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

دسته‌بندی روش‌های محاسبه میزان تشابه معنایی لغات و جملات با بهره‌گیری از شبکه واژگان

احمد استیری، محسن کاهانی و فاطمه پورغلامعلی
ahmad.estiri@stu.um.ac.ir, kahani@um.ac.ir, pourgholamali.ro_am@stu.um.ac.ir
آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد

چکیده

امروزه با رشد چشمگیر اسناد منتشر شده در وب و نیاز اساسی به نگهداری، دسته‌بندی، بازیابی و پردازش آنها، توجه به پردازش زبان طبیعی توسط رایانه، بیش از پیش احساس می‌شود. در بسیاری از مواقع در کاربردهای مختلف پردازش زبان طبیعی، نیازمند محاسبه‌ی تشابه معنایی بین جملات و متناظراً کلمات هستیم. این مبحث در کاربردهای متعددی نظیر رفع ابهام واژه‌ها، خلاصه‌سازی متن، تصحیح خودکار لغات، ارزیابی خلاصه‌سازها و مترجم‌های ماشینی و موارد مشابه به شکل قابل توجهی مورد نیاز خواهد بود. اندازه‌گیری میزان تشابه ظاهری کلمات، نتایج چندان مطلوبی را در بر نخواهد داشت. روش‌هایی که برای اندازه‌گیری ارتباط معنایی کلمات از یک منبع لغوی استفاده می‌نمایند، آن منبع لغوی را به عنوان یک شبکه یا گراف می‌بینند و ارتباط معنایی را بر اساس خصوصیات مسیرها در این گراف محاسبه می‌نمایند. در بین منابع موجود، شبکه واژگان به شدت مورد توجه قرار گرفته و روش‌های متعددی برای محاسبه‌ی ارتباط بین کلمات بر اساس شبکه واژگان پیشنهاد گردیده است. در زبان فارسی نیز با توجه به تولید و توسعه‌ی دو شبکه‌ی واژگان فارس‌نت و فردوس‌نت، می‌توان محاسبه‌ی شباهت معنایی لغات را به جای محاسبه‌ی شباهت املایی و ظاهری لغات در کاربردهای فوق جهت بهبود نتایج مد نظر قرار داد.

کلمات کلیدی

پردازش زبان طبیعی، زبان فارسی، معناگرایی، تشابه معنایی، شبکه واژگان

۱- مقدمه

پردازش زبان طبیعی از جمله مسائل اساسی در حوزه هوش مصنوعی است که با توجه به گسترش روز افزون اسناد و اطلاعات منتشر شده در سالیان اخیر، توجهات گسترده‌ای را در زمینه‌های گوناگون به خود معطوف کرده است.

شباهت جملات مبحثی است که در زمینه‌های مختلف پردازش زبان طبیعی، بسیار تاثیرگذار می‌باشد. سیستم‌های پرسش و پاسخ، نیازمند تعیین شباهت بین جفت‌های سوال-پاسخ و یا سوال-سول می‌باشند [1]. در زمینه خلاصه‌سازی مبتنی بر گراف برای وزن‌دهی به یال‌ها، به شباهت بین جملات نیاز است [2]. کاربردهای دیگری چون دسته‌بندی متن [3]، و ترجمه ماشینی [4] از جمله زمینه‌های دیگری هستند که از شباهت جملات استفاده می‌نمایند. فرآیند محاسبه شباهت بین جملات، فرایندی بسیار دشوار و پیچیده است.

علی رغم اینکه بسیاری از کاربردها از معیارهای شباهت استفاده می‌کنند، اما بیشتر روش‌ها جملات را فقط بر مبنای سطح ظاهری مقایسه می‌کنند تا بر اساس معنا. علاوه بر این در اکثر روش‌هایی که پردازش‌های زبانی جمله را نیز در نظر می‌گیرند، تنها سطوح پایینی از آن مانند برچسب‌های بخش‌های سخن را مد نظر قرار می‌دهند.

روش‌های اندازه‌گیری شباهت جملات به سه دسته عمده تقسیم‌بندی می‌شوند [1]: معیارهای هم‌پوشانی کلمات، معیارهای TFIDF و معیارهای زبانی.

در معیارهای هم‌پوشانی کلمات، شباهت دو جمله بر اساس فرکانس کلمات مشترک بین دو جمله محاسبه می‌گردد. در [5] دو معیار مبتنی بر اشتراک کلمات برای تعیین شباهت بین جملات معرفی شده است: اشتراک ساده و اشتراک IDF . تابع اشتراک ساده کلمات $sim_{overlap}$ به صورت نسبت تعداد کلمات مشترک دو جمله به طول جملات تعریف می‌شود. تابع اشتراک IDF هم به صورت تعداد کلمات مشترک دو جمله که توسط معکوس تکرار سندشان (IDF) وزن‌دهی شده باشند، تعریف می‌شود. یکی از مشکلات این روش‌ها این است که تفاوتی بین عبارات تک کلمه‌ای و چند کلمه‌ای قائل نیستند.