چکیده سازی چند نوشتاری زبان فارسی

امیر شهاب شهابی[1]
دانشگاه آزاد اسلامی
واحد علوم و تحقیقات
دکتر محمدرضا کنگاوری
دانشگاه علم و صنعت ایران

امروزه، به دلیل افزایش حجم اطلاعات درباره موضوعات مختلف، سامانه‌های استخراج اطلاعات از اهمیت خاصی برخوردارند، اما از آن مهمتر سامانه‌ای است که بتواند چکیده یا خلاصه‌ای از مجموعه اطلاعات بازیابی شده را به کاربر ارائه دهد. این مقاله، رهیافتی در زمینه تولید خلاصه از چندین نوشتار ارائه می‌کند، به طوری که بتوان با استفاده از اطلاعات چند مقاله یا متن و استخراج نکات مهم آن و برقراری ارتباط بین آنها، به یک چکیده واحد از میان آنها رسید و آن را در اختیار استفاده کننده قرار داد. یک سامانه خلاصه ساز چند نوشتاری، متفاوت از خلاصه ساز تک نوشتاری است و این تفاوت به عواملی از قبیل فشردگی، سرعت، عدم تکرار، خوانایی و مرتبط بودن جملات خلاصه تولیدی با یکدیگر مربوط است. هدف این مقاله ارائه الگو برای ایجاد چنین سامانه‌ای می‌باشد.

مقدمه

گسترش سریع اطلاعات تدارک سازوکارهایی پیشرفته برای یافتن متون مورد نیاز از میان انبوه نوشتارها را بسیار مهم می‌کند. سامانه‌های بازیابی اطلاعات (information retrieval systems) موجود، حاوی موتورهای جست‌وجوی (search engine) پیشرفته‌ای هستند که در یافتن و رتبه بندی نوشتارها (documents) بر اساس حداکثر ارتباط با درخواست کاربر نقش بسزائی دارند. اما نتیجه کار هنوز بدین صورت است که کاربر باید اطلاعات یافته شده را تماماً مطالعه نماید تا مطلب مرتبط و مورد نظر خود را بیابد. این بدان معنی است که سامانه‌های چکیده ساز (summarizing systems) و بازیابی اطلاعات هنوز با هم جمع نگردیده‌اند. وضعیتی را فرض کنید که در آن کاربر از یک سیستم بازیابی اطلاعات در شبکه جهان گستر (world wide web) جست‌وجوی موضوعی را درخواست می‌کند (مثلاً یک عنوان خبری)، در این صورت سامانه صدها متن نزدیک به آن موضوع را همراه با درجه مرتبط بودن آن به کاربر ارائه می‌دهد که احتمالاً بسیاری از آنها اطلاعات مشابه را تکرار کرده‌اند، در حالی که ممکن است در بخش‌های اصلی متفاوت باشند. حال با این شرایط، سامانه‌ای که بتواند از مجموعه اطلاعات ارائه شده خلاصه‌ای تولید نماید که زیاد طولانی نباشد، از ارتباط خوبی برخوردار باشد، اطلاعات تکراری در آن نباشد و بالاخره امکان دسترسی به اصل مقاله را از روی هر جمله چکیده بدهد، بسیاری از مشکلات استفاده کنندگان را حل خواهد کرد. این فعالیتی است که سامانه‌های خلاصه ساز چند نوشتاری باید بتوانند انجام دهند (گلدشتاین، میتال، کاربونل و کالان، ۲۰۰۰؛ اشتاین، استرز الکوفسکی،بودن و باگا، ۲۰۰۰).

