
بسم الله الرحمن الرحيم

دانشکده مهندسی
گروه مهندسی کامپیوتر

معرفی یک سیستم بازشناسی کلمات فارسی مبتنی بر روش‌های پردازشی هوشمند

پایان‌نامه کارشناسی ارشد

نگارنده
ایمان پورحسین

استاد راهنما
آقای دکتر علیرضا عصاره

استاد مشاور
خانم دکتر بیتا شادگار

دی‌ماه ۱۳۹۳

باسمه‌تعالی

دانشگاه شهید چمران اهواز

مدیریت تحصیلات تکمیلی

(نتیجه ارزشیابی پایان‌نامه دوره کارشناسی ارشد/دکتری)

بدین‌وسیله گواهی می‌شود پایان‌نامه آقای **ایمان پورحسن** دانشجوی رشته **کامپیوتر- هوش‌مصنوعی** از دانشکده **مهندسی** به شماره دانشجویی **۹۱۱۴۲۰۲** تحت عنوان: **معرفی یک سیستم بازشناسی کلمات فارسی مبتنی بر روش‌های پردازشی هوشمند** جهت اخذ درجه **کارشناسی ارشد** در تاریخ **۱۳۹۳/۱۰/۰۶** توسط هیأت داوران مورد ارزشیابی قرار گرفت و با درجه ..... تصویب شد.

۱- اعضا هیأت داوران:	مرتبه علمی	امضا
الف - استاد راهنما: **دکتر علیرضا عصاره**	**دانشیار**	
ب - استاد مشاور: **دکتر بیتا شادگار**	**استادیار**	
ج – داور۱: **دکتر احسان نامجو**	**استادیار**	
د – داور۲: **دکتر کریم انصاری فرد**	**استادیار**	
هـ – نماینده تحصیلات تکمیلی دانشگاه (استاد ناظر):		
۲- مدیر گروه: **دکتر مرجان نادران طحان**	**استادیار**	
۳- معاون پژوهشی دانشکده:		
۴- مدیر کل تحصیلات تکمیلی:		

تقدیم به پدر بزرگوارم که همیشه راهنمای من بود، روشنی بخش راهم بوده

تقدیم به مادر خوبم، برای مهربانی های بی پایانش

قدردانی

با قدردانی از اساتیدم جناب آقای دکتر عصاره و سرکار خانم دکتر شادگار

برای راهنمایی‌ها و یاری‌شان در تمام مراحل انجام تحقیق و همه‌ی دوران تحصیلم

چکیده

نام خانوادگی: پورحسین	**نام:** ایمان	**شماره دانشجویی :** ۹۱۱۴۲۰۴

عنوان پایان نامه:

معرفی یک سیستم بازشناسی کلمات فارسی مبتنی‌بر روش‌های پردازشی هوشمند

استاد/ اساتید راهنما: دکتر علی‌رضا عصاره

استاد/ اساتید مشاور: دکتر بیتا شادگار

درجه تحصیلی: کارشناسی ارشد	**رشته:** مهندسی کامپیوتر	**گرایش:** هوش مصنوعی
دانشگاه : شهید چمران اهواز	**دانشکده:** مهندسی	**گروه :** کامپیوتر

تاریخ فارغ التحصیلی : ۹۳/۱۰/۰۶ **تعداد صفحه: ۹۸**

کلید واژه ها : توصیفگر نقطه، توصیفگر بدنه مبتنی بر جعبه محیطی، توصیفگر بدنه مبتنی بر نواحی، ترکیب توصیفگرها، کاهش واژه نامه‌های فارسی و عربی، بازشناسی کلمات دستنویس

چکیده :

در این پایان‌نامه برای بازشناسی کلمات دستنویس فارسی و عربی برون خط، دو مدل ترکیبی جدید پیشنهاد شده است. با داشتن تعداد زیادی از کلمات، رویکردهای کاهش واژه‌نامه، روش‌های قدرتمندی برای هرس ابتدایی کلمات محسوب می‌شوند. دو مدل پیشنهادی، مبتنی بر روش‌های کاهش واژه‌نامه‌اند و مراحل پیش پردازش، قطعه بندی، استخراج ویژگی و طبقه‌بندی را در بر می‌گیرند.

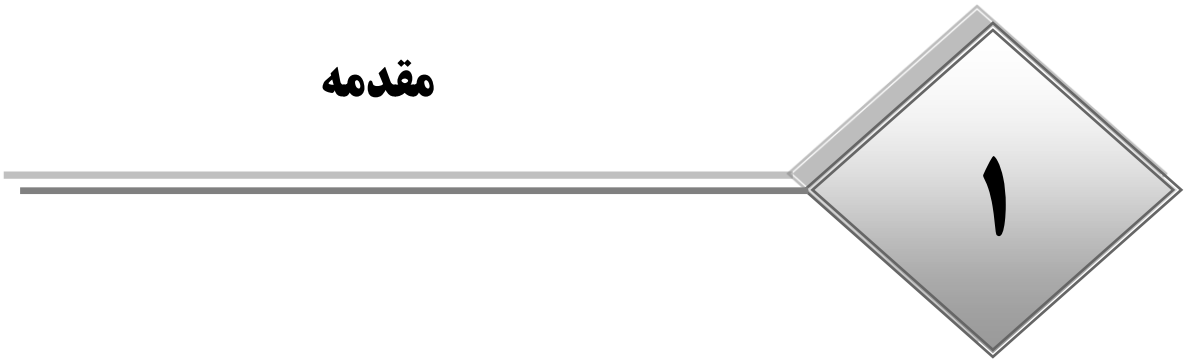
ویژگی‌هایی که از تصاویر کلمات استخراج می‌شوند را می‌توان به دو گروه خصوصیت نقاط کلمه و خصوصیت بدنه کلمه دسته‌بندی کرد. غالب مدل‌هایی که تاکنون ارائه شده است، تنها یک گروه از ویژگی‌ها را بکار می‌برند و توانایی کار با گروه دیگر را ندارند، اما مدل‌هایی که در این پایان‌نامه پیشنهاد شده، جهت افزایش کارایی، از هر دو گروه ویژگی استفاده می‌کنند. مدل‌های پیشنهادی بر روی پایگاه داده IFN/ENIT که متشکل از بیست و شش هزار تصویر کلمه دستنویس است آزمایش شده و با دقت ۹۵٪، نرخ کاهش ۸۹,۶٪ و ۹۶,۲٪ بدست آمده است.

فهرست مطالب

فهرست شکل‌ها و نمودارها

فهرست جدول‌ها

فصل اول



مقدمه

۱–۱ آشنایی با مفاهیم اولیه

تاکنون تلاش‌های بسیاری برای بازشناسی متون لاتین، چینی و ژاپنی صرف شده، اما برای بازشناسی متون فارسی و عربی، کارهای انجام‌شده نسبتا کم و پراکنده بوده است. از جمله علل این کم‌کاری می‌توان به سرمایه‌گذاری ناکافی، خصلت پیوسته بودن این خطوط و فقدان پایگاه داده‌های استاندارد و جامع برای متون فارسی و عربی اشاره کرد. برای تعریف مسأله و توضیح اهمیت آن، نیاز به آشنایی با برخی مفاهیم اولیه وجود دارد که در ادامه به معرفی مختصری از آن‌ها پرداخته می‌شود.

به طور کلی به تشخیص حروف، کلمات، متون و علائم نوشتاری دستنویس و تایپی توسط تصاویر اسکن شده‌ی آنها، بازشناسی نوری کاراکترها[۱] گفته می‌شود. تاکنون پژوهشگران متعددی بر روی روش‌های مختلف بازشناسی متن کار کرده‌اند و در زبان‌های رایج دنیا از قبیل انگلیسی، فرانسوی، چینی و زبان های مشابه سیستم‌های قدرتمندی در زمینه‌ی بازشناسی علائم، ارقام، حروف، کلمات و متون تولید و بسیاری از این سیستم‌ها کاربردی و تجاری شده‌اند.

