایش تعداد کلمات مورد بررسی در مدل ترایگرام دقت آن نسبت به مدل بایگرام افزایش یافته است. همچنین می‌توان گفت در مدل ترایگرام از تعداد جملات اشتباه تشخیص داده شده مدل بایگرام کاسته شده و به تعداد جملات به‌طور تقریبی صحیح تشخیص داده شده، افزوده شده است.

۶- نتیجه‌گیری و کارهای آتی

در این مقاله یک روش جدید برای بازشناسی متون فارسی ارائه شد. الگوریتم پیشنهادی در مرحلهٔ اوّل از یکی از روش‌های برچسب‌زنی اجزای متصل به‌منظور جداسازی زیرکلمات استفاده می‌کند. در مرحلهٔ دوم زیرکلمات استخراج‌شده کنار هم قرار گرفتند و همهٔ کلمات معنی‌دار و سپس جملات بالقوهٔ بامعنی تشکیل شدند. در آخرین مرحله از مدل‌های زبانی بایگرام و ترایگرام و همچنین از قواعد گرامری زبان فارسی به‌منظور تشخیص جملهٔ صحیح از میان

مجموعه‌ای از جملات استفاده شد. نتایج نشان داد که با اعمال الگوریتم پیشنهادی می‌توان غالب جملات فارسی را بازشناسی کرد. در عین حال، با توجه به استفادهٔ گسترده از کسره (در ارتباط بین صفت- موصوف و مضاف- مضاف الیه) که در شکل ظاهری کلمات مشخص نمی‌شود (عیسی پورو همکاران، ۱۳۸۶)، به‌نظر می‌رسد با استفاده از ساختارهای دقیق‌تر گرامر و همچنین ارتباط معنایی بین کلمات و کل جمله، بتوان خطاهای باقی‌مانده را رفع کرد.

(جدول- ۱): آمار جملات مورد بررسی

تعداد کل جملات	تعداد جملات با دو جملهٔ بالقوهٔ معنی‌دار	تعداد جملات با سه جملهٔ بالقوهٔ معنی‌دار	تعداد جملات با بیش از سه جملهٔ بالقوهٔ معنی‌دار
۴۲۰	۴۱	۱	۱

(جدول- ۲): نتایج حاصل از اعمال الگوریتم پیشنهادی

مدل‌های زبانی	تعداد جملات با چندین جملهٔ بالقوهٔ معنی‌دار	تعداد جملات صحیح تشخیص داده شده	تعداد جملات به‌طور تقریبی صحیح تشخیص داده شده	تعداد جملات اشتباه تشخیص داده شده
بایگرام	۴۳	۳۴	۶	۳
ترایگرام	۴۳	۳۵	۷	۱

منابع

ابراهیمی، افشین."استفاده از شکل کلی زیرکلمات چاپی در بازیابی تصویر مستندات و بازشناسی متون فارسی"، رسالهٔ دکتری بخش مهندسی برق، دانشگاه تربیت مدرس، تهران، ایران، ۱۳۸۴.

حسنی، حمید. پژوهشگر گروه فرهنگ‌نویسی فرهنگستان زبان و ادب فارسی."واژه‌های پرکاربرد فارسی امروز"، نشر کانون فرهنگی، ۱۳۸۴.

رضوی، سیّد محمّد. کبیر، احسان الله. "بازشناسی برخط حروف مجزای فارسی با شبکهٔ عصبی"، مجموعه مقالات سومین کنفرانس ماشین بینایی و پردازش تصویر ایران ص ۸۳-۸۹ ، دانشگاه تهران، اسفند ماه ۱۳۸۳.

رضوی، سیّد محمّد. کبیر، احسان الله. "بازشناسی برخط زیرکلمات فارسی با استفاده از نقاط و علائم حروف"، دومین کنفرانس فناوری اطلاعات و دانش، دانشگاه صنعتی امیرکبیر، تهران، خرداد ماه ۱۳۸۴.

عیسی‌پور، شهریار. همایون‌پور، محمّد مهدی. بی‌جن‌خان، محمود. "شناسایی محل کسرهٔ اضافه در زبان فارسی با استفاده از گرامر مستقل از متن احتمالاتی"، سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران، جزیرهٔ کیش، خلیج فارس، ایران ۱۹ الی ۲۱ اسفند ۱۳۸۶.

Adab, M., M.B.,"Simultaneous segmentation and recognition of Farsi/Latin printed texts with MLP".Inter-national Joint Conference on Neural Networks., 2002, pp.1534–1539.

AL-Muhtaseb, H., "Recognition of off-line printed Arabic text using hidden Marcov Model ".Signal Processing., 2008, pp.2902-2912.

