ارائهٔ یک سامانهٔ هوشمند و معناگرا برای ارزیابی سامانه‌های خلاصه‌ساز متون

راضیه طباطبائی[1]، محمدرضا فیضی درخشی[2] و سعید معصومی[3]

[1] باشگاه پژوهش‌گران جوان و نخبگان، واحد بناب، دانشگاه آزاد اسلامی، بناب، ایران

[1و2] دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

[3] پژوهشکدهٔ کامپیوتر، مجتمع ICT، دانشگاه صنعتی مالک اشتر، تهران، ایران

چکیده

امروزه با افزایش منابع متنی در شبکه جهانی وب، هر روز بر گسترهٔ اطلاعات قابل دسترس برای کاربران افزوده می‌شود؛ بنابراین جهت نگهداری و بازیابی و پردازش آنها از سامانه‌های خلاصه‌سازی خودکار متن، استفاده می‌کنیم. میزان کیفیت خلاصه‌سازهای ماشینی، توسط انسان‌ها مورد بررسی قرار می‌گیرد؛ اما این کار نیروی متخصص و زمان زیادی را می‌طلبد و هزینه‌بر خواهد بود؛ بنابراین برای حل این مشکل، در این مقاله سامانه‌ای به نام TabEval برای ارزیابی سامانه‌های خلاصه‌سازی خودکار متن ارائه شده است.

این سامانه با ایده و معماری جدید به محاسبهٔ میزان تشابه ظاهری و معنایی بین خلاصهٔ سامانه‌ای و خلاصه‌های انسانی (خلاصه‌های ایده‌آل) می‌پردازد. برای محاسبه میزان تشابه معنایی از شبکهٔ واژه‌ها استفاده می‌شود. خروجی حاصل از این سامانه توسط نیروهای متخصص در زمینهٔ ادبیات زبان فارسی مورد بررسی قرار گرفت. نتایج حاصل از بررسی‌ها حاکی از این بود که این سامانه همانند انسان، هوشمندانه عمل می‌کند.

واژگان کلیدی :ارزیابی هوشمند، خلاصه‌سازهای سامانه‌ای، پردازش زبان طبیعی، F-measure، معیار ارزیابی، پیوندهای هم‌رخداد، شبکهٔ واژه‌ها.

۱- مقدّمه

با توجه به افزایش حجم مستندات متنی، برای پاسخ‌گویی به نیازهای اطلاعاتی کاربران، دیگر روش‌های فنی بازیابی اطلاعات به‌تنهایی کارا نیستند. مطالعهٔ حجم زیاد متون برای کاربران بسیار سخت و زمان‌بر است و دراختیار داشتن خلاصه‌ای از مطالب مهم آنها، می‌تواند بسیار مفید باشد. فرآیند فشرده‌سازی یک منبع را به‌صورتی‌که حاصل حاوی اطلاعات مهم آن باشد، خلاصه‌سازی گویند. (شمس‌فرد، ۲۰۱۰) خلاصه‌سازی متون منجر به استفاده از منابع بیشتر با سرعت بالاتر و درنتیجه حاصل‌شدن اطلاعات غنی‌تر می‌شود.

با توجه به اهمیت بسیار زیاد خلاصه‌سازها، ارزیابی کیفیت خلاصه‌های تولیدشده توسط خلاصه‌سازها، موضوع مهمی است. به‌طور کلی، دو رهیافت برای ارزیابی خلاصه‌های ایجادشده توسط سامانه‌های خلاصه‌سازی خودکار وجود دارد: مقایسهٔ خلاصه با خلاصه‌های مرجع و یا با نظر مستقیم چند داور (نیروی متخصص).

برای ارزیابی سامانه‌های خلاصه‌ساز، میزان تشابه بین خلاصهٔ ماشینی و خلاصهٔ انسانی را می‌توان اندازه‌گیری کرد؛ به‌طوری‌که هر چقدر این تشابه زیاد باشد، نشان‌دهندهٔ این است که سامانهٔ خلاصه‌ساز بهتر عمل می‌کند. قبل از هر کاری بایستی پیش‌پردازش‌هایی روی متون انجام گیرد تا متون به شکلی استاندارد درآیند و قابل قیاس باشند. از آنجایی که هر شخص با توجه به نظر شخصی خود بخشی از متن را مهم‌تر تشخیص خواهد داد؛ بنابراین برای کاهش دخالت سلیقهٔ شخصی در تهیهٔ خلاصهٔ انسانی، چندین خلاصهٔ انسانی را با ایدهٔ جدیدی ترکیب کرده و یک خلاصهٔ

ایده‌آل به‌دست می‌آوریم و خلاصهٔ ماشینی را با این خلاصهٔ ایده‌آل مقایسه می‌کنیم. در مرحلهٔ پیش‌پردازش ابتدا متن مورد نظر را به جملات تشکیل دهنده‌اش می‌شکنیم سپس این جملات را هم به کلمات تشکیل دهنده‌شان می‌شکنیم و در نهایت کلمات بدست آمده را ریشه‌یابی می‌کنیم. درواقع این ریشهٔ کلمات دو متن خلاصهٔ ماشینی و خلاصهٔ ایده‌آل با هم مقایسه می‌شوند؛ ولی از آنجایی که این قیاس از نظر ظاهری می‌باشد، لذا از شبکهٔ واژه‌ها برای قیاس معنایی استفاده می‌کنیم. برای در نظر گرفتن ترکیبی از لغات در سطح معنا از پیوندهای هم‌رخداد استفاده می‌کنیم؛ بنابراین این سامانه علاوه‌بر خلاصه‌سازهای استخراجی، خلاصه‌سازهای چکیده‌ای را چه در سطح لغات و چه در سطح ترکیبی از لغات ارزیابی می‌کند. چون در خلاصه‌سازهای چکیده‌ای، کلمات خلاصهٔ تولیدشده می‌تواند تغییر کند؛ بنابراین بایستی از نظر معنایی نیز مقایسه شوند.