چکیده سازی چند نوشتاری (multidocument summarization) با تک نوشتاری (single document summarization) تفاوت‌های زیادی نظیر فشردگی مطلب، سرعت، عدم تکرار (antiredundancy) و انتخاب صحیح عبارات و بالاخره

چسبندگی و انسجام (cohesion) چکیده نهایی دارد. چکیده‌ها بر دو نوع هستند؛ عمومی (general) و وابسته به درخواست (query based) (گلدشتاین، کانترووویتز، میتال و کاربونل، ۲۰۰۱). که در آن به طور کلی از مقاله خلاصه تهیه می‌شود. اما در نوع دوم، بر اساس درخواست و سئوال کاربر و نکات مرتبط با موضوع درخواست، خلاصه مقاله تهیه می‌گردد. در هر دو نوع این چکیده، اختلاف بین تک نوشتاری و چند نوشتاری وجود دارد که اگر بخواهیم دقیقتر این تفاوت‌ها را مورد بررسی قرار دهیم بهتر است به پنج نکته زیر توجه کنیم (گلدشتاین و همکاران، ۲۰۰۰):

الف) روش‌هایی که باید از تکرار اطلاعات در چکیده‌های چند نوشتاری جلوگیری نمایند.

ب) مجموعه مقالات ممکن است بعد زمانی (temporal dimension) داشته باشند، بنابراین در چکیده‌های چند نوشتاری اطلاعات جدیدتر باید اطلاعات قدیمیتر را به روز کنند.

ج) طول چکیده و فشردگی کار در چکیده چند نوشتاری مهم است.

د) مسئله حل ارجاعات ضمایر (reference resolution) به موجودیت‌های تعریف شده در چند نوشتار پیچیده‌تر از یک نوشتار می‌باشد.

ه) در چکیده‌های چند نوشتاری این امکان باید وجود داشته باشد که از روی هر جمله بتوان به مقاله یا مقالات مرجع آن دست یافت.

پس از این مقدمه، در بخش‌های بعدی، ابتدا به تعریف چکیده می‌پردازیم و سپس رهیافت تولید خلاصه چند نوشتاری را بررسی می‌نمائیم.

تعریف چکیده

برای اینکه بتوان به سامانه چکیده ساز دست یافت، اول باید بتوان چکیده و سپس چکیده چند نوشتاری را تعریف نمود که ذیلاً به آنها اشاره می‌شود:

«یک متن خلاصه، اشتقاقی از متن و مقاله اصلی است که به وسیله انتخاب و / یا تعمیم نکات مهم آن فشرده شده است» (بوبن‌هوفر، ۲۰۰۲).

اما تعریف فوق در خصوص چکیده از یک نوشتار است. در صورتی که چکیده از چند نوشتار باشد با توجه به توضیحاتی که در مقدمه داده شد اضافاتی دارد که ارائه می‌گردد:

خلاصه‌ای از چند نوشتار، اشتقاقی خوانا از منابع است که برای هر یک از نوشتارها جداگانه تهیه و به وسیله انتخاب و / یا تعمیم نکات مهم آن، فشرده شده است و سپس عبارات مشابه آن خوشه‌بندی (clustering) گردیده‌اند. پس از آن، با توجه به اهمیت هر یک برای قرار گرفتن در چکیده نهایی رتبه بندی (ordering) و در نهایت با توجه به عامل فشرده بودن انتخاب می‌شوند و در چکیده نهایی قرار می‌گیرند.

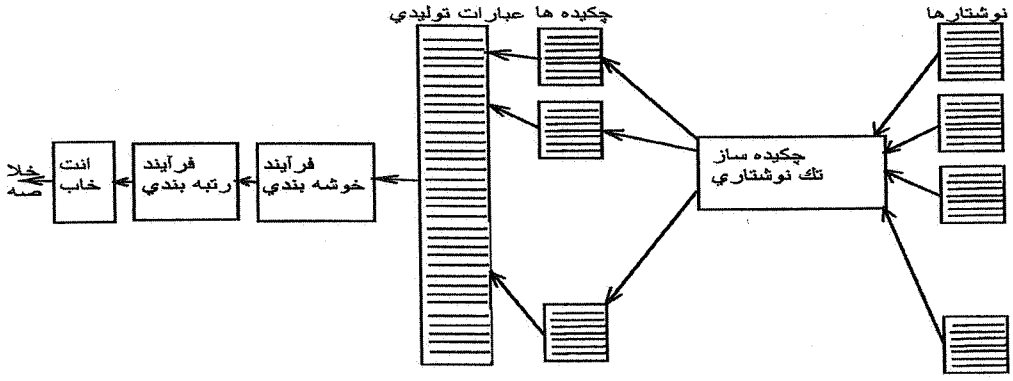
بنابراین از این تعریف نتیجه می‌گیریم که یک چکیده چند نوشتاری باید دارای مشخصات ذیل باشد:

الف) خوشه بندی؛ ب) پوشش؛ ج) عدم تکرار؛ د) مرتبط بودن جملات در خلاصه؛ ه) کیفیت؛ و) قابلیت تشخیص ناسازگاری‌ها در نوشتارها؛ ز) بهنگام کردن خلاصه برای متونی که بعد زمانی دارند و ح) نرم افزار محاوره‌ای (interactive software) مناسب و کارآمد که بتوان از هر جمله به اصل مقاله یا مقالات رسید و بر عکس.

حال با توجه به توضیحات فوق، به سراغ قالب کلی کار و اینکه یک سامانه چکیده ساز چند نوشتاری از چه بخش‌هایی باید تشکیل شود و در هر زمینه چه کارهائی صورت گرفته است، می‌رویم.

رهیافت تولید خلاصه چند نوشتاری از دید کلی

ساختار و سازمان کلی یک سامانه تولید چکیده چند نوشتاری (شکل ۱)، سامانه‌ای را پیشنهاد می‌کند که در آن ابتدا نوشتارها و مستندات هر یک، به طور جداگانه به یک زیرسامانه خلاصه ساز داده می‌شود و پس از طی فرآیندهای لازم، چکیدهٔ هر یک جدا تولید و تعدادی عبارت یا جمله ایجاد می‌گردد. سپس عمل خوشه‌بندی صورت می‌گیرد و خوشه‌هایی تشکیل می‌شود که در هر یک، عبارات و جملات مشابه وجود دارد. سپس به کمک معیارهای خاص عمل رتبه‌بندی آنها انجام می‌پذیرد و خوشه‌ها پشت سر هم با یک شماره ترتیبی قرار می‌گیرند تا چسبندگی و خوانائی متن حفظ گردد و نیز اطلاعات تکراری نداشته باشد و نهایتاً خلاصه چند



شکل ۱: سازمان سامانه چکیده ساز چند نوشتاری پیشنهادی

نوشتاری به دست می‌آید (مارکو و گربر، ۲۰۰۱). حال به شرح هر قسمت و فعالیت‌هایی که باید در هر زمینه صورت پذیرد، می‌پردازیم.

خلاصه ساز تک نوشتاری

این بخش از سیستم ابتدا یک واکافت نحوی (syntax parsing) و سپس یک واکافت کلامی (discourse parsing) (مارکو، ۲۰۰۱) انجام می‌دهد که در این راستا از گرامر زبان فارسی برای واکاوی استفاده می‌نماید. کاری که واکاو نحوی انجام می‌دهد، این است که جملات زبان را تجزیه و پس از آن یک مفسر (interpreter) با استفاده از درخت تجزیه (parse tree) تولید شده، منطق رتبه اول (first order logic)، گزاره‌هایی (predicates) تولید می‌نماید که مفهوم متن را ایجاد می‌کند (شهابی، ۱۳۷۶). از این کار برای تولید ساختار کلام (discourse structure) استفاده خواهد شد. سپس از روی درخت کلام (discourse tree) تولید شده که عبارات داخل کلام، برگ‌های آن می‌باشند به هر یک امتیاز می‌دهد، به طوری که عبارتی که به ریشه نزدیکتر باشد به دلیل تازگی ارزش بیشتری دارد و از اهمیت اطلاعاتی بیشتری برای قرار گرفتن در چکیده برخوردار خواهد بود (جورافسکی و مارتین، ۲۰۰۰).

اما گرامر زبان فارسی بررسی شده، از روی دستور زبان فارسی (ابومحبوب، ۱۳۷۵ و باطنی، ۱۳۸۰) و گرامر دو سطحی (two level grammer) (کرولی، ۱۹۹۱) ایجاد می‌شود که مورد استفاده واکاو نحوی قرار می‌گیرد. در این گرامر از فرامتغیرهائی (meta variables) نظیر شخص، شمار و زمان متناسب با آن فراقانون‌ها (meta rules) استفاده می‌شود. در هنگام عمل واکاو نحوی، با استفاده از SDT (Syntax Directed Translation) فعالیت‌هایی را در گرامر قرار می‌دهیم و وقتی به نقطه مورد نظر در درخت تجزیه رسیدیم عمل ذخیره و یا ساخت گزاره با منطق رتبه اول را انجام می‌دهیم آن را به پایگاه دانش (knowledge base) اضافه می‌کنیم (شهابی ۱۳۷۶).

اما در خصوص حل مسئله ارجاعات ضمایر، باید بتوانیم کلام فارسی را تحلیل و ساختار آن را ایجاد کنیم که قبل از ارائه راه حل برای آن باید قدری راجع به انواع ضمایر در زبان فارسی بحث کنیم.

انواع ارجاعات در کلام فارسی

انواع عبارات ارجاع در زبان فارسی به پنج نوع تقسیم می‌شوند: ۱) عبارات اسمی نامعین؛ ۲) عبارات اسمی معین؛ ۳) ضمایر؛ ۴) ضمایر اشاره؛ ۵) ارجاع یکی از چند.

اما عبارات ارجاع، خود به موجودیت‌های ارجاع شونده اشاره می‌کنند که این موجودیت‌ها بر سه نوع‌اند که ذیلاً به آنها اشاره می‌شود:

الف) موجودیت‌های قابل استنتاج و حدس زدن؛ ب) موجودیت‌های مجموعه‌های منقطع؛ ج) موجودیت‌های عمومی

که این موارد مسئله حل ارجاعات را پیچیده می‌گرداند، اما خوشبختانه محدودیت‌های نحوی و معنایی زبان فارسی در پیاده‌سازی الگوریتم حل مسئله ارجاع ضمایر کمک زیادی به ما می‌کند. این محدودیت‌ها عبارت‌اند از:

الف) توافق شماری: یعنی عبارات ارجاعی و مرجع آنها باید از نظر تعداد با هم موافقت داشته باشند.

ب) توافق شخص و نقش: یعنی عبارات ارجاعی و مرجع آنها باید از نظر شخص و نیز نقش دستوری با هم توافق داشته باشند.

ج) توافق جانداری: یعنی عبارات ارجاعی و مرجع آنها باید یا هر دو جاندار یا هردو غیر جاندار باشند.

د) محدودیت‌های نحوی: همان‌طور که می‌دانید تنها ضمایر انعکاسی می‌توانند با فاعل جمله هم مرجع باشند و هیچ ضمیر دیگری نمی‌تواند.

ه) محدودیت‌های انتخابی: گاهی اوقات محل قرار گرفتن ضمایر یا عبارات ارجاعی در جمله نقشی به آنها می‌دهد که مرجع آنها را محدود می‌نماید. به عنوان مثال جمله زیر را در نظر بگیرید:

«علی بنز خود را در گاراژ پارک کرد. او ساعت‌ها آن را رانده بود.» که در اینجا مرجع ضمیر می‌تواند بنز و یا گاراژ باشد، را داشته باشد، چون هر دو در جمله قبلی آمده‌اند و هر دو غیر جاندارند، اما چون محل قرار گرفتن ضمیر آن در جمله مفعول فعل راندن می‌باشد و این مفعول نیز نمی‌تواند گاراژ باشد و باید وسیله‌ای مانند ماشین، کامیون و غیره باشد لذا مرجع آن باید بنز انتخاب شود. بنابراین ما نیاز به یک دائره المعارف کلمات و معنی آنها داریم که البته در زبان انگلیسی چندین نمونه از آنها وجود دارد، ولی در زبان فارسی این کمبود کاملاً احساس می‌شود.

برخی از نکات که در انتخاب مرجع ضمایر به الگوریتم حل مسئله ارجاعات کمک می‌کنند، به قرار زیرند:

الف) تازگی: یعنی مرجعی که در جمله اخیر آمده است، انتخاب بهتری برای یک ضمیر می‌باشد.

ب) نقش دستوری: یعنی مراجعی که نقش فاعلی دارد، برجسته‌تر از مراجعی است که نقش مفعولی دارند و آنها نیز از مراجعی که در نقش‌های دیگری هستند، برجسته‌ترند.

ج) اشاره مکرر: اگر به یک مرجع در یک کلام تعداد دفعات زیادی اشاره شده باشد، حتی اگر تازگی نداشته باشد، برجستگی بیشتری برای انتخاب خواهد داشت.

د) موازی بودن: برای مراجعی که به طور موازی به کار می‌رود، اولویت بالایی وجود دارد.

ه) معنای افعال: گاهی اوقات معنای افعال موجود در جملات باعث اهمیت بیشتر یک مرجع نسبت به دیگری می‌شود. به مثال زیر توجه کنید:

«علی به احمد تلفن کرد. او دفترچه را در اتومبیل جا گذاشته بود.»

«علی احمد را نکوهش کرد. او دفترچه را در اتومبیل جا گذاشته بود.»

دو عبارت فوق تنها در فعل جمله اولشان تفاوت دارند. این چیزی است که به آن سبب شدن ضمنی می‌گویند، به طوری که سبب ضمنی عمل نکوهش متوجه مفعول است، در حالی که سبب ضمنی عمل تلفن کردن متوجه فاعل می‌باشد، بنابراین ضمیر او در عبارت اول به علی و در مثال دوم به احمد رجوع می‌کند.

ارائه الگوریتم حل مسئله ارجاع

در این پژوهش، هر موجودیتی که در متن یا کلام (discourse) دیده می‌شود در مدل کلام (discourse model) درج می‌شود و در همین زمان میزان برجستگی (salience factor) آن نیز محاسبه و در مدل ثبت می‌شود.

این میزان از روی یک سری عوامل به نام عوامل برجستگی محاسبه می‌شود. اگر یک مرجع مربوط به کلاس‌های قبلی باشد از آن می‌گذریم و میزان برجستگی آن را در مدل کلام نصف می‌کنیم. در ذیل، الگوریتم آن ارائه می‌شود (جورافسکی، ۲۰۰۰):

جدول ۱- *عوامل برجستگی*

تازگی جمله ۱۰۰
تأکید بر فاعل ۸۰
تأکید بر وجود داشتن ۷۰
تأکید بر مفعول بی واسطه ۵۰
تأکید بر مفعول با واسطه ۴۰
تأکید بر غیر قید بودن ۵۰
تأکید بر داخل عبارت بودن ۸۰

الف) کلیه مراجع بالقوه را تا چهار جمله قبل جمع‌آوری می‌کنیم.
ب) کلیه مراجعی را که بالقوه وجود دارند و توافقی در شمار یا جاندار بودن با ضمیر ندارند، حذف می‌کنیم.
ج) کلیه مراجعی را که محدودیت‌های نحوی هم مرجعی را بین جملات رعایت نمی‌کنند، حذف می‌نمائیم.
د) میزان برجستگی را با اضافه کردن مقادیر مناسب از جدول شماره ۱ به مقدار موجود محاسبه می‌کنیم.
ه) مرجع با بیشترین مقدار برجستگی را انتخاب می‌کنیم.
در مرحله بعد، باید جملات و عبارات را خوشه بندی کنیم.

خوشه‌بندی چکیده‌های نوشتاری

در این مرحله با انجام عملیات تشخیص مشابهت عبارات و متون، خوشه‌هایی تشکیل می‌گردد که داخل آن حاوی عبارات مشابه است. این کار در واقع یک کار بدون سرپرست (unsupervised) می‌باشد، یعنی بدون داده‌های آموزشی صورت می‌پذیرد و کلاً به دو صورت سلسله مراتبی و مسطح یا غیر سلسله مراتبی می‌تواند انجام پذیرد (مانینگ و شوتز، ۲۰۰۰). در روش مسطح، تعداد خوشه‌ها ثابت است و نمی‌تواند در هم ادغام شود، اما در روش سلسله مراتبی خوشه‌ها می‌توانند در هم ادغام شوند و خوشه بزرگتری تولید نمایند، فقط درجه مشابهت خوشه کمتر می‌شود.

برای تشخیص مشابهت، از روش مشابهت کسینوسی (cosine similarity) استفاده می‌شود (رادف، بلر گولدنسون و ژانگ، ۲۰۰۱؛ تاکو، اگاتا و آریکی، ۲۰۰۲). در این روش ابتدا تعداد کلمات هر جمله یا متن شمارش می‌گردد و کلماتی که شمارشان از حد آستانه بیشتر باشد در برداری درج می‌گردند، سپس بردارهای هر دو متن یا دو جمله در هم ضرب داخلی و کسینوس زاویه بین آنها به روش زیر محاسبه می‌شود. عدد حاصل، میزان مشابهت آن دو جمله را نشان می‌دهد که اگر از آستانه‌ای بیشتر باشد در خوشه مناسب خود قرار می‌گیرد.

$$Sim(I, J) = Cos(I, J) = (I.J) / (//I//^2 * //j//^2)$$

در مرحله بعد، رتبه بندی خوشه‌ها و عبارات مورد بررسی قرار می‌گیرد.

رتبه‌بندی خوشه‌ها و جملات

در این مرحله، جملات مهمتر باید در بالا و جملات دارای اهمیت کمتر در پایین فهرست قرار گیرند که برای این کار روش‌های زیر مورد استفاده قرار می‌گیرند:

الف) رتبه‌بندی از روی اکثریت (majority ordering).

در این روش، جملات و مفاهیمی که بیشتر در متن استفاده شده باشند، رتبه بالاتری خواهند داشت (برازیلی، الهداد و مک‌کون، ۲۰۰۱).

ب) رتبه بندی زمانی (chronological ordering).

در این روش، جملات و عباراتی که بعد زمانی دارند، ابتدا تمبر زمانی (time stamp) می‌گیرند و جملاتی که جدیدتر هستند اولویت بالاتری می‌یابند (برازیلی و همکاران، ۲۰۰۱؛ فیلاتواو هووی، ۲۰۰۱).

ج) رتبه‌بندی بر اساس معیار حداکثر بودن ارتباط حاشیه‌ای (relevance maximal marginal).

در این روش، بر اساس مشابهت درخواست کاربر با جملات موجود در متن و عدم تکرار جملات انتخابی عباراتی انتخاب می‌شوند که به آن روش حداکثر بودن ارتباط حاشیه‌ای یا maximal marginal relevance می‌گویند (گلدشتاین و همکاران، ۲۰۰۰).

انتخاب

در این مرحله، باید از میان خوشه‌های موجود انتخاب صورت پذیرد که معیارهایی نظیر جملاتی که تعداد کلمات بیشتری دارند و یا جملاتی که تعداد حروف اضافه یا تعریف آنها کمتر است و یا ضمیر کمتری دارند، امتیاز بیشتری می‌گیرند و لذا انتخاب می‌گردند. در نهایت، این جملات انتخابی پشت سر هم قرار می‌گیرند و با توجه به عامل فشردگی تا تعداد خاصی متوقف می‌شوند و چکیده نهایی را تولید می‌نمایند. اما کلام نهایی باید چسبندگی و انسجام لازم را داشته باشد که برای این کار باید محدودیت‌هایی در کلام قابل اعمال باشد. این کار به کمک استنتاج (inference) صورت می‌گیرد که متداولترین نوع آن استقراء می‌باشد که در واقع یک استنتاج قطعی (deterministic inference)

است. اما از آنجاکه سامانه‌های پردازش زبان طبیعی نمی‌توانند استنتاج قطعی نمایند (چرا که در استقراء از کل به جزء می‌رسیم، در حالی‌که چنین سامانه‌هایی که کل در آن همان معنای کلام باشد را نداریم) لذا از استنتاج دیگری که یک استنتاج غیر قطعی (undeterministic inference) است، به نام قیاس (induction) استفاده می‌نمائیم؛ به این ترتیب که بین جملاتی که می‌خواهند در خلاصه قرار گیرند قیاس که قانون آن modes tolen می‌باشد، انجام می‌دهیم و اگر عمل resolution در مجموعه گزاره‌ها انجام گرفت، در این‌صورت آن جملات با یکدیگر مرتبط اند و می‌توانند در خلاصه قرار گیرند.

نتیجه‌گیری

در نهایت می‌توان گفت که برای ایجاد یک خلاصه از چندین نوشتار ناچاریم ابتدا کل نوشتارها را جداگانه خلاصه، سپس آنها را خوشه بندی و رتبه بندی کنیم و بعد از میان آنها با توجه به عامل فشردگی تعدادی را انتخاب کنیم و طوری در چکیده نهایی قرار دهیم که ارتباط کلام حفظ گردد.

هنوز سامانه‌ای که برای خلاصه سازی زبان فارسی به صورت چند نوشتاری باشد، به طور صنعتی، تهیه نشده و تنها نمونه‌ای (prototype) از آن جهت پشتیبانی از یک زیر مجموعه از زبان فارسی تهیه گردیده است که ۷۰ درصد آن تکمیل شده است و در حال تحقیق و مطالعه روی یک نمونه واقعی به کمک متخصصان ادب پارسی می‌باشیم. بدیهی است که نمونه تهیه شده هنگام قرار گرفتن در محدوده بزرگتری از زبان پارسی قطعاً دستخوش تغییراتی خواهد گردید و تکامل خواهد یافت. نمونه تهیه شده دارای خطای بیش از ۳۰ درصد است که در حال بازنگری الگوریتم‌های تشخیص مشابهت جملات آن، از جمله با استفاده از منطق فازی می‌باشیم که در آینده نزدیک به نتیجه خواهد رسید.

منابع

ابومحبوب، ا. (۱۳۷۵). *ساخت زبان فارسی*. تهران: نشر میترا.

باطنی، م. ر. (۱۳۸۰). *توصیف ساختمان دستوری زبان فارسی*. تهران: انتشارات امیرکبیر.

شهابی، ا. ش. (۱۳۷۶). درک متن فارسی و تبدیل آن به پایگاه داده‌ای رابطه‌ای، امکان سنجی و ارائه الگو برای یک زیر زبان از زبان فارسی. پروژه کارشناسی ارشد دانشگاه آزاد اسلامی واحد تهران جنوب، دانشکده تحصیلات تکمیلی.

Brazily, R., Elhedad, N., & McKeown, K. (2001). Sentence ordering in multi document summarization. *Proceedngs of the 1st Human Language Technology Conference. San Diego, California.*

Bubenhofer, N. (2002). Text summarization. *English Seminar, WS ,2001-2002.*

Filatova, E., & Hovy, E. (2001). Assigning time stamps to event clauses. *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing.*

Goldstein, J., Mittal, V., Carbonell, J., & Calan, J. (2000). Creating and evaluating multi document sentence extract summaries. *Proceedings of 2000 ACMCIKM International Conference on Information and Knowledge Management. McLean, VA, USA,* 165-172.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (2001). Summarization text documents: *Sentence selection and evaluation metrics.*

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing* (pp. 240-285). Prentice Hall.

Krulee, G. K. (1991). *Computer Processing of Natural Language* (pp. 110-156). Prentice Hall.

Manning, C. D., & Schutze, H. (2000). *Foundation of Statistical Natural Language Processing* (pp. 230-300). The MIT Press.

Marcu, D., & Gerber, L. (2001). An iquiry into the nature of multi document abstracts, extracts and their evaluation. *Proceedings of Automatic Summarization Workshop.*

Marcu, D. (2001).
DUC-2001. *Workshop on Text Summarization (DUC-2001). NewOrleans.*

Radev, D., Blair-Goldensohn, S., & Zhang, Z.h. (2001). Experiments in Single and Multi-Document Summarization Using MEAD. *Workshop on Text Summarization (DUC-2001). New Orleans.*

Stein, G. C., Strzalkowski, T., Bowden, G., & Bagga, A. (2000). Evaluating Summaries for Multiple Documents in an Intractive Environment. *Second International Conference on Language Resourses and Evaluation.*

Takao, S., Ogata, J., & Ariki, Y. (2002). Topic segmentation of new speech using work similarity.