Amin, A., Kaced, A., Haton, J.P., Mohr, R.," Hand-written Arabic character recognition by the IRAC system". Proceedings of Fifth International Conference on Pattern Recognition., Miami Beach, FL, USA, 1980, pp.729-731.

Badie, K., Shimura, M., "Machine recognition of Arabic cursive scripts". Proceedings of International Workshop Pattern Recognition in Practice., Amsterdam, Netherlands, 1980, pp. 315-323.

Bahmani.Z and Alamdar.F, "Off-line Arabic/Farsi Handwritten Word Recognition Using RBF Neural Network and Genetic algorithm", *IEEE*,vol.2, no.3, pp.352-357, 2010.

Corazza, A., De Mori, R.,Gretter, R., Satta, G., "Language modeling using stochastic context-free grammars". Speech Communication., Vol.13(1-2), 1993, pp. 163-170.

Ebrahimi, A., Kabir, E.," A pictorial dictionary for printed Farsi sub-words".Pattern Recognition Letters.,2008, pp. 656-663.

Ebrahimpour.R and Davoudi.R, "Decision Templates with Gradient based Features for Farsi Handwritten word Recognition",
International Journal of Hybrid Information Technology, Vol.4, No.1,2011.

El-Abed, H., Margner, V.,"Arabic text recognition systems-state of the art and feature trends".International conference on Innovations in Information Technology., 2008, pp. 692-696.

Jacobs, C., Simard, P., Rinker, Viola and J.,"Text recognition of low-resolution document images".Eight International Conference on Document Analysis andRecognition(ICDAR'05)., Vol.2, 2005, pp. 695-699.

Khosravi, H., Kabir, E.,"Farsi font recognition based on Sobel-Roberts features". Pattern Recognition Letters,. 2008, pp.75-

Rosenfeld, R.,"Two decades of statistical language modeling: where do we go from here?".Proceedings of the IEEE,. Vol.88,.2000, pp.1270-1278.

Rosenfeld, R.,Chen, S. F. and Zhu, X.," Whole-sentence exponential language models: a vehicle for linguistic-statistical integration". Computer Speech & Language,. Vol.15 (1),.2001, pp.55-73.

Sarfaraz, M.,Nawaz, N., Al-Khuraidly, A.,"Offline Arabic text recognition system ".International Conference on Geometric Modeling and Graphics (GMAG'03)",.2003, pp.30.

Srihar, R.,Shetty, S., Srihar, S., "Use of language models in handwriting recognition".Center of excellence for document analysis and recognition (CEDAR),. 2007.

پریسا شیروانی دورهٔ کارشناسی خود را در سال ۱۳۸۷در رشتهٔ مهندسی برق (مخابرات) گذرانده و مدرک کارشناسی ارشدش را در سال ۱۳۹۰ در رشتهٔ مهندسی برق (مخابرات) از دانشگاه سمنان اخذ کرد. وی هم‌اکنون در دانشگاه علوم و فناوری سپاهان شهر اصفهان مشغول به فعالیت و زمینه‌های علمی مورد علاقهٔ وی پردازش سیگنال به‌ویژه پردازش متن است.

نشانی رایانامهٔ ایشان عبارت است از:
Shirvani.parisa@gmail.com

مهرداد وطن‌خواه خوزانی دورهٔ کارشناسی خود را در سال ۱۳۸۵ در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) در دانشگاه آزاد اسلامی واحد نجف‌آباد گذرانده و هم‌اکنون دانشجوی دورهٔ کارشناسی ارشد در رشتهٔ مهندسی نرم‌افزار گرایش شبکه در دانشگاه شفیلد هالام انگلستان است. زمینه‌های علمی مورد علاقهٔ وی شبکه‌های عصبی، طراحی و پیاده‌سازی نرم‌افزارهای یکپارچه و نهان‌نگاری اطلاعات است.

نشانی رایانامهٔ ایشان عبارت است از:
Mehrdad .vatankhah@gmail.com

خشایار یغمایی: دورهٔ کارشناسی خود را در سال ۱۳۶۲ در رشتهٔ مهندسی برق (الکترونیک) و دورهٔ کارشناسی ارشد خود را در رشتهٔ مهندسی برق (مخابرات) در سال ۱۳۶۴ در دانشگاه تهران گذرانده است. همچنین مدرک دکترای خود را در سال ۱۳۷۶ در رشتهٔ مهندسی برق (مخابرات) از دانشگاه surrey انگلستان اخذ کرد و به دانشکدهٔ مهندسی برق و کامپیوتر دانشگاه سمنان پیوست و در این دانشگاه مشغول به فعالیت است. زمینه‌های علمی مورد علاقهٔ وی پردازش سیگنال، پردازش صوت و تصویر و پردازش متن است.

نشانی رایانامهٔ ایشان عبارت است از:
khashayar.yaghmaie@gmail.com