شبکهٔ واژه‌ها درواقع شبکه‌ای معنایی از بیش از یکصد هزار مفهومی است که با روابط معنایی به هم مرتبطند. کلمات هم‌معنی در یک گروه قرار می‌گیرند و هر گروه نشان‌دهندهٔ یک معنی خاص است. این گروه‌ها می‌توانند با استفاده از روابط معنایی به همدیگر متصل شوند. (مانی، ۱۹۹۹)

۲- کارهای پیشین

سامانهٔ ارائه‌شده در این مقاله برای ارزیابی خلاصه‌سازهای زبان فارسی کاربرد دارد؛ لذا ابتدا مروری بر کارهای انجام‌شده در زمینهٔ خلاصه‌سازهای زبان فارسی خواهیم داشت و سپس کارهای پیشین سامانه‌های ارزیابی را مورد بررسی قرار می‌دهیم. سامانه‌های ارزیابی موجود برای زبان‌هایی غیر از زبان فارسی طراحی شده‌اند.

۲-۱- سامانهٔ FarsiSum

سامانهٔ FarsiSum (مزدک، ۲۰۰۴) یک ابزار خلاصه‌ساز تحت وب برای زبان فارسی است که بر پایهٔ روش SweSum (دالیانیس، ۲۰۰۰) که برای متون سوئدی مورد استفاده قرار می‌گیرد، ایجاد شده است. این سامانه قادر به خلاصه‌کردن متون روزنامه‌های فارسی با قالب HTML و متن کدشده با فرمت یونیکد[1] است. این سامانهٔ تحت ویندوز و به زبان پرل نوشته شده و برخی از ویژگی‌های SweSum را پیاده‌سازی نکرده است. Farsisum همان ساختار SweSum را استفاده کرده با این تفاوت که برخی از ماژول‌ها را به‌منظور استفاده برای محتوای یونیکد، اصلاح کرده است. این سامانه توسط نیما مزدک در سال ۲۰۰۴ ارائه شده و معروف‌ترین سامانهٔ خلاصه‌ساز فارسی است. در این سامانه به نخستین جمله وزن زیادی داده می‌شود.

۲-۲- سامانهٔ خلاصه‌ساز ایجاز[2]

سامانهٔ خلاصه‌ساز ایجاز، سامانه‌ای برای خلاصه‌سازی تک‌سندی و چندسندی متون خبری فارسی است. (پورمعصومی، ۱۳۹۳)

۲-۳- سامانهٔ خلاصه‌ساز TabSum

سامانهٔ TabSum یک ابزار خلاصه‌ساز برای زبان فارسی است که در سال ۹۲ در آزمایشگاه پردازش زبان دانشگاه تبریز توسط معصومی طراحی و پیاده‌سازی شده است. در این سامانه، ابتدا عملیات پیش‌پردازشی روی متن ورودی انجام می‌شود؛ سپس براساس مجموعه‌ای از ویژگی‌ها به جمله‌ها امتیازاتی داده می‌شود. امتیاز نهایی جمله ترکیب وزن‌داری از این ویژگی‌ها خواهد بود. در آخر نیز جملاتی با بالاترین امتیاز نهایی برای قرارگرفتن در خلاصه گزینش می‌شوند. در این سامانه از الگوریتم ژنتیک برای یافتن بهترین وزن ویژگی‌ها استفاده کرده‌اند. برای محاسبه امتیاز معنایی جملات از شبکه واژگان بهره برده شده است. در پیاده‌سازی این سامانهٔ خلاصه‌ساز با افزودن فاکتورهای جدید، بر دقت خلاصه‌ساز افزوده شده است. (معصومی، ۲۰۱۴)

۲-۴- ابزارهای ارزیابی خلاصه‌سازها

در زمینهٔ ارزیابی خلاصه‌سازهای فارسی، متأسفانه تاکنون کار قابل توجهی صورت نگرفته و سامانه خاصی وجود ندارد. در ادامه چند نمونه از ابزارها و سامانه‌هایی که برای ارزیابی خلاصه‌سازها در زبان‌های دیگر مورد استفاده قرار می‌گیرند، معرفی شده است.

۲-۴-۱- محیط ارزیابی خلاصه‌ها

محیط ارزیابی SEE، محیطی است که در آن، ارزیاب‌ها می‌توانند کیفیت یک خلاصه را در مقایسه با یک خلاصهٔ مرجع، مورد سنجش قرار دهند. متونی که درگیر ارزیابی هستند، با شکسته‌شدن به فهرستی از قطعات (عبارات،

[1] Unicode

[2] http://ijaz.um.ac.ir/

جملات و ...) مورد پیش‌پردازش قرار می‌گیرند. برای مثال هنگامی که یک سامانهٔ گزینشی با اندازهٔ قطعه جمله را ارزیابی می‌کنیم، ابتدا متون با شکسته‌شدن به جملات، آماده‌سازی می‌شوند و سپس در ادامه، قضاوت انسانی، تعیین‌کنندهٔ میزان کیفیت خلاصه‌ها خواهد بود. نسخهٔ خاصی از SEE در کنفرانس‌های DUC2001 [1] تا DUC2004 برای ارزیابی درونی متون خلاصهٔ اخبار، مورد استفاده قرار گرفت. (هاسل، ۲۰۰۴)

MEADeval -۲-۴-۲

MEADeval (رادِو، ۲۰۰۴) ابزاری برای ارزیابی خلاصه‌های استخراج‌شده با فرمت DUC و MEAD است که به زبان پرل پیاده‌سازی شده است و با مقایسهٔ خلاصهٔ سامانه‌ای با یک خلاصهٔ مرجع (یا خلاصهٔ ایده‌آل)، این سنجش را انجام می‌دهد. MEADeval به‌طور اساسی برای خلاصه‌سازهایی که خلاصه‌های استخراجی تولید می‌کنند، مناسب است؛ چون متن هر جمله در خلاصه‌های استخراجی، از تعدادی جملات در سند منبع، برگرفته شده است. با این حال می‌توان آن را به‌طور عمومی نیز به‌کار برد. این ابزار، تعدادی از معیارهای استاندارد را پشتیبانی می‌کند و با توجه به اینکه به زبان پرل پیاده‌سازی شده است، مستقل از پلت فرم عمل می‌نماید. از جمله مزایای این ابزار، قابلیت تطبیق و تبدیل بسیار راحت پیاده‌سازی آن به زبان‌های دیگر است. (لین، ۲۰۰۲) این ابزار درکل دو نوع طبقهٔ معیار برای ارزیابی دارد که عبارتند از: معیارهای co-selection و معیارهای content-based. معیارهای co-selection شامل precision، recall، Kappa و Relative Utility می‌باشد. معیارهای content-based نیز از cosine (که از TF*IDF استفاده می‌کند)، simple cosine، unigram و bigram-overlap تشکیل یافته است.

