نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

انتخاب ویژگی‌های مناسب برای توصیف شکل کلمات چاپی با استفاده از توصیفگر مکان‌های مشخصه

هما داودی[۱]، احسان اله کبیر[۲]
[۱] دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، h.davoudi@modares.ac.ir
[۲] دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، kabir@modares.ac.ir

چکیده

با توجه به پیوستگی حروف در کلمات فارسی، استفاده از روش‌های توصیف مبتنی بر شکل کلی کلمات در بهبود عملکرد سیستم‌های بازشناسی موثر است. با توجه به پیچیدگی شکل کلمات ، معمولا برای توصیف دقیق آن‌ها ویژگی‌های متعددی از هر تصویر استخراج می‌شود. حال آنکه با توجه به ساختار متنوع کلمات، ممکن است استفاده از تمام ویژگی‌ها برای توصیف همه‌ی آن‌ها نیاز نباشد. بنابراین، تعیین ویژگی‌های مناسب برای هر کلاس و بکارگیری این ویژگی‌ها در توصیف نمونه‌های متعلق به آن کلاس، می‌تواند در بهبود توصیف شکل کلمات موثر افتد. در این مقاله به بررسی روش‌های انتخاب ویژگی مبتنی بر اطلاعات کلاس نمونه‌ها می‌پردازیم و روشی برای انتخاب ویژگی‌های مناسب برای هر کلمه، با استفاده از توصیفگر مکان‌های مشخصه ارائه می‌کنیم. برای هر زیر-کلمه، زیر مجموعه‌ای از ویژگی‌ها که بیشترین تمایز را بین نمونه های یک کلاس با سایر نمونه ها ایجاد می کند، به عنوان ویژگی‌های مناسب برای آن زیر-کلمه در نظر گرفته می‌شود. روش ارائه شده بر مجموعه‌ای از شکل زیر-کلمات اعمال شده و با انجام آزمایش‌های مختلف کارایی آن ارزیابی شده است.

کلمات کلیدی

توصیفگر شکل، زیر-کلمات چاپی، مکان‌های مشخصه، انتخاب ویژگی، خط فارسی.

۱- مقدمه

بازشناسی کلمات، بر اساس دو رویکرد کلی مبتنی بر قطعه بندی و مبتنی بر شکل کلی انجام می‌شود. در رویکرد مبتنی بر قطعه بندی، هر کلمه ابتدا به مجموعه‌ای از زیر-تصاویر قطعه بندی می‌شود و سپس، براساس نتایج بازشناسایی حروف تشکیل دهنده آن شناخته می‌شود. حال آنکه در رویکرد مبتنی بر شکل کلی، ویژگی‌های بدست آمده از کل شکل کلمه برای بازشناسی آن بکار می‌روند و بازشناسی در سطح کلمه انجام می‌شود. در این روش، هر کلمه جزئی مستقل در نظر گرفته می‌شود و بنابراین نیازی به قطعه بندی اولیه تصویر نیست. علاوه براین، با توجه به اینکه برای هر کلمه یک کلاس در نظر گرفته می‌شود، ویژگی‌های منحصر به شکل کل یک کلمه نیز در بازشناسی وارد می‌شوند؛ این ویژگی‌ها ممکن است در بررسی مجزای حروف نادیده گرفته شوند.

کارایی روش‌هاي مبتنی بر توصیف شکل کلی در سیستم‌های بازشناسی کلمات در تحقیقات مختلف نشان داده شده است. استفاده از اطلاعات شکل کلی کلمه در بازیابی میان مجموعه محدود کلمات، حجم پردازش را به شکل قابل ملاحظه‌اي کاهش می‌دهد. همچنین، به کارگیری این ویژگی‌ها روشی کارامد براي نشان کردن کلمات پرس و جو در تصاویر اسناد به حساب می‌آید.

تحقیقات متعددی برای توصیف شکل کلمات از روش‌های رایج در توصیف شکل‌های عمومی استفاده کرده‌اند و توصیفگرهای مختلف سراسری را برای استخراج ویژگی از شکل کلمات ارائه داده‌اند. [۱-۵] . با توجه به پیوسته نویسی در خط فارسی، ویژگی‌های مبتنی بر شکل‌کلی برای توصیف کلمات فارسی مناسب است. از این رو در تعدادی از مقالاتی که برای بازشناسی کلمات فارسی ارائه شده، این رویکرد بررسی شده است. در [۶] از ۴۵ گشتاور زرنیکی برای توصیف زیر-کلمات چاپی و دستنویس فارسی استفاده شده است. در روشی دومرحله‌ای که در[۱] برای بازشناسی زیر-کلمات ارائه شده، از ویژگی‌های مکان مشخصه و توصیفگرهای فوریه کانتور استفاده شده است. در [۷] به منظور کاهش فضای جستجو در سیستم‌های بازشناسی، یک دیکشنری تصویری ارائه شده است که براساس خوشه بندی مجموعه زیرکلمات به دست آمده است. درآن تحقیق، توصیفگرهای مختلفی مانند توصیفگرهای فوریه، گشتاورها و ویژگی‌های مبتنی بر ناحیه بندی بررسی شده‌اندو در نهایت ویژگی‌های مکان مشخصه برای توصیف شکل کلمات انتخاب شده اند. با توجه به کارآمدی ویژگی‌های مکان مشخصه در توصیف شکل زیرکلمات چاپی فارسی، در این مقاله نیز برای توصیف شکل زیرکلمات از این توصیفگر استفاده می‌کنیم.

توصیفگر مکان‌های مشخصه توصیفگری مبتنی بر ناحیه است که بر اساس موقعیت مکانی هر نقطه پس زمینه، عددی (کدی) به آن نسبت می‌دهد. مجموعه نقاط با کدهای مشابه، نواحی مختلف را در شکل ایجاد می‌کنند. با کنار هم قرار دادن ویژگی‌های به دست آمده از هر کدام از این نواحی، بردار ویژگی نهایی به دست می‌آید.

با توجه به پیچیدگی شکل کلمات فارسی، برای توصیف دقیق‌تر کلمات نیاز است تا جزئیات شکل کلمات بررسی شود. از این رو در توصیفگر مکان‌های مشخصه، دامنه نسبتا بزرگی از اعداد برای کد کردن نقاط پس زمینه در نظر گرفته می‌شود (به