---

[۱] Optical Character Recognition (OCR)

تشخیص مبلغ چک و صحت امضا در بانک‌ها، تشخیص خودکار آدرس گیرنده در نامه‌های پستی و تشخیص پلاک خودرو در دوربین‌های کنترل سرعت جاده‌ها ازجمله کاربرد این سیستم‌هاست. با توجه به این که بازشناسی متون یکی از مهمترین بخش های دولت الکترونیک به شمار می‌رود، در دهه ی اخیر، در کشور ما نیز تقاضا برای یک سیستم بازشناسی متن فارسی، خاصه در سازمان‌های دولتی، به شدت افزایش یافته است. هر چند در ایران تحقیقات متعددی در زمینه‌ی بازشناسی متن گسسته و پیوسته صورت گرفته، لیکن به دلیل عدم هماهنگی و پشتیبانی مناسب، سیستم‌های تجاری قابل اعتماد برای بازشناسی متن فارسی بسیار کم است [۱]. با وجود الگوریتم‌های فراوانی که طی چند دهه برای مسئله‌ی شناسایی نوری حروف پیشنهاد شده، تاکنون روشی که قابل رقابت با انسان باشد ارائه نشده است و لذا این مسئله همچنان مورد بررسی همه‌جانبه است. در مورد زبان‌های فارسی و عربی که از پیچیدگی نوشتاری بیشتری نسبت به زبان‌های انگلیسی، فرانسوی، آلمانی و زبان‌های مشابه برخوردارند، الگوریتم‌ها و نرم افزارهای طراحی شده در این زمینه نیز دقت کمتری دارد و لذا نیاز به تحقیقات بیشتر در این زمینه احساس می‌شود.

هدف از این پژوهش، پیاده‌سازی یک سیستم خودکار بازشناسی کلمات دستنویس فارسی از روی تصاویر اسکن شده آنها است. تمرکز ما در طراحی این سیستم بر روی چالش‌های اساسی این زمینه تحقیقاتی و ارائه چندین راه حل جدید است. وجود سیستم تشخیص کلمات از دو جنبه اهمیت دارد. یکی این‌که هنوز متون بسیاری وجود دارند که نسخه‌ی الکترونیک آن‌ها موجود نیست و ایجاد نسخه‌ی الکترونیک این متون می‌تواند در مدیریت و استفاده از آن‌ها کمک نماید. از طرف دیگر بسبب سهولت نگارش با دست، هنوز افراد بسیاری تمایل دارند که بجای استفاده از رایانه، متون خود را دستنویس کنند. برای تحقق این مسئله، برنامه‌ای نیاز است که متون دستنویس را به متون الکترونیکی تبدیل کند.

۱-۲ خصوصیات خط فارسی

خصوصیات خط فارسی و عربی متفاوت از لاتین است که در ادامه توصیف مختصری از جنبه‌های مهم، ارائه می‌شود. حروف فارسی به‌واسطه آزادی عمل نویسنده در نگارش، یکی از مشکل ترین زبان‌ها برای تشخیص است [۳]. خط فارسی چه دستنویس و چه چاپی ذاتا پیوسته است. این خط بصورت افقی و از راست به چپ نوشته می شوند. خط فارسی از نظر نحوه حرکت قلم و ساختار آن بسیار شبیه به خط عربی است و یک تشخیص دهنده کلمات فارسی می‌تواند برای تشخیص کلمات عربی نیز استفاده شود. تنها تفاوت بین اسناد فارسی و عربی در مجموعه حروف الفبای آن‌هاست، مجموعه حروف الفبای فارسی شامل همه ۲۸ حرف الفبای عربی به اضافه ٤ حرف دیگر {پ،ژ،گ،چ} است. حروف فارسی معمولا بصورت یکجا نوشته می‌شوند و در اکثر موارد با حرکات مکمل دیگر از قبیل نقاط، علائم مد و سرکش و غیره کامل می‌شوند.