ISI ROUGE بستهٔ ارزیابی خودکار خلاصه -۳-۴-۲

بستهٔ ISI ROUGE که بعدها با نام ROUGE [2] (لین، ۲۰۰۴) معروف شد، تلاشی برای خودکارکردن ارزیابی خلاصه‌هاست. این ابزار که برای زبان انگلیسی پیاده‌سازی شده مبتنی بر فراخوانی n تایی‌های مشترک بین خلاصه‌های ماشینی و خلاصه‌های مرجع (خلاصه‌های انسانی) است. در حال حاضر، ROUGE معروف‌ترین و پرکاربردترین ابزار ارزیابی است. (لین، ۲۰۰۴)

۳- ابزار پیشنهادشده

هدف نهایی سامانه‌های خلاصه‌سازی، تولید خلاصه‌هایی باکیفیت نزدیک به خلاصه‌های انسانی است؛ بنابراین ابزاری به نام TabEval که برای ارزیابی خلاصه‌سازها پیشنهاد و طراحی کرده‌ایم، میزان شباهت خلاصهٔ سامانه‌ای به خلاصهٔ انسانی را محاسبه می‌کند و درنهایت اگر این میزان شباهت زیاد باشد، نشان‌دهندهٔ این است که عملکرد خلاصه‌ساز خوب است. جهت بررسی کیفیت خلاصه‌سازهای فارسی توسط سامانهٔ ارائه‌شده نیازمند پیکره بودیم. در زمینهٔ خلاصه‌سازی متون فارسی تاکنون پیکرهٔ رسمی و استانداردی تولید نشده است؛ بنابراین پیکره‌ای متشکل از سی سند که از اخبار همشهری تهیه کردیم و هر سند توسط پنج فرد خبره با نرخ فشرده‌سازی ۳۰٪ خلاصه شده و مورد استفاده قرار گرفت. در روش پیشنهادی برای اینکه مشخص شود در تولید خلاصه‌های انسانی، افراد خبره بر کدام جمله‌ها تأکید بیشتری داشته‌اند، خلاصه‌های انسانی را به مجموعه‌های N تایی ، N-1 تایی و ... و ۱ تایی تقسیم کردیم؛ به‌طوری‌که مجموعهٔ N شامل جمله‌های با N بار تکرار در خلاصه‌های انسانی است.

الگوریتم روش پیشنهادی شامل شش مرحلهٔ کلی است:

۱) تبدیل خلاصه‌های انسانی به مجموعه‌های N و N-۱ و ... تا ۱ عضوی
۲) تقسیم خلاصهٔ سامانه‌ای به دو قسمت مشترک و غیرمشترک
۳) امتیازدهی به جملات قسمت مشترک خلاصهٔ سامانه‌ای
۴) پیش‌پردازش مجموعه‌های خلاصهٔ انسانی و قسمت غیر مشترک خلاصهٔ سامانه‌ای جهت اندازه‌گیری میزان شباهتشان از نظر معنایی و ظاهری
۵) یکسری معیارهایی جهت امتیازدهی به جملات قسمت غیر مشترک خلاصهٔ سامانه‌ای
۶) محاسبهٔ امتیاز نهایی خلاصهٔ سامانه‌ای

در روش به‌کار گرفته شده، اگر فرض کنیم که برای هر سند N تا خلاصهٔ انسانی وجود داشته باشد، مجموعه‌هایی با اندیس ۱ تا N را تشکیل می‌دهیم. بدین صورت که اگر جمله‌ای از خلاصه ماشینی در N تا از خلاصه‌های انسانی حضور داشته باشد، عضو مجموعهٔ با

[1] http://duc.nist.gov
[2] Recall-Oriented Understudy for Gisting Evaluation

اندیس N و اگر در N-۱ تا خلاصهٔ انسانی حضور داشته باشد، عضو مجموعهٔ با اندیس N-۱ و ... است. درواقع توسط این مجموعه‌ها میزان اهمیت یک جمله تعیین می‌شود. جملات عضو مجموعهٔ با اندیس N در همهٔ خلاصه‌های انسانی حضور داشته‌اند، که نشان‌دهندهٔ این است که افراد بر این جملات تأکید داشته‌اند؛ پس دارای اهمیت زیادی هستند. به‌طورکلی اندیس مجموعه (N, N-۱, ..., ۱) تعداد تکرار جملات عضو آن مجموعه را در خلاصه‌های انسانی نشان می‌دهد.

خلاصهٔ سامانه‌ای را به‌صورت دو قسمت مشترک و غیر مشترک در نظر می‌گیریم. حال نوبت آن می‌رسد که جملات خلاصهٔ سامانه‌ای را با جملات مجموعه‌ها مقایسه کنیم. در نتیجهٔ مقایسه، جملاتی که به‌طور کامل به هم شباهت دارند در قسمت مشترک و بقیهٔ جملات در قسمت غیر مشترک قرار می‌گیرند. جملات موجود در قسمت مشترک براساس شمارهٔ مجموعه‌شان امتیازدهی می‌شوند؛ در مرحله بعدی مجموعه‌ها و قسمت غیر مشترک وارد مرحلهٔ پیش‌پردازش می‌شوند.

در مرحلهٔ پیش‌پردازش ابتدا متون خلاصهٔ انسانی و خلاصهٔ سامانه‌ای یکسان‌سازی می‌شوند و سپس ابزار تشخیص‌دهندهٔ جملات، تشخیص‌دهندهٔ کلمات، حذف‌کنندهٔ کلمات توقف (واژه‌های عمومی)، ریشه‌یاب و برچسب زن بایستی به‌ترتیب بر روی متون اعمال شوند.

