گروه کامپیوتر

عنوان:

خلاصه سازی خودکار متون فارسی مبتنی بر هستی شناسی

پایان نامه برای دریافت درجه کارشناسی ارشد رشته مهندسی کامپیوتر-گرایش هوش مصنوعی

استاد راهنما:

دکتر محمدرضا فیضی درخشی

استاد مشاور:

مهندس سعدالله سبحانی

پژوهشگر:

مجید رمضانی

تابستان ۱۳۹۱


بسم الله الرحمن الرحیم

تشکر و قدر دانی

با تقدیر شایسته از استاد فرهیخته و فرزانه جناب آقای دکتر محمدرضا فیضی درخشی، نهایت سپاس و قدردانی خود را از زحمات ایشان اعلام می دارم که در همه حال همراه و یاورم بوده اند. همچنین از همراهی استاد عالی قدر جناب آقای مهندس سعدالله سبحانی در تکمیل پژوهش نهایت تشکر را دارم.

کلیه حقوق مادی مترتب بر نتایج مطالعات، ابتکارات و نوآوری های

ناشی از تحقیق موضـــوع این پایان نامـــــه

متعلق به موسسه آموزش عالی نبی اکرم (ص) تبریز است.

تقدیم به آستان حقیقت و آنان که وصالش را می جویند،

و به آنانی که در راه اعتلای دانش و مبارزه با جهل و نادانی همواره کوشیده اند،

به ویژه بزرگ آموزگارمان پیامبر اکرم (ص) و امامان معصوم (علیهم سلام)

پیش گفتار

امروزه حجم اطلاعات در دسترس به طور وسیعی در حال گسترش است، به طوری که ادعا می شود انفجاری در حجم اطلاعات تولید شده رخ داده است. تعداد مقالات منتشر شده در زمینه های مختلف روز به روز در حال افزایش بوده و محققین و پژوهش گران را با حجم وسیعی از اطلاعات مواجه کرده است. در این میان ضرورت وجود روش هایی که قادر به تولید خلاصه هایی از متون مختلف بوده و جایگزین مناسبی برای مطالعه کامل اسناد باشند، هرچه بیشتر مورد تاکید قرار می گیرد.

یک خلاصه عبارت است از یک نسخه مختصر و کوتاه شده از یک (یا چند) سند که حاوی مفاهیم و نکات عمده مطرح شده در آن (آنها) بوده و جایگزینی برای مطالعه کامل متون منبع محسوب می شود. وجود سیستمی که به طور خودکار قادر به تولید خلاصه هایی از اسناد باشد، بسیار مفید خواهد بود. چنین سیستمی با جلوگیری از مواجهه با حجم انبوهی از اسناد که اهمیت و میزان ارتباط آنها با موضوع مورد نظر معین نیست، منجر به کاهش هزینه زمانی محققین علوم مختلف شده و تنها با مطالعه خلاصه اسناد، امکان انتخاب مهمترین و مرتبط ترین موارد را برای مطالعه بیشتر فراهم می کند. البته اغلب اسناد و مقالات علمی حاوی بخش چکیده هستند، اما بایستی توجه داشت که این بخش ها الزاماً شامل مهمترین مفاهیم و ایده های موجود در متون نیستند، بلکه تنها یک مرور کلی از موضوعات مطرح شده را در بر دارند. به تعبیر دیگر مفاهیم موجود در خلاصه های متون بسیار مفید تر و کامل تر از مفاهیم موجود در چکیده ها هستند.

به طور کلی خلاصه ها را می توان به دو بخش عمده تقسیم کرد؛ خلاصه های استخراجی و خلاصه های تولیدی. در خلاصه های استخراجی فرآیند خلاصه سازی عبارت است از شناسایی و استخراج مهمترین بخش های موجود در متن یا متون منبع. در این نوع از خلاصه های عبارت ها و بخش های مختلف موجود در خلاصه عیناً در متون منبع آمده اند. در مقابل در خلاصه های تولیدی خلاصه اسناد پس از درک مفاهیم ارائه شده در متون منبع و بیان آنها، در شکل عبارات جدید تولید می شوند. به تعبیر دیگر در این نوع از خلاصه ها، خلاصه نهایی علیرغم اینکه حامل مفاهیم موجود در متن منبع است، شامل عبارت هایی خواهد بود که در متون منبع وجود ندارد.

بر اساس یک تقسیم بندی دیگر خلاصه ها با توجه به تعداد اسناد مورد خلاصه سازی، به دو نوع تک سنده و چند سنده تقسیم می شوند. در سیستم های خلاصه ساز تک سنده در هر بار خلاصه سازی تنها یک سند واحد مورد خلاصه سازی قرار می گیرد، در حالی که در سیستم های چند سنده، در هر بار چندین متن مورد خلاصه سازی قرار می گیرد.

خلاصه ها از لحاظ موضوعات تحت پوشش نیز به دو دسته تقسیم می شوند؛ خلاصه سازهای مبتنی بر موضوع یا پرس و جو محور، تنها خلاصه هایی از اسناد ورودی را تولید می کنند که مرتبط با موضوع یا پرس و جوی عنوان شده باشد. در مقابل خلاصه سازهای عام خلاصه جامعی از متون منبع تولید می کنند که حاوی همه موضوعات مطرح شده در آنهاست.

هدف این پژوهش تولید یک سیستم خلاصه ساز خودکار استخراجی، تک سنده و عام است که اساس آن را مفاهیم هستی شناسی تشکیل می دهند. هستی شناسی یک پایگاه دانش شامل همه موجودیت های عالم هستی و روابط میان آنهاست. در این مطالعه از هستی شناسی فارس نت به عنوان پایگاه دانش هستی شناسی مرجع استفاده خواهد شد که مهمترین جملات و عبارات بر اساس مفاهیم موجود در آن برای قرار گیری در خلاصه نهایی انتخاب خواهند شد. طرح کلی سیستم پیشنهادی از این قرار است که در گام اول با تطبیق سند ورودی با پایگاه دانش هستی شناسی، موجودیت های آن و همچنین روابط میان آنها استخراج می شوند. سپس از این اطلاعات برای ساخت یک گراف تحت عنوان *گراف موضوعی* که در بر دارنده شمای کلی سند ورودی است، استفاده می شود. در مراحل بعدی با اعمال پردازش هایی روی این گراف، مهمترین بخش متصل آن انتخاب شده و بر این اساس مهمترین جملات موجود در متن منبع شناسایی شده و برای حضور در خلاصه انتخاب می شوند.

در بخش اول از این پژوهش به مرور مفاهیم خلاصه سازی خودکار متن پرداخته و با تبیین اهمیت و ضرورت آن، مراحل سنتی این فرآیند مورد توجه قرار می گیرد. همچنین با بررسی هستی شناسی که یکی از مفاهیم مربوط به علم فلسفه است، کاربرد آن در دانش هوش مصنوعی مطالعه خواهد شد. در فصل دوم با تشریح عوامل تاثیرگذار در فرآیند خلاصه سازی، روش های مختلف انجام آن را مد نظر قرار داده و چند نمونه از سیستم های خلاصه ساز تولید شده همراه با ویژگی های آنها مورد بررسی قرار خواهند گرفت. در فصل سوم نیز با تشریح جزئیات معماری سیستم پیشنهادی، مراحل مختلف تولید خلاصه گام به گام مورد مطالعه قرار گرفته و عملکرد سیستم تبیین می شود. فصل چهارم با هدف معرفی روش های مختلف ارزیابی سیستم های خلاصه ساز خودکار، به ارزیابی سیستم پیشنهادی پرداخته و نتایج حاصل از آن مورد مطالعه و بررسی قرار می گیرند. در انتها در فصل پنجم با ارائه یک نتیجه گیری، کارهای ممکن در این زمینه به عنوان کارهای آینده مورد بحث قرار خواهد گرفت.

