

سیستم خلاصه‌ساز استخراجی ترکیبی بر اساس طبقه‌بندی کننده‌های با نظارت برای متون فارسی

زینب رحیمی^۱، جهان‌شاه کبودیان^۲، طاهره امامی آزادی^۱، یاسر شکفته^۱، ایمان صراف رضایی^۱

^۱ گروه پردازش صوت و زبان‌های طبیعی، پژوهشکده پردازش هوشمند علائم (RCISP)، تهران.
{Rahimi, T.emami, Shekofteh, Sarraf}@rcisp.com

^۲ استادیار، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی دانشگاه رازی، کرمانشاه.
kabudian@razi.ac.ir

چکیده

با افزایش روزافزون حجم اطلاعات متنی در موضوعات مختلف، هر روز بر گستره اطلاعات قابل دسترس برای کاربران افزوده می‌شود. این رشد علاوه بر مزایای قابل توجهش، با ایجاد حجم بالای اطلاعات و در نتیجه سردرگمی کاربران در گزینش مطالب، ضرورت سامان‌دهی مطالب و ذخیره‌سازی آن‌ها بصورت بهینه و حذف اطلاعات افزوده از آن‌ها را ایجاد می‌کند. سیستم‌های خلاصه‌سازی خودکار متون می‌توانند به گزینش مطالب از بین انبوه اطلاعات موجود کمک شایانی نمایند. در این راستا یک سیستم خلاصه‌ساز استخراجی تک‌سندی با نظارت معرفی می‌گردد که در طراحی آن از ترکیب خروجی دو ابزار طبقه‌بندی (ماشین بردار پشتیبان) و SSVM (SVM ساختاری) استفاده شده است. ویژگی‌های مورد استفاده در این سیستم شامل مکان جمله، شباهت به اسامی یا عبارات مناسب، وجود عبارت نشانه، نقل قول‌ها، اطلاعات از پیش دانسته، طول جمله و شمارش کلمات است. نتایج حاصل از ارزیابی سیستم مذکور در مقایسه با سیستم‌های معرفی شده خصوصاً برای زبان فارسی بهبود قابل ملاحظه‌ای را نشان می‌دهد.

کلمات کلیدی

خلاصه‌سازی متن، خلاصه‌سازی تک‌سندی، خلاصه‌سازی استخراجی، SVM.

تولید خلاصه از یک یا چند سند، در بر گرفتن اطلاعات مهم منابع اصلی و در عین حال کوتاه بودن آن است.

۱- مقدمه

محتوای متن خلاصه به مواردی مانند شرایط متن ورودی، طبیعت متن، هدف خلاصه‌سازی، قصد و سلیقه خواننده و ... وابسته است. برخی عواملی که در خلاصه‌سازی اهمیت دارند عبارتند از [6]:
(۱) عوامل ورودی: شامل مواردی از قبیل فرم ورودی (شامل طول متن ورودی، ساختار متن)، زبان متن (انگلیسی، فرانسه...)، نوع متن (متن ادبی، داستانی و...) و موضوع متن است.
(۲) عوامل خروجی: شامل سه عامل محتوا، قالب و شیوه است. عامل محتوا مربوط به آن است که خلاصه عمومی یا پاسخ پرس و جو،

فرایند خلاصه‌سازی خودکار متون از جمله مسائل دشوار در پردازش زبان طبیعی^۱ به شمار می‌آید که در حالت ایده‌آل بر پایه فهم مطلب استوار است و باید از عملکرد شناختی مغز انسان در این کار پیروی نماید. در یک تعریف کلی، خلاصه‌سازی عبارت است از نمایش متن/متون در قالب یک متن، به طوری که در خروجی، اطلاعات مهم متن/متون اصلی حفظ شده باشد و طول چکیده یا خلاصه، نباید بیش از نصف متن یا متون اصلی باشد [5]. چنانچه از این تعریف استنباط می‌شود، ویژگی عمده در خلاصه‌سازی خودکار، قابلیت