در فاز پیش‌پردازش این سامانه از پیمانهٔ ریشه‌یاب و برچسب‌زن سامانهٔ ایجاز (پورمعصومی، ۱۳۹۳) استفاده شده است. همچنین برای شبکهٔ واژگان از نسخهٔ ۲ فارس نت دانشگاه شهید بهشتی استفاده شده است. (شمس‌فرد، ۲۰۱۰) تشخیص‌دهندهٔ جملات با استفاده از علامت‌های جداکنندهٔ جملات و با به‌کارگیری برخی قواعد گرامری زبان فارسی، مرز جمله‌ها را برای استفاده در مراحل بعدی تعیین می‌کند. تشخیص‌دهندهٔ کلمات با استفاده از علامت‌های جداکنندهٔ کلمات، واژه‌ها را شناسایی می‌کند. حذف‌کنندهٔ کلمات توقف نیز کلمات کم‌اهمیت را حذف می‌کند. حذف واژه‌های عمومی (ایست‌واژه‌ها[1]) در متون فارسی در اکثر موارد، باعث بهبود نتیجهٔ ارزیابی (معصومی، ۲۰۱۴) می‌شود. در این سامانه از سی‌صد ایست‌واژه که ترکیبی از واژه‌های توقف سامانه‌های مختلف پردازش بودند استفاده شد؛ سپس ریشه‌یاب نیز ریشهٔ واژه‌های شناسایی‌شده از مرحلهٔ قبل را بدست می‌آورد و برچسب‌زن نیز به تشخیص گروه‌های اسمی می‌پردازد؛ همچنین بعد از اینکه مرحلهٔ پیش‌پردازش اعمال شد، برای تعیین کیفیت خلاصه‌ها به‌صورت خودکار، با درنظرگرفتن معیارهایی و همچنین با استفاده از شبکهٔ واژگان، میزان شباهت ظاهری و معنایی جملات غیر مشترک با مجموعه‌ها را محاسبه می‌کنیم. این معیارها تعداد واحدهایی را که بین خلاصه‌های سامانه‌ای و مجموعهٔ خلاصه‌های انسانی هم‌پوشانی دارند، محاسبه می‌کند. نمودار (۳-۱) نشان‌دهندهٔ شش مرحلهٔ اصلی روش پیشنهادی است.

۳-۱- امتیازدهی قسمت مشترک خلاصهٔ سامانه‌ای

اگر تعداد کل جملات خلاصهٔ سامانه‌ای M باشد، حداکثر امتیاز یک جمله (S_i) عددی مابین صفر تا یک خواهد بود.

$$0 \le Score(S_i) \le 1$$ رابطهٔ (۱)

جملات خلاصهٔ سامانه‌ای با ابزار جداکنندهٔ جملات شکسته می‌شوند؛ سپس هر جمله‌اش (S_i) با جملات مجموعه‌های N و N-1 و... و ۱ (نحوهٔ تشکیل این مجموعه‌ها در بخش معماری روش پیشنهادی توضیح داده شد) مقایسه می‌شود اگر در مجموعه‌ها نیز تکرار شده باشد در قسمت مشترک خلاصهٔ سامانه‌ای قرار گرفته و امتیاز $\frac{j}{N}$ به آن جمله تعلق می‌گیرد که در آن j شماره مجموعهٔ شامل جمله و N تعداد کل خلاصه‌های انسانی است.

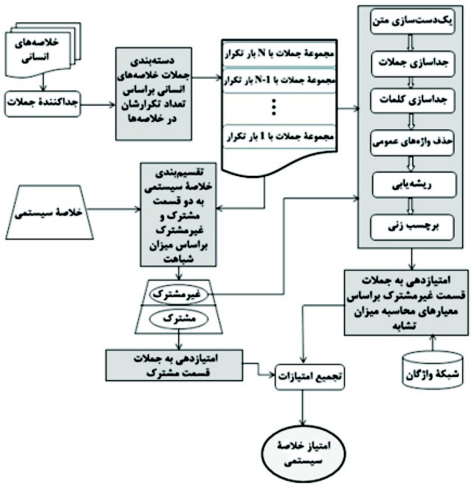
۳-۲- امتیازدهی قسمت غیر مشترک خلاصهٔ سامانه‌ای

همان‌طور که در بخش‌های قبلی مطرح شد، جملاتی از خلاصهٔ سامانه‌ای که در هنگام مقایسه در هیچ‌کدام از مجموعه‌های خلاصهٔ انسانی تکرار نشده‌اند، در قسمت غیر مشترک قرار می‌گیرند. جملات این قسمت از نظر ظاهری در خلاصه‌های انسانی نیامده‌اند؛ اما این دلیل نمی‌شود که امتیاز صفر به این جملات تعلق گیرد؛ چون این احتمال وجود دارد که از لحاظ معنایی شبیه جملات مجموعه‌های خلاصهٔ انسانی باشند. بنابراین برای محاسبهٔ امتیاز قسمت غیر مشترک مجبوریم که تک‌تک جملات غیر مشترک را با تک‌تک جملات مجموعه‌ها مقایسه کرده و براساس میزان تشابه به آنها امتیازی را اختصاص می‌دهیم؛ سپس امتیاز جمله‌ای را که دارای بیش‌ترین درجهٔ تشابه باشد، انتخاب می‌کنیم. این مقایسه توسط معیارهایی مانند طولانی‌ترین

[1] Stop words

زیررشتهٔ مشترک (LS)، تعداد واژگان مشترک (CWS)، وجود روابط معنایی (SRS) و تعداد کلمات هم‌رخداد(NCL) بین دوجمله صورت می‌گیرد که در ادامه شرح داده می‌شوند. لازم به ذکر است که برای اعمال این معیارها بایستی جملات پیش‌پردازش شوند؛ به‌طوری‌که جملات تنها شامل کلماتی با نقش اسم و صفت خواهد بود.