چکیده

با توجه به گسترش روزافزون اطلاعات در دسترس از طریق اینترنت، لزوم استفاده از روش های خلاصه سازی خودکار متن، بیش از پیش احساس می شود. روش هایی که با استخراج مهمترین مطالب موجود در اسناد مانع از مطالعه کامل حجم انبوه از آنها شوند. خلاصه سازی عبارت است از فشرده سازی متن (متون) منبع و تولید یک نسخه کوتاه تر از آن به نحوی که محتوای اطلاعاتی آن حفظ شود. اغلب سیستم های خلاصه ساز با استفاده از روش های سطحی و معیارهای آماری به استخراج مهمترین بخش های متن منبع پرداخته و خلاصه نهایی را شکل می دهند. هدف این پژوهش استفاده از یک روش مبتنی بر پایگاه دانش در فرآیند خلاصه سازی است. در این راستا از پایگاه دانش هستی شناسی فارس نت به منظور دستیابی به مفاهیم موجود در متون و تولید خلاصه آنها استفاده خواهد شد. هستی شناسی یکی از مباحث مربوط به علم فلسفه است که یک ساختار سلسله مراتبی از همه موجودیت های عالم هستی به همراه روابط حاکم بر آنها فراهم می کند. در این پژوهش ابتدا با نگاشت متن مورد خلاصه سازی با پایگاه دانش هستی شناسی، گرافی تحت عنوان گراف موضوعی شکل می گیرد که حامل شمای مفهومی متن منبع است. سپس با استفاده از معیارهای مختلف تعیین اهمیت گرافی، اهمیت نسبی هر یک از گره های گراف ارزیابی می شود. سرانجام از این مقادیر به منظور تعیین اهمیت جملات مختلف موجود در متن منبع و ساخت خلاصه نهایی استفاده خواهد شد. نتایج حاصل از ارزیابی خلاصه های تولید شده، حاکی از برتری روش پیشنهاد شده در این پژوهش نسبت به سیستم های خلاصه ساز موجود است.

کلمات کلیدی: خلاصه سازی خودکار متن، هستی شناسی، خلاصه سازی استخراجی

فهرست مطالب

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فهرست شکل ها

عنوان **صفحه**

فصل ۱

مقدمه

۱-۱ مقدمه

ایده ایجاد توانایی پردازش زبان بشری برای کامپیوتر، همزمان با پیدایش نخستین کامپیوتر ها در ذهن محققین شکل گرفت. فعالیت در این زمینه که مرز مشترک میان علوم زبان شناسی[۱] و هوش مصنوعی[۲] محسوب می شود، از سال ۱۹۵۰ با هدف ایجاد مدل های محاسباتی از زبان طبیعی برای تجزیه، تحلیل و تولید آن شدت گرفت و مبنی بر مورد استفاده، اسامی مختلفی به خود گرفت: پردازش زبان و گفتار[۳]، فناوری زبان بشری[۴]، پردازش زبان طبیعی[۵]، زبانشناسی محاسبه ای[۶]، تشخیص و سنتز گفتار[۷] . اولین کارها در زمینه پردازش زبان طبیعی روی ترجمه ماشین[۸] متمرکز بود. ترجمه خودکار از زبان روسی به انگلیسی در سال ۱۹۵۴ بصورت بسیار ابتدایی توسط شرکت IBM انجام گرفت.

علیرغم عمر بسیار طولانی بازیابی اطلاعات[۹] که مربوط به تعامل میان انسان و کامپیوتر[۱۰] بوده و در پی جستجوی اسناد و دستیابی به اطلاعات موجود در آن است، شاخه استخراج اطلاعات[۱۱] اخیرا مورد توجه مجامع علمی قرار گرفته است. البته دلایل دیگری نظیر رشد بی رویه میزان داده های متنی از طریق شبکه اینترنت و نیز برگزاری سمینارها و کنفرانس های متعدد در این زمینه، در این روند تاثیر گذار بوده اند. هدف این شاخه استخراج خودکار اطلاعات موجود در متون ساخت یافته و یا نیمه ساخت یافته است. می توان چنین اظهار داشت که این شاخه با استفاده از ابزار پردازش زبان طبیعی به دنبال پردازش متون مرتبط با زبان های بشری و دستیابی به اطلاعات موجود در آنهاست.

