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

مقایسه روش مارکوف مخفی و پردازش زبان‌های طبیعی در مدل‌سازی آماری زبان فارسی

عاطفه سهرابی ، سعید مظفری
atefeh_sohraby@yahoo.com
mozaffari@semnan.ac.ir
دانشکده برق و کامپیوتر، دانشگاه سمنان

چکیده

در این مقاله، احتمال وجود یا عدم وجود کلمات گوناگون در زبان فارسی، در قالب مدل‌سازی کلمات این زبان، مورد بررسی قرار گرفته است. برای انجام این مدل‌سازی وجود یک پایگاه داده الزامی است، که برای این منظور لغت‌نامه بیژن‌خان استفاده شده است. مدل‌سازی کلمات بر مبنای دو روش مدل مارکوف مخفی و روش‌های آماری پردازش زبان انجام گرفته اند. در این مقاله، مرتبه اول و مرتبه دوم از هر یک از این دو روش، بطور مجزا بررسی شده، و روابط و چگونگی پیاده‌سازی آن‌ها تشریح شده‌اند. مرتبه اول از این روش‌ها، شیوه‌ای است که تا کنون برای مدل‌سازی کلمات مورد توجه بوده است. در حالیکه مرتبه دوم از روش‌ها برای مدل‌سازی کلمات، بعنوان روشی جدید در این مقاله ارائه می‌شود. نتایج بدست آمده، دقت هر روش در مدل‌سازی و تشخیص کلمات را بیان می‌کند، و موفقیت چشم‌گیر مدل مارکوف مخفی مرتبه دوم را در تشخیص کلمات صحیح و غیرصحیح در مقایسه با سایر روش‌ها نشان می‌دهد.

کلمات کلیدی

مدل‌سازی زبان[۱]، لغت‌نامه بیژن‌خان[۲]، مدل مارکوف مخفی[۳] (hmm)، روش‌های آماری پردازش زبان[۴].

۱- مقدمه

مدل‌سازی یک زبان، عملاً با استفاده از مدل‌سازی اجزای کوچکتر آن، مانند حروف، کلمات و جملات قابل انجام است. در این مقاله، دو شیوه مناسب مدل‌سازی کلمات توصیف شده است. بطور کلی یک مدل آماری احتمال کلمات را توصیف می‌کند. یعنی با استفاده از آن می‌توان با اطلاعات جاری، احتمال حرف‌های بعدی را در یک کلمه پیش‌بینی کرد. بدین ترتیب با ساخته شدن یک کلمه و قرار گرفتن کلمات بدست آمده در پی هم، جملات بدست می‌آیند و از ترکیب جملات، یک زبان طبیعی مدل‌سازی می‌گردد.

از کاربردهای وسیع مدل‌سازی زبان، در سیستم‌های اتوماتیک تشخیص گفتار، و تشخیص متون تایپی و دست‌نویس، است. ازجمله سایر کاربردهای آن می‌توان به استفاده در سیستم‌های ترجمه متون، پردازش زبان‌های طبیعی (NLP)، پردازش و تصحیح خودکار لغات در پردازنده‌های متن، تبدیل متن به صدا (TTS) و بهبود نتایج بازیابی اطلاعات متنی در موتورهای جستجو اشاره کرد.

با توجه به کاربردهای ذکر شده، در سال‌های گذشته تحقیقات بسیاری برای مدل سازی زبان‌های پراستفاده جهان و بیش از همه زبان انگلیسی انجام شده است و این تحقیقات همچنان ادامه دارند. لازم به ذکر است که مدل‌های آماری زبان انگلیسی تقریبا از سال ۱۹۸۰ در سیستم‌های واقعی به کار رفته‌اند، اما برای مدل‌سازی زبان فارسی متأسفانه هنوز هیچ تحقیق گسترده و قابل اعتنایی صورت نگرفته است.

در راستای هدف مدل‌سازی زبان فارسی، در ادامه دو روش مدل مارکوف مخفی و مدل‌های آماری در قالب روش N-gram برای بررسی وضعیت کلمات، ارائه می‌گردد. سازماندهی این مقاله به شکل زیر است: پایگاه داده مورد استفاده، در بخش ۲ بطور کامل تشریح می‌شود. مدل مارکوف مخفی و مدل‌های آماری N-gram به ترتیب در زیربخش‌های بخش ۳ به تفصیل مورد بحث قرار می‌گیرند و در نهایت مقایسه‌ای از این دو روش در بخش ۴ ارائه می‌گردد.

۲- پایگاه داده

برای هر مدل‌سازی، به پایگاه داده‌ای برای آموزش یا در واقع تنظیم پارامترهای آزاد مدل نیاز است. برای تهیه یک مدل آماری، داده‌های آموزشی باید دارای توزیع احتمالی باشند که بعداً مدل با آنها سروکار دارد؛ بعبارتی مدل آماری بدست آمده، زمانی کارآمد است که پایگاه داده شامل انواع مختلف کلمات، از