اگر S_i جملهٔ غیر مشترک خلاصهٔ سامانه‌ای و S_j جمله‌ای از مجموعه‌های خلاصهٔ انسانی و $Sim(S_i, S_j)$ نشان‌دهندهٔ میزان تشابه آن دوجمله و J شمارهٔ مجموعهٔ شامل جملهٔ S_i باشد و N تعداد کل خلاصه‌های انسانی را نشان دهد، آنگاه امتیاز هر جملهٔ غیر مشترک خلاصهٔ سامانه‌ای به‌صورت زیر محاسبه می‌شود:



(نمودار ۳-۱): معماری کلی روش پیشنهادی

رابطهٔ (۲)

$$sim(S_i,S_j)=\frac{LS+CWS+SRS+NCL}{4}$$

رابطهٔ (۳)

$$score(S_i)=\operatorname{argmax}_{1\le j\le N}\left(sim(S_i,S_j)\right)\times\frac{J}{N}$$

۳-۲-۱- معیار ارزیابی تعداد واژگان مشترک

اگر S_i جملهٔ غیر مشترک خلاصهٔ سامانه‌ای و S_j جمله‌ای از مجموعه‌های خلاصهٔ انسانی باشد، این ویژگی هم‌پوشانی لغوی بین S_i و S_j را مدنظر قرار می‌دهد. اگر S_i کلمات مشترک بیشتری با S_j داشته باشد، فرض می‌شود که S_i با S_j مرتبط است. اگر X مجموعهٔ کلمات جملهٔ S_i و Y مجموعه کلمات جملهٔ S_j باشد، امتیاز این ویژگی برای جملهٔ S_i به‌صورت زیر حساب می‌شود: (منظور از علامت |.| همان تعداد اعضای مجموعه یا همان کاردینالیتی است.)

رابطهٔ (۴)

$$CWS^{1}(S_i)=\frac{|X\cap Y|}{|X\cup Y|}$$

۳-۲-۲- امتیاز تعداد دوکلمه‌ای مشترک

به‌منظور محاسبهٔ این ویژگی در مرحلهٔ نخست تمامی دو گرم‌های[2] موجود در بین تمامی جملات متن نخستین را یافته و فراوانی آن‌ها محاسبه می‌شود. اگر فراوانی دو گرمی بزرگ‌تر از یک باشد، آنگاه این دو گرم به‌عنوان یک هم‌رخداد در نظر گرفته می‌شود.

در مرحلهٔ دوم، کلمات جملهٔ S_i (جمله‌های قسمت غیر مشترک خلاصهٔ سامانه‌ای) با کلمات جملهٔ S_j (جمله‌های مجموعه‌های خلاصهٔ انسانی) مقایسه می‌شود. این روند مقایسه بررسی می‌کند که آیا کلمه‌ای از جملهٔ S_i با یک کلمه از جملهٔ S_j دارای رابطهٔ هم‌رخدادی است یا نه؟ درصورتی‌که دارای این رابطه باشد، یک پیوند هم‌رخدادی بین S_i و S_j وجود دارد.

بنابراین امتیاز این ویژگی برای جملهٔ S_i به‌صورت زیر محاسبه می‌شود:

رابطهٔ (۵)

$$NCL^{3}(s)=\frac{n}{a*b}$$

در فرمول بالا n تعداد پیوندهای هم‌رخداد جمله‌های S_i و S_j است. a طول جملهٔ S_i و b طول جملهٔ S_j است که منظور از طول جمله، تعداد یونیگرم‌های موجود در جمله یا همان کلمات محتوایی است.

۳-۲-۳- امتیاز رابطهٔ معنایی با استفاده از شبکهٔ واژگان

معیار ارتباط بین دو کلمه از دو جملهٔ متفاوت S_i و S_j به‌صورت زیر تعریف می‌شود:

- Sense[4] کلمهٔ جملهٔ S_i عضو Synset[5] ای باشد که Sense کلمهٔ جملهٔ S_j عضو آن است.

---

[1] Common Word Score
[2] Bigram
[3] Number of Co-ocurrence Link
[4] مفهوم
[5] هم‌معنای

- Synset کلمهٔ جملهٔ S_i یک رابطهٔ شمول با Synset کلمهٔ جملهٔ S_j داشته باشد.
- Synset کلمهٔ جملهٔ S_i یک رابطهٔ زیرشمول با Synset کلمهٔ جملهٔ S_j داشته باشد.
- Synset کلمهٔ جملهٔ S_i یک رابطهٔ Related-to با Synset کلمهٔ جملهٔ S_j داشته باشد.
- Synset کلمهٔ جملهٔ S_i یک رابطهٔ هم‌رخداد با Synset کلمهٔ جملهٔ S_j داشته باشد.

در شبکهٔ واژگان هر کلمه می‌تواند چندین Sense داشته باشد یعنی می‌تواند دارای چندین معنی باشد به‌عنوان مثال کلمهٔ شیر می‌تواند بیانگر معانی مختلف شیرنوشیدنی، شیرحیوان و شیرآب باشد. منظور از Synset مجموعه‌ای از کلمات مترادف است. به‌عناون مثال معلم، دبیر، آموزگار و مدرس را می‌توان یک Synset در نظر گرفت.