ایده خلاصه سازی خودکار متن نیز به اواخر دهه ۱۹۵۰ میلادی معطوف می شود. طی این سالها

---

[1].Linguistics
[2].Artificial Intelligence
[3].Speech and Language Processing
[4].Human-language Technology
[5].**N**atural **L**anguage **P**rocessing (NLP)
[6].Computational Linguistics
[7].Speech Recognition and Synthesis
[8].**M**achine **T**ranslation (MT)
[9].**I**nformation **R**etrieval (IR)
[10].**H**uman-**C**omputer **I**nteraction (HCI)
[11].**I**nformation **E**xtraction (IE)

تلاش های بسیاری برای رسیدن به این هدف انجام گرفت. در این راستا هم از اطلاعات سطح پایین (اطلاعات آماری) و هم از اطلاعات سطح بالای (مدل های دامنه ای) موجود در اسناد استفاده شد، اما حاصل یک نتیجه تقریبی از خلاصه یک متن بود که با نمونه های انسانی فاصله مشهودی داشت. البته امروزه نیز ساخت یک سیستم خلاصه ساز توانا که خلاصه های بسیار خوانا تولید کرده و نتایج بسیار قانع کننده ای داشته باشد، فرآیند بسیار پیچیده ای خواهد بود.

۱-۲ پردازش زبان طبیعی

پردازش زبان طبیعی (NLP) یک رویکرد ماشینی تحلیل متون است که با استفاده از مجموعه ای از تئوریها و فناوری ها به اهداف خود دست پیدا می کند. تعریف جامع و یکسانی از این مفهوم در دست نمی باشد، اما آنچه که به طور معمول میان تعاریف موجود مشترک است بدین قرار است: «پردازش زبان طبیعی عبارت است از یک مجموعه از تکنیک های محاسباتی برای تحلیل و نمایش متون طبیعی و نیز تجزیه و تحلیل آنها در یک یا چند سطح، به منظور دستیابی به پردازش زبان بشرگونه، در پی رسیدن به اهداف مختلف و انجام امور مربوط به زبان». این تعریف حاوی نکاتی است که نیاز به تفسیر و توضیح بیشتری دارد. بدین شرح که:

«متون طبیعی» منحصر به زبان خاصی نبوده و همه زبان های بشری را شامل می شود. حتی لزومی بر مکتوب بودن آن وجود ندارد و امکان اینکه ورودی سیستم های پردازش زبان طبیعی بصورت شفاهی باشد، وجود دارد. البته بایستی دقت کرد که متون مورد پردازش نباید متونی باشند که با هدف پردازش طراحی شده باشند، و بایستی از متون معمولی و طبیعی برای این منظور استفاده کرد.

عبارت «در یک یا چند سطح» مبین این مطلب است که پردازش زبان طبیعی در سطوح مختلفی امکان پذیر است. انسان نیز برای تولید و یا درک زبان، در چندین سطح به پردازش زبان طبیعی می پردازد. هر یک از این سطوح اقدام به انتقال بخش خاصی از معنا می نماید. سیستم های مختلف پردازش زبان طبیعی هر یک از بخش های مجزایی برای پردازش زبان در سطوح مختلف استفاده می کنند. تفاوت میان سیستم ها نیز ناشی از توانایی و قدرت هر یک از این بخش هاست.

عبارت «پردازش زبان طبیعی بشرگونه» مبین این امر است که، پردازش زبان طبیعی شاخه ای از دانش هوش مصنوعی است. تلاش برای رسیدن به عملکرد انسان گونه، خود مؤیّد این امر است.

بایستی توجه داشت که پردازش زبان طبیعی خود یک هدف و غایت مستقل نمی باشد، بلکه هدف اصلی استفاده از آن در تحقیقات و مباحث مربوط به هوش مصنوعی است. برای مثال از آن

می توان به عنوان ابزاری برای رسیدن به اهدافی نظیر بازیابی اطلاعات، ترجمه ماشین یا پاسخگویی سوالات[1] که هر یک جزئی از سیستم های پردازش زبان طبیعی می شوند استفاده کرد.

همان طور که پیشتر ذکر شد هدف پردازش زبان طبیعی، دستیابی به پردازش انسان گونه زبان طبیعی است. لغت «پردازش» به طور عمد برای این منظور انتخاب شده است و نباید با لغت «درک[2]» جایگزین شود. اگر چه این شاخه از هوش مصنوعی در اصل همان «درک زبان طبیعی[3]» (NLU) است، اما از آنجا که این مهم تا کنون حاصل نشده است استفاده از این واژه بیشتر مورد پسند خواهد بود. یک سیستم درک زبان طبیعی مناسب بایستی قادر باشد:

- ورودی متنی را تفسیر نماید.
- متن ورودی را به زبان دیگری ترجمه نماید.
- قادر به پاسخگویی به سوالاتی در مورد متن باشد.
- توانایی استفهام از متن را داشته باشد.

در حالی که پردازش زبان طبیعی تنها توانسته تا حدی به اهداف اول، دوم و سوم دست پیدا کند، اما همچنان در دستیابی به هدف آخر ناکام مانده است. البته به طور کلی اهداف عملی بسیاری را می توان با توجه به کاربرد، برای پردازش زبان طبیعی عنوان کرد.

با اینکه این شاخه از هوش مصنوعی تحت عنوان «پردازش زبان طبیعی» شناخته شده است، با این حال دو تاکید متفاوت در آن وجود دارد: پردازش زبان و تولید زبان[4]. در «پردازش زبان» با تحلیل زبان به دنبال دستیابی به یک نمایش پر معنی از آن هستیم، در حالی که در «تولید زبان»، زبان طبیعی از نمایش های مذکور تولید می شود. وظیفه «پردازش زبان» با وظیفه خواننده یا شنونده[5] مشابه است. وظیفه «تولید زبان» نیز با وظیفه نویسنده یا گوینده[6] یکسان است. با اینکه بسیاری از تئوری ها و تکنیک ها بین این دو بخش مشترک هستند، «تولید زبان طبیعی» از پیچیدگی های بیشتری برخوردار است.

تحقیقات در زمینه پردازش زبان طبیعی مربوط به دهه ۱۹۴۰ میلادی می شود. ترجمه ماشین

---

[1]. **Q**uestion **A**nswering (QA)
[2]. Understanding
[3]. **N**atural **L**anguage **U**nderstanding (NLU)
[4]. Language Generation
[5]. Reader or Listener
[6]. Writer or Speaker

اولین برنامه کامپیوتری مرتبط با پردازش زبان طبیعی بود. بوس و وی اِور[1] طی جنگ جهانی دوم اولـین پروژه ترجمه ماشین را برای رمز گشایی کدهای دشمن ارایه کردند که الهام بخش ایده های بسـیاری در ذهن محققین و پژوهشگران مختلف شد. پس از آن تحقیقات در بسیاری از موسسـات تحقیقـاتی آمریکـا آغاز شد.

۱-۳ خلاصه سازی خودکار متن

خلاصه سازی متن که یکی از شاخه های پردازش زبان طبیعی محسوب می شود عبارت اسـت از، فشرده سازی متن منبع به یک نسخه کوتاهتر به نحوی کـه محتـوای اطلاعـاتی آن حفـظ شـود[۱]. بـه عبارت دیگر نسخه کوتاه شده بایستی مهمترین مفاهیم متن منبع را در بر داشته باشد. با انجام این عمل توسط کامپیوتر، به عبارت دیگر انجام خودکار آن، به ایـن فرآینـد خلاصـه سـازی خودکـار مـتن اطـلاق می شود. البته قابل ذکر است که علیرغم اینکه فرآیند خلاصه سازی خودکار متن روی ورودی های متنی متمرکز است، ورودی فرآیند خلاصه سازی می تواند اطلاعات چنـد رسـانه ای، صـوت، ویـدئو، اطلاعـات آنلاین و یا ابر متن[2] باشد.

۱-۳-۱ اهمیت و لزوم

رشد و افزایش بی رویه اطلاعات در دسترس در اینترنت یا به تعبیر بهتر سرریزی اطلاعـات[3][۲] ، بیش از پیش بر لزوم وجود یک روش جایگزین برای نمایش و انتخاب محتواهای متنی و چنـد رسـانه ای تاکید می کند. روشی که بر اساس آن مهمترین بخش های متون بیان شده و کاربر بتواند با مطالعه آنهـا تصمیماتی در مورد مطالعه متون اتخاذ کند. به تعبیر دیگر در عصری کـه اطلاعـات در دسـترس خیلـی بیشتر از میزان اطلاعات مورد نیاز است، و همه پژوهشگران و محققین در عرصه هـای مختلـف بـا حجـم بسیار زیاد اطلاعات روبرو هستند، بهترین راهکار پیش رو چیست؟ آیا بایسـتی همـه اطلاعـات و اسـناد موجود در یک زمینه خاص مطالعه شده و مطالب مهم آنها استخراج شوند؟ آیا باید از چندین نفر کـه بـه آن مطلب آشنایی دارند، تقاضای همکاری کرد؟ در صورت کمبود منابع زمانی چه باید کرد؟ بعلاوه اینکه هیچ تضمینی از اینکه متن مورد مطالعه مطابق با نیاز های خواننده باشد، وجود ندارد! . بدون شک بـرای انسان مطالعه و خلاصه سازی حجم انبوهی از اسناد متنی که در اینترنت یا دیگر پایگـاه هـای اطلاعـاتی قابل دسترسی است، کار بسیار دشواری است و شاید با در نظر گرفتن محدودیت زمان، امری غیر ممکـن باشد. لذا چاره ای جز انجام خودکار این فرآیند توسط ماشین به ذهن انسان خطـور نمـی کنـد. بـا الگـو

---

[1].Booth and Weaver
[2].Hypertext
[3].Information Overload

گیری از انسان و مدلسازی فرآیند خلاصه سازی، استفاده از تکنیک های هوش مصنوعی به عنوان بهترین انتخاب برای دستیابی به این هدف پیشنهاد می شوند.

ذکر برخی از کاربردهای خلاصه سازی خودکار متن می تواند در تبیین اهمیت این فرآیند موثر باشد. برای مثال در شاخه پزشکی در بسیاری از مواقع دستیابی به موقع به اطلاعات و شرایط بیماران مشابه و آگاهی از شرایط خاص آنها برای درمان امری ضروری است. با وجود سرریزی اطلاعات موجود و حجم انبوه اسناد مرتبط، پزشک ملزم به مطالعه و جستجو میان همه اسناد و رکوردهای بیماران مختلف برای دستیابی به اطلاعات مورد نیاز است. در حالیکه وجود یک سیستم خلاصه سازی خودکار، بخصوص سیستم هایی که برای کاربردهای پزشکی خاص شده اند، می تواند بسیار مفید بوده و مهارت های پزشکی را با ذخیره منابع زمانی در دسترس بسیاری قرار دهد[۳].

یکی دیگر از کاربردهای خلاصه سازی خودکار متن می تواند در شاخه حقوق باشد. متخصصین این شاخه وظایف سخت و پر مسئولیتی دارند و نیز منابع آنها اغلب پراکنده بوده و مطالعه همه آنها نیازمند زمان بسیاری است. در اینجا نیز وجود سیستم هایی برای تولید خلاصه ای دقیق، به منظور در دسترس قرار دادن اطلاعاتی فشرده از اسناد قضایی مرتبط شامل قوانین، طرح های پیشنهادی و تصمیمات دادگاه های مربوطه و یا خلاصه فرآیند محاکمه، کاملا ضروری به نظر می رسد.

مزایای خلاصه سازی خودکار متن بسیار زیاد بوده و کار متخصصین و پژوهش گران مختلف را در همه زمینه ها، بسیار ساده تر کرده[۴] و تنها به یک شاخه علمی خاص تعلق ندارد. می توان گفت که هر جا که به نحوی نیازی به مطالعه و خواندن باشد، وجود چنین سیستم هایی منجر به حفظ منابع مختلف خواهد شد. به عبارت دیگر چنین سیستم هایی منجر به تسهیل امور اقشار مختلف جامعه می شود. به عنوان مثال برای افراد فعال در عرصه های اقتصادی و بورس اطلاع سریع و به موقع از اخبار و شرایط حاکم که دستخوش نوسانات بسیاری هستند، در تصمیم گیری های مربوطه تاثیر گذار خواهد بود. یا به هنگام خرید محصول خاصی می توان با گردآوری اطلاعات خلاصه از مشخصات و ویژگی های آن توسط تولید کننده های مختلف، مناسب ترین را انتخاب کرد.

علاوه بر این از این سیستم ها می توان برای طبقه بندی خودکار اسناد، فیلترینگ اسناد، جستجوی اطلاعات در اینترنت، سیستم های پیشنهاد محتوا ، پرتال های خبری و بسیاری از کاربردهای دیگر، استفاده کرد[۵].