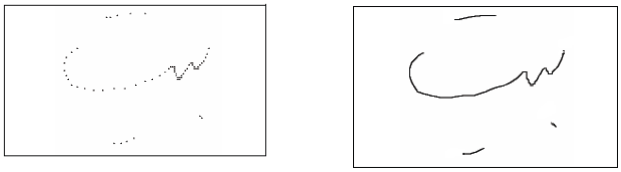
نقاط، کوچکترین و متنوع‌ترین بخش در زبان‌های فارسی و عربی هستند و برخلاف سایر زبان‌ها مانند انگلیسی که تنها دو حرف الفبای آن نقطه دارد، در زبان فارسی، بیش از نیمی از حروف (۸ حرف از ۳۲ حرف الفبا) نقطه‌دار هستند. نقاط یک کلمه در پایین و یا بالای خط زمینه قرار دارند و بصورت یک نقطه، دو نقطه و سه نقطه ظاهر می‌شوند. اگرچه اهمیت نقاط در زبان‌های فارسی و عربی غیر قابل انکار است اما از نقطه نظر آماری، نقاط نقش حیاتی ندارند و می‌توانند خطای تشخیص را افزایش دهند. علاوه بر این، ابهام در نوشتن نقاط گاهی اوقات سبب می‌شود تصویر یک کلمه به شکل‌ها و معانی کاملا متفاوتی خوانده شود، چرا که بسیاری از حروف، یک بدنه اصلی واحد دارند و تنها به‌وسیله حضور یا عدم حضور، محل و تعداد نقاط از یکدیگر متمایز می‌شوند. حروفی که دارای بدنه‌ی یکسانی هستند را می‌توان در مجموعه‌های جداگانه‌ای قرار داد که عبارتند از: {ب، پ، ت، ث، ن}، {ج، چ، ح، خ}، {د، ذ}، {ر، ز، ژ}، {س، ش}، {ص، ض}، {ط، ظ}، {ع، غ}، {ف، ق}.

برخلاف حروف انگلیسی، که به دو دسته حروف بزرگ و کوچک تقسیم می‌شوند، زبان فارسی این ویژگی را ندارد ولی در عوض حروف فارسی بسته به محل قرار گرفتنشان در کلمه می‌توانند چندین شکل مختلف داشته باشند.

۱–۳ انواع رویکردها در سیستم های بازشناسی خط

سیستم‌های بازشناسی خط را می‌توان از جنبه‌های مختلفی تقسیم‌بندی کرد. در ادامه چند نمونه از دسته‌بندی‌های رایج را شرح می‌دهیم.

- از لحاظ نوع الگوی ورودی، سیستم‌های بازشناسی می‌توانند به انواع، سیستم‌های برخط[۱] و سیستم‌های برون‌خط[۲] تقسیم شوند. در بازشناسی برخط، نوشتار، در همان زمان نگارش توسط سیستم تشخیص داده می‌شود و دستگاه ورودی این سیستم‌ها یک قلم نوری است. در این روش علاوه بر اطلاعات مربوط به قلم نوری، اطلاعات زمانی مربوط به مسیر قلم نیز در اختیار است که نقش مهمی را در این سیستم‌ها ایفا می کند. در بازشناسی برون‌خط، تنها از تصویر متن ورودی استفاده می‌شود و تفسیر داده‌ها مستقل از فرآیند تولید، صورت می‌گیرد [۴]، [۵]. شکل ۱-۱ تصویر کلمه‌ی "بیت" را به‌صورت برخط و برون‌خط نشان می‌دهد.



(ب) (الف)

شکل ۱-۱: تصویر کلمه‌ی "بیت". الف) برون‌خط ب) برخط

- بر اساس نوع نوشتار، سیستم‌های بازشناسی می‌توانند، به دو دسته‌ی سیستم‌های تشخیص نوشته‌های تایپی و نوشته‌های دستنویس تقسیم شوند. منظور از نوشته‌های

[۱] On-Line

[۲] Off-Line

تایپی (چاپی)، نوع نوشتاری است که توسط نرم‌افزارهای رایانه‌ای و با فونت‌های استاندارد نوشته می‌شود.

- بر اساس واحدهای زبان، این سیستم‌ها، به چهار سطح سیستم‌های تشخیص حروف و ارقام مجزا، سطح زیرکلمه، سطح کلمه و سطح جمله و متن تقسیم می‌شوند.
- بر اساس نوع ویژگی‌های استخراجی، روش‌های بازشناسی متن به چند گروه عمده تقسیم می‌شوند که عبارتند از: روش‌های مبتنی‌بر ویژگی‌های آماری[۱] و روش‌های مبتنی‌بر ویژگی‌های ساختاری[۲] و یا ترکیبی از هر دو روش [۲]. ویژگی‌های آماری به آن دسته از ویژگی‌ها اطلاق می‌گردد که بر اساس اطلاعات آماری که از یک کلمه استخراج می‌شود، به دست می‌آید. در مقابل ویژگی‌های ساختاری، به الگوی نوشتاری هر کلمه مربوط است و از روی نحوه‌ی نوشتن و شکل طبیعی حروف به دست می‌آیند.

۱-۴ اهداف تحقیق

بر اساس آنچه بیان شد، می‌توان اهداف این تحقیق را به‌صورت زیر بیان نمود:

۱. بهبود قابلیت بازشناسی کلمات فارسی و عربی با استفاده از تکنیک‌های رایج در هوش مصنوعی
۲. ارائه تکنیک‌های جدید جهت کاهش واژه نامه
۳. ارائه روشی کارآمد و جدید جهت توصیف کلمات فارسی
٤. به کارگیری خلاقیت در الگوریتم‌های موجود به‌منظور افزایش کارآیی
۵. پیشنهاد یک سیستم خودکار به‌منظور بازشناسی کلمات فارسی

[۱] Statistical Features

[۲] Structural Features

در این راستا این پایان‌نامه، یک سیستم خودکار به منظور بازشناسی کلمات فارسی و عربی پیشنهاد و ارائه می‌کند. اصول، مفاهیم و ویژگی‌های اصلی این چارچوب به‌طور تئوریک تشریح می‌شوند و سپس به‌منظور ارزیابی عملی سیستم، مدل پیشنهادی با دیگر روش‌های موجود مقایسه می‌شود.

۱–۵ ساختار پایان‌نامه

در این پایان‌نامه، سیستم پیشنهادی بازشناسی کلمات فارسی و عربی با کمک تکنیک‌های یادگیری ماشین مورد بررسی قرار گرفته است که در زیر شرح مختصر مباحث انجام گرفته در هر فصل آورده شده است:

- **فصل ۲:** در این فصل، مروری بر تحقیقات پیشین و چالش‌های موجود در این زمینه انجام شده است.
- **فصل ۳:** این فصل، دربرگیرنده مبانی روش تحقیق و جزئیات روش پیشنهادی به‌منظور بازشناسی کلمات است.
- **فصل ٤:** این فصل، چگونگی طراحی و پیاده‌سازی مدل پیشنهادی و نتایج حاصله از این پیاده‌سازی را شرح می‌دهد.
- **فصل ۵:** در این فصل، ویژگی‌های کلی روش‌های پیشنهادی بحث می‌شود. سپس نتیجه‌گیری کلی انجام شده و پیشنهادهایی نیز برای ادامه‌ی کار در آینده ارائه شده است.

فصل دوم

۲ پیشینه‌ی تحقیق

امروزه، حجم زیادی از اسناد کاغذی موجود، توسط اسکنرها یا دوربین‌ها، به اسناد تصویری دیجیتالی تبدیل می‌شوند. ذخیره‌سازی، بازیابی و مدیریت کارآمد این آرشیوهای تصویری در بسیاری از کاربردها نظیر برنامه‌های اتوماسیون اداری و کتابخانه‌های دیجیتالی اهمیت فراوان دارند. از این رو دستیابی به الگوریتم‌های موثر به‌منظور آنالیز تصویری اسناد یک نیاز اساسی به حساب می‌آید [۶]. بازشناسی متن، یکی از موضوعات بسیار جالب در شناسایی الگو[۱] است که در چند دهه‌ی اخیر، فعالیت‌های زیادی به خود اختصاص داده است. در حال حاضر بازشناسی نوری حروف به یکی از بخش‌های مهم و پرکاربرد این حوزه تبدیل شده است و روز به روز کاربردهای جدیدی در این زمینه شناسایی می‌شود.

۲–۱ تکنیک کاهش واژه نامه

در زمینه تشخیص نوری حروف، پیشرفت های قابل توجهی حاصل شده و بسیاری از کاربردها از قبیل خواندن خودکار آدرس های پستی، چک های بانکی و فرم ها پا به عرصه ظهور نهاده‌اند. اکثر کاربردهای منتشر شده برای بازشناسی کلمات لاتین و چینی است و بازشناسی

[۱] Pattern Recognition (PR)

کلمات فارسی و عربی به‌دلیل خصوصیات ویژه این زبان‌ها و حجم تحقیقات انجام شده پیشرفت بسیار کندی داشته است [۱۳].

بدلیل وجود برخی ابهامات و تنوع زیاد سبک‌های نوشتن، سیستم‌های تشخیص کلمه در بیشتر زبان ها عمدتا مبتنی‌بر مجموعه‌ای از کلمات هستند که واژه‌نامه[۱] نامیده می‌شوند. واژه‌نامه‌ها، فهرستی از کلمات مجاز را که انتظار می‌رود بوسیله سیستم، بازشناسی شوند، ارائه می دهند. بسته به نوع کاربرد، اندازه واژه‌نامه‌ها از ۳۰-۲۰ کلمه در خواندن مقادیر چک‌های بانکی شروع می‌شود و تا ۶۰،۰۰۰-۱۰،۰۰۰ کلمه برای بازشناسی متون انگلیسی نیز می‌رسد [۱۴]. استفاده از واژه‌نامه از جایی اهمیت یافت که تعداد کلمات تشخیصی افزایش یافت و سیستم‌ها باید تعداد کلمات بیشتری را بازشناسی می‌کردند. این موضوع سبب شد پیچیدگی محاسباتی مرحله‌ی بازشناسی افزایش یابد و کارایی عملکرد تشخیص، به شدت کم شود. پس از این مرحله، تلاش‌های بسیاری انجام گرفت تا پیچیدگی محاسباتی مرحله تشخیص، با روش‌های مختلف کاهش یابد. یکی از اصلی‌ترین تکنیک‌ها برای حل این مسئله، روش‌های کاهش واژه‌نامه[۲] است. این روش‌ها، سیستم بازشناسی را برای فائق آمدن بر تعداد زیاد الگوهای مرجع[۳] (کلمات واژه‌نامه)، یاری می‌دهند و روی دقت و سرعت فرآیند بازشناسی، تاثیرگذار هستند [۱۵].

هنگامی‌که در مرحله بازشناسی، یک شکل کلمه جستجو می‌شود، واژه‌نامه، پس از نگه‌داشتن تعدادی از اشکال، هرس می‌شود. این اشکال، تصاویری هستند که احتمال بیشتری برای تطبیق با کلمه مورد جستجو دارند. کاهش پیچیدگی محاسباتی، یکی از عمده ترین اهداف روش‌های کاهش واژه‌نامه است که می‌تواند به افزایش سرعت فرآیند بازشناسی بیانجامد [۱۶]. مسائل واژه‌نامه بزرگ، چندین مرتبه مشکل‌تر از واژه‌نامه‌های کوچک است. عملیات تشخیص در واژه‌نامه‌های بزرگ با روش‌های کاهش واژه‌نامه و حذف ابتدایی آن دسته از ورودی‌ها که

---

[۱] Lexicon

[۲] Lexicon Reduction Methods

[۳] Reference Patterns