برای اعمال این معیار، جملهٔ S_i موجود در قسمت غیر مشترک و جملهٔ S_j مجموعه‌ها که در مرحلهٔ پیش‌پردازش با ابزار جداکنندهٔ کلمات به کلماتش شکسته شده‌اند، با مقایسهٔ معنایی هر کلمه از S_i با کلمات S_j، تعداد واژه‌هایی را که دارای روابط معنایی می‌باشند به‌دست می‌آوریم. اگر X تعداد کلمات جملهٔ S_i و Y تعداد کلمات جملهٔ S_j و R تعداد روابط معنایی بین دو جملهٔ S_i و S_j باشد امتیاز رابطهٔ معنایی بین دو جملهٔ مذکور برابر است با:

رابطهٔ (۶)
$$SRS^{1}=\frac{R}{X*Y}$$

۳-۲-۴- معیار ارزیابی طولانی‌ترین زیررشتهٔ مشترک

در این معیار ارزیابی، از الگوریتم‌های محاسبهٔ طولانی‌ترین زیررشتهٔ مشترک بین دو رشته[2] استفاده می‌شود. اگر فرض کنیم X طول جملهٔ S_i و Y طول جملهٔ S_j باشد و طولانی‌ترین زیررشتهٔ مشترک بین دو جملهٔ S_i و S_j را $Lcs(S_i,S_j)$ بنامیم، امتیاز این معیار به‌صورت زیر محاسبه خواهد شد:

رابطهٔ (۷)
$$LS=\frac{Lcs(S_i,S_j)}{\min(X,Y)}$$

۳-۳- امتیازدهی خلاصهٔ سامانه‌ای

امتیاز نهایی خلاصهٔ سامانه‌ای از میانگین امتیازات جملات قسمت مشترک و غیر مشترک خلاصهٔ سامانه‌ای به‌دست می‌آید.

اگر S_i جملهٔ خلاصهٔ سامانه‌ای و M مجموعهٔ جملات خلاصهٔ سامانه‌ای باشد، امتیاز نهایی خلاصهٔ سامانه‌ای به‌صورت زیر محاسبه می‌شود:

رابطهٔ (۸)
$$\text{Final Score}=\frac{\sum_{S_i\in\text{Machine summary}} score(S_i)}{|M|}$$

۴- ارزیابی

برای ارزیابی دقیق روش پیشنهادی، احتیاج به یک مجموعه دادهٔ مناسب و استاندارد است؛ لذا پیکره‌ای متشکل از ۳۰ سند و هر سند توسط پنج فرد خبره که متخصص در زمینهٔ ادبیات فارسی بودند، خلاصه شد و مورد استفاده قرار گرفت.

تاکنون برای ارزیابی عملکرد یک سامانهٔ خلاصه‌ساز معیارهای دقت/فراخوانی و یا ترکیبی از آنها (F-measure) را برای آن سامانه محاسبه می‌کردند؛ فراخوانی(R)[3]، میزان انطباق جمله‌های خلاصهٔ مرجع (ایده آل) در خلاصهٔ سامانه‌ای را محاسبه می‌کند. دقت(P)[4]، بر عکس فراخوانی بوده و میزان انطباق جملات خلاصه‌های سامانه‌ای در خلاصهٔ مرجع را محاسبه می‌کند. در رابطه‌های زیر S برابر مجموعه جملات خلاصهٔ ماشینی و T برابر مجموعه جملات خلاصهٔ انسانی است.

رابطهٔ (۹)
$$P=\frac{|S\cap T|}{|S|}$$

رابطهٔ (۱۰)
$$R=\frac{|S\cap T|}{|T|}$$

رابطهٔ (۱۱)
$$\text{F-measure}=2\times\frac{P\times R}{P+R}$$

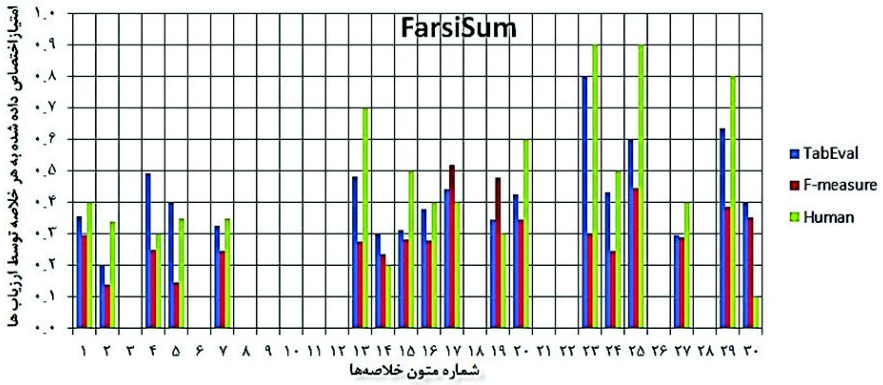
جهت ارزیابی میزان کارایی روش پیشنهادی، سامانه‌های خلاصه‌ساز FarsiSum، Ijaz و TabSum را توسط TabEval و F-measure و توسط چند خبره که متفاوت با افراد تولیدکنندهٔ خلاصهٔ انسانی هستند، مقایسه و ارزیابی می‌کنیم. به‌طوری‌که نرخ فشرده‌سازی برای هر کدام از سیستم‌های خلاصه‌ساز ۳۰٪ در نظر گرفته شده است. درواقع در این فرآیند، متون پیکره‌ای که درقبل تهیه کرده‌ایم، با سامانه‌های خلاصه‌ساز مذکور خلاصه می‌شوند؛ سپس برای کلیهٔ خلاصه‌های سامانه‌ای انتخاب‌شده از پیکره جهت ارزیابی و خلاصه‌های انسانی مرتبط با هر کدام، میزان دقت و فراخوانی را محاسبه می‌کنیم. ترکیبی از این دو معیار (F-measure)، امتیاز نهایی سامانه‌های خلاصه‌ساز را نشان

[1] Semantic Relation Score
[2] LCS (Longest Common Subsequence)
[3] Recall
[4] Precision

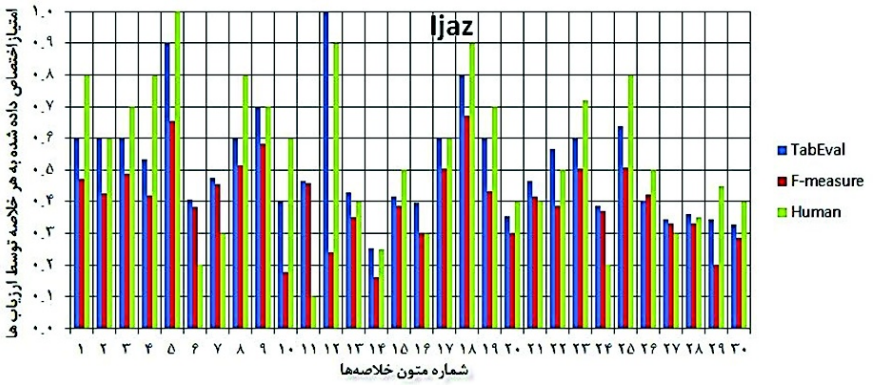
می‌دهد. F-measure مقداری مابین ۰ و ۱ است؛ به‌طوری‌که هرچقدر به یک نزدیک باشد، بهتر است. لازم به ذکر است به‌دلیل اینکه خلاصه‌ساز FarsiSum نتوانست خلاصه‌ای برای دوازده سند (سندهای ۸ الی ۱۲ و ۲۱ الی ۲۲ و ۲۶ و ۲۸) تولید کند، یعنی تعداد جملات تولیدی برای خلاصه صفر است؛ لذا ارزیابی یک خلاصهٔ خالی بی‌معنی است. بنابراین در نمودار امتیاز صفر برای آنها مدنظر قرار داده شده است.

معیار F-measure، سامانه‌های خلاصه‌ساز را فقط در سطح ظاهر ارزیابی می‌کند. درصورتی‌که TabEval علاوه‌بر سطح ظاهری، سطح معنایی (در حدی که تعریف شده است) را نیز در نظر می‌گیرد؛ چون طبق نمودارهای ۴- ۱و ۴-۵ و ۴-۶ ابزار TabEval نسبت به F-measure به خلاصهٔ انسانی نزدیک است. حال برای بررسی عملکرد انسان‌گونهٔ روش پیشنهادی به این ترتیب عمل می‌کنیم:

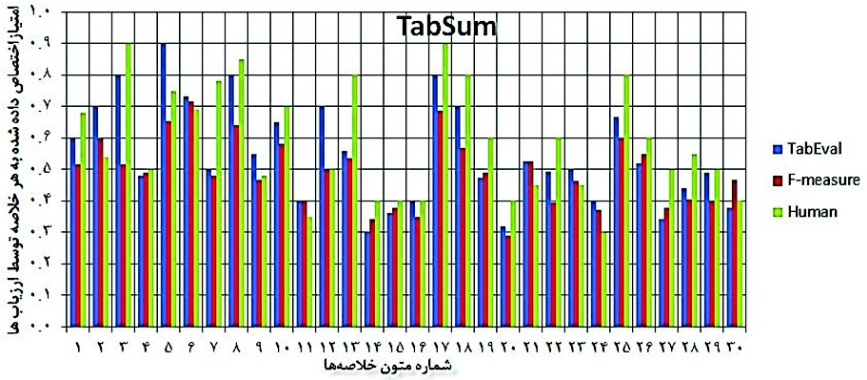
کیفیت خلاصه‌های تولیدی هر کدام از سامانه‌های خلاصه‌ساز بالا با درنظرگفتن تمامی معیارهای پیشنهادی، یک‌بار با سامانهٔ پیشنهادی (TabEval) و بار دیگر توسط افراد خبره مورد ارزیابی قرار گرفتند. ابزار پیشنهادی و همچنین افراد خبره بعد از بررسی سامانه‌های خلاصه‌ساز، مشخص کردند که TabSum نسبت به بقیهٔ سامانه‌های خلاصه‌ساز دارای عملکرد بهتری بود. خلاصه‌سازهای FarsiSum، Ijaz و TabSum توسط F-measure، TabEval و انسان ارزیابی شدند. نتایج حاصل از ارزیابی در نمودارهای (۴-۱)، (۴-۲) و (۴-۳) نشان داده شده است. برای نشان‌دادن عملکرد بهتر TabEval، میانگین اختلاف امتیاز اختصاص داده شده برای تک‌تک خلاصه‌سازها توسط F-measure و توسط TabEval از امتیازهای اختصاص داده شده به تک‌تک خلاصه‌سازها توسط انسان (Human) محاسبه شد. جدول (۴-۱) نتایج به‌دست‌آمده نشان داد که اختلاف امتیاز بین TabEval و انسان از اختلاف امتیاز بین F-measure و انسان کم‌تر است. که این حاکی از آن است که روش پیشنهادی (TabEval) بسیار نزدیک به ارزیابی افراد خبره بوده و دارای عملکردی انسان‌گونه و هوشمند است.



(نمودار ۴-۱): عملکرد روش پیشنهادی در ارزیابی FarsiSum



(نمودار ۴-۲): عملکرد روش پیشنهادی در ارزیابی Ijaz



(نمودار ۴-۳): عملکرد روش پیشنهادی در ارزیابی TabSum

(جدول ۴-۱): اختلاف کمی F-measure و TabEval از ارزیابی انسانی در خلاصه‌سازها

	میانگین اختلاف امتیاز F-measure از امتیاز ارزیاب انسانی	میانگین اختلاف امتیاز TabEval از امتیاز ارزیاب انسانی
خلاصه ساز FarsiSum	۰/۲۸۵۱۶۷۶۲۹	۰/۰۴۵۱۲۷۳۰۱
خلاصه ساز Ijaz	۰/۱۳۴۴۷۳۶۲۳	۰/۰۱۹۹۹۲۲۵۷
خلاصه ساز TabSum	۰/۰۹۳۹۹۹۸۸۸	۰/۰۳۶۱۹۳۴۴۱
میانگین اختلاف کل از ارزیاب انسانی در همه خلاصه سازها	۰/۱۷۱۲۱۳۷۱۳	۰/۰۳۳۷۷۱

۵- نتیجه‌گیری و پیشنهادها

در وب امروزی و با توجه به حجم گستردهٔ مطالب موجود و کمبود وقت و از طرفی ناکارآمدی سامانه‌های خلاصه‌سازی موجود، وجود یک سامانهٔ قدرتمند برای خلاصه‌سازی حجم انبوه کتب و مقالات و اخبار به‌شدت احساس می‌شود. سامانهٔ پیشنهادی در این مقاله میزان قدرتمندی این سامانه‌های خلاصه‌ساز را ارزیابی می‌کند. البته مشکل خلاصه‌سازی در زبان فارسی به مراتب بیشتر از زبان‌های دیگر است. سامانهٔ ارائه‌شده به‌دلیل مدنظر قراردادن معنا قادر به ارزیابی مترجم‌های ماشینی و سامانه‌های خلاصه‌ساز استخراجی و چکیده‌ای است. درواقع سامانه‌ای است که میزان تشابه دو متن را می‌تواند استخراج کند.

اگر دو سامانهٔ خلاصه‌ساز از نظر کیفیت خلاصهٔ تولیدیشان امتیاز یکسانی را کسب کرده باشند، سامانهٔ ارزیاب ارائه‌شده را می‌توان به‌گونه‌ای بهبود داد که بتواند سامانه‌های خلاصه‌ساز را از نظر اینکه تولیدی کدامیک قابلیت درک راحت‌تری برای انسان دارد، انتخاب کند. این کار با غنی‌سازی شبکهٔ واژه‌ها می‌تواند صورت بگیرد.

قدردانی و تشکر

لازم به ذکر است که هزینه‌های اجرای طرح (طباطبائی، ۱۳۹۲) توسط باشگاه پژوهش‌گران جوان تأمین و تخصیص یافته است. در این قسمت لازم می‌دانم از تمامی مسؤولان محترم باشگاه پژوهش‌گران جوان به‌دلیل حمایت‌های بی‌دریغشان در اجرای این طرح و تولید این سامانهٔ هوشمند و معناگرا کمال تشکر و قدردانی را به‌جا آورم.

۶- مراجع

پور معصومی آ.، کاهانی م.، طوسی س.ا.، استیری ا.، "ایجاز: یک سامانهٔ عملیاتی برای خلاصه‌سازی تک‌سندی متون خبری فارسی،" مجله پردازش علائم و داده‌ها، دوره ۱۱، شماره ۱، ص. ۳۳-۴۸، ۱۳۹۳.

طباطبائی ر.، معصومی س.، اولین سامانهٔ هوشمند معناگرا برای ارزیابی سامانه‌های خلاصه‌سازی، شمارهٔ قرارداد پژوهشی ۹۲۰۹۴، باشگاه پژوهش‌گران جوان، دانشگاه آزاد واحد بناب، ۹۳-۱۳۹۲.

Dalianis H., "SweSum - A Text Summarizer for Swedish, Technical report," TRITANA-P0015, IPLab-174, NADA, KTH, 2000.

Hassel M., "Evaluation of Automatic Text Summarization - A practical implementation, Licent-iate Thesis," KTH NADA, Sweden, 2004.

Lin C. -Y., "Looking for a Few Good Metrics: Automatic Summarization Evaluation - How Many Samples Are Enough?," In Proceedings of the NTC-IR Workshop 4, Tokyo, Japan, June 2 - June 4, 2004.

Lin C. -Y., "ROUGE: a Package for Automatic Evaluation of Summaries," In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

Lin C. -Y., Hovy E., "Manual and Automatic Evaluation of Summaries," USC Information Scie-nces Institute 4676 Admiralty Way, Marina del Rey, CA 90292, In Proceedings of the Workshop on Automatic Summarization post conference workshop of ACL-02, Philadelphia, PA, U.S.A., July 11-12, DUC, 2002.

Mani I., Maybury M., "Advances in Automatic Text Summarization," The MIT Press, 1999.

Masoumi, S., Tabatabaei, R., Feizi-Derakhshi, M.-R., "TabSum- A new Persian text summarizer," Journal of Mathematics and Computer Science, Volume 11, Issue 4, Pages 330-342, 2014.

Mazdak N., Hassel M., "FarsiSum - a persian text summarizer," Master thesis, Department of linguistics, Stockholm University, 2004.

Radev D., Allison T., Blair-Goldensohn S., Blitzer J., Celebi A., Drabek E., Lam W., Liu D., Otterbacher J., Qi H., Saggion H., Teufel S., Topper, Winkel A., and Zhang Z., "MEAD - A platform for multidocument multilingual text summarization," In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, May 2004.

Shamsfard M., Hesabi A., Fadaie H., Famian A., Bagherbeigi S., Mansoory N., Fekri E., Monshizadeh M., Assi S.M., Semi Automatic Development of FarsNet; The Persian WordNet, Accepted at Global WordNet Conference (GWA2010), India, 2010.

راضیه طباطبائی مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر (گرایش نرم‌افزار) در سال‌های ۱۳۹۰ و ۱۳۹۲ از دانشگاه تبریز دریافت کرده است. زمینهٔ پژوهشی مورد علاقهٔ ایشان شامل داده‌کاوی، پردازش متن، موازی‌سازی و پردازش‌های توزیع شده است.

نشانی رایانامهٔ ایشان عبارت است از:

tabatabaey_ac@yahoo.com

محمّدرضا فیضی درخشی دارای دکترا و کارشناسی ارشد مهندسی کامپیوتر در گرایش هوش مصنوعی از دانشگاه علم و صنعت ایران بوده و مدرک کارشناسی خود را نیز در رشتهٔ مهندسی کامپیوتر، گرایش نرم‌افزار از دانشگاه اصفهان اخذ کرده است. ایشان هم‌اکنون عضو هیأت علمی گروه مهندسی کامپیوتر دانشگاه تبریز است. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش زبان‌های طبیعی، پردازش معنایی وب و اسناد، الگوریتم‌های بهینه‌سازی و پایگاه داده است.

نشانی رایانامهٔ ایشان عبارت است از:

mfeizi@tabrizu.ac.ir

سعید معصومی مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر (گرایش نرم‌افزار) در سال‌های ۱۳۸۹ و ۱۳۹۲ از دانشگاه تبریز دریافت کرده است. وی هم اکنون دانشجوی دکترای مهندسی کامپیوتر (گرایش نرم‌افزار) در دانشگاه صنعتی مالک اشتر تهران است. زمینهٔ پژوهشی مورد علاقهٔ ایشان شامل داده‌کاوی، پردازش متن، موازی‌سازی، پردازش‌های توزیع شده و بازپیکربندی نرم‌افزار است.

نشانی رایانامهٔ ایشان عبارت است از:

masoumi_ac@yahoo.com