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

بهبود کارایی سیستم خلاصه‌ساز متون فارسی با استفاده از الگوریتم هرس در شبکه‌های عصبی

نوشین ریاحی۱، فاطمه غزالی۲ و محمد علی غزالی۳
۱عضو هیئت علمی دانشگاه الزهرا(س) ، nriahi@alzahra.ac.ir
۲ دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه الزهرا(س) ، Fatemeh.Ghazali@Student.alzahra.ac.ir
۳ دانشجوی کارشناسی ارشد نرم افزار دانشگاه علم تحقیقات خراسان رضوی ، mohammadali.ghazali@gmail.com

چکیده

با رشد روز افزون منابع و اسناد در دسترس نیاز به سیستم‌های خلاصه‌سازی محسوس‌تر شده است، به همین خاطر در سال‌های اخیر سیستم‌های خلاصه‌سازی زیادی ارائه شده‌اند. سیتم های خلاصه ساز بر اساس نوع خروجی به دو دسته گزینشی و چکیده ای تقسیم می گردند. در سیتم های خلاصه ساز گزینشی انتخاب جملات خلاصه با توجه به برخی ویژگی های آن جمله صوررت می گرفت که در گذشته میزان تأثیر این ویژگی ها یکسان در نظر گرفته می شده است و یا با استفاده از تکنیک-های سعی و خطا ضرایب متغیری به هر یک از این پارامتر ها اختصاص داده می شد. در این مقاله علاوه بر معرفی ویژگی های جدید مبتنی بر پاراگراف (نظیر: پاراگراف‌های آغاز شده با کلمات استفهامی، پاراگراف های آغاز شده با انواع بالت، پاراگراف‌های حاوی کلمات کلیدی، طول پاراگراف) در فاز امتیاز دهی به جملات، با استفاده از شبکه عصبی نیز میزان اهمیت هر یک از پارامترها تعیین می شود. همچنین خلاصه نهایی با استفاده از تکنیک هرس کردن در شبکه عصبی تولید می گردد. در نهایت بر اساس ارزیابی انجام شده دریافتیم که استفاده از این رویکرد در سیستم‌های خلاصه‌سازی علاوه بر افزایش پیوستگی و بهبود مقیاس‌های دقت و فراخوانی باعث افزایش سرعت نسبت به سیستم های تقریباً مشابه موجود نیز شده است.

کلمات کلیدی

خلاصه سازی متن، شبکه عصبی، خلاصه سازی گزینشی(استخراجی)،الگوریتم هرس کردن، امتیاز دهی

۱ – مقدمه

افزایش روز افزون میزان اطلاعات متنی در دسترس افراد از یک سو و اهمیت صرفه جویی در زمان از سویی دیگر، نیاز به راهکار-های تسریع کننده دستیابی به اطلاعات را بارزتر نموده است، خلاصه‌سازی متون یکی از بهترین راه حل‌های ارائه شده در این زمینه می‌باشد.

خلاصه‌سازی متن فرآیند شناسایی برجسته‌ترین اجزای منبع و گردآوری آن در حجمی کمتر می باشد، که این اجزا بیشترین تطابق و تعلق به مفهوم و موضوع مطرح شده در سند را داشته باشد[1].

از این رو سیستم‌های خلاصه ساز متن را می توان بر اساس سه فاکتور ورودی، هدف و خروجی به انواع مختلفی طبقه‌بندی نمود، لیکن با توجه به اهمیت فاکتور خروجی می‌توان گفت سیستم مذکور به دو نوع چکیده‌ای و استخراجی تقسیم می گردد. هدف از خلاصه چکیده‌ای، استخراج و درک مفهوم اصلی متن می‌باشد. بنابراین عیناً از جملات موجود در متون ورودی استفاده نمی‌گردد و در اکثر موراد به تولید متن پرداخته می شود. اما در خلاصه استخراجی تولید متنی صورت نمی گیرد، بلکه هر جمله از متن موجود در خلاصه تولید شده رو نوشتی عینی از جمله‌ای در متن یا متون ورودی می‌باشد. تا کنون بیشتر پژوهش‌های انجام شده در این زمینه منجر به تولید سیستم‌های خلاصه‌ساز نوع استخراجی شده است.

عملیات کلی فرآیند خلاصه‌سازی متن در سیستم‌های استخراجی شامل سه مرحله اصلی پیش‌پردازش، پردازش و تولید خلاصه می-باشد. در مرحله پیش‌پردازش کلیه عملیات یکدست سازی متون از قبیل: حذف اطلاعات غیر ضروری، ریشه‌یابی کلمات متن اصلی، تعیین مرز واحدهای متنی و... صورت می‌پذیرد، سپس متون آماده پردازش در مرحله بعد بسته به نوع خلاصه ساز(چکیده ای-استخراجی) مورد تحلیل، پردازش و امتیازدهی قرار گرفته و در آخرین مرحله با توجه به جملات تولید شده در فاز قبل و امتیازات تخصیص یافته به هریک از واحدهای متنی(جمله، کلمه و...) همچنین درصد فشرده‌سازی، خلاصه مورد نظر تولید می‌گردد.

خلاصه سازی خودکار متون یکی از قدیمی‌ترین کاربردهای پردازش زبان طبیعی است که آغاز آن به دهه ۵۰ میلادی باز می گردد[2]. لیکن با توجه به مورد بحث بودن خلاصه‌سازی متون فارسی به معرفی پیشینه برخی از مهمترین کارهای انجام شده در زبان فارسی پرداخته می شود: