
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

رساله دکتری مهندسی هوش مصنوعی

ارائه یک مدل برای ارزیابی انسجام متن با استفاده از ویژگی‌های آماری

نگارنده: محمد عبدالهی

استاد راهنما

دکتر مرتضی زاهدی

استاد مشاور

دکتر هدی مشایخی

دی ماه ۱۳۹۸

دانشکده: مهندسی کامپیوتر

گروه: هوش مصنوعی

رساله دکتری آقای محمد عبدالهی

تحت عنوان: ارائه یک مدل برای ارزیابی انسجام متن با استفاده از ویژگی‌های آماری

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک رساله دکتری ارزیابی گردید و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی: دکتر مرتضی زاهدی		نام و نام خانوادگی: دکتر هدی مشایخی	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی:		نام و نام خانوادگی:	
نام و نام خانوادگی:			
نام و نام خانوادگی:			

تقدیم اثر

خداوندا مددی کن تا دانش اندکم نه نردبانی باشد برای افزایش تکبر و نه حلقه ای برای اسارت و نه وسیله ای برای تجارت، بلکه گامی باشد برای تجلیل از تو و متعالی ساختن زندگی خود و دیگران. پروردگارا کمکم کن تا بتوانم دانش اندکم را از دیگران دریغ نکرده و فقط برای رضای تو در اختیار دیگران قرار دهم.

اگر در این ناچیز عظمتی است تقدیم به:

استادان و آموزگارانی که تا به این لحظه مشوق و همراهم بودند و...

همسر وفادارم که با تمام کاستی ها و کمبودهای زندگی دانشجویی همیشه همراه و مشوقم بود و هیچگاه لب به شکوه باز نکرد..

تشکر و قدردانی

خداوند را سپاس می گویم به خاطر همه نعمت هایی که به من عطا نمود، و زمان این بزرگترین نعمتی که همیشه سعی کردم از آن قافل نبوده، شاکرش باشم و از ذره ذره اش بهترین استفاده را ببرم

و استادانی را راهنمایم قرار داد تا چشمم را به زیبایی های دنیا باز نمایند

بر خود واجب دانستم از استاد بزرگم جناب آقای دکتر مرتضی زاهدی قدردانی کنم که قبل از علم به من اخلاق و افتادگی آموختند. از استاد محترم سرکار خانم هدی مشایخی جهت راهنمایی ها و مشاوره های

ارزشمندشان سپاسگذاری کرده،

و از تمامی عزیزان و همکاران آزمایشگاه پردازش زبان طبیعی دانشگاه صنعتی شاهرود، آقایان دکتر مهدی یعقوبی، دکتر مهدی حسینی، حامد زرگری، ایمان فیروزیان و خانم ها دکتر مرضیه رحیمی و سمیرا

حور علی که با همکاری ها، راهنمایی ها و همیاری هایشان همواره مشوق من بودند قدردانی نمایم.

تعهدنامه

اینجانب محمد عبدالهی دانشجوی دوره دکتری رشته **مهندسی کامپیوتر گرایش هوش مصنوعی** دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **ارائه یک مدل برای ارزیابی انسجام متن با استفاده از ویژگی‌های آماری** تحت راهنمائی دکتر مرتضی زاهدی متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود بوده و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

چکیده

ارزیابی انسجام متن از حوزه‌های پژوهشی مهم در پردازش زبان طبیعی بوده و به یکی از موضوعات مورد علاقه اما چالش برانگیز در تمامی رویکردهای پردازش متن تبدیل شده است. به طور کلی انسجام متن در دو بخش محلی و عمومی مورد ارزیابی و سنجش قرار می‌گیرد. انسجام محلی به معنای ارتباط مفهومی بین جملات متوالی با توجه به پیوستگی واژگانی آنان و انسجام عمومی به پیوستگی موضوعی سرتاسر متن و پاراگراف‌های آن خواهد بود. انسجام اغلب متن‌ها در اثر اعمال الگوریتم‌های پردازشی مانند خلاصه‌سازی، تولید متن، ساده سازی و غیره کاهش می‌یابد. از این رو تمام سیستم‌های ماشینی تمایل داشته تا پس از اعمال رویکرد پردازشی خود انسجام خروجی خود را سنجیده تا در صورت نامطلوب بودن، الگوریتم پردازشی خود را بهبود دهند.

اغلب راه حل‌های ارائه شده درگیر با مفاهیم معنایی واژگان و الگوهای زبان‌شناسی بوده و بزرگ‌ترین چالش آنان محدودیت به یک حوزه خاص، نداشتن قابلیت گسترش به سایر زبان‌ها، الگوریتم‌های پیچیده و ارزیابی انسجام محلی فقط در محدوده چند جمله متوالی بوده است. این روش‌ها با محدود کردن خود به همرخدادی واژگان در بخش کوچکی از متن اغلب آنان در سنجش انسجام عمومی بویژه در متون بلند از دقت بالایی برخوردار نبوده‌اند. تا به حال تعداد کمی از رویکردهای ارائه شده اقدام به ارزیابی همزمان انسجام محلی و عمومی کرده و در متن‌های بلند دقت خوبی نداشته‌اند.

این رساله با استفاده از رویکردهای آماری، بکارگیری دانش پنهان واژه‌های موجود به ارزیابی انسجام جملات در کل متن پرداخته است. مدل پیشنهادی با ارتقای انسجام محلی از سطح جملات متوالی به سطح پاراگراف و انسجام عمومی به سطح وابستگی موضوعی پاراگراف‌های متوالی ارزیابی دقیق‌تری را پیشنهاد داده است. رویکرد پیشنهادی با استفاده از بردارهای واژگانی word2vec، تبدیل واژه‌ها به بردار عددی و ایجاد ماتریس‌های فاصله گذر جملات، مدلی ساده و کارا با نام مدل ارزیابی انسجام مبتنی بر تعبیه کلمه ارائه داده است. مهمترین ویژگی‌های مدل ارائه شده توانایی ارزیابی همزمان انسجام محلی و عمومی در متن‌های بزرگ و با تعداد جملات زیاد، عدم وابستگی به موضوع متن و مفهوم واژه‌ها و قابلیت گسترش و اعمال بر روی سایر زبان‌ها هستند. مدل پیشنهادی در متن‌های کوتاه از دقت کمتری در مقایسه با روش‌های موجود برخوردار بوده اما برتری آن در مواجهه با متن‌های بلند و با تعداد جملات بیشتر نمایان می‌شود. این بهینه‌سازی در متن‌های با بیش از دویست و پنجاه جمله ۲,۴ درصد افزایش یافته که در مجموع برای متون بین صد و پنجاه تا سیصد جمله برتری ۱,۹۵ درصدی را نشان می‌دهد. با وجود دقت کمتر مدل در مقایسه با سایر روش‌های پیشین در متن‌های کوتاه باز هم میانگین دقت آن در کل متن‌های مورد ارزیابی (کوتاه، متوسط، بلند) ۰,۴۱ درصد بهبود را نمایش می‌دهد.

کلمات کلیدی: انسجام متن، انسجام محلی، انسجام عمومی، فضای بردار واژه، مدل‌های زبانی.

لیست مقالات مستخرج از پایان نامه

1- Abdolahi, M. and Zahedi, M., *A New Model for Text Coherence Evaluation Using Statistical Characteristics.* Journal of Electrical and Computer Engineering Innovations, 2018. 6(1): pp. 15-24**.**
2- Abdolahi, M. and Zahedi, M., *Textual Coherence Improvement of Extractive Document Summarization Using Greedy Approach and Word Vectors*, International Journal of Modern Education and Computer Science (IJMECS), 2019**.** 11(4): pp. 23-31.
3- Abdolahi, M. and Zahedi, M., *Sentence Matrix Normalization Using Most Likely N-grams Vector*, International Journal of Mechatronics, Electrical and Computer Technology (IJMEC), 2018. 8(30): pp. 4018-4028.
4- Abdolahi, M. and Zahedi, M., *A New Method for Sentence Vector Normalization Using Word2vec*. International Journal of non-linear analysis and applications (IJNAA), 2019.
5- Abdolahi, M. and Zahedi, M., *An overview on text coherence methods.* in 2016 Eighth International Conference on Information and Knowledge Technology (IKT). 2016. IEEE.
6- Abdolahi, M. and Zahedi, M., *Text coherence new method using word2vec sentence vectors and most likely n-grams*. in 2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS). 2017. IEEE.
7- Abdolahi, M. and Zahedi, M., *Sentence matrix normalization using most likely n-grams vector*. in 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI). 2017. IEEE.
8- Abdolahi, M. and Zahedi, M., *Text summarization using graph theory and machine translation techniques.* 1st conference on innovation in electrical and computer engineering (IECT), Tehran, Iran, 2017.
9- Abdolahi, M. and Zahedi, M., *An Overview on Persian Text Summarization.* 4th International Conference new study on Computer and IT, Sajad University of technology, Mashhad. Iran, 2018.

۱۰- عبدالهی محمد و زاهدی مرتضی. *بهبود روش‌های ارزیابی انسجام متن با ترکیب مزایای سه رویکرد مبتنی بر موجودیت، گراف و آنتروپی*. سومین کنفرانس بین المللی بازشناسی الگو و تحلیل تصویر ایران (IPRIA). ۱۳۹۶.

فهرست مطالب

فهرست جداول

فهرست اشکال

فهرست علائم

ردیف	عبارت علمی	معادل انگلیسی	علامت اختصاری
۱	فراوانی- عکس فراوانی	Term Frequency Inverse Document Frequency	TF-IDF
۲	بسته واژگان	Bag Of Words	BOW
۳	مدل مبتنی بر گراف‌های دوقسمتی و شبکه‌های موجودیت	Bipartite Graph Structure of Entity Grids	BGSEG
۴	پیش‌بینی واژه هدف با توجه مجموعه بزرگ واژگان	Continuous Bag Of Words	CBOW
۵	معیار شباهت کسینوسی	Cosine Similarity criterion	CS
۶	مدل ارزیابی انسجام مبتنی بر تعبیه کلمه	Embedding-based Coherence Evaluation Model	ECEM
۷	الگوریتم بیشینه سازی مورد انتظار	Expectation Maximization Algorithm	EM
۸	جایگذاری‌های متفاوت واژه‌ها در جمله ترجمه شده	Model generates a number of different translations for a sentence	IBM
۹	معکوس فاصله منهتن	Inverse Manhattan distance	IMD
۱۰	معکوس فاصله انتقال واژه	Inverse Word Mover’s Distance	IWMD
۱۱	ان_گرام‌ها با فاصله بیش از یک	Long distance n-grams	LDN
۱۲	پیش‌بینی عبارت با توجه به واژه حذف شده	Skip-Gram Negative Sampling	SGNS
۱۳	تعبیه کلمه	Word Embedding	WE
۱۴	فاصله انتقال واژه	Word Mover’s Distance	WMD

فصل۱ : مقدمه

۱-۱ مقدمه

متن قطعه‌ای از زبان نوشتاری بوده که برای ایجاد یک ارتباط معین از طریق نویسنده بکار گرفته شده و تعبیری است که خواننده از آن بدست می‌آورد. معمولا درک کامل متن بدون توجه به بافت و لحنی که متن در آن ظاهر می‌شود ناممکن است. متن می‌تواند تنها شامل یک واژه باشد. مثلا واژه "خطر" که بر روی یک تابلوی هشدار دهنده قرار دارد و شامل ساختار و لحن است. متن همچنین می‌تواند شامل یک نوشته طولانی مانند یک داستان باشد. از این رو کیفیت یک متن ربطی به اندازه آن ندارد. در واقع از دید عموم متن عبارت از زنجیره‌ای از جملات است. اما جمله‌ها تنها تولید کننده متن نبوده و تجسمی از متن را ایجاد می‌کنند [۱]. به عبارتی دیگر متن یک واحد نوشتاری معنایی است که از جمله‌ها تشکیل نمی‌شود، بلکه در جمله‌ها تجسم می‌یابد. از دیدگاه بسیاری از محققان زبان‌شناسی یک متن باید دارای هفت ویژگی زیر باشد [۲]:

- انسجام[۱]
- پیوستگی[۲]
- پیام‌رسانی و آگاهی‌بخشی[۳]
- وابستگی به یک موضوع خاص[٤]
- بینامتنی (وابستگی یک متن به متن‌های دیگر)[٥]
- قابلیت پذیرش از سوی خواننده[٦]
- هدفمندی[۷]

از بین ویژگی‌های ذکر شده، تمرکز این رساله بر ارزیابی انسجام متن است. یک متن زمانی دارای انسجام است که اجزای آن با هم یک پیوند منطقی داشته و موضوع اصلی آن قابل درک باشد. متن منسجم یک احساس را خلق می‌کند. از این رو یک متن منسجم با نوشتاری که از تعدادی جمله که به

---

[1] Coherence
[2] Cohesion
[۳] Informative
[4] Ontextually
[5] Inter textuality
[6] Acceptability
[7] Intentionality

صورت اتفاقی در کنار هم قرار گرفته و هیچگونه مفهوم کلی را نمی‌رسانند تفاوت دارد. متن می‌تواند دارای جملاتی مرتبط به هم باشد اما ممکن است دارای انسجام نباشد. به متن زیر دقت کنید:

I am a **teacher**. The **teacher** was late for **class**. **Class** rhymes with **grass**. The **grass** is always greener on the other side of the fence. But it wasn't.

در این متن هر جمله به صورت مفهومی به جمله بعد خود متصل شده است. این اتصال از نظر معنای واژگانی و دستوری کاملا صحیح است. اما جملات آن دارای هیچگونه حس مشترکی نیستند. یعنی انسجامی بین آنها وجود ندارد. گاهی ممکن است نتوان بین جملات یک متن هیچگونه اتصال واژگانی و دستوری برقرار کرد. اما می‌توان یک موضوع را از آن استنباط کرد. در این حال متن منسجم و پیوسته فرض خواهد شد. متن زیر نمونه‌ای از جملات متوالی کاملا متفاوت اما با انسجامی موضوعی هستند:

A: There's the phone.

B: I'm in the bath.

A: OK

یکی از آثار بسیار مهم در معرفی ویژگی انسجام در متن، کتاب معروف هالیدی و حسن [۲] با عنوان "انسجام در متون انگلیسی" است. نامبردگان انسجام را یک مفهوم معنایی دانسته که به روابط معنایی موجود در متن اشاره دارد. این روابط لزوما دستوری نبوده و بر اساس دانش مشترکی که بین نویسنده و خواننده موجود است شکل می‌گیرند. به عنوان مثال متن یک پاراگراف در صورتی پیوستگی دارد که جملات آن بتوانند موضوع مربوط به آن پاراگراف را به خوبی پوشش دهند. از نظر هالیدی و حسن موضوع انسجام متن برخاسته از این سئوال کلیدی است، "چه چیز یک متن گفتاری یا نوشتاری را از یک مجموعه جملات بدون ربط به هم متمایز می‌سازد؟". لذا بدیهی است که اغلب تعاریف در این خصوص برگرفته از ایده آنها است. در ادامه به توصیفی کوتاه در مورد سایر ویژگی‌های یک متن پرداخته می‌شود.

انسجام و پیوستگی مفاهیمی نزدیک به یکدیگر بوده و گاهی به جای یکدیگر بکار رفته، اما تفاوت‌هایی نیز با یکدیگر دارند. پیوستگی متن عبارت از شکل صحیح قرارگیری واژه‌های یک متن به دنبال هم و تشکیل ساختار صحیح جمله است، ولی انسجام عاملی است که نشان می‌دهد متن مجموعه‌ای از جملات مرتبط به هم بوده و یا اینکه جملات آن با هم ارتباطی ندارند. طبق این نظریه پیوستگی یک متن عبارت از ویژگی‌های ظاهری بوده که می‌توان آنرا توسط عناصری متنی مانند واژگان، حروف ربط و دستور زبان مشخص کرد. اما انسجام عبارت از بازنمایی ذهنی خواننده از متن بوده و به طور کلی اشاره به وجود یا عدم وجود نشانه‌هایی صریح در متن دارد که خواننده را قادر ساخته تا بین ایده‌های متن ارتباط برقرار کند. یک متن پیوسته ممکن است حاوی هیچ الگوی انسجامی نبوده اما جملات آن به صورت ظاهری به هم متصل باشند. انسجام، در مقایسه با پیوستگی، عبارت از مطابقت متن با انتظاراتی است که خواننده از متن داشته و نتیجه‌ای است که خواننده از متن می‌گیرد [۳].

منظور از پیام‌رسانی و آگاهی‌بخشی متن، برآورده کردن انتظار خواننده در مورد موضوعی است که عنوان متن مطرح کرده است. هر متن باید حاوی اطلاعات جدیدی باشد. افزون بر این پیام درون متن حتما باید برای خواننده قابل درک بوده و اگر خواننده نتواند اطلاعات نهفته درون آن را بیابد به این مجموعه جملات متن گفته نمی‌شود [۲].

هیچ متنی به تنهایی نوشته نشده و برگرفته از سایر متون مشابه بوده و در بافت سایر متون قرار دارد. به این ویژگی عامل بینامتنی گفته می‌شود. هر متن از یک جهت موجب ارتباط بین نویسنده و خواننده بوده و از جهتی دیگر موجب اتصال خود با سایر متون می‌شود. از این رو مجموعه نوشته‌ای را می‌توان واجد شرایط متن بودن دانست که به لحاظ معنایی به متن دیگری مرتبط باشد. به عبارت دیگر هر گاه بخشی از یک متن به صورت عینی یا مفهومی در متن دیگری حضور داشته باشد رابطه بینامتنی بین آن دو برقرار است.

قابلیت پذیرش متن به نگرش خواننده مربوط شده تا متنی منسجم و پیوسته را دریافت کند. به عبارت ساده‌تر یک زنجیره از جملات را زمانی می‌توان متن نامید که خواننده آن را درک کرده و یک مفهوم را در جملات متوالی آن دنبال کند. به صورتی که با بر هم زدن این توالی درک موضوع مشکل شده و یا اصلا امکان پذیر نباشد.

قصد یا هدفمندی متن ناشی از علایق، نیازها و انتظارات کاربران است. عامل هدفمندی تابع نگرش و موضع پدید آورنده آن نیز بوده و بر آن است که بتواند منظور خود را طی جملاتی منسجم و پیوسته به خواننده منتقل کند.

یک متن دارای ویژگی‌های دیگری در مقایسه با سایر حوزه‌ها مانند گفتار و تصویر است. به عنوان مثال یکی ازمهمترین مشکلات در حوزه پردازش گفتار و پردازش تصویر کشف، کاهش و یا حذف نویز بوده، در حالی که بزرگ‌ترین چالش در حوزه پردازش متن اطلاعات گمشده و ابهامات معنایی موجود در متون است. بعلاوه در پردازش یک تصویر سیستم پردازش کننده بیشتر به اطلاعات نهفته در خود تصویر تکیه داشته و کمتر نیازمند بررسی دانش موجود در پس‌زمینه آن و یا اطلاعات خارجی است، در حالی که در پردازش متن اطلاعات خارجی پس زمینه و دانش خارجی موجود در آن کمک بسیار بیشتری در تشخیص برخی از ابهامات آن می‌کنند. به عنوان مثال در تشخیص مفهوم یک جمله سایر جملات موجود در همان متن و یا حتی متن‌های دیگر تاثیر زیادی دارند. متن را می‌توان داده‌ای زماندار نامید. به این معنی که هر جمله در زمانی خاص رخ داده که توالی جملات آن بر اساس همین زمان تشکیل می شوند. از این رو نمی‌توان جایگاه آنان را در کل متن تغییر داد. همین مفهوم معنای اصلی انسجام متن است.

از مشکلات ذاتی موجود در ارائه و نمایش واژه‌ها در هر روش می‌توان به محل قرارگیری آنان در جمله و عبارات اصطلاحی متشکل از چند واژه اشاره کرد. به عنوان مثال هر یک از دو واژه Iran و Air دارای معنی مخصوصی برای خود هستند. اما عبارتIran Air مفهوم یک شرکت هواپیمایی را در خود

داشته و به مفهوم تکی واژه‌های تشکیل دهنده خود ارتباط کاملی ندارد. در این حالت بردار تشکیل دهنده مربوط به آن باید به مفهوم کلی عبارت توجه داشته باشد [۴]. همچنین می‌توان ارتباط انسجامی آن را با سایر جملاتی که حاوی دو واژه Iran و Air هستند در نظر گرفت و محاسبه کرد.

با سیل عظیم و افزایش روز افزون داده‌های متنی در وب، وجود کتابخانه‌های متنی الکترونیکی و مقادیر عظیم داده‌های متنی موجود در سیستم‌های کامپیوتری، ثروت عظیمی از اطلاعات خارجی و پس‌زمینه برای متن در هر زمینه و هر زبانی وجود داشته و براحتی قابل دسترس است. به زبانی ساده‌تر داده‌های متنی تنها داده‌هایی هستند که قدمت ذخیره سازی آنها برابر با قدمت تاریخ تمدن بشر بوده و براحتی می‌توان به بخش‌های با ارزشی از آن در هر حوزه، زبان، زمان، فرهنگ و ملیتی دست یافت. این چنین گنجینه داده‌ای را در هیچکدام از حوزه‌های پردازشی دیگر مانند پردازش صوت و تصویر نمی‌توان یافت. ویژگی بارزتر متن این بوده که حجم ذخیره‌سازی داده‌های آن در مقایسه با سایر داده‌های ذکر شده بسیار پایین بوده، لذا الگوریتم‌های پردازشی آن دارای سرعت و دقت بسیار بالاتری هستند. متن ساخته دست انسان بوده و این تولید اغلب توسط افراد با دانش بالاتر و با توجه به قواعد گرامری و مورفولوژیکی انجام می‌شود. از این رو بسیار ساخت یافته‌تر از سایر داده‌ها بوده و در نتیجه دانش بیشتری را می‌توان از آن استخراج کرد. دانش بافت‌شناسی[۱] موجود در متن بیشتر به نحوه تشکیل اجزای متن پرداخته که از آن جمله می‌توان به سیلاب‌ها، ریشه واژه و پیشوند و پسوندهای یک واژه اشاره کرد. دانش نحوی موجود در متن نیز شامل مواردی مانند برچسب‌گذاری واژگان، صفت‌های تفضیلی و عالی، مفرد یا جمع بودن افعال و شکل جمع اسامی و غیره است.

۱-۲ تعریف واژه‌ها

مدل‌های زبانی[۲]: یکی از مهمترین بخش‌ها در حوزه پردازش زبان طبیعی بوده که با استفاده از مفاهیم آماری احتمال ادای یک دنباله از واژه‌های به دنبال هم را محاسبه می‌کند.

مدل ان_گرام: ساده‌ترین و پرکاربردترین مدل زبانی آماری بوده که احتمال رخداد یک واژه را پس از دنباله‌ای از n-1 واژه بیان می‌کند.

بسته واژگان[۳]: این مدل عبارت است از دریافت یک مجموعه متنی، شمارش و تعیین فرکانس تکرار ظاهر شدن هر واژه در یک مجموعه متنی است.

---

۱ Morphologic

2 Language models

3 Bag of words

انسجام متن[1]: حوزه‌ای پژوهشی بوده که به تشخیص و ارزیابی انسجام موضوعی متن‌های نوشته شده توسط عامل انسانی و سیستم‌های تولید کننده متن، خلاصه‌سازی، ساده‌سازی، امتیازدهی خودکار مقالات و ترجمه ماشینی می‌پردازد.

انسجام محلی[2]: انسجام محلی به وابستگی موضوعی بخشی محلی از یک متن مانند دو یا چند جمله متوالی اشاره داشته و این انسجام زمانی اتفاق می‌افتد که ارتباط بین اجزای موجود در این بخش‌ها موجب ایجاد حسی مشترک در خواننده شود.

انسجام عمومی[3]: انسجام عمومی به پیغام اصلی موجود در کل متن اشاره داشته و به نوعی وابسته به انسجام محلی بخش‌های مختلف متن است.

عوامل انسجامی[4]: عواملی قابل استخراج از متن که بتوان با استناد به آنان رای بر منسجم بودن متن حاوی آنها داد. این عوامل می‌توانند شامل واژگان، عوامل ارتباطی بین آنها و روابط گرامری موجود در متن باشند.

فضای بردار واژه[5]: یک مدل محاسباتی از معنای واژگان بوده که از الگوهای توزیع شده واژه‌ها در مجموعه‌های بزرگ داده‌های متنی استفاده کرده و شباهت‌های معنایی آنان را در یک فضای برداری نشان می‌دهد.

تعبیه کلمه[6]: این تکنیک در ابتدا برای بکارگیری در حوزه معنایی متن معرفی شد. اما خیلی زود مشخص شد که می‌توان از آن در حوزه‌های نحوی و آماری نیز استفاده کرد. در این تکنیک شباهت‌های معنایی و ارتباط واژگان موجود هنگام قرارگیری در متن مورد توجه قرار گرفته و واژه‌ها در یک فضای برداری نگاشت می‌شوند. این فضا می‌تواند فضای اقلیدسی بوده و این نگاشت تعبیه کلمه نیز نامیده می‌شود.

همرخدادی واژگان[7]: در زبان‌شناسی، همرخدادی واژه‌ها به تکرار دو واژه یا عبارت بیش از یک حد احتمال در یک متن، در کنار یکدیگر و با یک نظم خاص اشاره می‌کند. همرخدادی می‌تواند به عنوان یک شاخص از نزدیکی معنایی این دو بخش تفسیر شود.

پیکره واژگان[8]: یک مجموعه بزرگ از واژگان و ارتباط آنها در ساختار متون است. پیکره واژگان در حقیقت یک شبکه معنایی از هزاران واژه و روابط بین آنان بوده که می‌تواند یک زبانه یا چند زبانه باشد. به عنوان مثال گربه یک پستاندار است، پستاندار یک حیوان است و حیوان یک جاندار است. از پیکره

[1] Text coherence
[2] Local coherence
[3] Global coherence
[4] Coherence factors
[5] Word space vector
[6] Word embeddings
[7] Word co-occurrence
[8] Wordnet

واژگان برای متن کاوی، تحلیل معنایی، اعتبارسنجی نظرات و بررسی درستی قواعد زبانی استفاده می‌شود.

۱-۳ هدف‌های کلی پژوهش

در سال‌های اخیر پژوهش‌های زیادی بر روی ارزیابی انسجام متون انجام شده است. همچنین در این سال‌ها تلاش‌های زیادی برای تولید سیستم‌هایی با توانایی تولید متن با قابلیت درک و خوانایی بالا و نزدیک به متون تولید شده توسط انسان نیز انجام شده است. از سوی دیگر پژوهش در مورد ایجاد سیستم‌هایی که بتوانند مقدار انسجام یک متن تولید شده را ارزیابی کرده و یا در صورت امکان آن را بهبود دهند بسیار مورد توجه قرار گرفته است. این پژوهش‌ها در اغلب حوزه‌های متفاوت پردازش متن مانند تولید متن[۱]، تحلیل و بررسی احساس از روی متن[۲]، سیستم‌های پرسش و پاسخ[۳]، امتیاز دهی به مقالات[۴] [۵-۷]، ترجمه آماری ماشینی[۵] [۸-۱۰] و خلاصه‌سازی متن[۶] [۱۱-۱۳] انجام پذیرفته است. پژوهش‌های صورت گرفته در بخش ارزیابی انسجام از دو دیدگاه معنایی[۷] و نحوی[۸] مورد بررسی قرار گرفته‌اند که موارد موجود در بخش معنایی نیازمند درک کامل مفاهیم زبان‌شناسی هستند. ارزیابی انسجام متن به دو حوزه کلی ارزیابی انسجام محلی و ارزیابی انسجام عمومی تقسیم می‌شود. انسجام محلی به معنای ارتباط مفهومی بین جملات متوالی با توجه به پیوستگی واژگانی آنان است. اما انسجام عمومی به پیوستگی موضوعی سرتاسر متن و تمامی پاراگراف‌های آن خواهد بود. یکی از مهمترین چالش‌های این حوزه تحقیقاتی ارزیابی همزمان انسجام محلی و عمومی است. تا به حال کمتر پژوهشی بر روی هر دو حوزه به طور همزمان کار کرده و روش‌های ترکیبی ارائه شده لااقل در بخش ارزیابی انسجام عمومی از دقت پایینی برخوردار بوده و یا یا انسجام عمومی بخشی از متن را ارزیابی کرده‌اند. از سویی دیگر بررسی وابستگی عمومی اجزای تشکیل دهنده در متن‌های بزرگ بسیار مشکل بوده و در رویکردهای پیشنهاد شده قبل از دقت کمی برخوردار بوده است. این دقت به دلیل فاصله زیاد بخش‌های ابتدای متن با بخش‌های با فاصله زیادتر، عدم وجود و یا الگوهای انسجامی بسیار کم بین بخش‌های اول و سایر بخش‌های با فاصله زیاد و یا عنوان متن با تمامی بخش‌های متن بوده است. در این پژوهش سعی گردیده تا با بکارگیری روش‌هایی آماری رویکردی برای ارزیابی انسجام یک متن در هر دو حوزه محلی

[1] Text generation
[2] Sentiment analysis
[3] Question answering systems
[4] Essay scoring
[5] Machine translation
[6] Text summarization
[7] Semantic
[8] Syntactic

و عمومی و به طور همزمان ارائه شود. توانایی روش ارائه شده در متن‌های بزرگ از دقت بیشتری در مقایسه با روش های قبلی است.

۱-۴ اهمیت و ضرورت پژوهش

با افزایش متن‌های موجود در وب، متون ترکیبی و تالیفی، و معرفی رویکردهای متفاوت ماشینی پردازش متن لزوم ایجاد روش‌هایی ماشینی برای ارزیابی انسجام، وابستگی موضوعی و خوانایی خروجی اینگونه متن‌ها و در صورت امکان بهبود این پارامتر هر چه بیش از پیش احساس می‌شود. گسترش دانش داده کاوی، علوم ریاضی و آمار، و دانش‌های زیر مجموعه هوش مصنوعی مانند یادگیری ماشین، شناسایی الگو، شبکه‌های عصبی و یادگیری عمیق موجب تولید و پیشنهاد رویکردهایی کارا در زمینه تولید، ارزیابی و بهبود متون منسجم شده است. این پژوهش با تکیه بر دانش هوش مصنوعی راهکاری عملیاتی برای ارزیابی انسجام متن در دو حوزه محلی و عمومی فراهم کرده است. با توجه به اینکه بررسی، تشخیص و ارزیابی انسجام عمومی نوشته‌های متنی در هر دو حالت تولید توسط عامل انسانی و سیستم‌های پردازشگر متن از اهمیت بالایی برخوردار است، لزوم پیشنهاد روشی بدون وابستگی به زبان، موضوع و اندازه بیش از پیش احساس می‌شود. با توجه به دستاوردهای این پژوهش و در راستای ارتقای سیستم‌های ارزیاب انسجام عمومی، روش ارائه شده در این رساله، انسجام خروجی تمامی سیستم‌های پردازش متن اعم از خلاصه‌سازی، ساده‌سازی، ترجمه ماشینی، تولید کننده متن و غیره را مورد ارزیابی قرار می‌دهد.

۱-۵ کارهای انجام شده در پایان‌نامه

در این رساله روشی برای ارزیابی همزمان انسجام محلی و عمومی متن ارائه شده است. رویکرد پیشنهاد شده علاوه بر سادگی الگوریتم، از نتایج بهتری در مقایسه با روش‌های پیشنهاد شده قبل در بررسی وابستگی اجزای متن بویژه در متن‌های بلند برخوردار است. دلیل این بهینه‌سازی در نظر گرفتن پاراگراف‌های موجود به عنوان زیر متن‌هایی مستقل از هم بوده که هر کدام مفهوم خاصی را شروع، دنبال و نتیجه‌گیری می‌کنند. با ارزیابی پیوستگی موضوعی و توالی صحیح جملات موجود در هر پاراگراف انسجام محلی و بررسی وابستگی مفهومی و توالی درست پاراگراف‌های موجود در کل متن انسجام عمومی آن اندازه‌گیری می‌شود.

در روش پیشنهادی ما پس از انجام پیش‌پردازش‌های اولیه و با استفاده از روش‌های مبتنی بر بردارهای واژگانی و اهمیت مکانی آنان در جمله، واژه‌های متن به بردارهایی عددی تبدیل و سپس گذر معنایی آنان در جملات متوالی و غیر متوالی و با فاصله زیادتر ارزیابی شده است. روش پیشنهادی بر

روی دادگان استاندارد Earthquakes & Accident و پایگاه داده ایجاد شده داستانی اعمال شده است. دلیل استفاده از پایگاه داده استاندارد مقایسه روش با رویکردهای قبلی بکار گیرنده آنان و دلیل استفاده از پایگاه داده ایجاد شده بررسی توانایی روش معرفی شده بر روی متن‌های داستانی و روایی بزرگ‌تر با تعداد جملات زیاد است.

۱-۶ اهمیت و ارزش پژوهش

عملیات پیش‌پردازش در مدل ارائه شده بسیار ساده، سریع و کارا است. پیش‌پردازش متن ورودی معمولا با توجه به نوع متن، زبان، الگوریتم و نوع عملیات پردازشی بعدی متفاوت بوده و برای هر حوزه پردازش متن مدلی خاص پیشنهاد می‌شود. یکی از چالش‌های مهم روش‌های پیش‌پردازش برای هر متن محدودیت‌های موجود در هر حوزه پردازشی بوده که منحصر به همان حوزه خواهد بود. لذا نمی‌توان روش ارائه شده در یک حوزه را به سایر حوزه‌ها تعمیم داده و بکار برد. انتخاب نوع عملیات پیش‌پردازش و درصد اعمال آن تاثیر بسیار بالایی بر دقت و سرعت عملیات اصلی پردازشی بعد خواهد داشت. پیش‌پردازش‌های انجام شده در این تحقیق عبارت از پاک‌سازی، جداسازی اجزای متن و نرمال سازی واژگان باقیمانده بوده که با روش‌هایی ساده انجام شده است.

با توجه به ماهیت و شکل نگارش متون داستانی و روایی (بویژه در نمونه‌های با حجم بیشتر و طول زیاد)، جداسازی اجزای آنان اغلب با چالش‌هایی همراه بوده و موجب تولید خروجی‌هایی با درصدی خطا می‌شود. در روش پیشنهادی با استفاده از ابزارهای موجود در NLTK پایتون نشانه‌گذاری[۱] متناسب برای اینگونه متون آموزش داده شده که موجب تولید نتایجی بهتر شده است.

۱-۷ نوآوری پژوهش

در این رساله رویکردی ساده برای نمایش و ارزیابی انسجام متن پیشنهاد شده است. تمرکز این رویکرد بر بکارگیری تعبیه کلمه و تبدیل واژه‌ها به بردارهای عددی و جملات به ماتریس‌های ایجاد شده از این بردارها است. تاکنون از تبدیل جملات متن به ماریس‌های عددی در حوزه سیستم‌های پرسش و پاسخ استفاده شده است [۱۴]. اما در این رساله از این ایده برای تشخیص و ارزیابی انسجام متن استفاده شده است. با وجود اینکه اغلب روش‌های پیشنهاد شده قبلی انسجام یک متن را در یک محدوده محلی ارزیابی کرده و کمتر به ارزیابی انسجام عمومی پرداخته‌اند، روش معرفی شده هر دو رویکرد ارزیابی انسجام محلی و عمومی را به طور همزمان مورد توجه قرار داده است. در اغلب رویکردهای معرفی شده مفهوم انسجام محلی پیوستگی موضوعی چند جمله متوالی و در یک محدوده محلی بوده است. اما

[۱] Tokenizer

رویکرد پیشنهاد شده مفهوم انسجام محلی را از چند جمله متوالی به سطح یک پاراگراف ارتقا داده است. در روش‌های ارائه شده قبل مفهوم انسجام عمومی وابستگی موضوعی تک تک جملات متن با عنوان و موضوع اصلی بوده است [۱۵]. در عمل بررسی ارتباط موضوعی یک به یک جمله‌ها با عنوان متن عملی پر هزینه بوده و این عمل در متن‌های بزرگ بسیار مشکل‌تر می‌شود. اما در روش ارائه شده ارتباط موضوعی پاراگراف‌ها معیار ارزیابی انسجام عمومی است. لذا با توجه به روش معرفی شده ارزیابی انسجام عمومی به دلایل ذیل عملی ساده بوده و با دقتی بالا انجام پذیر است. نخست انسجام هر پاراگراف به عنوان یک واحد منسجم محلی قبلا ارزیابی شده و اغلب یک پاراگراف بخش بزرگی از یک متن را تشکیل می‌دهد. ثانیا تعداد پاراگراف‌های موجود در یک متن بسیار کمتر از تعداد جملات آن هستند. عمل ارزیابی انسجام موضوعی تعداد محدودی پاراگراف عملی ساده‌تر بوده و با دقتی بهتر انجام‌پذیر است. این عمل همان ارزیابی انسجام عمومی کل متن است.

یکی از بزرگ‌ترین چالش‌ها در رویکردهای تبدیل کننده واژه به بردار و جمله به ماتریس هم اندازه نبودن ماتریس‌های تولید شده است. به دلیل تفاوت در اندازه جمله‌های موجود در هر متن، ماتریس‌های جملات در اندازه‌های مختلف تولید شده و موجب اشکال در اعمال الگوریتم‌های پردازشی و مقایسه‌ای خواهد شد. تا به حال رویکردهای متفاوتی برای رفع این چالش پیشنهاد شده است. یکی از مهمترین روش‌ها تبدیل ماتریس جمله به بردار جمله بوده که این عمل با میانگین‌گیری ستون‌های ماتریس انجام می‌شود. با وجود اینکه روش‌های مبتنی بر میانگین گیری از سادگی الگوریتم و سرعت بالایی برخوردار هستند اما دارای مشکلاتی از قبیل احتمال بالای نزدیک بودن بردارهای جملات، کاهش شدید اطلاعات قابل استخراج از جمله و دقت بسیار پایین در ارزیابی انسجام عمومی می‌شوند. روش‌های جدیدتر برای کاهش ابعاد و تبدیل ماتریس‌های جملات متن به یک بردار هم اندازه از یادگیری عمیق استفاده کرده‌اند [۱۶]. در رویکرد پیشنهادی دیگری (روش ارائه شده در مقاله ژورنالی استخراجی شماره ۳) برای هم سایز کردن ماتریس جمله‌ها از بردار ان_گرام‌های با احتمال بالا در متن اصلی استفاده شده است. مدل زبانی ان_گرام ساده‌ترین و پرکاربردترین مدل زبانی آماری بوده که احتمال رخداد یک واژه را پس از دنباله‌ای از n-1 واژه بیان می‌کند. این روش از دقت بسیار بالاتری نسبت به میانگین گیری بردارهای واژگان برخوردار بوده و منجر به تبدیل متن به ماتریس‌هایی نرمال و هم سایز برابر با تعداد جملات متن می‌شود. ماتریس‌های نرمال شده آماده برای اعمال هر گونه اعمال پردازشی بوده که نمونه‌ای از این بکارگیری در روشی ارائه شده برای ایجاد خلاصه‌های استخراجی منسجم بکار گرفته شده است (روش ارائه شده در مقاله ژورنالی استخراجی شماره ۲). اما در رویکرد ارائه شده در این رساله به هیچ عنوان نیازی به هم سایز کردن ماتریس‌های تولیدی نبوده و الگوریتم پیشنهادی از روشی ساده‌تر ماتریس فاصله جملات را تشکیل داده که برای مقایسه آنان نیازی به هم بعد بودن آنان نیست.

برخلاف رویکردهای مبتنی بر کیسه واژگان[1] و مدل‌های استفاده کننده از معیار TF-IDF[2] این روش اصلا متکی به شکل نگارش واژه و مفهوم ظاهری آن نبوده، به راحتی واژه‌های با شکل نگارش متفاوت ولی مفهوم نزدیک به هم و همچنین شکل نگارش نزدیک به هم ولی مفهوم متفاوت در یک محدوده محلی را تشخیص داده و ارتباط مفهومی جملات حاوی آنها را کشف می‌کند.

تقریبا تمامی رویکرد پیشنهادی قبل در ارزیابی انسجام محلی مشکلی نداشته و این عمل را با دقت قابل قبول انجام داده‌اند. اما اغلب در ارزیابی و سنجش انسجام عمومی دارای مشکل بوده و این عمل را با دقت پایینی انجام داده‌اند. یکی از بزرگ‌ترین چالش‌های آنان در برخورد با متن‌های بزرگ و با تعداد جملات بالا بوده است. محاسبه وابستگی مفهومی جملات متوالی و نزدیک به هم، به مرور و با طولانی شدن متن دارای خطای محاسبات شده، به صورتی که هر چه به انتهای متن نزدیک‌تر می‌شویم این فاصله بیشتر شده، تا حدی که ممکن است هیچگونه الگوی انسجامی بین جملات انتهایی و ابتدایی متن نبوده، اما از نظر الگوریتم منسجم فرض شوند. رویکرد پیشنهادی مبتنی بر پاراگراف تا حدود زیادی این مشکل را کاهش داده و در متن‌های بزرگ از دقت و سرعت بالاتری نسبت به رویکردهای قبلی برخوردار است.

ویژگی بینامتنی تقدم و تاخر چند متن را در یک حوزه مشخص می‌کند [۱۷-۱۸]. به این معنی که با بررسی متن‌های مربوط به یک نویسنده می‌توان تا حدودی ترتیب انتشار آنان را مشخص کرد. این ویژگی حتی در مورد متن‌های نگارش شده در مورد یک موضوع اما توسط افراد مختلف نیز صدق می‌کند. خصوصیت بینامتنی زمانی مورد توجه است که تعدادی از متن‌های موازی مربوط به یک فرد یا یک موضوع مورد ارزیابی باشد. تا به حال از ویژگی بینامتنی یک سند متنی برای تشخیص و ارزیابی انسجام آن استفاده نشده است. اما با توجه به اینکه در رویکرد پیشنهادی واحد پردازش برای انسجام محلی و همچنین عمومی پاراگراف است، از این ویژگی برای ارزیابی انسجام عمومی استفاده شده است. وقتی که خروجی سطح قبل پاراگراف‌های منسجم باشند، انسجام بینامتنی می‌تواند ارتباط موضوعی هر پاراگراف را با پاراگراف‌های قبل و بعد و همچنین با موضوع اصلی متن که همان عنوان آن است را ارزیابی کند.

رویکرد پیشنهادی مراحل ارزیابی انسجام محلی و عمومی را با استفاده از تبدیل متن به ماتریس‌های عددی و با کمک الگوریتم word2vec انجام می‌دهد. این الگوریتم در اغلب حوزه‌های دیگر پردازش متن بکار برده شده و انعطاف پذیری آن در متن‌ها، حوزه‌ها و موضوعات مختلف بررسی شده است. در این روش عملیات ارزیابی در سه سطح انجام می‌شود. در سطح نخست بر روی متن ورودی پیش‌پردازش‌های پیش فرض انجام می‌شود. سطح دوم ارزیابی انسجام محلی در یک پاراگراف مورد توجه قرار گرفته و در سطح آخر با بررسی ارتباط موضوعی پاراگراف‌ها به هم، انسجام عمومی کل متن

[1] Bag of words
[2] Term frequency inverse document frequency

مورد ارزیابی قرار می‌گیرد.

۱-۸ دلایل انتخاب روش

قرار گیری مناسب جملات به دنبال هم، انتقال مفاهیم مرتبط از یک جمله به جملات دیگر، القای یک موضوع، رسیدن به هدفی مشخص و درک درست خواننده از مفهوم اصلی متن، هدف اصلی تمام تولید کنندگان متن است. این انتقال و گذر صحیح مفاهیم و آشکار شدن هدف متن، انسجام نامیده شده و تمامی رویکردهای تولید کننده و پردازش کننده متنی مایل به ارزیابی کیفیت آن هستند. اغلب رویکردهای پیشین برای ارزیابی انسجام یک متن از عوامل معرفی شده توسط هالیدی و حسن [۲] کمک گرفته‌اند. این عوامل ابتدا در حوزه زبان‌شناسی مطرح شده و تمامی آنان دارای دیدگاهی معنایی هستند.

ارزیابی شباهت واژگان موجود در بخش‌های متفاوت یک متن و یا متن‌های موازی در بسیاری از رویکردهای پردازش متن بکار گرفته شده است. از جمله مهمترین رویکردهای استفاده کننده از ارزیابی شباهت واژگان می‌توان به بازیابی اطلاعات[۱]، ترجمه ماشینی[۲]، خوشه‌بندی اسناد متنی[۳] و سیستم‌های تشخیص سرقت ادبی[۴] اشاره کرد. این رابطه شباهت می‌تواند مربوط به دو واژه نزدیک به هم مانند (هندوانه، طالبی) و یا مرتبط با هم مانند (بوته، طالبی) باشد. با توجه به اینکه سیستم‌های تشخیص سرقت ادبی میزان شباهت بخش‌های متفاوت دو متن را از نظر واژگانی و وابستگی مفهومی جملات ارزیابی می‌کنند، می‌توان از الگوهای بکار گرفته شده در آنان برای ارزیابی انسجام و وابستگی مفهومی دو متن و یا بخش‌های متفاوت یک متن استفاده کرد. از این رو ارزیابی شباهت واژگان راه حل مناسبی برای تشخیص و ارزیابی انسجام و وابستگی جملات متوالی (انسجام محلی) و حتی با فاصله (انسجام عمومی) در یک متن است.

برای ارزیابی شباهت واژگان از دو روش الگوریتم‌های مبتنی بر تزاروس (روش‌های مبتنی بر WordNet) و الگوریتم‌های توزیعی (روش‌های مبتنی بر بردار واژگان) استفاده می‌شود. در الگوریتم‌های مبتنی بر تزاروس دو واژه زمانی به هم شبیه هستند که در سلسله مراتب تزاروس به هم نزدیک باشند. در این روش طول مسیر هر واژه با خودش برابر با یک است (۱-۱).

$$Similarity(w_1, w_2) = \frac{1}{MinPath(w_1, w_2)} \qquad (۱-۱)$$

در این فرمول میزان شباهت دو واژه w_1 و w_2 برابر با معکوس مسیر آنان $(MinPath\ (w_1, w_2))$ در ساختار درختی wordnet است. تصویر (۱-۱) بخشی از پیکره wordnet بوده که این فرمول را توصیف می‌کند.

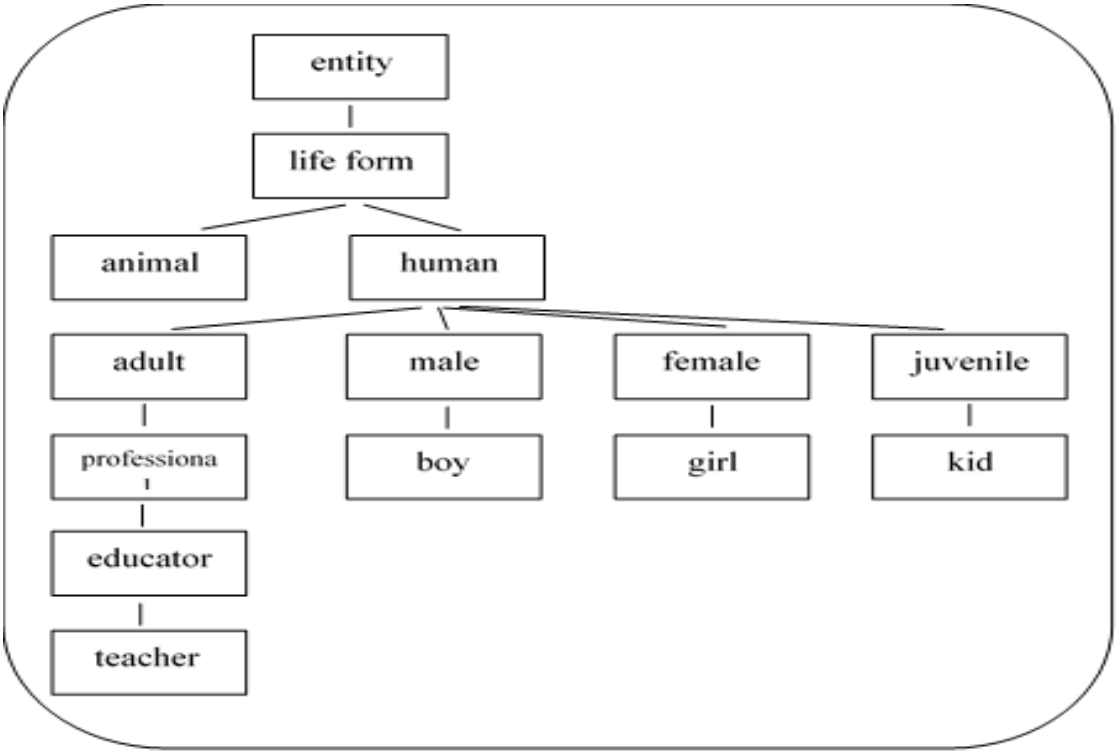
---

[۱] Information retrieval

[۲] Machine translation

[۳] Document clustering

[۴] Plagiarism detection



شکل ۱-۱: بخشی از پیکره wordnet

در این پیکره شباهت کلمه boy با خودش برابر یک شده، شباهت کلمه boy با girl برابر ۰,۲۵ شده شباهت کلمه boy با professional برابر ۰,۲۰ می‌شود. گاهی اوقات بین دو کلمه بیش از یک مسیر وجود داشته که از بین آنها کوتاه‌ترین مسیر انتخاب می‌شود.

$$\begin{aligned} &Similarity\ (boy, boy) = 1 / minPath\ (1, 1) = 1/1 = 1 \\ &Similarity\ (boy, girl) = 1 / minPath\ (1, ۴) = ۰/۲۵ \\ &Similarity\ (boy, professional) = 1 / minPath\ (1, 5) = ۰/۲۰ \end{aligned} \qquad (۲-۱)$$

اما الگوریتم‌های توزیعی که نام دیگر آنان روش‌های مبتنی بر بردار[۱] است واژگان بیشتری را شامل شده و در ابتدا در زبان‌هایی که دارای پایگاه داده تزاروس نبودند مورد استفاده قرار گرفت. اما خیلی زود و با معرفی رویکردهای مبتنی بر یادگیری عمیق در اغلب زبان‌ها بویژه انگلیسی از توجه بیشتری برخوردار شد. این الگوریتم‌ها بر این مفهوم استوار هستند که اگر دو واژه A و B در محیط‌های یکسانی دیده شوند دارای شباهت بیشتری هستند. یکی از مهمترین الگوریتم‌های این حوزه الگوریتم word2vec (WE) بوده که در بسیاری از روش‌های پردازشی متن سال‌های اخیر مورد استفاده قرار گرفته است. روش پیشنهادی این رساله از بردار واژگان (WE) و تعبیه کلمه‌ای آنان در متن برای ارزیابی وابستگی موضوعی و انسجام جملات متن استفاده کرده است. از این رو روش معرفی شده درگیر مفاهیم معنایی واژه‌ها نشده و از رویکردهای غیر معنایی جهت ارزیابی انسجام استفاده کرده است.

مقالات و روش‌های پایه‌ای که این پژوهش با استناد به آنان انجام شده است عبارت از روش پیشنهادی

[۱] Vector space

مت کسنر و همکارانش [۱۹] و رویکرد پیشنهادی اچ لی [۱۴] هستند. مت کسنر روش جدیدی برای تعیین فاصله بین دو بخش متن ارائه داده و اچ لی [۱۴] برای ارزیابی ارزش یک جمله از تبدیل آن به بردارهای واژگانی مبتنی بر یادگیری عمیق استفاده نموده است. ارزیابی و مقایسه مدل نیز با روش ارائه شده توسط کریستینا لیوما و همکارانش [۲۰] انجام شده است. نامبردگان با ترکیب رویکرد شبکه موجودیت و گراف‌های دو قسمتی سه پارامتر جدید برای ارزیابی انسجام معرفی کرده‌اند که موجب افزایش کارایی و سادگی الگوریتم مقایسه جملات شده است.

در روش ارائه شده این رساله واژه‌های موجود در جمله به بردار و فاصله گذر واژه‌ها بین جمله‌ها نیز به ماتریس‌هایی عددی تبدیل شده‌اند. سپس با ترکیب این ماتریس‌ها، ماتریس شباهت پاراگراف ایجاد شده که با ارزیابی ضریب همبستگی سطرهای آن انسجام محلی پاراگراف اندازه‌گیری می‌شود. در نهایت با اعمال روش بر روی فاصله تمام پاراگراف‌های متن انسجام و وابستگی موضوعی پاراگراف‌های متوالی و در نتیجه انسجام عمومی متن اندازه‌گیری می‌شود. عمل تبدیل جملات متن به ماتریس‌های عددی قبلا هم در پردازش متن بکار گرفته شده است. آلکسی سورین و آلساندرو موسچیتی (۲۰۱۶) با استفاده از بردارهای واژگانی Word2vec و تبدیل جملات به ماتریس‌هایی متشکل از این بردارها رویکردی نوین را بر سیستم‌های پرسش و پاسخ ارائه دادند. نامبردگان با ایجاد ماتریس‌های جمله سوال و مجموعه‌ای از جملات پاسخ، متناسب‌ترین پاسخ را برای سوال مربوطه انتخاب کرده‌اند [۱۴] [۲۱].

تاکنون اغلب رویکردهای پیشنهاد شده در این حوزه از عوامل واژگانی موجود در روش‌های مبتنی موجودیت استفاده کرده‌اند. این رویکردها از بین واژگان موجود در یک جمله فقط اسم‌ها، عبارات اسمی و ضمیرها را استخراج کرده و فقط در جایگاه‌های فاعلی یا مفعولی مورد پردازش قرار داده‌اند. ضمنا تقریبا در تمامی این روش‌ها (به جز رویکردهایی که در ترکیب با تئوری گراف بکار گرفته شده‌اند) ارتباط بین این واژگان فقط در جملات مجاور مورد بررسی قرار گرفته که در نتیجه تمرکز انسجام ارزیابی شده بیشتر در یک محدوده محلی بوده است. اما بکارگیری WE می‌تواند در تشخیص ارتباط موضوعی تمام جملات متن بکار گرفته شده و علیرغم تشخیص انسجام محلی، انسجام عمومی متن را نیز ارزیابی کند.

چالش مهم تمامی رویکردهای پردازش متن جایگذاری صحیح واژه‌ها در جمله و جمله‌ها در پیکره اصلی متن است. این چالش در رویکردهایی مانند خلاصه‌سازی استخراجی، تولید متن و ساده سازی متن که خروجی آنان با ایجاد تغییراتی در متن ورودی بوجود می‌آید بیشتر نمایان می‌شود. این ویژگی موجب شده تا اکثر روش‌ها و الگوریتم‌های معرفی شده متکی به زبان و موضوعی خاص بوده، امکان اعمال آن بر روی سایر زبان‌ها و حوزه‌ها ممکن نبوده و یا در صورت امکان به شدت دقت و کارایی روش کاهش یابد. یک ویژگی مهم در رویکردهای مبتنی بر فضای بردار و بویژه WE عدم توجه آنان به ترتیب واژه‌ها در متن است. این رویکردها فارغ از محل قرارگیری واژه در جمله، بردار آن را ایجاد و ویژگی‌های مورد نظر را استخراج می‌کنند.

یک متن به اجزای کوچک‌تری تقسیم می‌شود. در اغلب روش‌های پیشنهاد شده در پردازش متن

واژه به عنوان کوچک‌ترین بخش یک سند متنی در نظر گرفته شده است. البته یک واژه را می‌توان به اجزای کوچک‌تری نیز تقسیم کرد. بسیاری از رویکردها با توجه به هدفی که داشته‌اند از این اجزا برای عملیات پردازشی خود استفاده کرده‌اند. اجزای قابل تفکیک یک واژه عبارت از کاراکتر، سیلاب، ریشه و وند هستند. بکارگیری این اجزا نقش مهمی در ارزیابی صحت و درستی واژه و نقش آن در جمله دارد. توجه به این اجزا در هرگونه عملیات پردازشی ارزیابی درستی یک واژه، نقش گرامری آن در جمله و میزان تشابه و وابستگی آن به سایر واژه‌ها از اهمیت بالایی برخوردار است. به دلیل عدم نیاز به تشخیص صحت واژه و پردازش واژگانی، پژوهش حاضر به این اجزا توجه نداشته و کوچک‌ترین واحد مورد پردازش را جمله در نظر گرفته است.

عبارت جزء بعدی یک متن بوده که شامل چند واژه است. در برخی از رویکردهای پردازشی عبارات بسیار مورد توجه قرار گرفته‌اند. مدل‌های زبانی ان_گرام نوع خاصی از عبارات بوده که در بسیاری از رویکردهای متفاوت پردازش متن مورد استفاده قرار گرفته‌اند.

اما مهمترین جزء در تشکیل یک متن منسجم، جمله‌های مرتبط و منظم هستند. جمله را می‌توان مهمترین جزء در هر عملیات پردازش متن در نظر گرفت. زیرا جمله نخستین جزء معنی‌دار در هر سند متنی بوده و به تنهایی می‌تواند حامل یک پیام باشد. جمله می‌تواند از حداقل دو واژه (فاعل و فعل) تشکیل شده که حاوی پیامی بوده و حتی بسیار بزرگ‌تر (تا چند خط) و شامل سایر اجزا مانند اسم، قید، صفت، حروف اضافه تا عبارات ترکیبی مانند عبارات اسمی، عبارات قیدی، عبارات وصفی و غیره باشد. وقتی که مفهوم انسجام در یک متن مورد توجه باشد نخستین جزئی که باید مفهوم انسجام در آن رعایت شود واحد جمله است. قابلیت پیام‌رسانی یک جمله به مفهوم انسجام آن بوده و این انسجام با ترکیب واژه‌های درست و جایگذاری صحیح آنان تضمین و در غیر این صورت کاهش خواهد یافت.

ارزیابی انسجام متن به دو بخش عمده ارزیابی محلی و عمومی تقسیم می‌شود. تا به حال اغلب رویکردهای معرفی شده به تشخیص و ارزیابی انسجام محلی پرداخته‌اند. مهمترین روش معرفی شده در مورد انسجام محلی رویکرد مبتنی بر موجودیت بوده است [۲۲]. اما این نکته قابل ذکر بوده که تقریبا تمامی رویکردهای قبلی نیز ارتباط چند جمله متوالی را در نظر گرفته و به انسجام درونی جمله توجهی نکرده‌اند. آنان نیز یک جمله را به عنوان یک واحد متنی دارای انسجام پذیرفته، ارتباط آن را با جملات قبلی و بعدی مورد ارزیابی قرار داده و در صورت وجود ارتباطی مطمئن و قابل قبول بین آنها، منسجم بودن سند متنی مورد بررسی را فقط در یک حوزه محلی تایید کرده‌اند.

یکی دیگر از اجزای مهم هر سند متنی پاراگراف است [۲۳]. یک پاراگراف بخشی از یک متن تشکیل شده از چند جمله بوده که با در نظر گرفتن موضوع اصلی متن، بر روی زیر موضوعی واحد تمرکز دارد [۲۳]. علاوه بر استقلال موضوعی نسبی، پاراگراف دارای ساختار درونی بوده و جملات تشکیل دهنده آن دارای ارتباط هستند. پاراگراف دارای سر، بدنه و بخش انتهایی بوده و هر نوع به‌هم‌ریختگی این نظم سبب می‌شود که انسجام عمومی متن کاهش یافته و درک موضوع اصلی کم شود. موضوع اصلی هر متن به صورت کلی در عنوان متن مشخص شده است. این موضوع به چند زیر موضوع تقسیم شده که

توصیف مربوط به هر زیر موضوع در قالب یک پاراگراف در متن قرار می‌گیرد. از این رو موضوع هر پاراگراف تا حدودی مستقل از پاراگراف‌های دیگر بوده و فقط در مسائل کلی‌تر با آنها ارتباط دارد. اگر هر پاراگراف به صورت جداگانه بررسی شود، مانند این است که یک متن جداگانه مورد ارزیابی قرار گرفته است [۲۴]. این زیر متن باید کلیه خصوصیات یک متن کامل از قبیل انسجام، پیوستگی، پیام‌رسانی و آگاهی‌بخشی، وابستگی به یک موضوع خاص، بینامتنی (وابستگی یک متن به متن‌های دیگر)، قابلیت پذیرش از سوی خواننده و هدفمندی را داشته باشد. استخراج این ویژگی‌ها از یک متن کوتاه به نام پاراگراف به مراتب ساده‌تر از کل سند متنی است. حال اگر بتوان بجای انسجام محلی چند جمله متوالی محدود، انسجام موضوعی یک پاراگراف را انسجام محلی فرض کرد، پیوستگی موضوعی ارزیابی شده از دقت بالاتری برخوردار خواهد بود. انسجام محلی در سطح پاراگراف به معنی ارتباط و پیوستگی موضوعی جملات موجود در یک پاراگراف است. یک پاراگراف منسجم یک متن منسجم محلی بوده که در درون خود قانون انسجام عمومی را رعایت کرده است [۲۵].

روش‌های ارزیابی مبتنی بر موضوع[۱] با یافتن جمله موضوعی و ارتباط آن با سایر جملات انسجام یک متن را ارزیابی می‌کنند [۲۶]. یافتن موضوعیت مرکزی یک متن نیازمند پردازش‌های پیچیده‌ای است، اما یافتن موضوعیت مرکزی یک پاراگراف ساده بوده و براحتی امکان پذیر است. آنچه یک پاراگراف را از سایر واحدهای یک سند متنی متمایز می‌کند سادگی استخراج جمله موضوعی[۲] و جمله هدف[۳] آن است. نما و طرح کلی پاراگراف توسط جمله اول آن مشخص می‌شود. جمله آخر پاراگراف جمله هدف بوده و یک نتیجه‌گیری کلی از موضوعیت آن را به تصویر می‌کشد [۲۷]. از این رو برای ارزیابی انسجام موضوعی جملات هر متن به صورت محلی و در سطح یک پاراگراف می‌توان ارتباط موضوعی جملات متوالی را با جمله موضوعی متن سنجید. حتی امکان ارزیابی هدفمندی موضوعی پاراگراف نیز با مقایسه موضوعی جملات مجاور با جمله هدف وجود دارد. ارتباط انسجامی بین جملات از نزدیک‌ترین جمله تا جملات با فاصله تا یک حدود مشخص از یک الگوی خاصی پیروی می‌کند. یعنی جملات نزدیک‌تر دارای الگوهای انسجامی بیشتر و جملات با فاصله بیشتر دارای الگوهای کمتری هستند. روند افزایش اختلاف بین دو جمله با فاصله آنها نسبت مستقیم داشته و اگر این افزایش از یک روند ثابتی پیروی کند نشانه انسجام بیشتر پاراگراف مربوطه است.

روزنفلد در سال ۱۹۹۶ با معرفی ایده‌ای در ان_گرام‌های با فاصله[۴] ثابت کرد که اطلاعات موجود در تاریخچه واژه‌های متوالی با افزایش فاصله بین آنها کاهش یافته و این کاهش از یک الگوی خاصی پیروی می‌کند [۲۸]. اما با افزایش فاصله بیشتر از پنج واژه میزان اطلاعات موجود ثابت خواهد ماند. نامبرده ارزیابی سرگشتگی ارتباط[۵] بین واژه از فاصله شش تا ده را اندازه گرفته و برای اثبات نظریه خود این

---

[۱] Topic based coherence evaluation models

[۲] Topic sentence

[۳] Goal topic sentence

[۴] Long distance n-grams

[۵] Perplexity

مقدار را در فاصله هزار واژه نیز مشخص کرده است (جدول ۳-۱). وی در پژوهش انجام شده مشخص کرد که این مقدار از فاصله شش به بعد تغییر محسوسی نداشته و حتی در فاصله هزار نیز تقریبا با فاصله شش تا ده برابر است. این نظریه در مورد اطلاعات مشترک و انسجام معنایی تمام اجزای یک متن، بویژه جملات متوالی نیز قابل قبول است. انگرام‌های با فاصله بیشتر از یک[۱] نیز در واژه‌های یک متن به همین مفهوم اشاره داشته و ایده‌ای برای ارزیابی انسجام عمومی متن است. بر همین اساس جملات متوالی تا فاصله پنج از روند کاهش الگوهای انسجام به صورت ثابت پیروی کرده و این کاهش الگوهای انسجامی از آن فاصله به بعد تقریبا ثابت می‌شود. حال اگر این کاهش انسجامی در پنج جمله‌های متوالی پاراگراف محاسبه شود مقادیر حاصل تقریبا نباید تغییر محسوسی داشته باشند. هر چه این میزان برابری نزدیک‌تر به هم باشد پاراگراف منسجم‌تر است.

جدول ۱-۱: اندازه پیچیدگی ارتباط بدست آمده توسط روزنفلد با LDB [۲۸]

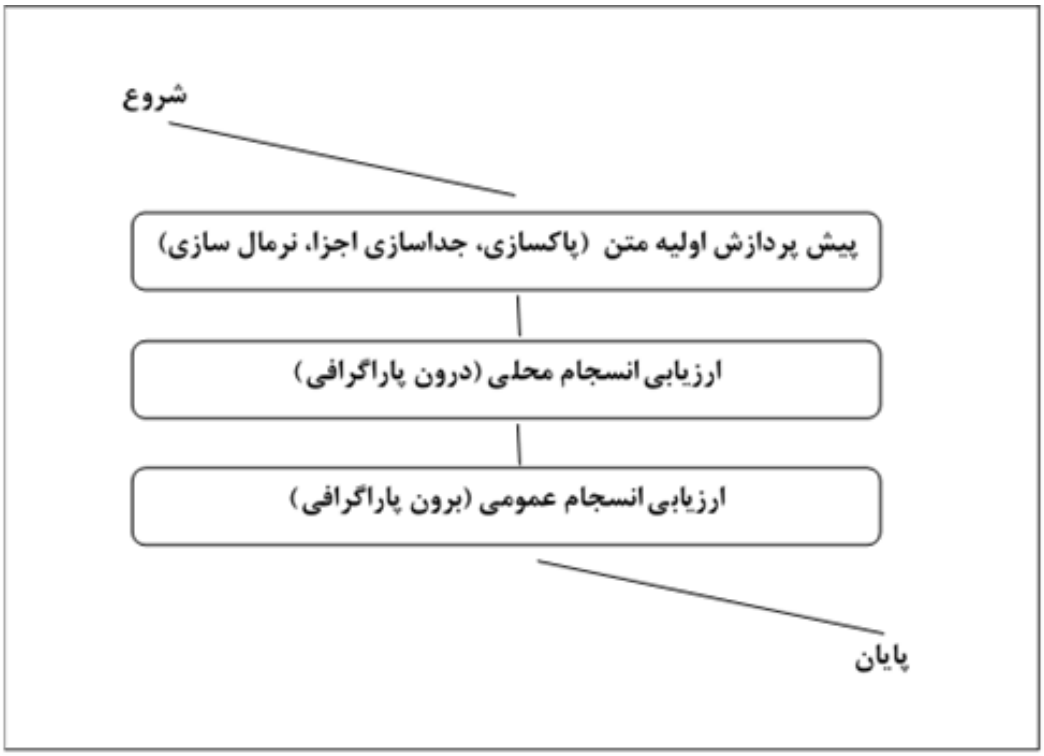
فاصله بین دو واژه	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۰۰۰
پیچیدگی ارتباط	۸۳	۱۱۹	۱۲۴	۱۳۵	۱۳۹	۱۳۸	۱۳۸	۱۳۹	۱۳۹	۱۳۹	۱۴۱

در روش پیشنهادی این رساله ابتدا با استفاده از رویکرد پیشنهادی روزنفلد انسجام درونی یک پاراگراف ارزیابی شده و سپس به ارتباط انسجامی بین پاراگراف‌های سند متنی پرداخته می‌شود. معمولا پاراگراف‌های متوالی با بکارگیری مجدد واژه‌ها و یا حتی واژه‌های با مفهوم بسیار نزدیک به هم مرتبط می‌شوند [۸۱]. یکی از ویژگی‌های مهم هر سند متنی ویژگی بینامتنی و یا به عبارتی ساده‌تر وابستگی یک متن به متن‌های دیگر منتشر شده توسط یک فرد یا موجود در یک حوزه است. ویژگی بینامتنی تقدم و تاخر تولید یک متن را نیز مشخص کرده و پوشش می‌دهد. به این معنی که با بررسی متن‌های مربوط به یک نویسنده می‌توان تا حدودی ترتیب انتشار آنان را مشخص کرد. این ویژگی حتی در مورد متن‌های نگارش شده در مورد یک موضوع اما توسط افراد مختلف نیز صدق می‌کند. ویژگی بینامتنی زمانی مورد توجه است که تعدادی از متن‌های موازی مربوط به یک فرد یا یک موضوع مورد ارزیابی باشد. تا به حال از ویژگی بینامتنی یک سند متنی برای تشخیص و ارزیابی انسجام آن استفاده نشده است. با توجه به اینکه در رویکرد پیشنهادی واحد پردازش برای انسجام محلی و همچنین عمومی پاراگراف است، ویژگی بینامتنی بین پاراگراف‌ها مورد ارزیابی قرار گرفته و مبنای انسجام عمومی متن خواهد بود.

وقتی که خروجی سطح نخست پاراگراف‌های منسجم باشند، انسجام بینامتنی می‌تواند ارتباط موضوعی هر پاراگراف را با پاراگراف‌های قبل و بعد و همچنین با موضوع اصلی متن (عنوان متن) ارزیابی کند. همخوانی و نزدیکی جملات موضوعی هر پاراگراف با عنوان متن، تشخیص تقدم و تاخر پاراگراف‌های تشکیل دهنده و ارتباط بینامتنی بین پاراگراف‌ها می‌تواند ابزار مناسبی برای ارزیابی انسجام عمومی

[۱] Long distance n-gram (LDN)

باشد. رویکرد پیشنهادی یک پاراگراف مجازی متشکل از عنوان سند متنی به جای جمله موضوعی پاراگراف و جملات موضوعی هر پاراگراف به عنوان سایر جملات تشکیل دهنده آن ایجاد می‌کند. اعمال روش‌های تعریف شده ارزیابی محلی پاراگراف بر روی پاراگراف مجازی موجب تشخیص و تعیین اندازه انسجام عمومی کل متن خواهد شد. تصویر (۱-۲) نمایی کلی از روش پیشنهادی را نشان می دهد.



شکل ۱-۲: نمایی کلی از روش پیشنهادی

همانطور که در شکل (۱-۲) نشان داده شده است ابتدا پیش‌پردازش‌های اولیه بر روی متن ورودی انجام می‌شود. این عملیات شامل پاک‌سازی و آماده سازی اولیه متن، جداسازی بخش‌های متفاوت مانند واژگان، جملات و پاراگراف‌ها، نرمال سازی جملات و واژگان است. پس از پیش‌پردازش وابستگی موضوعی جملات موجود در هر پاراگراف به عنوان ارزیابی انسجام محلی اندازه‌گیری می‌شود. این وابستگی شامل فاصله گذر واژگان هر جمله به جملات بعدی است که این مقادیر درون ماتریسی به نام ماتریس فاصله گذر قرار داده می‌شود. در نهایت ارتباط منطقی و موضوعی پاراگراف‌های متوالی بررسی می‌شوند. این ارتباط تعیین کننده انسجام عمومی متن است.

۱-۹ ساختار پایان نامه

این رساله شامل چهار فصل دیگر به شرح زیر است:

فصل دوم (مرور ادبیات): در این فصل در ابتدا مروری بر تاریخچه ارزیابی انسجام متن، کاربردها، حوزه‌ها و رویکردهای متفاوت آن پرداخته می‌شود. سپس به مدل‌های معروف پیشنهادی، روند پیشرفت و طبقه‌بندی رویکردهای متفاوت در حوزه پژوهشی تشخیص یک متن منسجم بررسی شده و در نهایت رویکرد پیشنهادی در این رساله و با روش‌های پیشین مقایسه شده و نوآوری‌های آن ذکر می‌شود.

فصل سوم (روش پیشنهادی): این فصل به چند بخش اصلی تقسیم می‌شود. ابتدا مدل پیش‌پردازش پیشنهادی معرفی شده، سپس مدل ارائه شده جهت ارزیابی انسجام توصیف و بردارهای انسجام معرفی می‌شوند. در نهایت به معرفی نحوه محاسبه و ارزیابی شباهت معنایی و میزان انسجام دو جمله پرداخته می‌شود.

فصل چهارم (ارزیابی و نتایج): در این فصل ابتدا پایگاه داده مورد استفاده معرفی شده و سپس مدل پیشنهادی ارزیابی می‌شود.

فصل ۲ : مروری بر ادبیات موضوع و کارهای انجام شده قبل

۲-۱ مقدمه

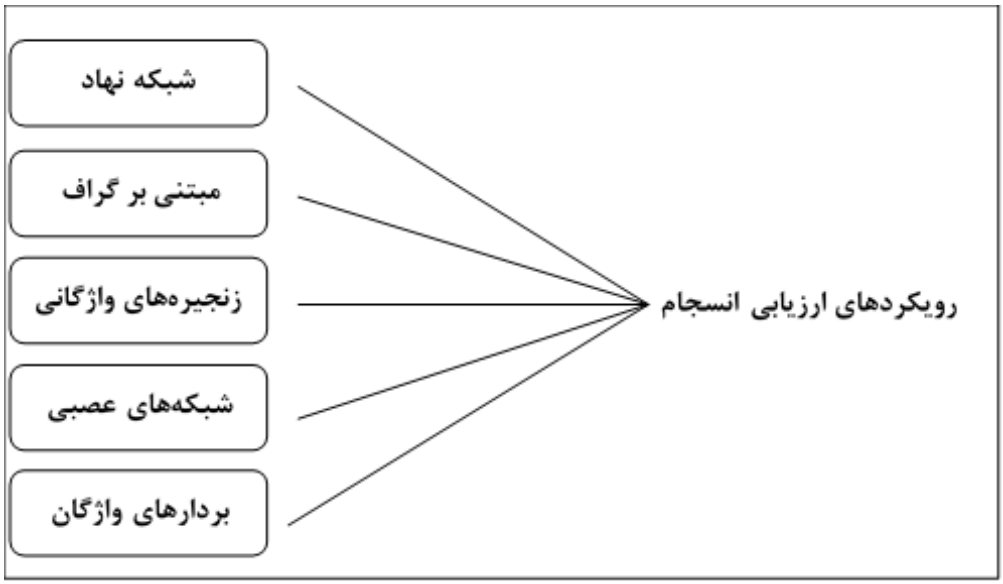
متن‌های غیر منسجم اغلب خروجی سیستم‌های پردازش متن هستند. می‌توان رویکردهای متفاوت ارزیابی انسجام را بر اساس آنان دسته‌بندی کرد. معمولا یک متن غیر منسجم خروجی سیستم‌های تولید کننده متن، خلاصه‌سازی، ساده‌سازی، امتیازدهی خودکار مقالات، متن تولید شده توسط یک سیستم پرسش و پاسخ و یا حتی متن تولید شده توسط یک فرد اما با دانش نگارشی پایین باشد. از این رو تمامی سیستم‌های ذکر شده سعی در ارزیابی انسجام و پیوستگی موضوعی خروجی تولید شده خود داشته تا در صورت لزوم آن را بهبود بخشند.

تولید متن منسجم، از همان ابتدای معرفی رویکردهای ماشینی خلاصه‌سازی متن توسط لون در سال ۱۹۵۸ مورد توجه قرار گرفت [۲۹]. بزرگ‌تر شدن متن‌ها، امکانات جستجو قوی‌تر، دستیابی و ترکیب مجموعه بزرگ منابع متنی در وب، پیشرفت سریع در تمامی حوزه‌های پردازش زبان طبیعی و متن، موجب شد که انسجام و پیوستگی متون تولید شده به یکی از مهمترین دغدغه‌های پژوهشگران این حوزه‌ها تبدیل شود. اما بررسی نظری ویژگی‌های یک متن منسجم و روش‌های بهبود آن نخستین بار توسط هالیدی و حسن [۲] در کتاب "انسجام در انگلیسی" مطرح گردید. لذا بدیهی است که اغلب تعاریف در این خصوص برگرفته از ایده آنها است. نامبردگان انسجام را یک مفهوم معنایی دانسته که به روابط معنایی موجود در متن اشاره دارد. این روابط لزوما دستوری نبوده و بر اساس دانش مشترکی که بین نویسنده و خواننده موجود است شکل می‌گیرند. یکی از نخستین و مهمترین مطالعات بر روی ارزیابی ماشینی انسجام متن به کار انجام شده توسط فولتز و همکارانش در سال ۱۹۹۸ باز می‌گردد [۳۰]. از نظر نامبردگان متن منسجم نوشته‌ای است که ارتباطی معنایی بین جملات متوالی آنان وجود داشته باشد. نامبردگان همچنین یک نمایش برداری را برای این مدل معرفی کرده که از معنای لغوی برای محاسبه ارتباط معنایی بین جملات متوالی استفاده می‌کرد. از آن زمان به بعد رویکردهای دیگری معرفی شدند که پایه اصلی پژوهش آنان نظریه فولتز بود. بیشتر روش‌های معرفی شده به محاسبه مقدار ارتباط موضوعی جملات متوالی پرداخته و مبتنی بر روش‌های با ناظر بودند.

در ادامه به معرفی و مروری کوتاه بر مهمترین رویکردهای ارائه شده، روش‌های پیشنهادی، روند پیشرفت، طبقه‌بندی یافته‌های پژوهش‌های دیگر پژوهشگران در سطح دنیا و تعیین و شناسایی ابهام‌های موجود در حوزه تشخیص یک متن منسجم پرداخته می‌شود. این بخش با هدف مروری کلی بر رویکردهای گذشته سعی بر این داشته تا دانسته‌های موجود، پیش‌زمینهٔ تاریخی و وضعیت کنونی موضوع را چنان بیان کند که خواننده بدون مراجعه به منابع پیشین، نتایج حاصل از مطالعات قبلی را درک و ارزیابی کند. سپس پژوهش‌های مشابه با رویکرد معرفی شده در این رساله معرفی شده و تطابق‌ها و تفاوت‌های آنان بررسی می‌شوند. در نهایت با توجه به بررسی انجام‌شده بر روی مراجع پژوهش، بخش‌های قابل گسترش و چشم‌اندازهای آینده مورد بررسی قرار می‌گیرند.

۲-۲ دسته بندی رویکردهای ارزیابی انسجام متن بر اساس مدل‌های معرفی شده

در این بخش به دسته بندی مدل‌های مهم معرفی شده ارزیابی انسجام متن بر اساس موضوع پرداخته می‌شود. البته مسلم است که این بخش تمامی مدل‌های موجود را در بر نگرفته و فقط به آنهایی پرداخته می‌شود که از نظر این مطالعه مهم است. در شکل (۲-۱) این مدل‌ها معرفی شده‌اند.



شکل ۲-۱: دسته بندی حوزه‌های مهم معرفی شده ارزیابی انسجام متن بر اساس موضوع

۲-۲-۱ مدل‌های مبتنی بر شبکه موجودیت[1]

مدل مبتنی بر شبکه موجودیت یکی از مهمترین و پر استفاده‌ترین رویکردهای پیشنهاد شده است. در این مدل علاوه بر ویژگی‌های معرفی شده روش‌های قبلی، نقش گرامری اسم‌ها و عبارات اسمی موجود در جملات متوالی در نظر گرفته شده و به عنوان یک ویژگی انسجام دو جمله مورد ارزیابی قرار می‌گیرد [۱۵]. این مدل نخستین بار توسط بارزیلی و لاپاتا پیشنهاد شد [۳۱] [۲۲]. اما خیلی زود توسط افراد دیگر نیز بکار گرفته شده و به صورت ترکیبی با سایر روش‌ها مورد استفاده قرار گرفت. ایده اصلی این رویکرد بر این است که تغییرات ایجاد شده بین عوامل انسجامی جملات مجاور هم در یک متن منسجم دارای الگوهای منظم و با قاعده‌ای هستند. در این روش هر متن با یک ماتریس دو بعدی که به آن شبکه موجودیت نیز گفته می‌شود نمایش داده می‌شود. در ماتریس ایجاد شده سطرها نشان دهنده

[1] Entity grid models

جملات و ستون‌ها نشان دهنده عوامل انسجامی موجود در جمله بوده و سطرهای متوالی مشخص‌کننده جملات متوالی هستند. به علاوه درایه مشخص‌کننده وجود عامل انسجامی، شامل اطلاعات نقش آن در جمله است. این اطلاعات می‌توانند به صورت‌های متفاوتی نمایش داده شوند. به این صورت که اگر عامل مورد نظر فاعل یا عبارت فاعلی بود با (S)، مفعول یا عبارت مفعولی با (O) و در غیر این صورت با (X) نمایش داده می‌شود. در صورت عدم وجود عامل مورد نظر در جمله درایه مربوط به آن با (-) مشخص خواهد شد. در روش فوق در صورت تکرار یک عامل در یک جمله با بیش از یک نقش گرامری، نقش گرامری دارای رتبه بالاتر انتخاب می‌شود. به عنوان مثال نقش فاعلی برای یک عامل دارای رتبه بالاتری نسبت به نقش مفعولی است. فرضیه اساسی در رویکرد فوق بر این اصل استوار است که شکل توزیع و پیکربندی عوامل انسجامی در ماتریس تولیدی مشخص‌کننده برخی قواعد در نوع انسجام ایجاد شده است. به عنوان مثال متن‌های منسجم دارای ستون‌هایی با چگالی بالا بوده و یا اینکه ستون‌هایی که عوامل موجود در آنها بیشتر فاعل یا مفعول هستند بیشتر موجب ایجاد انسجام در متن می‌شوند.

رویکردهای مبتنی بر شبکه موجودیت جزو محبوب‌ترین روش‌های ارزیابی انسجام متن بوده و پژوهش‌های زیادی بر روی آن انجام و پژوهشگران متعددی از آن بهره جسته و در بهبود آن کوشیده‌اند. اما این رویکردها دارای محدودیت‌هایی نیز بودند و این محدودیت‌ها موجب شد در سال‌های اخیر توجه پژوهشگران به سوی معرفی روش‌های بهبود یافته آن، سایر حوزه‌ها مانند تئوری گراف، شبکه‌های عصبی، یادگیری عمیق و چندین حوزه و الگوریتم دیگر و یا بکارگیری روش به صورت ترکیبی با سایر الگوریتم‌ها جلب شود. کاستی بزرگ روش اولیه مبتنی بر شبکه موجودیت استفاده از تکرار عینی اسم‌ها و عبارات اسمی در مقایسه جملات متوالی بود. اما در روشی که توسط ژانگ و همکارانش در سال ۲۰۱۵ معرفی شد این کاستی تا حدودی برطرف شده و ایده استفاده از اسامی هم خانواده و مشابه به جای تکرار عینی اسم یا عبارت اسمی به جای فاعل و مفعول معرفی شد [۳۲]. در تحقیقات نامبردگان مشخص شد که در ۴۲٪ حالات دو جمله همسایه دارای یک موجودیت اسمی مشترک نبوده که این حالت موجب کاهش چهل درصدی دقت ارزیابی انسجام می‌شود. از این رو نامبردگان استفاده از واژه‌هایی که دارای ارتباطات معنایی هم‌خانوادگی مانند (car, Automobile) یا ارتباطات حوزه‌ای مانند (car, petrol station) را نیز پیشنهاد کرده‌اند. رویکردهای مبتنی بر شبکه موجودیت جزو محبوب‌ترین روش‌های ارزیابی متن بوده و پژوهش‌های زیادی بر روی آن انجام و پژوهشگران متعددی از آن بهره جسته و در بهبود آن کوشیده‌اند. این روش‌ها دارای محدودیت‌هایی نیز بودند و این محدودیت‌ها موجب شد در سال‌های اخیر توجه پژوهشگران را به سوی استفاده از سایر حوزه‌ها مانند تئوری گراف، شبکه‌های عصبی، یادگیری عمیق و غیره جلب شده و یا از ترکیب آنان با روش‌های مبتنی بر شبکه موجودیت استفاده کنند. تسنیم محیودین و همکاران در رویکردی به ارائه روشی برای ارزیابی انسجام بین متن‌های غیر همزمان ارسال شده توسط افراد به صورت ایمیل، بلاگ، متن‌های ارسال شده در شبکه‌های اجتماعی، پاسخ سوالات مطرح شده در ارتباطات قبلی و یا سایت‌های به اشتراک گذاری تجربیات شخصی پرداخته‌اند [۳۳]. نامبردگان نشان داده‌اند که می‌توان از مدل‌های انسجامی موجود در رویکردهای مبتنی

بر شبکه موجودیت و ترکیب آنان با شبکه‌های عصبی برای پیش بینی ساختار موضوعی گفتگو بخصوص اطلاعات حیاتی برای ساخت سیستم‌های مکالمه موثر و سیستم‌های پاسخگو به پرسش‌های افراد استفاده کرد. در ادامه به برخی از مهمترین محدودیت‌های روش اولیه مبتنی بر موجودیت پرداخته می‌شود:

- اغلب آنان فقط امکان ارزیابی انسجام را در حوزه‌ای محلی پرداخته‌اند. تشخیص ارزیابی انسجام عمومی در آنان از دقت بالایی برخوردار نیست.
- فقط وابستگی موضوعی جملات مجاور در نظر گرفته می‌شود. ممکن است جملات با فاصله بیشتر نیز دارای شواهدی مبتنی بر تایید انسجام داشته باشند. اما روش فوق قادر به استخراج آنان نیست.
- امکان توسعه روش جهت ارائه مدلی که بتواند علاوه بر ارزیابی انسجام، راهی را برای بهبود انسجام پیشنهاد دهد وجود ندارد.
- وابستگی شدید به زبان و موضوع متن داشته و با اعمال یک روش بر روی زبان و موضوعی متفاوت به شدت از دقت روش کاسته می‌شود.
- از جایگاه گرامری واژه‌ها استفاده می‌کنند (فاعل و مفعول).
- وابستگی شدید به مفاهیم معنایی و قواعد گرامری منجر به درگیر شدن مدل با پیچیدگی ابعاد می‌شود. به عنوان مثال اگر K حالت گذر و R قاعده گرامری وجود داشته باشد، شبکه‌ای با KR گذر ایجاد می‌شود [۳٤]. به همین دلیل تمام مدل‌های قبلی از K<=3 استفاده کرده‌اند.
- روش‌های انتخاب و استخراج موجودیت‌ها (فاعل و مفعول) کاملا معنایی بوده و در مقایسه با روش‌های آماری از پیچیدگی محاسباتی بیشتری برخوردار هستند.
- ویژگی‌های استخراجی از موجودیت‌ها (فاعل، مفعول، غیره) محدوده جستجو را کوچک می‌کنند. مدل ارائه شده راهی برای بررسی سایر موجودیت‌ها (افعال، صفات، قیدها، ...) ارائه نداده است [۳٤].
- پراکندگی داده‌ها در آن بشدت احساس می‌شود. موجودیت‌های قابل استخراج به صورت پراکنده در سرتاسر متن وجود داشته که ممکن است با هم فاصله زیادی هم داشته باشند. روش ارائه شده پاسخی برای یافتن انسجام جمله‌های حاوی موجودیت‌ها اما با فاصله زیاد ارائه نداده است.
- این رویکرد برای متن‌های کوتاه و یا در حد متوسط مفید بوده، اما در برخورد با متن‌های بزرگ و با تعداد جملات بالا به شدت از سرعت و دقت آن کاسته می‌شود.

برای پوشش برخی از کاستی‌های ذکر شده رویکردهایی ترکیبی ارائه شده است. از جمله روش‌های ترکیبی برای پوشش برخی از کاستی‌های ذکر شده می‌توان به ترکیب آن با شبکه‌های عصبی [۳۵] و ترکیب با گراف‌های دوقسمتی [۲۰] اشاره کرد. در روشی معرفی شده توسط بورستین از ویژگی‌های معرفی شده در مدل مبتنی بر موجودیت و ترکیب آنان با برخی از مفاهیم زبان شناسی مانند غلط‌های گرامری و واژگانی مدل بهینه برای ارزیابی مقالات و نوشته‌های دانشجویی پیشنهاد شده است [٦].

ترکیب مدل شبکه موجودیت و شبکه‌های عصبی بازگشتی برای ارزیابی انسجام محلی توسط دت تین بکار گرفته شده است [۳٤].

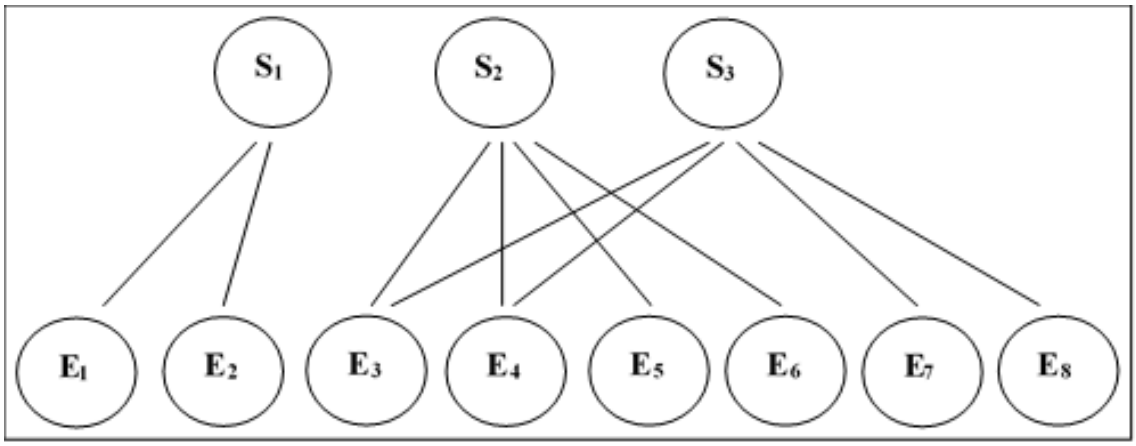
۲-۲-۲ مدل‌های مبتنی بر گراف[۱]

گراف‌ها یکی از مناسب‌ترین ساختمان‌های داده برای ذخیره سازی، نمایش اطلاعات، نمایش ارتباط بین داده‌ها، پردازش اطلاعات و امکانات پیمایش به شکل‌های متفاوت هستند. با توجه به ماهیت ویژه یک متن و ارتباط غیر خطی جملات با فاصله‌های متفاوت استفاده از نظریه گراف‌ها در تمامی حوزه‌های پردازش متن رایج بوده و همیشه به عنوان یکی از مهمترین و شاید نخستین گزینه برای پژوهشگران حوزه پردازش متن مورد توجه قرار گرفته است [۳۶]. برخی از کاستی‌ها و مشکلات موجود در مدل مبتنی بر موجودیت و ویژگی‌ها و امکانات خاص گراف‌ها عامل اصلی این توجه بوده و ترکیب این دو ایده موجب تولید رویکردهایی با کارایی بسیار بهتر شده است [۱۵]. یکی از کاستی‌های مدل اولیه مبتنی بر موجودیت ارزیابی انسجام دو یا سه جمله مجاور بوده و جملات با فاصله بیشتر را پوشش نمی‌دهد. این مدل در متن‌های با تعداد جملات بالا موجب کاهش شدید دقت می‌شود. در این مدل انسجام محلی بین جملات متوالی ارزیابی می‌شود و این تشخیص همیشه با درصد خطایی پذیرش می‌شود. اما به مرور تا به انتهای متن خطاهای مربوط با هم جمع شده، ممکن است که جملات ابتدا و انتهای متن هیچگونه ارتباطی با هم نداشته اما متن منسجم فرض شود. استراب و گیوندئو با معرفی مدلی ترکیبی مبتنی بر موجودیت و گراف تا حدودی دامنه ارزیابی انسجام را از جملات متوالی فراتر برده و توانستند ارتباط موضوعی جملات با فاصله بیشتر را نیز مشخص و ارزیابی کنند [۳۷]. ترکیب روش شبکه موجودیت با تئوری گراف یکی از بهترین رویکردهایی بوده که موجب شده علاوه بر ارزیابی انسجام محلی انسجام عمومی یک متن نیز ارزیابی شود. در روش پیشنهادی نامبردگان تعاملات بین جملات و عوامل انسجامی موجود در آنها در یک گراف دو قسمتی[۲] و به صورت $G = (V_s, V_e, L, w)$ مدل می‌شود. در گراف حاصل V_s مجموعه جملات، V_e مجموعه عوامل انسجامی، L پیوندهای بین آنان و W وزن اختصاص داده شده به هر پیوند است. بخش نخست گراف دو قسمتی ایجاد شده حاوی جملات و بخش دوم حاوی عوامل انسجامی استخراج شده از جملات است (تصویر ۲-۲). هر نود جمله به یک نود عامل پیوند برقرار کرده، اگر و فقط اگر در جمله عامل مورد نظر وجود داشته باشد. گراف مورد نظر جهت‌دار بوده و همیشه جهت آن از بخش جملات به بخش ها عوامل انسجامی است. وزن هر پیوند با توجه به نقش گرامری عامل مورد نظر در جمله تعیین می‌شود. سپس گراف تولید شده تبدیل به یک ماتریس شده که درایه‌های مورد نظر در صورت ارتباط بین جمله و موجودیت موجود در آن عدد ارتباطی و در غیر این صورت مقدار صفر را دریافت می‌کنند. در این روش برای کاهش گراف و کوچک کردن ماتریس اسپارس ایجاد

[۱] Graph based models

[۲] Bi-partial graph

شده نگاشتی در گراف صورت می‌گیرد. در گراف نگاشت شده نودها فقط شامل جملات متن بوده که در صورت وجود موجودیت مشترک بین آنان پیوندی بین آنان برقرار می‌شود. پیوندها جهت دار بوده و ترتیب ارتباط بین نودها با توجه به ترتیب قرار گیری جمله‌ها در متن خواهد بود. این پیوندها وزن‌دار بوده و وزن هر پیوند مشخص‌کننده تعداد موجودیت مشترک بین دو جمله است. اگر بین دو جمله موجودیت مشترکی وجود نداشت هیچگونه پیوندی نیز بین آنها وجود نخواهد داشت.



شکل ۲-۲: گراف دو قسمتی ایجاد شده مدل استراب [۳۷]

محدودیت مدل مبتنی بر موجودیت این بود که فقط توانایی مشخص کردن انسجام بین جملات همسایه را داشت. اما این رویکرد با قرار دادن عوامل استخراجی در گراف موجب شد که دامنه مقایسه از جملات همسایه فراتر رفته و عمل ارزیابی انسجام در متن‌های کوتاه در کل جملات متن و در متن‌های بلند در محدوده وسیع‌تری انجام شود. فاز آموزشی مدل پیشنهادی بسیار ساده‌تر بوده و برخی مشکلات موجود در مدل مبتنی بر موجودیت مانند ماتریس‌های تنک[۱] و پراکندگی داده‌ها را تا حد زیادی بهبود یافته است. وزن یال‌های مربوطه به صورت معادله (۲-۱) (۲-۲) محاسبه می‌شوند [۳۷]. در این معادله E_{ik} عبارت از مجموعه موجودیت‌های مشترک بین جمله S_i و S_k ، $w(e, S_i)$ وزن بین موجودیت مربوطه e و جمله S_i و N تعداد تعداد لینک‌های خروجی از یک نود است.

$$W_{ik} = \sum_{e \in E_{ik}} w(e, S_i) \times w(e, S_k) \qquad (۲-۱)$$

$$LocalCoherence(T) = AvgOutDegree(P) \qquad (۲-۲)$$
$$= 1/N \sum_{i=1..N} OutDegree(s_i)$$

پترسون و سیمونسن نیز روشی ترکیبی از مدل مبتنی بر موجودیت، تئوری گراف و آنتروپی برای ارزیابی ارتباط موضوعی جملات یک متن پیشنهاد کردند [۳۸]. در این رویکرد نیز موجودیت‌های استخراج شده از متن به عنوان رئوس گراف و از ارتباط بین آنان به عنوان یال‌ها استفاده شده است.

[۱] Sparse matrix

ایده اصلی روش پیشنهادی آنان این بود که با افزایش واژه‌های شرکت کننده در متن اطلاعات جانبی بیشتری وارد متن می‌شود. با افزایش اطلاعات جانبی، تمرکز بر روی یک موضوع کاهش یافته که موجب پایین آمدن انسجام و پیوستگی عمومی متن می‌شود. روش مورد اشاره با محاسبه آنتروپی انسجام متن را اندازه‌گیری کرده است. برای این کار، ان_گرام های موجود در جملات با توجه به ترتیب وقوع در متن و در هر جمله از متن استخراج شده و سپس با استفاده از معیار بیشینه احتمال استاندارد[۱] مقدار احتمال آنها محاسبه می‌شود (۲-۳) (۲-۴) (۲-۵) [۳۸]. در این معادلات $f(e_i)$ فرکانس تکرار واژه e_i در متن d، $|E|$ نیز مجموع تکرارهای تمام موجودیت‌های استخراج شده در متن d است. به همین ترتیب احتمال وقوع ان_گرام‌ها نیز محاسبه شده که در آن $p(e_{i-1}, e_i)$ موجودیت e_i بوده و پس از موجودیت e_{i-1} آمده است (۲-۴). در نهایت آنتروپی موجودیت‌های استخراج شده محاسبه می‌شود (۲-۵). در این معادله $H(X)$ مقدار آنتروپی متغیر X در فضای نمونه Ω است.

$$p(e_i) = f(e_i) \Big/ |E| \qquad (۲-۳)$$

$$p(e_i \mid e_{i-1}) = f(e_{i-1}, e_i) \Big/ f(e_i) \qquad (۲-۴)$$

$$H(X) = -\sum_{x \in \Omega} p(x) \log_2 p(x)$$
$$p(x) = \Pr(X = x)_{x \in \Omega} \qquad (۲-۵)$$

نامبردگان علاوه بر معیار درجه خروجی هر نود که در روش‌های پیشنهادی پیشین مورد استفاده قرار گرفته شده، معیارهای مهم دیگری مانند رتبه صفحه[۲]، ضریب خوشه‌بندی[۳]، میانی[۴]، فاصله موجودیت[۵]، جریان موضوعی مجاور[۶]، جریان موضوعی مجاور وزن‌دار[۷] و معیارهای جریان موضوعی غیر مجاور[۸] را بکار گرفته‌اند.

مسگر و استراب در رویکرد دیگری از زیرگراف‌های تکراری برای استخراج الگوهای انسجامی استفاده کرده‌اند [۳۹]. ایده اصلی نامبردگان بر این موضوع استوار بود که متن‌های منسجم دارای الگوهای خاصی در زیرگراف‌های استخراجی خود هستند. از این رو پس از ایجاد گراف متن، تمامی زیرگراف‌های القایی سه نودی آن استخراج کردند. سپس تعداد زیرگراف‌ها را با حذف زیرگراف‌های دارای پیوندهای دوطرفه و معکوس کاهش داده و در نهایت همانند رویکردهای قبل انسجام متن مورد نظر محاسبه کردند.

---

۱ Standard maximum likelihood

۲ PageRank

3 Clustering coefficient

4 Betweenness

۵ Entity distance

6 Adjacent topic flow (ATF)

7 Adjacent weighted topic flow (AWTF)

8 Non adjacent topic flow metrics

معیارهایی مانند تعداد اجزا[۱]، زیرگراف‌های تکراری[۲]، زیرگراف‌های ساده[۳]، امضای گراف[۴]، حداقل زیرگراف پشتیبان[۵] و میانگین درجه خروجی نود[۶] از معیارهای سنجش انسجام در رویکرد پیشنهادی هستند. روش فوق پیشنهادی در مورد استفاده از زیرگراف‌های بزرگ‌تر و بیش از سه نود را نیز داده اما توضیحاتی در مورد روش انجام این کار نداده است. در معیار میانگین درجه خروجی نود (۲-۶) و زیرگراف‌های تکراری (۲-۷) *OutDegree(s)* مجموعه وزن‌های خروجی از نود s و $||S||$ تعداد جملات موجود در متن بوده و sg زیرگراف، $count(sg_i,,G)$ تعداد زیرگراف‌های sg در گراف G و ϕ زیرگراف‌های مرتبط و استخراجی است [۳۹].

$$AvgOutDegree(P) = \sum_{s \in S} OutDegree(S) \Big/ \|S\| \qquad (۲-۶)$$

$$\phi(sg_i, G) = count(sg_i, G) \Big/ \sum_{sg_i \in (sg_1, \ldots, sg_m)} count(sg_j, G) \qquad (۲-۷)$$

استفاده از گراف دوقسمتی و نگاشت آن به گراف‌های فشرده‌تر چالش‌هایی را نیز به همراه دارد. تقریبا در اغلب مدل‌های معرفی شده یک گراف دو قسمتی تولید شده به یک گراف دیگر نگاشت شده تا گراف و ماتریس تولید شده کوچک‌تر شود. اگر چه این روش در اغلب رویکردهای مبتنی بر گراف استفاده شده است اما گراف تولید شده قدرت کافی برای نمایش تمام اطلاعات انسجامی بین جملات را ندارد. زیرا برخی از اطلاعات و نشانه‌های انسجامی بین جملات که در گراف دوقسمتی اصلی وجود داشتند در نگاشت مربوطه نادیده گرفته می‌شوند. کریستینا لیما و همکارانش روشی جدید را برای حل این مشکل پیشنهاد داده‌اند. نامبردگان گراف دوقسمتی را نگاشت نکرده، اما سه ویژگی مهم از آن را استخراج کرده و بر اساس آنان انسجام متن را ارزیابی کرده‌اند [۳۶]. ویژگی‌های استخراج شده در این رویکرد ضریب خوشه‌بندی مبتنی بر فاصله دو طرفه[۷] *(bipDCC)*(۲-۸)، ضریب خوشه‌بندی نامتقارن دو طرفه[۸] *(bipACC)* (۲-۹) و ضریب همبستگی دو طرفه[۹] *(bipLC)* (۲-۱۰) هستند [۳۶]. در این معادلات N_T زیرمجموعه‌ای از نودهایی هستند که دارای پیوند بوده، i و j موقعیت جمله s_i و s_j است.

---

[۱] Number of components
[۲] Frequent subgraphs
[۳] Basic subgraphs
[۴] Graph signature
[۵] Minimum support
[۶] Average out degree
[۷] Bipartite distance-based clustering coefficient
[۸] Bipartite asymmetric clustering coefficient
[۹] Bipartite linkage coefficient

$$bipDCC\left(s_i, s_j\right) = \frac{1}{|j-i|} \times \frac{\left|NT\left(s_i\right) \cap NT\left(s_j\right)\right|}{\left|NT\left(s_i\right) \cup NT\left(s_j\right)\right|} \quad (۸-۲)$$

$$bipACC\left(s_i, s_j\right) = \frac{1}{|i-j|} \times \frac{\left|NT\left(s_i\right) \cap NT\left(s_j\right)\right|}{\left|NT\left(s_i\right)\right|} \quad (۹-۲)$$

$$bipLC\left(s_i\right) = \frac{\sum_{ek,el \in NT(si)} \frac{1}{d_{si}\left(ek,el\right)}}{\frac{\left|\left(NT\left(s_i\right)\right)\right| \times \left(\left|NT\left(s_i\right)\right| - 1\right)}{2}} \quad (۱۰-۲)$$

در یکی از مقالات استخراج شده از این رساله (مقاله استخراجی کنفرانسی ۹) با ارائه روشی ترکیبی از مزایای سه رویکرد مبتنی بر موجودیت، گراف و آنتروپی استفاده شده و روشی جدید که کمبود و مشکلات روش‌های قبلی را برطرف کرده و بهبود داده معرفی شده است. یکی از بزرگ‌ترین کاستی‌های رویکرد نخستین شبکه موجودیت محدودیت آن در تشخیص انسجام محلی و عدم امکان در نظر گرفتن ارتباط جملات با فاصله دورتر بود. در رویکرد پیشنهادی در این مقاله با ایجاد یک ماتریس بالا مثلثی و قرار دادن ویژگی‌های انسجامی استخراجی در آن امکان سنجش انسجام عمومی جملات ایجاد شده است. نوآوری‌های روش پیشنهادی در مقاله ذکر شده عبارت از موارد اشاره شده زیر هستند. نوع موجودیت‌ها (فاعل، مفعول، غیره) در ماتریس فوق مشخص شده و در امتیاز دهی وابستگی دو جمله تاثیر داده می‌شوند. در صورتی که امتیاز نوع موجودیت‌ها در رویکرد نخستین با هم برابر بود. از طرفی دیگر در رویکرد نخستین اگر یک موجودیت در دو نقش گرامری بود همیشه نقش غالب مد نظر قرار می‌گرفت. اما در رویکرد پیشنهادی هر دو نقش در نظر گرفته شده و موجب افزایش امتیاز وابستگی دو جمله می‌شوند. در مدل شبکه موجودیت اگر یک موجودیت با بیش از یک بار تکرار و در یک نقش وجود داشت، یک ارتباط در نظر گرفته شده و یک امتیاز انسجام به دو جمله داده می‌شد. در روش پیشنهادی این موارد در نظر گرفته شده و امتیاز بیشتری به انسجام دو جمله دارای این شرایط داده می‌شود. در نهایت در روش‌های قبلی جملات دارای موجودیت مشترک با فاصله بیشتر تاثیری در افزایش انسجام نداشته، ولی در روش پیشنهادی این تاثیر در نظر گرفته شده است.

۲-۲-۳ مدل‌های مبتنی بر زنجیره‌های واژگانی[۱]

بکارگیری و استفاده از زنجیره‌های واژگانی نیز در سال‌های اخیر در بسیاری از حوزه‌های مرتبط با پردازش متن بویژه در ارزیابی انسجام مورد توجه قرار گرفته است. یکی از مهمترین استفاده‌های زنجیره‌های واژگانی استخراج واژگان کلیدی بوده که استخراج این واژگان نقش بسیار مهمی در ارزیابی

[۱] Lexical chains-based models

بخش‌های متفاوت متن و انسجام آنان دارد [۴۰]. این مدل‌ها در ابتدا یک پایگاه داده کوچک از واژگانی که با هم ارتباط معنایی داشته و در متن مورد پردازش موجود هستند ایجاد می‌کنند. سپس با استفاده از این زنجیره واژگان به تعیین ارتباط موضوعی، مفهومی و دنباله‌ای یک بخش از متن با بخش‌های قبلی و بعدی آن می‌پردازند. در یک مدل پیشنهاد شده توسط شانگ و دینگ از زنجیره‌های واژگانی برای ارزیابی انسجام متن خروجی در ترجمه آماری ماشینی استفاده شده است [۹]. در روش پیشنهادی نامبردگان ابتدا یک مجموعه از زنجیره‌های واژگانی در متن مبدا ایجاد شده و سپس با توجه به آن مجموعه زنجیره واژگانی متن مقصد ایجاد می‌شود. سپس با مقایسه این دو مجموعه مقدار انسجام متن مقصد در مقایسه با متن مبدا محاسبه می‌شود. در مدل پیشنهاد شده دیگری نیز سامسندرن و همکارانش از زنجیره‌های واژگانی برای اندازه‌گیری میزان انسجام و کیفیت مقالات استفاده کرده‌اند [۴۱]. در این تحقیق نامبردگان از دو مجموعه ویژگی‌های استخراجی زنجیره واژگانی استفاده کرده‌اند. اولین مجموعه ویژگی‌ها چگونگی آدرس دهی زمینه‌های موضعی و اجزای انسجامی در متن بوده و دومین مجموعه ویژگی چگونگی ارتباط این زمینه‌های موضعی با نشانه‌های صریح متغیرهای زبانی هستند. هتیت میت لین و همکاران در مدلی پیشنهادی از زنجیره‌های واژگانی برای ایجاد متن‌های خلاصه منسجم استفاده کرده‌اند [۴۲]. نوآوری روش نامبردگان در استفاده از ارتباط معنایی واژه‌های کلیدی و بکارگیری آنان در ایجاد یک زنجیره واژگانی کارآمد است. استخراج واژه‌های کلیدی از متن با استفاده از ویژگی جدید توزیع احتمال انتقال[۱] انجام شده و پس از ایجاد زنجیره واژگانی کارآمد از آن برای استخراج جملات کاملا مرتبط و ایجاد خلاصه‌ای منسجم استفاده شده است.

۲-۲-۴ مدل‌های مبتنی بر شبکه‌های عصبی[۲]

برای چیرگی بر کاستی‌های ویژگی‌های معنایی، رویکردهای نوین به سوی مدل‌هایی مبتنی بر شبکه‌های عصبی روی آوردند. این مدل‌ها راه حل‌های جدیدی جهت استخراج ویژگی‌ها پیشنهاد نموده و همچنین نمایشی ساده‌تر از مفهوم انسجام را به ارمغان آوردند. شبکه‌های بزرگ با حافظه کوتاه مدت[۳] یکی از موثرترین رویکردهای ارائه شده در این حوزه بود [۴۳]. در مدلی توسط لی و جورفسکی پیشنهاد شد و از دو مدل مجزا استفاده می‌کرد. در روش ارائه شده دیگری توسط لاگس واران و لی مدلی جهت تشخیص و ارزیابی انسجام و بهینه‌سازی مشکل ترتیب قرار گیری جملات پیشنهاد شد. نامبردگان برای این ارزیابی از شبکه‌های عصبی بازگشتی[۴] استفاده کردند [۴۴]. یادگیری عمیق نیز رویکردی نوین بوده که در سال‌های اخیر گام‌های بسیار موثری در ارزیابی انسجام محلی و عمومی متن برداشته است. نامبردگان در رویکرد دیگری با استفاده از شبکه‌های عصبی بازگشتی و یادگیری عمیق روشی را در

---

[۱] Transition probability distribution generator (TPDG)

[۲] Neural networks models

[۳] Long short-term memory networks (LSTM)

[۴] Recursive neural networks (RNN)

ارزیابی ترتیب صحیح جملات پیشنهاد داده‌اند [۴۵]. کلویی کیدون و همکاران با ارائه مدلی مبتنی بر چک لیست و شبکه‌های عصبی بازگشتی راه حلی برای تولید و ارزیابی انسجام عمومی متن پیشنهاد داده‌اند [۴۶]. در مدل دیگری معرفی شده توسط یون کیم از مدلی بدون ناظر و مبتنی بر مدل‌های زبانی شبکه‌های عصبی استفاده شده است. مدل آموزشی تک لایه استاندارد استفاده شده است. نامبرده با استفاده از تکنیک اعمال فیلترهای لایه‌ای یک مدل آموزشی تک لایه استاندارد[۱] برای آموزش بردار واژگان معرفی کرده است [۴۷]. مدل معرفی شده دیگری در این حوزه روش پیشنهاد شده توسط ژانگ و همکارش است [۴۸]. نامبردگان نیز از شبکه‌های عصبی استاندارد[۲] برای طبقه‌بندی متن استفاده کرده‌اند.

۲-۲-۵ مدل‌های مبتنی بر بردار واژگان[۳]

واژه‌ها می‌توانند به بردارهایی در یک فضای برداری نگاشت شوند. این نگاشت همان تعبیه کلمه بوده و یکی از حوزه‌های مورد توجه در اغلب رویکردهای پردازش متن است. به عبارتی ساده‌تر تعبیه کلمه معنی هر واژه را در ارتباط با سایر واژه‌های موجود در متن مشخص می‌کند. این تکنیک سعی در تشخیص ارتباطات و شباهت‌های معنایی واژه‌ها به هم داشته و تا به حال رویکردهای متفاوتی از آن نیز پیشنهاد شده است. اما آنچه بین همه آنها مشترک است این است که در این فضای برداری واژگان نزدیک به هم دارای معنی و مفهوم نزدیکی به هم هستند [٤٩]. ایده استفاده از سایر واژه‌های موجود در یک متن برای درک مفهوم واقعی هر واژه ابتدا توسط فریت در سال ۱۹۵۷ معرفی شد. نظریه معرفی شده مبتنی بر این واقعیت است که معنی واقعی یک واژه در ارتباط مستقیم با واژگان همسایه بوده و نمی‌توان درک صحیحی از مفهوم یک واژه بدون در نظر گرفتن سایر واژه‌ها موجود در همسایگی آن داشت. این همسایگی می‌تواند در حد سایر واژگان موجود در یک عبارت، جمله، پاراگراف و یا حتی کل متن باشد. در تکنیک‌های مبتنی بر تعبیه کلمه علاوه بر واژه‌ها، اطلاعات خارجی نیز نقش مهمی داشته و از آنها استفاده می‌شود. این اطلاعات در سایر واژه‌ها مجاور، سایر جملات و حتی بخش‌های دیگر متن پراکنده شده‌اند. به عبارتی ساده‌تر دانشی را که می‌تونان از خود واژه استخراج کرد فقط بخشی از اطلاعات موجود بوده و اغلب دارای ابهاماتی نیز هست [۵۰]. مثال‌های (۲-۱۱) نمونه‌هایی از بکارگیری بردارهای واژگانی هستند. در این مثال مشخص شده که بردار حاصل از تفریق دو بردار *woman* و *man* تقریبا برابر با بردار حاصل از تفریق دو بردار *aunt* و *uncle* است:

---

[۱] Canonical single layer training model

[۲] Convolutional neural networks

[۳] Word embeddings

$$man \rightarrow woman$$
$$uncle \rightarrow aunt \qquad (۲-۱۱)$$
$$vec(woman) - vec(man) \approx vec(aunt) - vec(uncle)$$

تا به حال رویکردهای زیادی از بردارهای واژگانی در حوزه‌های متفاوت پردازش متن استفاده کرده‌اند. شبکه‌های عصبی استاندارد در این زمینه پاسخ بسیار خوبی داده و در برخی از رویکردهای پیشنهاد شده اخیر مورد استفاده قرار گرفته‌اند [۵۱-۵۲]. شبکه‌های استاندارد در سایر حوزه‌های پردازش متن مانند خلاصه‌سازی، سیستم‌های پرسش و پاسخ و تشخیص موضوع هم بکار گرفته شده‌اند [۲٤، ۵۳-٥٤] یکی از مهمترین روش‌های ارائه شده تعبیه کلمه استفاده از روش‌های مبتنی بر انرژی است. روش word2vec از خانواده این روش‌ها بوده و در سال ۲۰۱۳ توسط تیم گوگل و میکولوف و همکارانش معرفی شده است [۴]. این روش با الهام گیری از مدل‌های مبتنی بر شبکه عصبی در پردازش متن ایجاد شده و یک شبکه عصبی دو لایه بوده که قادر به حدس و تشخیص مفهوم یک واژه با دقت بسیار بالا بر پایه حضورهای قبلی آن در متن است. هدف اصلی و مزیت word2vec گردآوری و کنار هم قرار دادن بردارهای واژگان شبیه به هم در یک فضای برداری است. این عمل موجب کشف شباهت‌ها با استفاده از ریاضیات و آمار می‌شود. روش نامبرده بردارهایی را بدون مداخله انسان ساخته که ویژگی‌های یک واژه را به صورت عددی در خود جای داده است. به عبارت ساده تر این مدل می‌تواند واژگان با مفاهیم نزدیک به هم را در یک فضای برداری در نزدیک هم قرار داده و فاصله بین آنان را مشخص کند[۱]. هر واژه در مدل نامبرده شده یاد می‌گیرد تا لگاریتم احتمال واژگان همسایه را بیشینه کند. در این معادله w_i و w_t واژگان همسایه T فاصله بین دو واژه در یک متن است [۴].

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{j \in nb(t)} \log p(w_i \mid w_t) \qquad (۲-۱۲)$$

پنینگتون و همکارانش نیز مدتی بعد مدل جدید اما متفاوت دیگری مبتنی بر بردارهای واژه با نام Glob2vec را ارائه کردند [۵۵]. در ادامه موضوع انسجام متن خروجی از دیدگاه حوزه‌های متفاوت پردازش زبان طبیعی مورد بررسی قرار می‌گیرد.

۲-۲-۶ رویکردهای ترکیبی

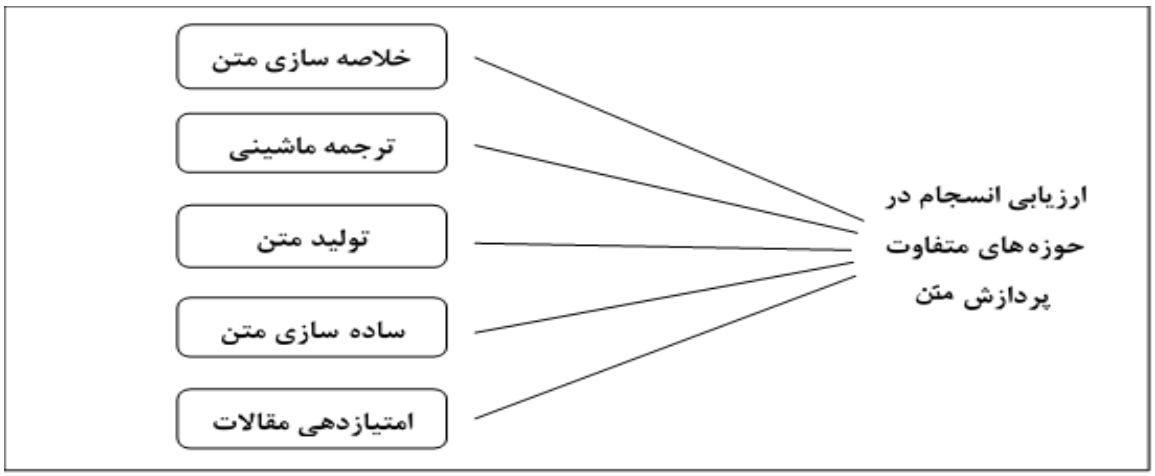
هر کدام از رویکردهای معرفی شده دارای مزایا و کاستی‌هایی بوده که بسیاری از پژوهشگران از روش‌های ترکیبی برای افزایش این مزایا و کاهش کاستی‌ها استفاده کرده‌اند. با توجه به اینکه روش مبتنی بر موجودیت یکی از بهترین مدل‌های ارائه شده برای ارزیابی محلی بوده اغلب روش‌های ترکیبی پیشنهاد شده این مدل را به عنوان روش پایه و اصلی خود انتخاب کرده و با ادغام آن با الگوریتمی تکمیلی سعی

[۱] https://code.google.com/p/word2vec/.

در کاهش و پوشش معایب آن داشته‌اند. از جمله روش‌های ترکیبی می‌توان به ترکیب آن با گراف‌های دوقسمتی [۲۰] [۳۶-۳۷]، شبکه‌های عصبی [۳۳-۳۵] ، رویکردی ترکیبی قابل استفاده در ترجمه ماشینی و خلاصه سای متن [۸]، رویکردهای ترکیبی شبکه موجودیت، گراف و آنتروپی [۳۸] و (مقاله استخراجی کنفرانسی ۹) اشاره کرد. در روشی معرفی شده توسط بورستین از ویژگی‌های معرفی شده در مدل مبتنی بر موجودیت و ترکیب آنان با برخی از مفاهیم زبان شناسی مانند غلط‌های گرامری و واژگانی مدل بهینه برای ارزیابی مقالات و نوشته‌های دانشجویی پیشنهاد شده است [۶].

۲-۳ دسته بندی رویکردهای ارزیابی انسجام متن بر اساس حوزه‌های متفاوت در پردازش متن

اغلب متن‌های غیر منسجم خروجی سیستم‌های پردازش متن هستند. از این رو در تمامی پژوهش‌های مرتبط با هر حوزه پردازش متن، بخش مهمی از تحقیق انجام شده ارزیابی و در صورت امکان بهبود انسجام و پیوستگی موضوعی متن خروجی خواهد بود. در این بخش به دسته بندی مدل‌های مهم معرفی شده ارزیابی انسجام بر اساس حوزه‌های متفاوت در پردازش متن پرداخته می‌شود (تصویر۲-۳). البته مسلم است که این بخش تمامی مدل‌های موجود را در بر نگرفته و فقط به آنهایی پرداخته می‌شود که از نظر این مطالعه مهم است.



شکل ۲-۳: ارزیابی انسجام بر اساس حوزه‌های متفاوت در پردازش متن

۲-۳-۱ رویکردهای مورد استفاده در خلاصه‌سازی متن

تولید خلاصه‌های منسجم، از همان ابتدای معرفی اولین رویکرد ماشینی خلاصه‌سازی متن توسط لون

در سال ۱۹۵۸ مورد توجه قرار گرفت [۲۹]. نامبرده از اطلاعات آماری قابل استخراج از متن مانند فرکانس واژگان در جملات متوالی برای انتخاب جملات مرتبط به هم استفاده کرده و خلاصه‌ای منسجم‌تر را تولید کرد. اغلب متن‌های غیر منسجم نیز خروجی‌های متون خلاصه شده توسط انسان یا ماشین هستند. در تمامی رویکردهای خلاصه‌سازی ماشینی در هر دو بخش استخراجی[۱] و چکیده‌ای[۲]، بخش‌هایی از متن حذف می‌شوند. از این رو کاهش انسجام در متون خلاصه شده اجتناب ناپذیر بوده، ارزیابی میزان انسجام و یا در صورت امکان بهبود آنان یکی از بزرگ‌ترین دغدغه‌های تمام رویکردهای خلاصه‌سازی متن است. بیشترین پژوهش‌های انجام شده بر روی ارزیابی و ایجاد متن‌های منسجم تولید شده نیز د این حوزه انجام شده است [۵۶-۵۷]. یکی از مشکلات بزرگ در خلاصه‌سازی ماشینی متن، بویژه در رویکرد استخراجی، از بین رفتن انسجام و پیوستگی متن خلاصه شده است. این مشکل در خلاصه‌سازی‌های چند سندی به مراتب بیشتر می‌شود. یکی از راهکارهای رفع این مشکل، تشخیص جملاتی است که موجب افزایش انسجام در متن می‌شوند و قرار دادن آنان در خلاصه خروجی است. این رویکرد دارای دو چالش است. نخست اینکه با چه معیاری این جملات تشخیص داده شوند و دوم اینکه با برگزیدن و اضافه کردن آنها به متن خروجی خلاصه تولیدی حجیم‌تر شده و موجب کمرنگ‌تر شدن مفهوم خلاصه‌سازی می‌شود. راهکار دیگر کشف الگوهای ارتباطی بین جملات منسجم و استفاده از آنان برای استخراج این جملات مهم است. مشکل دیگر خلاصه‌سازی استخراجی، انتخاب جملات بسیار شبیه به هم بوده که در عمل یک مفهوم را می‌رسانند [۵۸]. این چالش موجب تکرار و پرگویی یک مفهوم در جملات متعدد شده که باعث افزونگی یک مطلب خواهد شد و از طرفی دیگر به دلیل مشخص شدن طول خلاصه ایجاد شده توسط کاربر، مطالب مهم اما با درجه اهمیت کمتر در متن نادیده گرفته خواهند شد [۵۶] نادیده گرفتن این مطالب نیز موجب کاهش شدید انسجام در جملات خروجی خواهد شد.

یکی از راهکارهای بسیار کارا کشف الگوهای ارتباطی بین جملات منسجم با استفاده از الگوهای آماری و استفاده از آنان برای استخراج این جملات مهم است. در یکی از قدیمی‌ترین رویکردهای ارائه شده ار. بارزیلای و همکارش در سال ۲۰۰۱ با استفاده از شباهت‌های آماری و ابزارهای خوشه‌بندی رویکردی کارا برای ایجاد خلاصه‌های منسجم در رویکردهای خلاصه‌سازی چند سندی ارائه دادند [۵۹]. نامبردگان ابزار SIMFINDER را معرفی کردند که از ترکیب انتخاب ویژگی‌های زبانی و روش‌های خوشه‌بندی استفاده می‌کرد. این ابزار واحدهای متشابه متنی مانند جملات و پاراگراف‌ها را از متن‌های مورد پردازش انتخاب و یک خلاصه کلی از همه آنان ایجاد می‌کرد. در این ابزار از رگرسیون لگاریتم خطی[۳] برای تبدیل ویژگی‌های متشابه در بخش‌های انتخابی و تبدیل آنان به یک ویژگی استفاده شد.

$$R = e^{\eta} \Big/ 1 + e^{\eta} \qquad (۱۳-۲)$$

[۱] Extractive summarization

[۲] Abstractive summarization

[۳] Log-linear regression

در این عبارت η مجموعه وزن‌دار و R نتیجه نهایی است [۵۹]. در روش پیشنهادی نامبردگان از تکنیک خوشه‌بندی غیر سلسله مراتبی[۱] استفاده کرده‌اند. تکنیک بهینه‌سازی معرفی شده سعی در کمینه کردن تابع ϕ داشته که در آن عدم شباهت داده‌های موجود در یک خوشه اندازه‌گیری می‌شود [۵۹].

$$\Phi(\rho) = \sum_{i=1}^{k} \left(\frac{1}{|Ci|} \sum_{x,y \in Ci \& x \neq y} d(x,y) \right) \qquad (۱۴-۲)$$

$$\rho = \{C1, C2, \cdots, Ck\}$$

از رویکردهای اولیه دیگر مدل معرفی شده توسط ال. آلونسو و ام. فیونتز (۲۰۰۳) است. نامبردگان مدلی ترکیبی را ارائه دادند که همزمان انسجام و پیوستگی موضوعی متن را اندازه‌گیری می‌کرد [۶۰]. مدل آنان از زنجیره‌های واژگانی و ساختارهای بلاغی[۲] برای تشخیص جملات مرتبط و تولید خلاصه منسجم استفاده می‌کرد. جی کریستن سن و اس سودرلند با استفاده از گراف‌هایی ویژه با نام گراف‌های G-FLOW رویکردی جدید را برای تولید خلاصه‌های منسجم در خلاصه‌سازی چند سندی معرفی کردند [۵۶]. روش نامبردگان ترکیبی از انتخاب زیر مجموعه‌هایی از جملات مهم و مرتبط در متن‌های متفاوت و رعایت ترتیب قرار گیری آنان در متون اصلی بود که منجر به تولید خلاصه‌هایی منسجم تر شد. در گراف تولید شده در این رویکرد w_{G+} معرف یال‌های مثبت، w_{G-} معرف یال‌های منفی، λ ضریب همبستگی برای وزن مثبت و منفی و x_i و x_{i+1} جملات متوالی در رئوس متوالی هستند [۵۶]. مقدار وزن صفر یک یال مشخص‌کننده دو جمله غیر مرتبط است.

$$Coh(X) = \sum_{i=1..|X|-1} \left(w_{G+}(xi, xi+1) + \lambda w_{G-}(xi, xi+1) \right) \qquad (۱۵-۲)$$

ار. ژانگ و دبلیو. لی (۲۰۱۵) در رویکردی از رویکردهای شناختی برای تولید خلاصه‌هایی منسجم از متن‌های روایی استفاده کردند [۵۷]. در روش پیشنهادی نامبردگان الگوریتم معرفی شده به دنبال ارتباط بین هر جمله و جمله مرتبط قبلی آن است. روش نامبردگان از سه ویژگی بکار گرفته شده توسط حافظه انسانی یعنی حافظه بلند مدت[۳]، حافظه کاری[۴] و حافظه بخش بندی[۵] شده استفاده می‌کند.

در یکی از مطالعات انجام شده در ایجاد خلاصه‌های چند سندی تحقیق ام. انجانیول و همکارانش است [۶۱]. نامبردگان با ترکیب و دنبال هم قرار دادن متن‌های موجود و تبدیل آنان به یک متن ابتدا با بکارگیری ویژگی‌های نحوی و معنایی جملات آن را امتیاز دهی کرده و سپس به ترتیب امتیاز آنان را

---

[۱] Non-hierarchical clustering technique

[۲] Rhetorical structure

[۳] Long term memory

[۴] Working memory

[۵] Episodic memory

مرتب کردند. در مرحله بعد با توجه به درصد خلاصه‌سازی جملات با امتیاز بالاتر را استخراج و در نهایت برای بهبود انسجام خلاصه تولید شده ترتیب قرار گیری جملات را برابر با ترتیب حضور جمله در متن اصلی خودش کردند. ویژگی‌های نحوی مورد استفاده بزرگترین زیردنباله مشترک[1] (۲-۱۶)، مدل‌های زبانی مشترک[2] (۲-۱۷) و شباهت در ترتیب قرار گیری لغات[3] بوده و برای استخراج ویژگی‌های معنایی از تبدیل جملات به بردارهای عددی[4] (۲-۱۸) استفاده نموده‌اند [۶۱].

$$Lcs\left(s1, s2\right) = len\left(lcws\left(s1, s2\right)\right) \Big/ \min\left(len\left(s1\right), len\left(s2\right)\right) \qquad (۲-۱۶)$$

$$J\left(X, Y\right) = \left|X \cap Y\right| \Big/ \left|X \cup Y\right| \qquad (۲-۱۷)$$

$$s\left(wi, wj\right) = e^{\alpha l} \times \left(e^{\beta h} - e^{-\beta h}\right) \Big/ \left(e^{\beta h} + e^{-\beta h}\right) \qquad (۲-۱۸)$$

در این عبارات $lcws$ بزرگترین زیر دنباله مشترک لغات در دو جمله، len طول جمله کوتاه‌تر، X و Y مجموعه انگرام‌های دو جمله، l کوتاه‌ترین فاصله بین دو لغت W_i و W_j در وردنت و آلفا و بتا نیز مقادیر ثابت هستند [۶۱]. می‌توان روش‌های معرفی شده در این حوزه را به سایر حوزه‌های پردازش متن مانند سیستم‌های پرسش و پاسخ، سیستم‌های مترجم و یا هر متن تولید شده با سایر روش‌ها تعمیم داده و استفاده کرد. در روشی جدید پتر جی لیو و محمد صالح مدلی برای تولید خلاصه‌های چند سندی از موضوعات موجود در ویکی پدیا ارائه داده‌اند [۶۲]. با وجود اینکه تولید خلاصه‌های دارای انسجام در رویکردهای چند سندی عملی مشکل و چالش برانگیز بوده، نامبردگان با بکارگیری ویژگی‌هایی مانند tf-idf, TextRank, SumBasic توانسته‌اند خروجی‌های منسجم‌تری از متن‌های موجود در ویکی پدیا تولید کنند.

۲-۳-۲ رویکردهای مورد استفاده در ترجمه آماری ماشینی

انسجام در ترجمه‌های انجام شده توسط یک سیستم ترجمه ماشینی نیز از اهمیت بالایی برخوردار است [۶۳]. تا به حال، تقریبا در تمامی روش‌های ارائه شده ترجمه ماشینی عمل ترجمه به صورت جمله به جمله انجام شده است. در این حالت ترجمه جملات به صورت مستقل از هم پذیرفته که در اغلب حالات منجر به ایجاد متنی غیر منسجم می‌شود. از این رو، در سال‌های اخیر ایجاد ترجمه‌هایی منسجم

[1] Longest common sequence (LCS)
[2] Common n-gram features
[3] Word order similarity
[4] Sent2Vec features

درتمامی حوزه‌های پژوهشی بسیار مورد توجه قرار گرفته و کارهای زیادی در این خصوص انجام شده است. اهمیت انسجام در بخش ترجمه ماشینی در مقایسه با سایر رویکردهای ذکر شده مانند خلاصه‌سازی متن بسیار بیشتر بوده و از بسیاری جهات مورد توجه است. این اهمیت از آن جهت است که در رویکردهای قبل عمل ارزیابی و یا ایجاد انسجام در یک زبان مورد بررسی بوده و ورودی و خروجی متن در سیستم یاد شده درگیر با یک زبان و یک ادبیات بودند. اما عملیات ترجمه ماشینی یک عمل دو زبانه بوده و انسجام یک متن در دو زبان مورد بررسی قرار گرفته و تاثیر هر کدام بر روی دیگری کاملا مشخص نیست [٦٤].

اچ. فوکس (۲۰۰۲) یکی از پیشگامان بهبود انسجام در متن‌های ترجمه شده است [۱۰]. نامبرده معتقد است اغلب متن‌های ترجمه شده از انسجام پایینی برخوردار هستند. تا قبل از معرفی روش وی تمامی رویکردهای ارائه شده در فاز IBM2 از جابجایی واژگان استفاده می‌کردند. اما او در این فاز از جابجایی عبارات استفاده کرده و نتیجه‌ای بسیار بهتر را دریافت کرد. از روش‌های معرفی شده موفق ارزیابی انسجام متن ترجمه شده رویکرد پیشنهادی ژانگ و همکارانش است. نامبردگان استفاده از زنجیره لغوی را برای ارزیابی یک ساختار منسجم لغوی در متن‌های ترجمه شده پیشنهاد کرده‌اند [۹]. فرض این نظریه بر این است که زنجیره‌های لغوی متن ترجمه شده به طور مستقیم با زنجیره‌های لغوی متن مبدا مطابقت دارد. این فرض کاملا معقولانه بوده، زیرا شرط نخست ترجمه یک متن امانتداری واژگانی، مفهومی و ساختاری آن است. در روش پیشنهادی ابتدا زنجیره‌های لغوی موجود در هر متن قبل از ترجمه بدست آمده و سپس با استفاده از مدل طبقه بندی ماکزیمم آنتروپی هر واژه موجود در زنجیره به معنی معادل خود در زبان مقصد ترجمه شده و در جایگاه خود با توجه به مدل انسجام لغوی زبان مبدا نگاشت می‌شود. نامبردگان در روشی دیگر نیز از مدل‌های مبتنی بر موضوع برای تعیین انسجام متن استفاده کرده‌اند [٦٥]. در رویکرد پیشنهادی ابتدا به شناسایی مفاهیم کلیدی موجود در متن پرداخته شده و روابط موجود بین مفاهیم کلیدی را در قالب یک مدل سلسله مراتبی ترسیم کرده‌اند. در نهایت با مقایسه متن مورد تحلیل با ساختار سلسله مراتبی ترسیم شده و میزان انطباق آن با این ساختار انسجام متن موجود ارزیابی شده است. در رویکردی دیگر اسمیت و همکارانش روشی مبتنی بر آموزش خطا را پیشنهاد کرده‌اند [٦٦]. نامبردگان خطاهای ممکن را به یک متن اعمال کرده و سپس با آموزش سیستم از آن برای ارزیابی انسجام متن‌های ترجمه شده خروجی استفاده کرده‌اند.
در برخی از رویکردهای ارائه شده از الگوریتم‌های EM و IBM معرفی شده در روش‌های موجود در ترجمه آماری ماشینی جهت ارزیابی انسجام متن استفاده شده است [۸]. این رویکردها بر این عقیده هستند که همانطور که در هنگام ترجمه برای یک واژه از زبان مبدا چندین واژه در زبان مقصد وجود داشته و پیشنهاد می‌شود، یک واژه می‌تواند موجب پیوند شدن یک جمله به چند جمله شده و الگوریتم جملات با احتمال بالاتر را انتخاب کند. لذا در ارزیابی انسجام نیز هر واژه موجب اتصال موضوع جمله به چندین واژه در سایر جملات خواهد شد که می‌توان محتمل‌ترین واژه و در نتیجه جمله مرتبط را یافت. همچنین با استفاده از مدل IBM۱ فرمولی ارائه شده که توسط آن مشخص می‌شود برخی از واژگاه

موجود در جمله S_{i+1} توسط برخی از واژگان موجود در جمله S_i ایجاد شده‌اند. مدل‌های مبتنی بر موضوع[1] نیز در ارزیابی انسجام متن بکار گرفته شده‌اند. در رویکردی معرفی شده توسط ژانگ از این نظریه برای سنجش انسجام ترجمه ماشینی استفاده شد [٦٧]. در این نظریه برای بررسی میزان پیوستگی و انسجام متن ابتدا مفاهیم کلیدی موجود در متن شناسایی و سپس روابط میان آنها در قالب یک مدل سلسله مراتبی ترسیم می‌گردد. سپس متن مورد تحلیل با ساختار ترسیم شده مقایسه و هر چه درجه انطباق بیشتر باشد میزان انسجام بیشتر است. روش‌های پیاده سازی پیرو این نظریه در گروه رویکردهای نحوی قرار می‌گیرند. انسجام لغوی[2] رویکرد دیگری است که در برخی از پژوهش‌ها بویژه در حوزه ترجمه ماشینی به آن پرداخته شده است. این رویکرد به مفهوم انسجام در سطح واژگان پرداخته و شامل ارتباطاتی از قبیل تکرار، هم آیی، تضاد، جزء به کل و شمول معنایی است. وانگ و همکارش نیز مدلی مبتنی بر انسجام لغوی برای ارزیابی انسجام یک متن ترجمه شده معرفی کرده‌اند [٦٨]. در برخی از مدل‌های پیشنهادی برای ارزیابی انسجام ترجمه خروجی از ترکیب سایر رویکردهای قبلی استفاده شده است. کارین سیم اسمیت و همکارانش با ترکیب سه روش شبکه موجودیت، معیارهای شباهت در شبکه‌های مبتنی بر گراف و مدل‌های مبتنی بر الگوهای نحوی مدلی جدید و کارا برای ارزیابی انسجام متن ترجمه شده ارائه کردند [٦٩]. اغلب روش‌های معرفی شده قبل بیشترین تلاش خود را برای بالا بردن کیفیت هر جمله ترجمه شده به تنهایی کرده و امکان تاثیر مفهوم سایر جملات قبل و بعد را نادیده گرفته‌اند. اچ. ژانگ و همکاران با معرفی روشی جدید و کارا راه حلی برای این کاستی یافته‌اند. نامبردگان پس از ترجمه هر جمله تاثیر آن را بر جملات ترجمه شده قبل بررسی کرده و با تعریف مدلی مبتنی بر گفتمان محتویات و پاداش مجدد به تغییر و بهینه‌سازی ترجمه‌های قبلی می‌پردازند [٧٠].

۲-۳-۳ رویکردهای مورد استفاده در تولید متن

تولید متن یکی از مهمترین حوزه‌های پژوهشی پردازش زبان طبیعی بوده و ارزیابی انسجام متن تولید شده نیز مهمترین هدف در این حوزه است. در سال‌های اخیر مطالعات زیادی در حوزه تولید ماشینی متن انجام شده که تلاش همه آنها بر ایجاد یک خروجی با کیفیت هر چه نزدیک‌تر به متن تولید شده توسط انسان بوده است. اما تولید متن توسط یک برنامه کامپیوتری دارای چالش‌هایی بوده که ایجاد متنی منسجم‌تر از مهمترین آنها است. از دیگر چالش‌ها می‌توان به محدودیت یک سیستم تولید متن به یک حوزه خاص و عدم قابلیت اعمال و گسترش در حوزه‌های دیگر اشاره کرد. سازندگان سیستم‌های تولید متن بسیار مایل هستند که راهی برای ارزیابی دقت متن تولید شده و وابستگی موضوعی آن یافته تا بتوانند سیستم خود را طوری آموزش دهند که در مسیر بهبود این پارامترها گام بردارند. ای. نات و

[1] Topic-based

[2] Lexical cohesion

ار دیل روشی با قاعده مبتنی بر مجموعه روابط عبارات نشانه[۱] ارائه کردند [۷۱]. نامبردگان همچنین مجموعه‌ای طبقه بندی شده بزرگ از این عبارات نشانه ایجاد کردند. کیدون و همکارانش با معرفی نوعی شبکه عصبی بازگشتی با نام مدل عصبی چک لیست[۲] روشی را برای ارزیابی عمومی متن‌های تولید شده ارائه داده‌اند [۵۵]. نامبردگان در این رویکرد از الگوریتم جستجوی پرتو[۳] برای انتخاب محتمل‌ترین دنباله مورد نظر در تولید متن استفاده کرده‌اند. این الگوریتم از سرعت و دقت بیشتری برای رمزنگاری شبکه‌های عصبی بازگشتی برخوردار است.

با توجه به اینکه متن‌های تولید شده توسط انسان از روانی متن و وابستگی موضوعی بیشتری با عنوان برخوردار است تی سی فریرا و همکارانش رویکردی را پیشنهاد کرده‌اند که برای تولید متن منسجم از الگوهای انسانی تولید متن تقلید می‌کند [۷۲]. در این روش از تنوع در استفاده از واژگان و ساختار بکارگیری آنان استفاده شده است. مدل معرفی شده از دو الگوریتم بیزین ساده[۴] و شبکه‌های عصبی بازگشتی[۵] استفاده برای تولید و انتخاب بخش‌های متن استفاده کرده است. در این مدل هر بخش از متن شامل یک دوتایی (x, y) بوده که x مجموعه ویژگی‌های مرتبط را مشخص کرده y بخش‌هایی که x به آنها اشاره می‌کند است. در روش فوق از معیار تفاوت جانسون شانون[۶] برای ارزیابی شباهت دو بخش y و Y استفاده شده است [۷۲].

$$JSD(y \| Y) = \tfrac{1}{2} D(y \| m) + \tfrac{1}{2} D(Y \| m) \qquad (۱۹-۲)$$

شبکه ای عصبی انتقالی[۷] نقش بسیار مهمی در تولید متن‌های مصنوعی منسجم ایفا کرده‌اند. وای ژانگ و همکارانش در رویکردی پیشنهادی این شبکه‌ها را برای تولید متن‌های مصنوعی استفاده بکار برده‌اند [۷۳]. در این روش از شبکه های بزرگ با حافظه کوتاه مدت[۸] به عنوان تولید کننده و شبکه‌های استاندارد[۹] برای تفکیک کننده استفاده شده است.

اغلب روش‌های تولید متن در قبل سعی در آموزش سیستم توسط مجموعه‌ای بزرگ از متون تولید انسان کرده‌اند که این یکی از مهمترین کاستی‌های رویکردهای قبل بود. زیرا در این روش‌ها هیچ الگویی برای بهینه‌سازی انسجام متن تولیدی در هنگام تولید وجود نداشت. دبلیو سونگ و همکاران با معرفی روشی که از دو تشخیص دهنده عصبی استفاده کرده است این کاستی را بهبود بخشیده‌اند [۷٤] مدل معرفی نامبردگان مدل آموزشی انتقاد منفی متوالی[۱۰] نام داشته که در تولید متن سیگنال‌هایی در سطح

---

۱ Cue phrases
۲ Neural checklist model
۳ Beam search
۴ Naïve Bayes
۵ Recurrent neural networks
۶ Jensen-Shannon divergence (JSD)
۷ Generative adversarial network (GAN)
۸ Long short- term memory (LSTM)
۹ Convolutional neural network
۱۰ Negative-critical sequence training

جمله و پاراگراف برگردانده که توسط این سیگنال‌ها انسجام درون جمله‌ای و درون پاراگرافی اندازه گرفته شده و با اجرای دوباره بهبود داده می‌شود.

۲-۳-۴ رویکردهای مورد استفاده در ساده‌سازی متن

شکل دیگر متن‌های غیر منسجم، خروجی متن‌های ساده شده است. ساده‌سازی نحوی متن یکی از حوزه‌های پردازش زبان طبیعی بوده که سعی در کاهش پیچیدگی گرامری و لغوی متن داشته به صورتی که اطلاعات و مفهوم اصلی متن به طور کامل حفظ شود [۷۵]. سیستم‌های ساده‌ساز چون سعی در قابل فهم کردن پیچیدگی‌های لغوی و گرامری برای استفاده افراد با دانش پایین‌تر در آن حوزه را دارند، به ندرت می‌توانند مفهوم نویسنده اصلی را برسانند. درک سلیقه‌ای، استنباط غلط و یا حذف بخش غیر قابل فهم برای شخص یا ماشین ساده‌ساز از جمله عواملی هستند که منجر به تولید یک متن ساختگی و غیر منسجم می‌شوند. هدف اصلی ماشین‌های ساده‌ساز افزایش قابلیت درک یک متن برای خواننده (شامل ساده‌سازی مطالب پیچیده گرامری، آماده سازی متن برای افراد با دانش پایین‌تر، تبدیل اصطلاحات فنی و کنایه‌ای به معادل‌های ساده‌تر، ...) و یا آماده سازی متن برای پردازش توسط یک برنامه کامپیوتری مانند سیستم‌های ترجمه ماشینی، خلاصه‌سازی، پرسش و پاسخ، ... (شامل کوتاه کردن یا تقسیم جملات بزرگ، یافتن معادل برای کنایه‌های ادبی،...) است [۷۶-۷۷].

اولین تلاش برای ساده سازی متن و ایجاد یک خروجی منسجم مربوط به مدل معرفی شده توسط ار. کاندرسکار و بی. سرینواس (۱۹۹۶) است [۷۸]. مدل درختی معرفی شده توسط نامبردگان از یک نمایش مبتنی بر قائده برای ترکیب عبارات و اطلاعات وابستگی بین آنان استفاده کرده است. ای. سیدارتان در رویکردی پیشنهادی چگونگی استفاده از تجزیه و تحلیل دقیق کم عمق[۱] برای ساده سازی نحوی متن و ایجاد یک خروجی منسجم را معرفی کرد [۷۹]. نامبرده با ایجاد یک مجموعه از قوانین ساده سازی دستی و تجزیه و تحلیل دقیق از جنبه‌های سطح گفتمان موفق به بازنویسی متن اولیه به زبانی ساده‌تر شد. مدل معرفی شده توسط وی به سه بخش کلی تجزیه و تحلیل، تغییر شکل واژه‌های مشکل به ساده و تولید متن جدید و منسجم تقسیم شده بود و مواردی مانند ترتیب حضور جملات در متن، انتخاب کلید واژه‌ها، ارجاع بین بخش‌های متن و انتخاب دقیق و تعیین کننده بخش‌های مختلف برای تولید یک خروجی منسجم استفاده شد.

اصطلاحات پزشکی در متن‌ها و مقالات مربوط به سلامت و درمان یکی از بزرگ‌ترین مشکلات سد راه عموم برای استفاده از این متون و مقالات است. تا به حال راه حل‌هایی برای ساده سازی این متون ارائه شده و افرادی با ساده سازی دستی این عمل را انجام داده‌اند. اما ایجاد رویکردی ماشینی برای تولید این متون به صورت ساده و منسجم خواسته بسیاری از محققین این حوزه بوده است. جی. لروی و همکارش روشی مبتنی بر ساده سازی واژگانی این متون ارائه دادند [۸۰]. اس. ما و ایکس. سان با

[۱] Shallow robust analysis

بکارگیری دو رویکرد ساده سازی و خلاصه‌سازی متن روشی ترکیبی برای ایجاد خلاصه‌هایی منسجم و ساده شده در زبان چینی معرفی کرده‌اند [۷۶]. هدف نامبردگان بهبود ارتباط معنایی بین متن اصلی و خلاصه ساده شده تولیدی است. آنان برای این منظور روشی با نام ارتباط معنایی مبتنی بر شبکه‌های عصبی[۱] را پیشنهاد داده‌اند. با توجه به اینکه اغلب رویکردهای ارزیابی انسجام متن با متن‌های بزرگ مشکل دارند نامبردگان از رویکردی با نام واحد انکدر خود توجه[۲] برای ایجاد حافظه در متن ورودی استفاده کرده‌اند. رویکرد واحد انکدر سعی در اندازه‌گیری اهمیت کلمه و میزان اطلاعاتی را که به آن بخش از متن ارائه می‌کند دارد. در این مرحله هر کلمه x_t به سلول LSTM تزریق شده که خروجی آن بردار h_t است. در این عبارت f تابع LSTM است [۷۶].

$$h_t = f\left(x_t, h_{t-1}\right) \quad (۲۰-۲)$$

ال برکن و همکاران نیز رویکردی جدید را برای مقابله با متون مشکل و فنی پزشکی ارائه داده‌اند [۸۱]. نامبردگان با ایجاد یک پایگاه داده از متون ساده شده و واژگان عمومی در علوم پزشکی راه حل ساده اما بسیار کارا برای حل مشکل بکارگیری واژگان بسیار تخصصی در متون پزشکی ارائه داده تا درک این متون را برای افراد عادی فراهم کنند. پایگاه داده ایجاد شده شامل جملاتی ساده تولید شده توسط افراد خبره بوده که توسط یک مترجم ماشینی مبتنی بر شبکه‌های عصبی به زبان مقصد ترجمه شده است.

۲-۳-۵ رویکردهای مورد استفاده در امتیازدهی خودکار مقالات

مهمترین متن‌های غیر منسجم متن‌های ترکیبی هستند. منظور از متن‌های ترکیبی نوشته‌هایی هستند که کل آن نگارش توسط یک فرد انجام نشده و بخش‌های مختلف آن از منابع مختلف جمع‌آوری و در کنار هم قرار گرفته‌اند. بخش‌های مختلف اینگونه متن‌ها شاید از نظر موضوعی به هم مرتبط باشند، اما خواننده را دچار سردرگمی در درک مفهوم می‌کنند. با توجه به اینکه هر نویسنده از ادبیات لغوی و گرامری منحصر به فردی برای بیان نظریه خود استفاده می‌کند، می‌توان با بررسی این نوشته‌ها به وضوح به مفهوم عدم انسجام پی برد. بیشترین تولید این گونه متن‌ها خروجی نوشته‌هایی است که بخش‌های مختلف آن از منابع مختلف و با سلیقه‌های نگارشی متفاوت گردآوری شده‌اند. متن‌های علمی و ادبی دریافت شده از افراد تحت عنوان مقاله یا تالیف از نمونه‌های مشخص متون ترکیبی هستند. چون یکی از موارد مهم برای پذیرش یک نگارش علمی یا پژوهشی انسجام و یکپارچگی متن نگارش شده و بازنویسی تمام بخش‌های آن توسط یک نگارنده بوده، امتیاز دهی به محوریت موضوع و انسجام جملات موجود در یک متن علمی و ادبی از مهمترین شاخه‌های این حوزه است.

---

[۱] Semantic relevance based neural network

[۲] Self-gated attention encoder

وی سونگ و همکاران در پژوهشی با بکارگیری ویژگی‌های ساده متنی رویکردی ساده اما کارا را در ارزیابی وابستگی موضوعی و انسجام مقالات دانشجویی ارائه داده‌اند [۸۲]. روش پیشنهادی آنان در عین سادگی ارزیابی انسجام مقالات چینی را در هر دو حوزه محلی و عمومی انجام داده است. ویژگی‌های ساده‌ای که نامبردگان از آن استفاده کرده‌اند عبارت از جایگاه جملات در پاراگراف (ابتدا، میانه، آخر)، واژگان و عبارات نشانه، زنجیره‌های واژگانی، ویژگی‌های ساختاری متن مانند تعداد عبارات موجود در جمله و پاراگراف، و وابستگی مقاله به عنوان آن با معیار شباهت کسینوسی است. جی هوانگ و همکاران رویکردی کارا مبتنی بر مدل شبکه موجودیت برای ارزیابی انسجام مقالات دانشجویی به زبان چینی ارائه داده‌اند [۸۳]. در این رویکرد از تکرار واژگان و ویژگی همرخدادی آنان برای تشخیص میزان شباهت جملات متوالی استفاده شده است. در روش پیشنهادی نامبردگان ابتدا یک ماژول همرخدادی با عملکردی بالا ایجاد کرده و سپس آن را توسط پیکره COLEN آموزش دادند. در معادله (۲-۲۱) انسجام متن مورد بررسی محاسبه شده که در آن T نشان دهنده متن، c_i زنجیره و S_i جملات موجود در متن هستند [۸۳].

$$P\,coherence(T) = P(C_i\ , S_i) \qquad (۲۱-۲)$$

با توجه به اینکه تشخیص صحیح انسجام در متون با طول بیشتر ممکن است کاهش بیابد از یک معیار نرمال (بین صفر و یک) متناسب با طول متن *Link Score* استفاده می‌شود (۲-۲۲). در این معادله $r_{i,j}$ نقش گرامری موجودیت زنجیره j در جمله S_i است. مقدار *Max(n)* و *Min(n)* کمترین و بیشترین امتیاز موجودیت زنجیره با طول n در مجموعه آموزشی است [۸۳].

$$LinkScore = \log p\left(r\,i,j \mid r_{(i-h),j} \mid \cdots \mid r_{(i-1),j}\right) - Min(n) \Big/ Max(n) \qquad (۲۲-۲)$$

در نهایت انسجام متن مورد بررسی ارزیابی شده (۲-۲۳) که در آن m تعداد زنجیره‌های همرخدادی، n تعداد کلمات جمله S_i و *weight* تعداد موجودیت های متفاوت در یک زنجیره همرخدادی است [۸۳].

$$P\,coherence(T) = 1/m \sum_{i=1}^{m}\left(LinkScore + \frac{weight}{n} \times LinkScore\right) \qquad (۲۳-۲)$$

۲-۳-۶ ارزیابی همزمان انسجام محلی و عمومی

تا به حال تعداد کمی از رویکردهای ارائه شده اقدام به ارزیابی همزمان انسجام محلی و عمومی کرده‌اند. هر چند این روش‌ها در بخش ارزیابی محلی از دقت قابل قبولی برخوردار بودند، اما در بخش ارزیابی عمومی ضعیف عمل کرده و نتوانسته‌اند دقت خوبی را ارائه دهند. مهمترین چالش این روش‌ها مواجهه با متن‌های بزرگ و تعداد جملات زیاد است. به دلیل بزرگی متن و تعداد جملات زیاد، این رویکردها

نمی‌توانند با تکیه بر مفاهیم معنایی موجود در جملات متوالی انسجام عمومی را در سطح تمام متن و با دقت بالا تشخیص داده و ارزیابی کنند. روش‌های مبنی بر گراف جزو اولین روش‌هایی بودند که سعی در ارزیابی وابستگی مفهومی جملات موجود در متن با فاصله‌های بیشتر کردند . این روش‌ها با ترکیب ویژگی‌های گراف‌ها با سایر روش‌ها مانند روش مبنی بر موجودیت دامنه مقایسه را به محدوده‌ای بزرگ‌تر از جملات همسایه و حتی کل متن ببرند. اما بزرگترین مشکل آنان در ارتباط با متن‌های بزرگ است [۲۰][۳۶-۳۷] [۳۹].

ارزیابی همزمان محلی و عمومی انسجام با استفاده از ویژگی‌های بسیار ساده توسط وی سونگ و همکاران نیز انجام شده است. نامبردگان با ارائه رویکردی ساده اما کارا ارزیابی وابستگی موضوعی و انسجام مقالات دانشجویی در زبان چینی را در هر دو حوزه انجام داده‌اند [۸۲]. لویس و ننکوا (۲۰۱۲) با معرفی رویکردی ترکیبی روشی را برای ارزیابی همزمان انسجام محلی و عمومی متن پیشنهاد داده‌اند [۸۴]. نامبردگان برای تحلیل و ارزیابی پیوستگی جملات متوالی مانند رویکردهای پیشین از احتمال جفت آیتم‌های موجود در جملات متوالی استفاده کرده (۲-۲۴) و برای ارزیابی انسجام عمومی مدل پنهان مارکو را بکار گرفته‌اند (۲-۲۵) [۸۲].

$$P(T) = \prod_{i=2}^{n} \prod_{i=1}^{|Si|} \Big/ {}_{|Si-1|} \sum_{k=1}^{|Si-1|} p\left(S_i^j \mid S_{i-1}^k\right) \qquad (۲-۲۴)$$

$$pm\left(h_j \mid h_i\right) = d\left(hi,hj\right)+\partial m \Big/ d(hi)+\partial m*C \qquad (۲-۲۵)$$

در مرحله ارزیابی عمومی با خوشه‌بندی جملات متن و قرار دادن جملات وابسته در خوشه‌های متفاوت احتمال گذر هر بخش از متن را به بخش بعد محاسبه نموده‌اند. در معادلات بکارگرفته شده T متن مورد ارزیابی، S_i جملات موجود در متن، h_i حالات موجود در مدل پنهان مارکو و pm احتمال گذر از یک حالت به حالت بعدی است. $d(h_i)$ تعداد متن‌هایی که جمله h_i در آن تکرار شده و $d(h_i, h_j)$ تعداد متن‌هایی است که جمله h_i بلافاصله قبل از جمله h_j در آن قرار گرفته است [۸۲]. یکی از مشکلات رویکردهای مبتنی بر شبکه موجودیت عدم امکان ارزیابی وابستگی جملات با فاصله زیاد و مشکل دیگر رویکردهای مبتنی بر گراف درگیر شدن با پیچیدگی‌های موجود در گراف‌های بزرگ و پیوندهای زیاد بین نودهای آنان است. از این رو این رویکردها توانایی ارزیابی انسجام عمومی و محلی را به طور همزمان نداشته و برای این کار باید از ترکیب آنان با سایر الگوریتم ها سود برد. در روش ترکیبی دیگری ما از مزایای سه رویکرد مبتنی بر موجودیت، گراف و آنتروپی استفاده کرده‌ایم تا کمبودها و مشکلات روش‌های قبلی بهبود داده شوند [مقاله کنفرانسی استخراجی ۱۰]. اسکات ای کروسلی و همکاران در سال ۲۰۱۶ ابزاری برای ارزیابی همزمان انسجام عمومی و محلی ارائه دادند [۳]. این ابزار TAACO نامیده شد و قابلیت اجرا بر روی اغلب سیستم‌های عامل را داشته است. ابزار نامبرده قابلیت ارزیابی انسجام در سطح محلی (جملات متوالی)، سطح عمومی (پاراگراف) و سطح کل متن به طور همزمان را

داشته است. این ابزار از پنج ویژگی اتصالات لغوی و معنایی[۱] واژگان در جملات متوالی برای ارزیابی انسجام محلی، همپوشانی واژگانی[۲] و همپوشانی معنایی[۳] برای ارزیابی محلی (جملات متوالی) و ارزیابی عمومی (سطح پاراگراف)، نوع نسبت نشانه‌ها به هم[۴] برای ارزیابی تمام متن و اطلاعات مرتبط با بخش‌های قبل متن[۵] (مانند ضمایر فاعلی و مفعولی) برای ارزیابی انسجام کل متن استفاده کرده است.

در روش پیشنهادی دیگری ما با ارتقای انسجام محلی از سطح جملات متوالی به سطح پاراگراف و ارزیابی انسجام عمومی با مقایسه وابستگی مفهومی پاراگراف‌های متوالی رویکردی ترکیبی را ارائه دادیم [مقاله ژورنالی استخراجی ۱]. این روش بدون توجه به معنی واژگان و ویژگی‌های انسجامی معرفی شده در رویکردهای معنایی قبل به طور همزمان انسجام محلی و عمومی را ارزیابی می‌کند. روش نامبرده با بکارگیری بردارهای واژگانی word2vec گوگل و تبدیل جملات به ماتریس‌های نرمال عددی ارزیابی وابستگی بخش‌های متفاوت متن را از مفاهیم معنایی واژگان به وابستگی‌های ماتریسی تبدیل کرده که نتیجه‌ای بهتر را حاصل کرده است. روش معرفی شده بر روی متن‌های بزرگ و با تعداد جملات بالا از دقت بالاتری برخوردار است.

۲-۴ تعبیه کلمه[۶] و الگوریتم word2vec

واژه‌ها می‌توانند به بردارهایی در یک فضای برداری نگاشت شوند. این نگاشت همان تعبیه کلمه یا تعبیه کلمه (WE) بوده و یکی از حوزه‌های مورد توجه در اغلب رویکردهای پردازش متن است. ایده استفاده از سایر واژه‌های موجود در یک متن برای درک مفهوم واقعی هر واژه ابتدا توسط فریت در سال ۱۹۵۷ معرفی شد. نامبرده معتقد بود که معنی واقعی یک واژه در ارتباط مستقیم با واژگان همسایه بوده و نمی‌توان درک صحیحی از مفهوم یک واژه بدون در نظر گرفتن سایر واژه‌ها موجود در همسایگی آن داشت [۴۹]. این همسایگی می‌تواند در حد سایر واژگان موجود در یک عبارت، جمله، پاراگراف و یا حتی کل متن باشد. در تکنیک‌های مبتنی بر WE علاوه بر واژه‌ها، اطلاعات خارجی نیز نقش مهمی داشته و از آنها استفاده می‌شود. این اطلاعات در سایر واژه‌ها مجاور، سایر جملات و حتی بخش‌های دیگر متن پراکنده شده‌اند. به عبارتی ساده‌تر دانشی که از خود واژه می‌توان استخراج کرد فقط بخشی از اطلاعات موجود بوده و اغلب دارای ابهاماتی نیز هست [۵۰]. قبل از بکار گیری تعبیه کلمه بیشتر الگوریتم‌های پردازش متن اغلب از متدهای موجود و معرفی شده در حوزه پردازش صوت و پردازش تصویر استفاده می‌کردند. یک متن دارای ویژگی‌های متفاوتی در مقایسه با سایر حوزه‌ها مانند گفتار و تصویر است. به

---

[۱] Words connectives

[۲] Lexical overlap

[۳] Semantic overlap

[۴] Type–token ratio

[۵] Givenness

[۶] Word embeddings

عنوان مثال یکی از مهمترین مشکلات در حوزه پردازش گفتار و پردازش تصویر کشف، کاهش و یا حذف نویز بوده، در حالی که بزرگ‌ترین چالش در حوزه پردازش متن اطلاعات گمشده و ابهامات معنایی موجود در متون است. یک سیستم پردازش بیشتر به اطلاعات نهفته در خود تصویر تکیه داشته و کمتر نیازمند بررسی دانش موجود در پس زمینه و یا اطلاعات خارجی خواهد بود، در حالی که در پردازش متن اطلاعات خارجی پس زمینه و دانش خارجی موجود در آن بسیار بیشتر در تشخیص برخی از ابهامات موجود کمک کننده هستند. به عنوان مثال در تشخیص مفهوم یک جمله سایر جملات متن و یا حتی متن‌های دیگر بسیار کمک کننده هستند. اغلب روش‌های پیشنهاد شده تا قبل از بکارگیری WE به ویژگی‌های خاص موجود در متن توجه نداشته و آنرا از دیدگاه سایر رویکردها مانند پردازش گفتار و تصویر مورد بررسی قرار داده‌اند. واژه‌ها و ارتباطات آنان در معادله (۲-۲۶) مفهوم این تکنیک را به وضوح مشخص می‌کند:

$$\begin{aligned}
&\text{man} \longrightarrow \text{woman}\\
&\text{uncle} \longrightarrow \text{aunt}\\
&\text{king} \longrightarrow \text{queen}\\
&\text{vec (woman)} - \text{vec (man)} \simeq \text{vec (aunt)} - \text{vec (uncle)}\\
&\text{vec (woman)} - \text{vec (man)} \simeq \text{vec (queen)} - \text{vec (king)}\\
&\text{vec (Berlin)} - \text{vec (Germany)} + \text{vec (France)} \simeq \text{vec (Pairs)}\\
&\text{vec (Einstein)} - \text{vec (scientist)} + \text{vec (Picasso)} \simeq \text{vec (painter)}
\end{aligned} \qquad (۲-۲۶)$$

تقریبا اکثر الگوریتم‌های پیشنهاد شده در حوزه پردازش آماری متن[۱] واژه را به عنوان واحد متنی در نظر گرفته‌اند. استفاده از واژه به عنوان واحد متن موجب تولید ماتریس تنک[۲] با درایه‌های صفر و یک شده که درایه‌های متناظر با یک به معنای موقعیت واژه مورد نظر در جمله است. ماتریس ایجاد شده حاوی اطلاعات جامعی از واژه نبوده و نمی‌تواند واژگان با شکل نوشتاری مشابه و معنی متفاوت و همچنین با شکل نوشتاری متفاوت ولی مفهوم مشابه را از هم تشخیص دهند. در اغلب موارد واژگان با شکل نگارشی مشابه در جایگاه‌های متفاوت در جمله مفهومی متفاوت را ارائه داده ولی در مدل فوق یکسان فرض می‌شوند. به عنوان مثال تشخیص تمام شباهت‌ها و تفاوت‌های دو واژه hotel و motel در این ماتریس امکان پذیر نیست. یکی از مهمترین روش‌های ارائه شده WE استفاده از روش‌های مبتنی بر انرژی[۳] است. روش word2vec از خانواده این روش‌ها بوده و در سال ۲۰۱۳ توسط تیم گوگل و میکولوف و همکارانش معرفی شده است [۴]. این روش با الهام‌گیری از مدل‌های مبتنی بر شبکه عصبی در پردازش متن ایجاد شده و یک شبکه عصبی دو لایه بوده که قادر به حدس و تشخیص مفهوم یک واژه با دقت بسیار بالا بر پایه حضورهای قبلی آن در متن است. هدف اصلی و مزیت word2vec گردآوری و کنار هم قرار دادن بردارهای واژگان شبیه به هم در یک فضای برداری است. این عمل موجب کشف شباهت‌ها با استفاده از ریاضی و آمار می‌شود. روش نامبرده بردارهایی را بدون مداخله انسان ساخته که

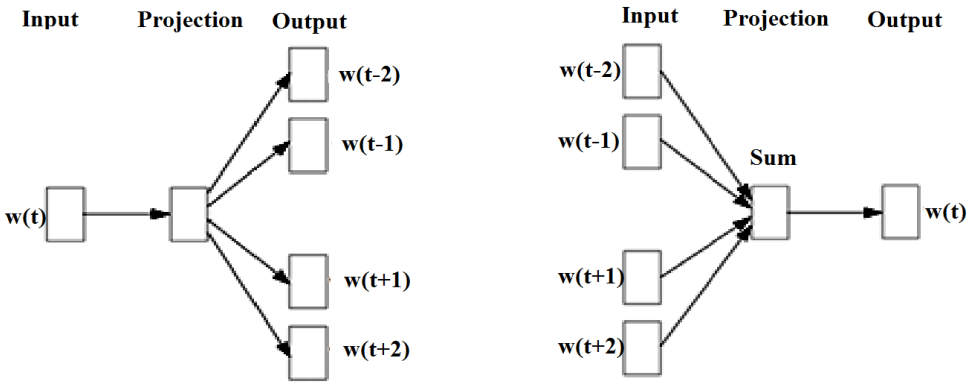
---

[۱] Text processing algorithms

[۲] Sparse matrix

[۳] Energy-based models

ویژگی‌های یک واژه را به صورت عددی در خود جای داده است. به عبارت ساده تر این مدل می‌تواند واژگان با مفاهیم نزدیک به هم را در یک فضای برداری در نزدیک هم قرار داده و فاصله بین آنان را مشخص کند. این عمل با اعمال مدل یادگیری و تکنیک‌های استانداردی مانند skip-gram در دیتابیس‌های بزرگی شامل میلیون‌ها واژه انجام می‌شود. در WE مبتنی بر شبکه عصبی[۱] هر واژه را با برداری از اعداد نشان داده که معنی آن را در ارتباط با سایر واژگان همسایه‌اش در متن ورودی نشان می‌دهد. این عمل به دو صورت پیش‌بینی واژه هدف با توجه مجموعه بزرگ واژگان[۲] و پیش‌بینی عبارت با توجه به واژه حذف شده[۳] انجام می‌شود. از لحاظ الگوریتمی این دو روش شبیه هم هستند با این تفاوت که CBOW واژه‌های هدف را از روی واژگان موجود در متن ورودی پیش‌بینی می‌کند ولی SGNS به صورت برعکس از روی واژه‌های مرجوعه هدف، واژگان ورودی را پیش‌بینی می‌کند. هر واژه در مدل نامبرده شده یاد می‌گیرد تا لگاریتم احتمال واژگان همسایه را بیشینه کند. پنینگتون و همکارانش نیز مدتی بعد مدل جدید اما متفاوت دیگری مبتنی بر بردارهای واژه با نام Glob2Vec را ارائه کردند [۵۵]. روش Word2vec عمل تبدیل واژه را به بردار اعداد با توجه به واژه حذف شده انجام می‌دهد. تصویر (۲-۴) این دو مدل را نشان می‌دهد:



شکل ۲-۴: Continuous Bag Of Words (راست) ، Skip-Gram Negative Sampling (چپ) [۳۳]

۲-۵ نتیجه‌گیری

مهمترین پایه‌گذار نظریه انسجام به صورت تئوری دو محقق هالیدی و حسن بوده‌اند که مفاهیم عمده انسجام متن و پارامترهای مهم قابل استخراج آن را معرفی کرده‌اند. اما تمام پژوهش‌های نامبردگان در حوزه زبان شناسی بوده و هیچگونه الگویی ریاضی یا کامپیوتری برای آن معرفی نکرده‌اند. انسجام متن با نخستین مطالعات در حوزه خلاصه‌سازی متن توسط لون در سال ۱۹۵۸ مورد توجه قرار گرفت [۲۹] و

[۱] Neural word embeddings
[۲] Continuous bag of words (CBOW)
[۳] Skip-gram negative sampling (SGNS)

بعدها فولتز و همکارانش در سال ۱۹۹۸ روشی مبتنی بر الگوریتم‌های ماشینی را برای ارزیابی انسجام یک متن پشنهاد دادند [۳۰]. مهمترین رویکردهای کامپیوتری برای ارزیابی انسجام روش مبتنی بر موجودیت بارزیلای و لاپاتا بوده که مفاهیم نخستین آن توسط بارزیلای و لاپاتا در سال ۲۰۰۵ معرفی شد و نامبردگان در گزارش تکنیکی خود در سال ۲۰۰۸ آنرا تکمیل کرده‌اند [۲۲]. از آن تاریخ به بعد اکثر روش‌های پیشنهادی از ویژگی‌های معرفی شده توسط آنان استفاده کرده‌اند. این رویکرد دارای کاستی‌هایی بود که در ابتدا برای بهبود آن از روش‌های ترکیبی با سایر حوزه‌ها استفاده شده و سپس اغلب پژوهشگران به سوی رویکردهای آماری روی آوردند.

بیشترین پژوهش‌های انجام شده ارزیابی متون منسجم بر روی متن‌های تولید شده توسط سیستم‌های خلاصه‌سازی انجام شده است. در سال‌های اخیر استفاده از بردارهای واژگان و الگوریتم word2vec بسیار مورد توجه قرار گرفته است. اغلب روش‌های پیشنهادی متمرکز بر روی یکی از دو حوزه ارزیابی متن (محلی یا عمومی) بوده و رویکردهای بسیار کمی به ارزیابی همزمان در هر دو حوزه پرداخته‌اند. مهمترین روش‌های معرفی شده که به طور همزمان در ارزیابی محلی و عمومی گام برداشته‌اند رویکردهای مبتنی بر گراف بوده، که باز هم در حوزه عمومی بسیار موفق نبوده، درگیر مفاهیم و پیچیدگی‌های الگوریتم‌های مبتنی بر تئوری گراف شده و در متن‌های بزرگ کارایی خود را از دست داده‌اند. جدول‌های (۲-۱) و (۲-۲) خلاصه از مقایسه رویکردهای معرفی شده را به تصویر می‌کشد. روش ارائه شده در این رساله با استفاده از بردارهای واژگان word2vec ، تبدیل جملات به ماتریس و استخراج ویژگی‌های آماری از این ماتریس‌ها انسجام محلی و عمومی یک متن را به سادگی ارزیابی کرده است. روش ارائه شده وابستگی زیاد به موضوع متن، معنی واژه‌ها نداشته و در متن‌های بزرگ نتیجه‌ای مطلوب‌تری را ارائه می‌دهد.

جدول ۲-۱: مقایسه رویکردهای مهم ارزیابی انسجام متن

رویکردهای ارزیابی انسجام	مزایا	کاستی‌ها
شبکه موجودیت	- مهمترین و پرکاربردترین رویکرد معرفی شده - بکارگیری نقش گرامری اسم‌ها و عبارات اسمی - سرعت بالا و الگوریتم ساده	- تکرار شکل کامل اسم‌ها و عبارات اسمی - ارزیابی انسجام فقط در حوزه محلی - وابستگی شدید به مفاهیم معنایی - مناسب متن‌های کوتاه و متوسط - فقط ارزیابی انسجام جملات مجاور
مبتنی بر گراف	- ارزیابی انسجام در جملات با فاصله بیشتر - اندازه وابستگی دو جمله با توجه به وزن یال - وزن دهی یال متناسب با نقش گرامری عامل انسجامی - بسیاربهینه در ترکیب با رویکردهای مبتنی بر موجودیت	- نیاز به نگاشت گراف در یک ماتریس - پیچیدگی گراف در متن های بزرگ - مناسب برای ارزیابی محلی و متن‌های کوتاه
زنجیره‌های واژگانی	- استفاده از واژگان بیشتر با نقش‌های گرامری متفاوت	- نیاز به تولید پایگاه داده واژگان
شبکه‌های عصبی	- بکارگیری رویکردهای هوشمند - نزدیک بودن به شکل ارزیابی انسانی - معرفی رویکردهایی کارا مانند یادگیری عمیق	- پیچیدگی بیشتر
بردار واژگان	- رویکردی جدید - امکان ارزیابی انسجام عمومی با دقت بیشتر	- پبچیدگی الگوریتم اولیه تبدیل واژگان به بردار

جدول ۲-۲: ارزیابی انسجام خروجی در حوزه‌های متفاوت پردازش متن

حوزه‌های متفاوت پردازش متن	ویژگی‌ها	برخی از ابزارهای مورد استفاده
خلاصه‌سازی متن	- اولین رویکرد درگیر با ارزیابی انسجام متن - مهمترین چالش در خلاصه‌سازی استخراجی - بسیار مورد توجه در خلاصه‌سازی چند سندی - مهمترین ویژگی در تولید جملات در خلاصه‌سازی چکیده‌ای	- رویکرهای خوشه‌بندی - زنجیره‌های واژگانی - رویکردهای شناختی - مدل‌های زبانی - تئوری گراف‌ها
ترجمه آماری ماشینی	- خروجی غیر منسجم به دلیل ترجمه جمله به جمله - نیاز به بررسی انسجام در دو زبان	-الگوریتم‌های EM و IBM - زنجیره‌های واژگانی - انسجام واژگانی - تئوری گراف‌ها
تولید متن	- مهمترین هدف در این حوزه - عدم قابلیت اعمال و گسترش در حوزه‌های دیگر	- شبکه عصبی بازگشتی - شبکه ای عصبی انتقالی - شبکه‌های بیزین
ساده‌سازی متن	- به هم خوردن انسجام با ساده سازی متن	- مدل درختی و نمایش مبتنی بر قائده - ساده سازی واژگانی - بکارگیری پایگاه داده وردنت - شبکه عصبی
امتیازدهی خودکار مقالات	-مشکل اغلب متن‌های تالیفی و گردآوری شده - مهمترین مورد برای پذیرش یک نگارش علمی پژوهشی	- جایگاه جملات - زنجیره‌های واژگانی - سطح واژگان -پایگاه داده وب

فصل۳ روش تحقیق

۳-۱ مقدمه

در این فصل به معرفی روش پیشنهادی این رساله برای ارزیابی انسجام متن پرداخته می‌شود. تمرکز این رویکرد بر بکارگیری تعبیه کلمه برای تبدیل واژه‌ها به بردارهای عددی و جملات به ماتریس‌های است. تبدیل جملات به ماتریس‌های عددی در برخی از حوه‌های پردازش متن از سیستم‌های پرسش و پاسخ استفاده شده است [۱۴]. اما در این رساله به جای ایجاد ماتریس‌های شباهت جملات از ماتریس‌های گذر واژگان از یک جمله به جمله بعدی استفاده شده که در جای خود مهمترین نوآوری روش پیشنهادی است. با توجه به اینکه این فاصله گذر در تمام متن محاسبه شده روش معرفی شده هر دو رویکرد ارزیابی انسجام محلی و عمومی را به طور همزمان مورد توجه قرار داده است. نوآوری مهم دیگر روش ارتقای ارزیابی انسجام محلی از سطح جملات متوالی به سطح پاراگراف است. در این روش نخست انسجام هر پاراگراف به عنوان یک واحد منسجم محلی ارزیابی شده و سپس وابستگی موضوعی پاراگراف‌های متوالی به عنوان بخشی از عمل ارزیابی انسجام عمومی مورد سنجش قرار می‌گیرد.

در این بخش ابتدا پیش‌پردازش‌های اولیه بر روی متن ورودی انجام می‌شود. این عملیات شامل پاک‌سازی و آماده سازی اولیه متن، جداسازی بخش‌های متفاوت مانند واژگان، جملات و پاراگراف‌ها، نرمال سازی جملات و واژگان است. پس از پیش‌پردازش وابستگی موضوعی جملات موجود در هر پاراگراف به عنوان ارزیابی انسجام محلی اندازه‌گیری می‌شود. این وابستگی شامل فاصله گذر واژگان هر جمله به جملات بعدی است که این مقادیر درون ماتریسی به نام ماتریس فاصله گذر قرار داده می‌شود. در نهایت ارتباط منطقی و موضوعی پاراگراف‌های متوالی بررسی می‌شوند. این ارتباط تعیین کننده انسجام عمومی متن است

۳-۲ پیش‌پردازش متن

واژه‌هایی که در یک متن پدیدار می‌شوند اغلب دارای تفاوت‌های ساختاری هستند. از این رو مهم‌ترین بخش در هر عملیات پردازشی، آماده سازی متن ورودی برای اعمال عملیات پردازشی اصلی خواهد بود. عملیات پیش‌پردازش با توجه به نوع متن، زبان، الگوریتم و نوع عملیات پردازشی بعدی متفاوت بوده و برای هر حوزه پردازش مدلی خاص پیشنهاد می‌شود. یکی از چالش‌هایی که در تعریف روش پیش‌پردازش وجود دارد محدودیت‌هایی است که در هر حوزه پردازشی موجود بوده و منحصر به همان حوزه خواهد بود. لذا نمی‌توان روش ارائه شده در یک حوزه را به سایر حوزه‌ها تعمیم داده و بکار برد. معروف‌ترین پیش‌پردازش‌های قابل انجام بر روی یک متن عبارت از واحدساز[۱]، حذف ایست‌واژه‌ها[۲]،

[۱] Tokenizer

[۲] Stop word removal

ریشه‌یابی[۱]، پیراسته‌سازی[۲] و برچسب زنی بخش‌های سخن[۳] هستند [۸۶-۸۷]. اما انتخاب نوع عملیات پیش‌پردازش و درصد اعمال آن تاثیر بسیار بالایی بر دقت و سرعت عملیات اصلی پردازشی بعد خواهد داشت. به طور کلی پیش‌پردازش متن به سه بخش کلی پاکیزه سازی[۴]، نرمال سازی[۵] و جداسازی اجزای متن[۶] تقسیم می‌شود. پاکیزه‌سازی متن عبارت از حذف کاراکترهای اضافی و واژگان غیر ضروری از متن است. کاراکترهای اضافی شامل کاراکترهای افزوده شده در متن‌های محاوره‌ای (looove .. love) و کاراکترهایی که مفهومی را به متن اضافه نمی‌کنند (" ، }، [، ...) هستند. واژه‌های اضافی موجود در هر متن ایست واژه‌ها بوده که بار معنایی خاصی را در متن منتقل نمی‌کنند. این عمل موجب افزایش سرعت پردازش و الگوریتم های اصلی عمل ارزیابی انسجام می‌شود. نرمال سازی عمل تبدیل واژه‌ها به یک فرم استاندارد برای پردازش ساده‌تر بوده که شامل ریشه‌یابی، پیراسته‌سازی، تبدیل کاراکترهای لهجه [۸۷] و برچسب گذاری اجزای سخن بوده و جداسازی اجزای متن، عمل تبدیل متن به اجزای آن شامل واژه‌ها، جملات و پاراگراف‌ها است. در ادامه عملیات مهم پیش‌پردازش معرفی می‌شوند:

نشانه‌گذار: نشانه‌گذار همان واحدساز بوده که مرز واژگان در متن را تشخیص داده و آن را به دنباله‌ای از واژه‌ها تبدیل می‌کند. به عبارتی ساده‌تر این ابزار برای شکستن یک متن بر اساس واحدهای با معنی آن مانند واژه، جمله، پاراگراف و موجودیت‌های با معنی مانند فاصله خالی و تب است. در اغلب پیش‌پردازش‌ها واحدسازی بیشتر در سطح واژه انجام شده و واحدهای استخراج شده به عنوان ورودی ابزارهای دیگر مانند ریشه‌یاب، پیراسته‌ساز و برچسب گذار استفاده می‌شود.

حذف ایست‌واژه‌ها: ایست‌واژه به واژگانی گفته شده که با وجود تکرار بسیار زیاد و حضور آنان در اغلب اسناد متنی مفهوم خاصی را منتقل نکرده و نقشی را در یافتن الگوهای پنهان آنان ایفا نمی‌کند. تمامی حوزه‌های پژوهشی پردازش متن به دنبال واژگانی بوده که حاوی اطلاعات مفیدی هستند و سیستم را در رسیدن به یک مدل طبقه بندی یاری کنند. از این رو بهتر است که اینگونه واژگان در مرحله پیش‌پردازش حذف شوند. در نگاه اولیه واژگان ربطی و تعریف جزو ایست‌واژه‌ها محسوب می‌شوند. ولی ایست‌واژه‌ها شامل کلمات بسیار دیگری شده که گاهی شامل افعال، افعال کمکی، اسم‌ها، قیدها و حتی صفات نیز می‌شوند. تقریبا در تمامی کاربردهای پردازش متن حذف این واژه‌ها نتایج پردازش را به شدت بهبود داده و سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. از جمله این واژه‌ها در زبان انگلیسی می‌توان به to, for, about و صدها نمونه دیگر اشاره کرد.

ریشه‌یابی: ریشه‌یابی عبارت از حذف پسوندها و پیشوندهای یک واژه و استخراج ریشه آن است. در هر زبان بیشتر واژه‌ها با توجه به نقش معنایی و نحوی خود در جمله با شکل‌های متفاوتی ظاهر می‌شوند.

---

[۱] Stemming
[۲] Lemmatizing
[۳] POS tagging
[۴] Text cleaning
[۵] Text normalization
[۶] Text segmentation

شکل‌های ظاهری متفاوت از جهتی نشان دهنده معنای متفاوت آن واژه در جمله بوده، اما با توجه به اینکه تمامی آنها از یک ریشه مشتق شده‌اند از نظر معنایی نزدیکی بسیار زیادی با هم دارند. از این رو در بسیاری از کاربردهای پردازش زبان طبیعی و بازیابی اطلاعات بهتر است که همه مشتقات یک واژه به شکل ساده ریشه آن تبدیل شوند (powerful → power). سه رهیافت رایج برای ریشه‌یابی وجود دارد که این سه رهیافت عبارت از رهیافت ساختاری، رهیافت جدول مراجعه و رهیافت آماری هستند.

- **رهیافت ساختاری:** الگوریتم‌های مربوط به رهیافت ساختاری (مبتنی بر قاعده) وابسته به تحلیل ساخت واژه زبان هستند. در این الگوریتم‌ها، با توجه به یک سری قواعد از پیش تعریف شده، به حذف برخی وندها جهت استخراج ریشه پرداخته می‌شود. الگوریتم پورتر مثالی از این دسته از الگوریتم‌ها بوده که از پنج مرحله تشکیل شده است. در طی این مراحل قواعدی بر روی واژگان اعمال شده و بزرگ‌ترین پسوند آن حذف می‌شود.
- **رهیافت جدول مراجعه:** در این رهیافت هر واژه و ریشه مربوط به آن در یک ساختار داده ذخیره شده‌اند. در نتیجه، ریشه هر لغت ذخیره شده را براحتی می‌توان یافت. از مشکلات این رهیافت نیاز آن به فضای حافظه زیاد بوده و همچنین برای هر واژه جدید، بایستی جدول به طور دستی به روز رسانی شود.
- **رهیافت آماری:** در این رهیافت بر مبنای آماره‌های موجود در یک پیکره متنی و با توجه به ساختمان واژه، قواعدی برای استخراج ریشه ایجاد می‌شود. برخی از این روش‌های آماری عبارت از تعداد رخداد، مدل‌های زبانی ان_گرام، تحلیل پیوند و مدل مخفی مارکوف هستند. مهمترین ویژگی رهیافت‌های آماری عدم نیاز آنان به دانش زبان‌شناسی است.

برچسب گذاری اجزای سخن: برچسب گذاری عمل انتساب برچسب‌های واژگانی به کلمات و نشانه‌های یک متن بوده، به صورتی که این برچسب‌ها نشان دهنده نقش واژگان و نشانه‌ها در جمله باشند. در واقع با استفاده ابزار برچسب گذار نقش واژه در جمله از نظر فعل، فاعل، نوع اسم و غیره مشخص می‌شود. برچسب گذاری اجزای سخن و پیکره‌های برچسب خورده در بسیاری از حوزه‌های دیگر پردازش زبان طبیعی مانند تبدیل متن به گفتار، سیستم‌های تشخیص خودکار گفتار، خطایاب و ترجمه ماشینی استفاده می‌شوند. از نمونه‌های انگلیسی آن می‌توان به Illinois Part Of Speech Tagger و Stanford POS Tagger اشاره کرد.

پیراسته‌سازی: پیراسته‌سازی از بسیاری جهات شبیه به ریشه یاب بوده ابزار بسیار مهمی برای تبدیل واژه‌ها به فرم پایه و استاندارد است. اما با این تفاوت که ریشه یاب این عمل را بر اساس شکل ظاهری واژه‌ها انجام داده و اغلب منجر به تولید کلمات بی‌معنی می‌شود (believes … believ) ولی پیراسته‌سازی همین عمل را با توجه به معنی و مفهوم کلمه انجام می‌دهد (believes… belief).

۳-۳ پیش‌پردازش‌های پیشنهادی

در این رساله نیز ابتدا پیش‌پردازش‌های لازم بر روی متن انجام شده تا متن مورد نظر برای استخراج بردارهای واژگان و تشکیل ماتریس‌های عددی آماده شود. پیش‌پردازش‌های انجام شده شامل سه مرحله بوده که در تصویر شماره (۳-۱) نمایش داده شده و الگوریتم پیش‌پردازش پیشنهادی به صورت زیر است:

۱. شروع
۲. پاک‌سازی متن
    ۲,۱. حذف کاراکترهای نقطه گذاری (: ، , ؛ " ...)
    ۲,۲. حذف سایر نمادها ({[|!< ...)
    ۲,۳. حذف اعداد به سبک نگارش ریاضی (۰ .. ۹)
    ۲,۴. تبدیل کاراکترهای بزرگ به کوچک
۳. بخش بندی متن
    ۳,۱. جداسازی پاراگراف‌ها
    ۳,۲. آموزش نشانه‌گذار
    ۳,۳. جداسازی جملات
    ۳,۴. جداسازی واژه‌ها
۴. نرمال سازی متن
    ۴,۱. حذف ایست واژه‌ها
    ۴,۲. استانداردسازی واژه‌های باقیمانده (پیراسته‌سازی[۱]، اصلاح واژگان محاوره‌ای، اصلاح کاراکترهای لهجه)
۵. پایان

در مرحله نخست ابتدا عملیات لازم برای پاک‌سازی و آماده سازی اولیه متن انجام می‌شوند. در بخش پاکیزه‌سازی متن ابتدا تمام کلمات با حروف بزرگ به حروف کوچک تبدیل می‌شوند. سپس کاراکترهای اضافی متن شامل کاراکترهای نقطه گذاری (, ; :)، کاراکترهای نمادین و سیمبول‌ها مانند ({ | \ [...)، اعداد به فرم نگارشی ریاضی صفر تا ۹ و کاراکترهای اضافی موجود در برخی از واژه‌های محاوره‌ای (looove → love) حذف می‌شوند. برای اصلاح واژه‌های محاوره‌ای و حذف کاراکترهای اضافی آنان از پیکره واژگان[۲] استفاده شده است.

در مرحله دوم بخش‌های متفاوت متن از هم جدا می‌شوند. به دلیل اهمیت پاراگراف‌ها در تحقیق جاری و انتخاب آنها به عنوان کوچک‌ترین واحد متن منسجم محلی ابتدا هر پاراگراف به عنوان متنی مستقل در نظر گرفته شده و پاراگراف‌های متن از هم جدا می‌شوند. سپس جملات و واژگان هر جمله از هم جدا شده و آماده برای پردازش‌های بعدی می‌شوند. برای جداسازی جملات و واژه‌های متن از نشانه‌گذار NLTK در پایتون استفاده شده است. نشانه‌گذار پیش فرض NLTK دارای محدودیت‌هایی

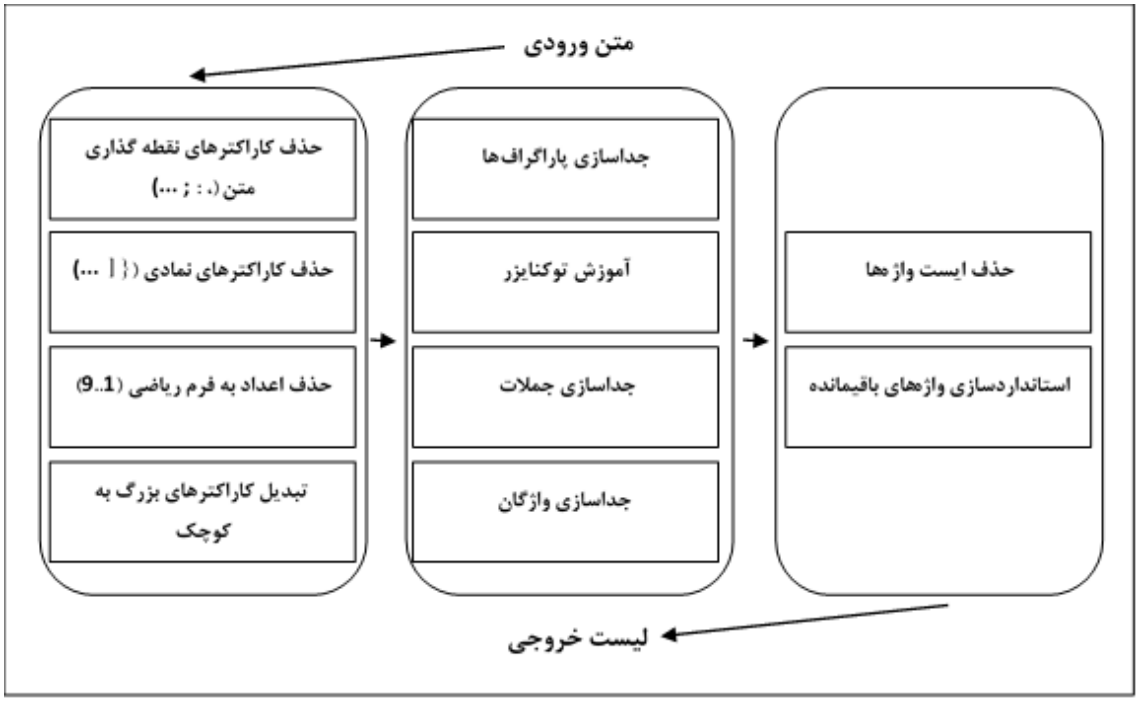
---

[۱] Lemmatizing
[۲] Wordnet

در متن‌های روایی و داستانی است. این نشانه‌گذار یک نشانه‌گذار عمومی بوده در تمامی متون در حوزه‌های متفاوت استفاده می‌شود. با توجه به ماهیت واژگان، جملات و ترکیب جملات در متون داستانی این نشانه‌گذار از دقت بالایی در این حوزه برخوردار نیست. نوآوری مهم دیگر روش پیشنهادی در بخش پیش‌پردازش آموزش نشانه‌گذاری متناسب برای اینگونه متون است. از این رو جهت افزایش دقت نشانه‌گذار NLTK، یک نشانه‌گذار متناسب با متن‌های روایی آموزش داده شده است. برای آموزش این نشانه‌گذار از ابزار PunktSentenceTokenizer و پیکره webtext استفاده شده است. پیکره webtext یک پیکره استاندارد شامل متن‌های روایی و نمایشنامه‌ای مناسب برای این آموزش است. در این آموزش ابزار WordPunctTokenizer برای آموزش نشانه‌گذار واژگان و sent_tokenize برای آموزش نشانه‌گذار جملات بکار گرفته شده است.

مرحله سوم نرمال سازی جملات و واژگان است. جهت نرمال سازی ابتدا ایست واژه‌های موجود در متن حذف می‌شوند. برای حذف ایست واژه‌ها از پایگاه داده ایست واژه‌های انگلیسی و ابزار موجود در NLTK کمک گرفته شده است. برای استاندارد سازی واژه‌های باقیمانده از پیراسته‌سازی استفاده می‌شود. پیراسته‌سازی و ریشه‌یاب دو ابزار بسیار مهم برای تبدیل واژه‌ها به فرم پایه و استاندارد بوده، با این تفاوت که ریشه‌یاب‌ها این عمل را بر اساس شکل ظاهری واژه‌ها انجام داده که اغلب منجر به تولید واژگان بی‌معنی می‌شوند (believes … believ). اما پیراسته‌سازها همین عمل را با توجه به معنی و مفهوم واژه انجام داده که کلمه تولیدی همیشه دارای معنی و مفهوم بوده و در پایگاه داده واژگان موجود خواهد بود (believes…belief). با توجه به استفاده از بردارهای واژگانی word2vec و نیاز آنان به واژگان با معنی پیش‌پردازش جاری از پیراسته‌سازی استفاده می‌کند. بانک اطلاعاتی[1] بردارهای واژگان مورد استفاده در این رساله حاوی ۱۳۲۴۳۰ بردار صد درایه‌ای برای ۱۳۲۴۳۰ کلمه است. اما هنوز هم ممکن است در متن‌های مورد پردازش واژگانی یافت شوند که در این پایگاه داده موجود نباشند. در روش پیشنهادی برای حل این مشکل رجوع به نزدیک‌ترین بردار موجود در بانک اطلاعاتی توصیه شده است. منظور از نزدیک‌ترین بردار شبیه‌ترین واژه موجود در پایگاه داده بوده که بتوان کلمه مورد پردازش را به آن نسبت داد. برای بدست آوردن شبیه‌ترین واژه به دنبال کلمه‌ای در پایگاه داده بوده که بزرگ‌ترین زیردنباله مشترک را با آن دارد.

[1] https://developer.syn.co.in/tutorial/bot/oscova/pretrained-vectors.html.



شکل ۳-۱: دیاگرام پیش‌پردازش‌های پیشنهادی

به عنوان مثال جملات زیر چهار جمله متوالی یک داستان هستند که عملیات پیش‌پردازش پیشنهادی بر روی آنها انجام می‌شود. مثال کامل‌تر به همراه کد برنامه در بخش پیوست آمده است:

"There was once a woman who wished very much to have a little child. She went to a fairy and said I should so very much like to have a little child. Can you tell me where i can find one? Oh that can be easily managed said the fairy."

اعمال عملیات پیش‌پردازش بر روی متن هر جمله را تبدیل به یک لیست کرده که درایه‌های هر لیست شامل واژگانی حاصل از نتیجه اعمال پیش‌پردازش های اولیه شامل جداسازی، حذف ایست‌واژه‌ها و پیراسته‌سازی بر روی این جملات هستند. این واژگان حاوی بیشترین اطلاعات انسجامی موجود در جملات اصلی بوده که مناسب‌ترین گزینه‌ها برای محاسبه فاصله گذر بین جملات متن هستند.

['there', 'woman', 'wish', 'much', 'little', 'child']

['she', go, 'fairy', say, 'I', 'much', 'like', 'little', 'child']

['tell', 'find', 'one']

['oh', 'easily', 'manage', say, 'fairy']

۳-۴ ارزیابی انسجام محلی درون پاراگرافی

در نمایش هر متن به صورت بردارهای عددی، هر واژه می‌تواند به یک بردار عددی تبدیل شود. در این روش هر واژه به برداری عددی به فرم زیر تبدیل شده که این بردار حاوی بیشترین اطلاعات مانند مفهوم

آن، جایگاه واژه در جمله و میزان شباهت آن به سایر واژه‌ها است. عبارت (۳-۱) نشان دهنده این بردار و k مشخص‌کننده ابعاد بردار واژه ایجاد شده آن و جدول (۳-۱) نمونه‌ای از بردار ایجاد شده از واژه there است. در این عبارت E_w واژه مورد نظر و e^i_w درایه‌های عددی تولید شده از آن واژه با الگوریتم word2vec هستند. این درایه‌ها مقادیری نرمال شده در محدوده یک و منهای یک هستند که مشخص کننده واژه و جایگاه آن در فضای بردار تمام واژه‌ها در آن زبان هستند. به دلیل عدم امکان نمایش یک بردار صد درایه‌ای، نمایش بردار در یک ماتریس ۸ در ۱۳ انجام شده است. منظور از ابعاد تعداد درایه‌های موجود بردار ایجاد شده بوده که با افزایش آن دقت بردار افزایش یافته و با کاهش آن سرعت پردازش الگوریتم‌های بعدی افزایش می‌یابد.

$$E_w = \left\{e^1_w, e^2_w, \ldots, e^k_w\right\} \qquad (۳-۱)$$

جدول ۳-۱: بردار عددی واژه there حاصل از الگوریتم word2vec

there							
-0.30479	0.000034	0.308977	-0.11234	-0.15212	-0.18277	-0.14013	0.292745
0.161065	0.039325	-0.19387	0.008936	-0.08029	-0.03834	0.062037	0.198288
0.107156	-0.24854	-0.25375	0.262419	0.205807	0.167403	-0.08231	-0.03385
-0.14518	0.239784	0.006883	0.124801	-0.09716	0.062771	0.076302	0.124304
-0.17317	0.038865	0.095185	0.216123	-0.29153	0.11469	-0.23899	-0.46213
-0.0655	-0.07283	-0.01881	0.149167	0.087325	-0.29789	-0.08947	0.000155
0.018536	0.0642	0.071892	-0.15039	-0.10499	-0.23895	-0.04366	0.045542
-0.29072	-0.26218	-0.10535	-0.26465	0.133357	-0.4406	0.35927	0.312932
-0.07736	-0.03436	-0.25794	-0.20849	-0.07258	0.241113	-0.21375	-0.26106
0.280875	0.053323	-0.03931	-0.14616	-0.36232	0.050909	-0.11909	-0.24229
-0.10817	-0.24898	-0.0803	-0.04238	-0.0447	-0.11746	-0.22794	0.205627
0.062454	-0.27596	-0.00388	-0.45579	0.094545	0.252821	0.187544	0.035061
0.060031	-0.05743	0.279065	0.01465				

یک متن زمانی منسجم فرض می‌شود که تمام اجزای آن با هم ارتباط منطقی داشته و این ارتباط در جملات متوالی کاملا قوی بوده و مشهود باشد. در روش پیشنهادی جمله به عنوان کوچک‌ترین واحد منسجم در یک متن پذیرفته شده است. در نخستین مرحله جملات از هم تفکیک شده و برای ارزیابی انسجام یک پاراگراف پیوستگی و وابستگی موضوعی جملات متوالی آن ارزیابی می‌شوند. در این رویکرد انسجام محلی عبارت از ارتباط موضوعی جملات متوالی یک پاراگراف بوده و انسجام و پیوستگی موضوعی پاراگراف‌های متوالی به عنوان انسجام عمومی در نظر گرفته می‌شود. در اغلب روش‌های پیشنهاد شده قبل برای ارزیابی وابستگی موضوعی و ارتباط جملات، با بکارگیری بردارهای تولید شده توسط الگوریتم word2vec برای هر جمله یک ماتریس تولید می‌شود. در حالت کلی عمل پیش بینی جمله بعدی بستگی به (n-i) جمله قبلی آن دارد (۳-۲):

$$p(t) = p(S_1)p(S_2,|S_1)p(S_3|S_1,S_2)\cdots p(S_n|S_1\cdots S_{n-1}) = \prod_{i=1}^{n} p(S_n|S_1\cdots S_{n-i}) \quad (۳-۲)$$

در این معادله محاسبه احتمال بروز یک جمله پس از جمله‌های قبلی $p(t)$ از ضرب احتمالات عبارت‌های $p(S_i \mid S_{i-1})$ و با مقایسه ویژگی‌های استخراجی از ماتریس‌های جملات تولید شده توسط روش word2vec انجام می‌شود (۳-۳):

$$p(S_i|S_{i-1}) = p((a_{(i-1)}\cdots a_{i-n})|a_{(i-1,1)},a_{(i-1,2)}),\ldots,a_{(i-1,m)}) \quad (۳-۳)$$

در این معادله $a(i-1), a(i-2), \ldots, a(i-n)$ ویژگی‌های مربوط به جمله S_i و ویژگی‌های مربوط به جملات S_{i-1} مقادیر $a(i-1,1), a(i-1,2), \ldots, a(i-1,m)$ هستند. رویکرد بکار گرفته شده این بخش برگرفته از روش معرفی شده توسط روزنفلد بوده و از ان_گرام با فاصله[۱] استفاده می‌کند [۲۸]. از این رویکرد در مقالات استخراجی از رساله استفاده شده است (مقالات ۱، ۷). یکی از مهمترین روش‌های ارائه شده در تبدیل واژه‌ها به بردارهای عددی استفاده از روش‌های مبتنی بر انرژی[۲] است. روش word2vec از خانواده این روش‌ها بوده و در سال ۲۰۱۳ توسط تیم گوگل، میکولوف و همکارانش معرفی شده است [۴]. این روش با الهام‌گیری از مدل‌های مبتنی بر شبکه عصبی در پردازش متن ایجاد شده و یک شبکه عصبی دو لایه بوده که قادر به حدس و تشخیص مفهوم یک واژه با دقت بسیار بالا بر پایه حضورهای قبلی آن در متن است (۳-٤). هدف اصلی و مزیت word2vec گردآوری و کنار هم قرار دادن بردارهای واژگان شبیه به هم در یک فضای برداری است. عبارت (۳-۴) فرمول کلی محاسبه این بردارها بوده که در آن $nb(t)$ مجموعه‌ای از واژگان همسایه و $p(w_i \mid w_t)$ ماکزیمم احتمال گذر از واژه w_i به واژه w_t است [۴].

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{j\in nb(t)}\log p(w_i|w_t) \quad (۳-٤)$$

۳-۴-۱ ارزیابی فاصله گذر جملات متوالی

در روش معرفی شده جمله به عنوان کوچکترین واحد منسجم پذیرفته شده است. از این رو برای ارزیابی انسجام موضوعی موجود در یک پاراگراف یا یک متن وابستگی موضوعی جملات متوالی ارزیابی می‌شود. اغلب روش‌های پیشنهاد شده برای ارزیابی ارتباط دو جمله از بسته واژگان و یا از معیار فراوانی وزنی TF-IDF آنان استفاده کرده‌اند. معیار TF-IDF مخفف فراوانی- عکس فراوانی بوده که در آن به هر واژه بر اساس فراوانی آن در متن یک وزن داده می‌شود. در واقع این روش وزن دهی نشان دهنده اهمیت یک واژه برای آن متن است. بزرگ‌ترین اشکال دو معیار فوق وابستگی کامل ویژگی‌های استخراجی به شکل ظاهری و املای واژه‌های موجود در متن است. در صورت وجود دو واژه با معنی یکسان ولی املای

[۱] LD bigrams

[۲] Energy-based Models

متفاوت این دو معیار نمی‌توانند شباهت و یا تفاوت آنان را به طور دقیق بیان کنند. این مشکل در اغلب موارد حتی برای واژگان با املای یکسان نیز رخ می‌دهد. زیرا املای یکسان گواهی بر یکسان بودن مفهوم آنان نبوده، عواملی زیادی مانند جایگاه آنان در جمله، نوع جمله (خبری، سوالی، ...) و حتی حوزه موضوعی متن در تعیین میزان شباهت آنان دخالت دارند. از این رو بدست آوردن فاصله یا شباهت دو واژه مشابه یا مترادف نیاز به معیارهایی قوی‌تر داشته، و روش‌های مبتنی بر این دو معیار قدرت زیادی برای ارزیابی این موارد را ندارند. جهت درک بهتر به دو جمله متوالی زیر اما با واژگان متفاوت دقت کنید:

my master got a hard test
but my teacher did not give me a good score on this exam

با توجه به اینکه این دو جمله هیچ واژه مشترکی ندارند اما یک نوع اطلاعات را منتشر کرده، و دارای نوعی انسجام و وابستگی موضوعی هستند. واژگان (my, me)، (mater, teacher)، (got, give) و (test, exam) در این دو جمله با هم تشابه معنایی داشته، موجب وابستگی و پیوستگی بین دو جمله شده و دنبال هم بودن آنان را نمایش می‌دهند. اما به هیچ عنوان امکان یافتن الگویی مشترک با معیارهای BOW و TF-IDF بین آنان وجود ندارد. برخی از روش‌های پیشنهادی از سایر معیارها مانند شباهت کسینوسی استفاده کرده‌اند. شباهت کسینوسی معیار خوبی برای ارزیابی وابستگی و شباهت دو جمله است. این معیار به خوبی می‌تواند انسجام و وابستگی موضوعی دو جمله مجاور را در سطح محلی محاسبه کند. اما در محاسبه انسجام عمومی مشکلاتی دارد.

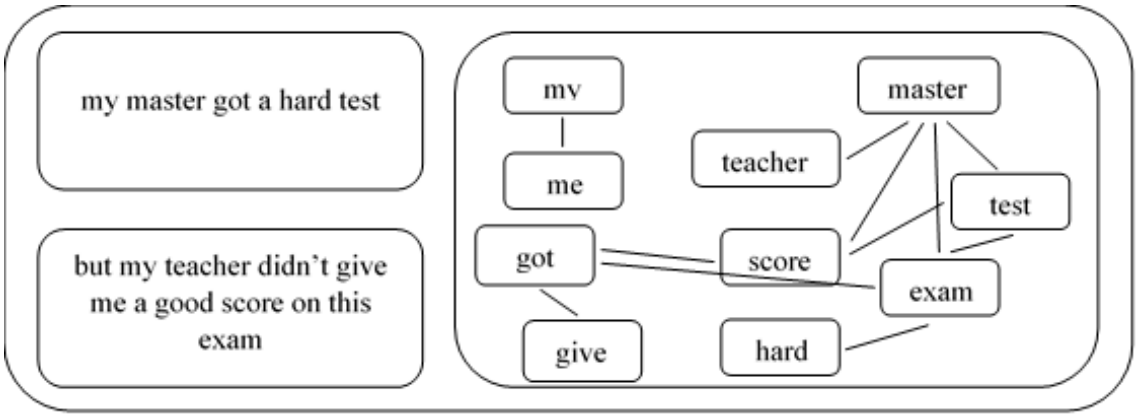
برخی از روش‌های پیشنهادی قبل برای ارزیابی وابستگی موضوعی دو جمله میانگین بردارهای لغات آنان را بدست آورده و با بررسی میزان شباهت این دو بردار اندازه انسجام دو جمله را تعیین کرده‌اند [۸۶-۸۷]. ارماکووا و همکارانش از این راه حل برای این مسئله استفاده کرده اند [۸۸]. نامبردگان برای تخمین و ارزیابی انسجام عمومی میانگین تمام شباهت‌های جملات متوالی را بدست آورده‌اند (۳-۵). در این رویکرد $|S|$ تعداد جملات متن S_{i-1} و S_i بردارهای متوالی تولید شده از میانگین سطرهای ماتریس‌های تولیدی هر ماتریس جمله هستند [۸۸].

$$Coherence(d) = \frac{1}{|s|-1}\sum_{i=2}^{|s|} SC(S_{i-1}, S_i) \qquad (۳-۵)$$

این روش‌ها نیز دارای کاستی‌هایی بوده که از آن جمله می‌توان به مواردی مانند عدم در نظر گرفتن برخی از ویژگی‌های مهم انسجام مانند موقعیت واژه در جمله، طول جمله و شباهت یا فاصله معنایی دو واژه منفرد در دو جمله اشاره کرد. در برخی از اوقات نیز به دلیل وجود مقادیر مثبت و منفی در درایه های هر بردار، بردارهای میانگین دارای فاصله کمی از هم شده که معیار خوبی برای ارزیابی وابستگی دو جمله نخواهند بود. در یک رویکرد پیشنهادی ما با معرفی روشی ساده از مدل‌های زبانی برای نرمال سازی ماتریس‌های جملات استفاده کرده‌ایم (مقاله کنفرانسی استخراجی شماره ۷).

راه حل دیگری که از آن می‌توان برای ارزیابی انسجام مفهومی دو جمله استفاده کرد فاصله انتقال

یک واژه از جمله مبدا به سایر واژگان در جمله مقصد است. فاصله انتقال واژه[1] معیار مناسب دیگری برای ارزیابی انسجام و وابستگی مفهومی دو جمله بوده که توسط مت جی کوسنر (۲۱۰۵) پیشنهاد و ارائه شد [۱۹]. این معیار عبارت از انتقال شباهت معنایی بین دو واژه از یک بخش به بخش دیگری از متن بوده و مشخص می‌کند که محتوای معنایی مجموعه واژگان موجود در یک جمله با چه فاصله‌ای به جمله بعدی انتقال پیدا کرده‌اند. هر چه این فاصله کمتر باشد ارتباط انسجامی دو جمله بیشتر است. معکوس فاصله انتقال واژه[2] عکس فاصله انتقال واژه بوده که در این رساله معرفی شده است. مقدار بیشتر این معیار نشان دهنده ارتباط انسجامی بیشتر دو جمله است. روش پیشنهادی ما با بکارگیری بردارهای عددی word2vec که مدلی آموزش داده شده بر روی میلیون‌ها واژه موجود در متون وب توسط گوگل، رویکرد پیشنهادی میکولوف و همکارانش [۴]، تبدیل کلمات به بردارهای عددی و معکوس فاصله انتقال واژه روشی ساده برای ارزیابی میزان انسجام و وابستگی دو جمله معرفی کرده است. شکل (۳-۲) نمایی از مفهوم فاصله انتقال واژه را نشان می‌دهد:

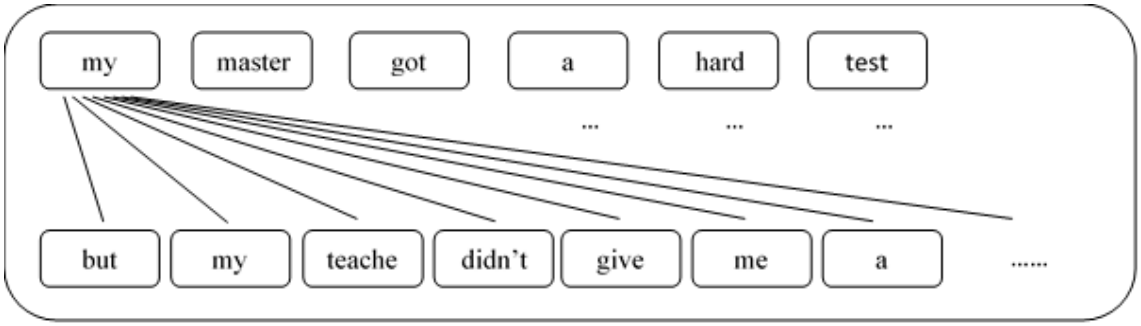


شکل ۳-۲: نمایش فاصله انتقال واژه (WMD) در چند واژه با مفهوم نزدیک به هم [۷۶]

فاصله انتقال واژه در دو جمله ارائه شده عبارت از انتقال شباهت معنایی بین دو واژه از یک بخش به بخش دیگر آنان مانند master و teacher است. در این پژوهش هزینه انتقال بین هر واژه از جمله مبدا را با تمام واژگان جمله مقصد در فضای برداری بردارهای word2bec محاسبه می‌شود. در این روش ابتدا هر واژه i در جمله d به تمام واژه‌های جمله d' منتقل شود. شکل (۳-۳) نمایی گرافیکی از این انتقال بین واژه‌ها متفاوت دو جمله را نشان می‌دهد:

[1] Word mover distance

[2] Inverse word mover distance



شکل ۳-۳: نرخ انتقال بین هر واژه از جمله d با جمله 'd [۹۱]

در این بخش به معرفی مدل ارائه شده برای تشخیص و ارزیابی میزان ارتباط موضوعی (انسجام) اجزای یک سند متنی با استفاده از WE پرداخته می‌شود. ما این مدل را "مدل ارزیابی انسجام مبتنی بر تعبیه کلمه" (ECEM)[۱] نامگذاری کرده‌ایم. مدل پیشنهادی دارای سه مرحله است:

- مرحله نخست: سند متنی به واحدهای منسجم آن (جملات و پاراگراف‌ها) تقسیم شده و برای هر جمله بردارهای واژگان آن با توجه به روش word2vec تولید می‌شود.
- مرحله دوم: ماتریس‌های فاصله جملات متوالی تشکیل می‌شوند.
- مرحله سوم: ارتباط موضوعی جملات با توجه به ارتباط ماتریسی آنان ارزیابی می‌شوند.

مدل ارائه شده فاصله گذر[۲] هر واژه در جمله مبدا را با تمام واژه‌های موجود در جمله دوم محاسبه کرده و مقادیر حاصل را درون ماتریسی به نام ماتریس فاصله واژه[۳] قرار می‌دهد. ماتریس حاصل چکیده‌ای از میزان شباهت یا تفاوت انسجامی دو جمله است. دو لیست نتیجه اعمال پیش‌پردازش های اولیه بر روی دو جمله از مثل‌های قبل را در نظر بگیریم:

A= ['there', 'woman', 'wish', 'much', 'little', 'child']
B= ['she', go, 'fairy', say, 'I', 'much', 'like', 'little', 'child']

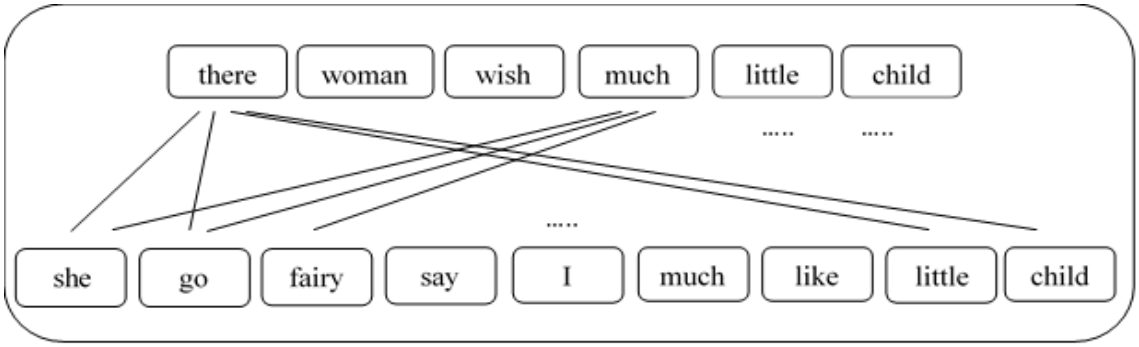
این لیست‌ها کلمات مشترک کمی با هم دارند. اما دارای اطلاعات مشترک بسیار بیشتری در مقایسه با سایر جملات موجود در متن هستند. برای تعیین فاصله گذر دو جمله بردار عددی تولید شده با روش word2vec هر واژه موجود در جمله اول (لیست یک) با تمامی واژه‌های جمله دوم (لیست دو) مقایسه شده و بر اساس معیارهای فاصله تعریف شده فاصله آنها محاسبه شده و درون ماتریس‌هایی با نام ماتریس فاصله گذر دو جمله قرار می‌گیرد. شکل (۳-۴) فاصله هر واژه در لیست نخست را با یک به یک واژه های لیست دوم به تصویر کشیده است.

---

[۱] Embedding-based coherence evaluation model

[۲] Travel cost

[۳] Word distance matrix



شکل ۳-۴: فاصله گذر واژه‌های دو لیست

جهت محاسبه فاصله هر دو واژه ابتدا بردارهای ایجاد شده word2vec آنان استخراج می‌شود. پایگاه داده استفاده شده برای هر واژه یک بردار صد درایه‌ای ایجاد کرده که جدول (۳-۲) چند نمونه از واژگان موجود در مثال قبل را به تصویر کشیده است. به دلیل تعداد زیاد ستون‌های جدول (۱۰۱ ستون) فقط هفت ستون از مقادیر عددی نمایش داده شده است.

جدول ۳-۲: بردار واژگان ۱۰۰ درایه‌ای word2vec مربوط به کلمات لیست‌های مورد مقایسه

Words	1	2	3	5	5	6	…	100
there	-0.30479	0.000034	0.308977	-0.11234	-0.15212	-0.18277	...	0.01465
woman	-0.082	0.232859	0.205318	-0.00608	-0.14159	0.201174	...	-0.049251
wish	0.13765	-0.12273	0.156674	0.217279	0.047827	0.086564	...	0.063056
much	-0.02115	-0.21978	0.130156	0.067502	0.103235	-0.01573	...	-0.004625
little	0.013614	-0.04969	0.257467	0.027133	0.116117	-0.13542	...	-0.060049
child	-0.03115	0.147306	0.440526	-0.18087	-0.04667	0.009171	...	-0.119888
...	...	...	...	...	...	...	...	...

در این روش معرفی شده در این رساله پس از پیش‌پردازش متن مورد پردازش و تولید لیست تمام جملات جدولی شامل تمام واژگان موجود در این لیست‌ها (واژگان بدون تکرار) و بردارهای آنان ایجاد می‌شود. این جدول از روی پایگاه داده اصلی استخراج شده که فقط شامل واژگان پردازش شده و انتخاب شده موجود در متن مورد پردازش بوده، حجم بسیار کمی داشته و کاملا محلی است. در اختیار قرار دادن این جدول (پایگاه داده بسیار کوچک از واژگان مورد استفاده در متن مورد پردازش) موجب افزایش بسیار شدید سرعت و دقت پردازش و ارزیابی الگوریتم پیشنهادی می‌شود. در روش پیشنهادی ابتدا تمام متن‌های مورد آموزش و تست (کلیه متون داستانی هانس کریستین آندرسن و پایگاه داده webtext) مورد پیش‌پردازش قرار گرفته و تمامی واژگان پیش‌پردازش شده درون این جدول قرار گرفته است. چون این عمل یکبار انجام شده و بارها استفاده می‌شود موجب افزایش سرعت ارزیابی های بعدی بر روی متون انتخابی خواهد شد. در مراحل عملی و ارزیابی‌های واقعی انسجام متون مختلف داستانی فقط واژگان و بردارهایی به این جدول افزوده خواهند شد که در متن مورد ارزیابی وجود داشته ولی در

این جدول نیستند. البته در تمامی آزمایش‌های انجام شده فقط تعداد بسیار محدود و کمی سطر به این جدول اضافه شده‌اند. تولید این پایگاه داده محلی بردارهای واژگانی نیز یکی از نوآوری‌های رویکرد پیشنهادی در این رساله است.

در تحقیق حاضر برای محاسبه فاصله گذر واژگان از لیست یک به لیست دو از سه معیار فاصله اقلیدسی[۱] (۳-۶)، معکوس شباهت کسینوسی[۲] (۳-۷) و فاصله هلینگر[۳] (۳-۸) استفاده شده است. در این معادلات $d(p, q)$ فاصله دو واژه بوده که pi درایه‌های بردار واژه مبدا و qi درایه‌های واژه مقصد هستند. در این روش به دلیل بردارهای صد درایه‌ای مقدار n=100 خواهد بود.

$$d(p,q)=\sqrt{\sum_{i=1}^{n}\left(p_i-q_i\right)^2} \qquad (۳-۶)$$

$$d(p,q)=1-\frac{\sum_i p_i \times q_i}{\sqrt{\left(\sum_i p_i^2\right)\left(\sum_i q_i^2\right)}} \qquad (۳-۷)$$

$$d(p,q)=\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2} \qquad (۳-۸)$$

اگر i و j واژگان موجود در دو لیست متوالی و n و m طول هر لیست باشند، ابتدا با استفاده از سه معیار فاصله گذر تمام واژه‌های موجود در دو لیست محاسبه شده و سه ماتریس فاصله تولید می‌شود. سپس در هر ماتریس کمترین مقدار موجود در هر سطر به عنوان نزدیک‌ترین واژه‌ها در دو جمله انتخاب و میانگین این مقدار در تمام سطرها به عنوان انسجام دو جمله با توجه به آن معیار بدست می‌آید. در نهایت میانگین سه مقدار فاصله در سه ماتریس به عنوان اندازه گذر واژه‌ها از جمله اول به جمله دوم در نظر گرفته می‌شود. معادله (۳-۹) شکل محاسبه مقادیر مینیمم از هر سطر و میانگین فاصله هر معیار به صورت را نشان می‌دهد.

$$d(S,S+1)=\min\left[d\left(w_{i=1..n},w_{j=1..m}\right)\right] \qquad (۳-۹)$$

در این عبارت $w_{i=1..n}$ واژه‌های موجود در لیست نخست و $w_{j=1..m}$ واژه‌های موجود در لیست بعدی و $d(S, S+1)$ فاصله جمله هستند. جدول‌های (۳-۳)، (۳-۴) و (۳-۵) به ترتیب ماتریس‌های فاصله گذر ایجاد شده دو لیست A و B بر اساس سه معیار فاصله اقلیدسی، معکوس شباهت کسینوسی و هلینگر هستند. در این جدول‌ها سطرها فاصله گذر هر واژه از لیست A به تمام واژه‌های موجود در لیست B هستند. در این ماتریس‌ها کمترین مقدار موجود در هر سطر مشخص‌کننده واژه‌های با کمترین فاصله در دو لیست بوده و میانگین این مقادیر نشان دهنده فاصله گذر دو لیست (دو جمله) هستند. لیست

---

[۱] Euclidean distance

[۲] Inverse cosine similarity

[۳] Hellinger distance

(۱) مینیمم مقادیر سطرهای ماتریس فاصله اقلیدسی، لیست (۲) مینیمم مقادیر سطرهای ماتریس معکوس شباهت کسینوسی و لیست (۳) مینیمم مقادیر سطرهای ماتریس هلینگر هستند. مقادیر (۴)، (۵)، (۶) میانگین هر لیست و مقدار (۷) نیز فاصله گذر محاسبه شده دو جمله است (۳-۱۰).

1) Minimum_List_ ED A&B= [0.64456, 0.67731, ۰/۷۱۶۸۹, 0, 0, 0]
2) Minimum_List_ICS A&B= [0.64288, 0.64165, 0.71689, 0, -0.23238, 0]
3) Minimum_List_H A&B= 0.66003, 0.47775, 0.71638, 0, 0, 0]

4) Mean_Value_ ED A&B = 0.30653
5) Mean_Value_ ICS A&B= 0.23981 (۳-۱۰)
6) Mean_Value_ H A&B = 0.30902

7) Two_Sentences_Travel_Cost=0.28512

جدول ۳-۳: ماتریس فاصله گذر بر اساس معیار فاصله اقلیدسی بر روی دو لیست

	she	go	Fairy	say	I	much	like	little	child
there	۰/۸۴۱۴۸	۰/۷۰۴۹۳	۰/۸۶۴۴۴	۰/۷۷۹۶۳	۰/۷۳۵۹۲	۰/۸۹۳۱۹	۰/۷۰۲۳۲	۰/۶۴۴۵۶	۰/۸۴۴۹۰
woman	۰/۷۵۰۲۵	۰/۷۵۴۹۹	۰/۸۲۱۳۴	۰/۸۲۰۶۶	۰/۷۴۰۷۲	۰/۸۱۳۲۹	۰/۶۹۷۲۱	۰/۶۷۷۳۱	۰/۴۷۷۷۵
wish	۰/۸۹۹۰۷	۰/۷۴۵۸۲	۰/۹۷۶۸۸	۰/۷۸۱۱۳	۰/۸۱۹۳۰	۰/۹۰۷۲۶	۰/۷۱۶۸۹	۰/۷۹۵۰۲	۰/۷۹۶۵۴
much	۰/۸۸۵۹۹	۰/۸۱۸۳۵	1	۰/۹۰۴۴۶	۰/۸۱۸۱۰	0	۰/۸۳۷۵۸	۰/۵۶۶۱۸	۰/۸۳۱۵۶
little	۰/۷۲۹۹۸	۰/۶۳۷۸۸	۰/۷۹۴۹۴	۰/۷۵۶۱۲	۰/۶۸۲۶۷	۰/۵۶۶۱۸	۰/۵۹۴۸۷	0	۰/۷۱۱۶۹
little	۰/۸۱۵۶۱	۰/۷۶۴۳۸	۰/۸۲۷۰۰	۰/۸۵۶۴۸	۰/۸۰۰۵۱	۰/۸۳۱۵۶	۰/۷۵۹۳۵	۰/۷۱۱۶۹	0

جدول ۳-۴: ماتریس فاصله گذر بر اساس معیار معکوس شباهت کسینوسی بر روی دو لیست

	she	Go	fairy	say	I	much	like	little	child
there	0.85323	0.72659	۰/۸۶۰۰۰	۰/۷۸۶۲۴	۰/۷۳۱۸۳	۰/۹۲۶۸۷	۰/۷۴۷۱۵	۰/۶۴۲۸۸	۰/۹۱۶۱۰
woman	۰/۶۷۸۹۵	۰/۸۳۶۸۱	۰/۷۷۷۲۹	۰/۸۷۳۶۰	۰/۷۴۳۵۲	۰/۷۶۸۸۵	۰/۶۴۱۶۵	۰/۸۴۴۸۱	۰/۶۹۹۳۱
wish	۰/۸۸۰۵۲	۰/۷۱۵۱۹	۰/۹۹۷۸۵	۰/۷۰۷۲۱	۰/۸۰۸۵۷	۰/۸۶۷۳۹	۰/۶۷۸۶۷	۰/۳۸۶۷۱	۰/۷۷۱۲۴
much	۰/۸۳۵۳۲	۰/۸۴۰۳۹	۱/۰۲۲۴۹	۰/۹۲۵۶۹	۰/۷۸۴۸۶	0	۰/۹۰۹۸۰	۰/۲۵۹۳۶	۰/۸۱۴۸۹
little	۰/۷۳۰۱۷	۰/۷۰۸۱۳	۰/۸۲۴۷۰	۰/۸۵۷۸۲	۰/۷۳۴۹۲	۰/۴۰۰۵۴	۰/۶۴۴۰۴	۰/۶۹۱۹۹	-0.2323
little	۰/۷۷۴۳۶	۰/۸۱۸۸۴	۰/۷۶۱۳۰	۰/۹۱۴۲۲	۰/۸۳۲۶۵	۰/۷۷۶۳۹	۰/۸۳۶۲۷	۰/۷۵۰۰۷	0

جدول ۳-۵: ماتریس فاصله گذر بر اساس معیار هلینگر بر روی دو لیست

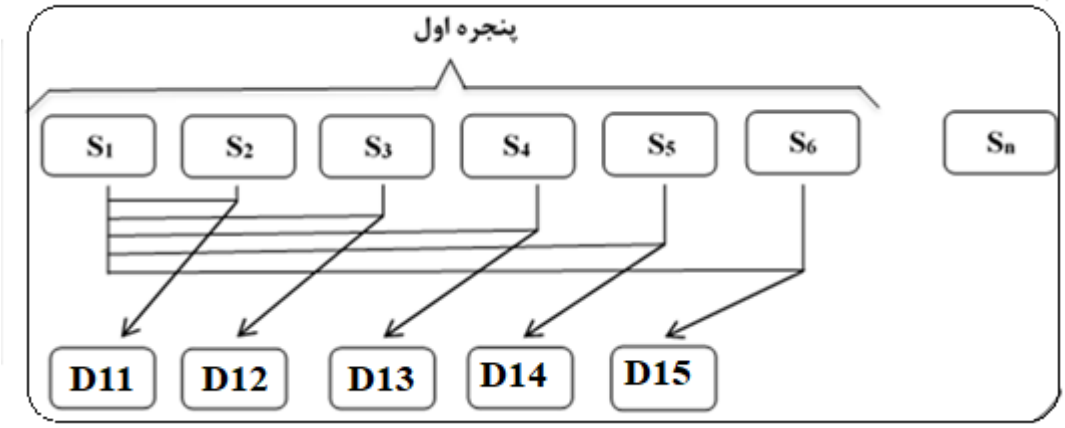
	she	Go	fairy	say	I	much	like	little	child
there	0.85884	0.71634	۰/۸۷۴۷۸	۰/۸۵۶۴۱	۰/۷۵۴۰۶	۰/۹۰۸۴۳	۰/۷۱۲۵۵	۰/۶۶۰۰۳	۰/۸۵۵۶۲
woman	۰/۷۵۰۲۵	۰/۷۵۴۹۹	۰/۸۲۱۳۴	۰/۹۱۴۶۹	۰/۷۴۰۷۲	۰/۸۱۳۲۹	۰/۶۹۷۲۱	۰/۶۷۷۳۱	۰/۴۷۷۷۵
wish	۰/۸۹۷۹۴	۰/۷۴۳۶۰	۰/۹۷۸۲۰	۰/۹۳۱۸۹	۰/۸۱۷۵۱	۰/۹۰۸۸۱	۰/۷۱۶۳۸	۰/۷۹۶۳۲	۰/۷۹۶۸۹
much	۰/۸۸۵۹۹	۰/۸۱۸۳۵	1	۰/۹۸۳۶۵	۰/۸۱۸۱۰	0	۰/۸۳۷۵۸	۰/۵۶۶۱۸	۰/۸۳۱۵۶
little	۰/۷۲۹۹۸	۰/۶۳۷۸۸	۰/۷۹۴۹۴	۰/۸۶۵۴۷	۰/۶۸۲۶۷	۰/۵۶۶۱۸	۰/۸۵۶۴۱	0	۰/۷۱۱۶۹
little	۰/۸۱۵۶۱	۰/۷۶۴۳۸	۰/۸۲۷۰۰	۰/۹۴۱۹۴	۰/۸۰۰۵۱	۰/۸۳۱۵۶	۰/۷۵۹۳۵	۰/۷۱۱۶۹	0

در سه جدول فوق فاصله گذر هر واژه از لیست مبدا (جمله مبدا) با با واژه‌های موجود در لیست مقصد (جمله مقصد) محاسبه شده و به عنوان یکی از درایه‌های جدول قرار می‌گیرد (مثال: فاصله گذر واژه there و she بر اساس معیار اقلیدسی برابر با ۰,۸۴۱۴۸ است). جدول حاصل نمایشگر ماتریس فاصله

گذر دو جمله با معیار فاصله اقلیدسی بوده و دو جدول دیگر نیز به صورت مشابه ماتریس‌های فاصله گذر دو جمله را بر اساس دو معیار دیگر معکوس شباهت کسینوسی و فاصله هلینگر محاسبه کرده‌اند. در نهایت مینیمم مقدار موجود در هر سطر ماتریس نشان دهنده کمترین فاصله بین دو واژه بوده که با میانگین‌گیری سه مینینمم بدست آمده توسط سه معیار فاصله گذر از جمله مبدا به جمله مقصد محاسبه شده است (۳-۱۰).

۳-۴-۲ ارزیابی انسجام محلی در سطح پاراگراف

روزنفلد در سال ۱۹۹۶ با استفاده از بایگرام‌های با فاصله ثابت کرد که واژه‌های متوالی موجود در یک متن با هم ارتباط مفهومی داشته و این پیچیدگی ارتباط تا فاصله پنج کاهش یافته، اما از آن به بعد ثابت می ماند [۲۸]. این الگو در مورد انسجام مفهومی تمام اجزای یک متن، بویژه جملات متوالی نیز صادق بوده و در پژوهش‌های قبلی (مقاله مستخرج شماره ۱) و پژوهش حاضر از همین الگو برای ارزیابی انسجام عمومی متن استفاده شده است. در این تحقیق با در نظر گرفتن یک پنجره پنج جمله‌ای در ابتدای متن و حرکت آن بر روی تمام متن، ارتباط انسجامی جملات موجود در هر پنجره از فاصله صفر (دو جمله دنبال هم) تا فاصله پنج محاسبه می‌شوند. مقادیر بدست آمده از هر پنجره درون یک بردار پنج درایه‌ای قرار گرفته و فاصله انسجامی هر پنج جمله متوالی تبدیل به یک بردار با پنج درایه می‌شود (شکل‌های (۳-۵) (۳-۶)). در نهایت با حرکت پنجره بر روی کل متن و اعمال الگوریتم بر روی آن ماتریس فاصله انسجام متن با پنج ستون و n-۵ سطر (شکل (۳-۷)) تولید می‌شود. با توجه به ذخیره سازی این فاصله‌ها در یک بردار و تشکیل ماتریس متن، هرچه مقادیر موجود در سطرهای ماتریس از فاصله کمتری برخوردار باشند انسجام محلی متن مورد نظر بیشتر است.

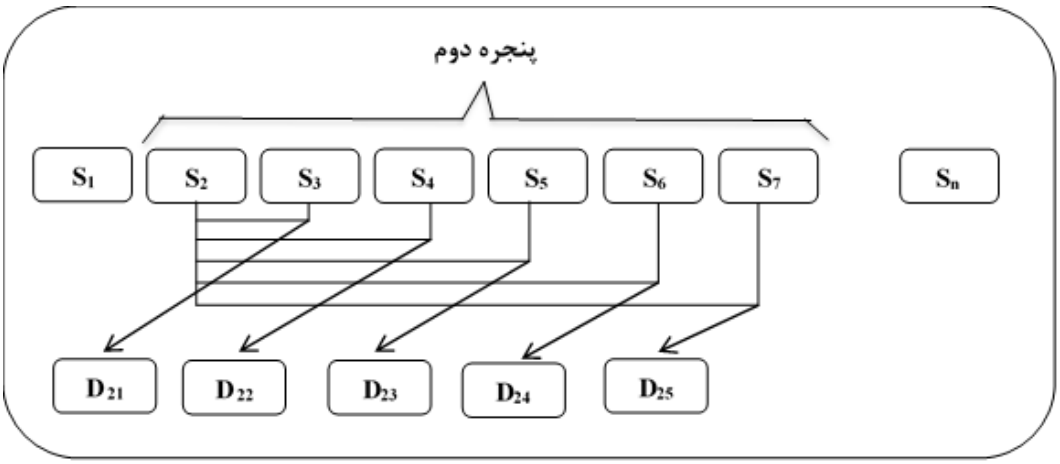


شکل ۳-۵: بردار فاصله انسجامی پنجره پنج جمله اول

در تصویر (۳-۵) S_1 تا S_6 جملات متوالی در پنجره اول و D_{11} تا D_{15} فاصله انسجامی آنان و در تصویر (۳-۶) S_2 تا S_7 جملات متوالی در پنجره اول و D_{21} تا D_{25} فاصله انسجامی آنان است.

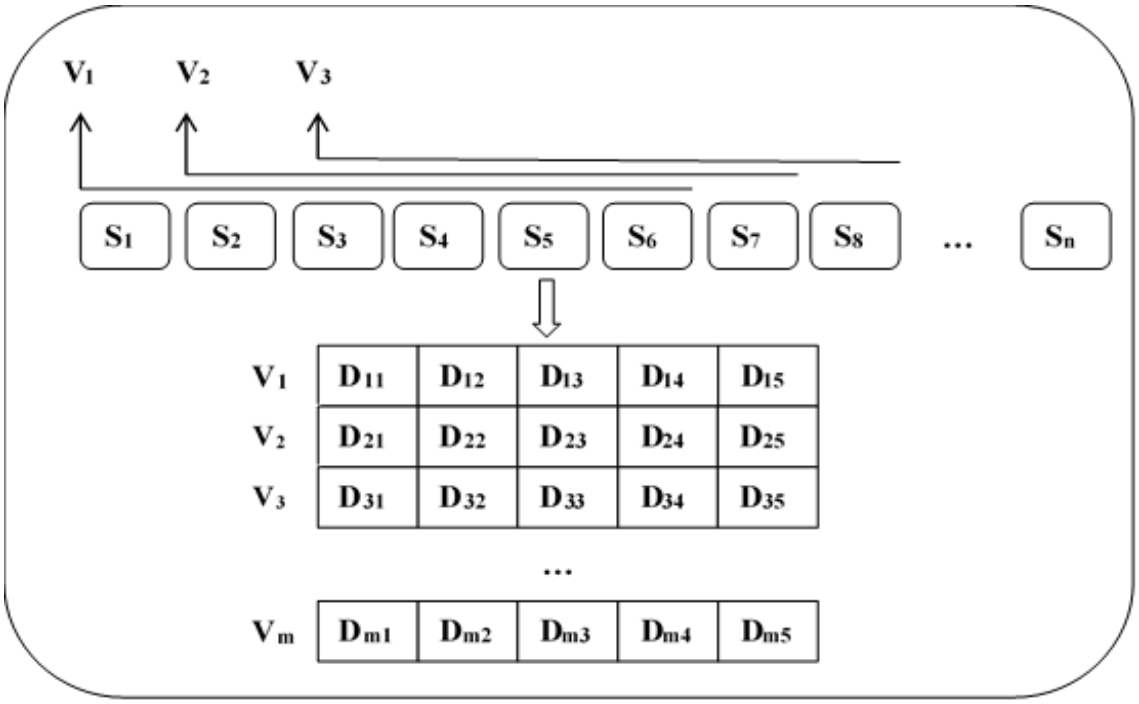
D_{11}=dist (S_1, S_2)
D_{12}=dist (S_1, S_3)
D_{13}=dist (S_1, S_4)
D_{14}=dist (S_1, S_5)
D_{15}=dist (S_1, S_6)

D_{21}=dist (S_2, S_3)
D_{22}=dist (S_2, S_4)
D_{23}=dist (S_2, S_5)
D_{24}=dist (S_2, S_6)
D_{25}=dist (S_2, S_7)



شکل ۳-۶: بردار فاصله انسجامی پنجره پنج جمله دوم

با فرض وجود n جمله در پاراگراف این عمل n-5 با انجام شده و منجر به تولید ماتریس فاصله انسجام پاراگراف می‌شود (شکل ۳-۷).



شکل ۳-۷: ماتریس فاصله انسجام متن

در نهایت جهت ارزیابی انسجام پاراگراف ضریب همبستگی بردارهای تولید شده هر پنجره (سطرهای ماتریس فاصله انسجام) نسبت به هم و به صورت متوالی اندازه‌گیری می‌شود (۳-۱۱). ضریب همبستگی شدت رابطه و نوع مستقیم یا معکوس بودن آنرا نسبت به هم در دو مجموعه داده مشخص کند. مقدار

این ضریب بین ۱ و ۱- بوده که هرچه رابطه دو مقدار با هم بیشتر شود مقدار به یک نزدیک‌تر، مقدار صفر نشان دهنده عدم وجود رابطه بین دو متغیر و مقدار منفی نشان دهنده رابطه معکوس آنان است. ضریب همبستگی بالاتر در متن مورد ارزیابی نشان دهنده انسجام محلی بیشتر پاراگراف است.

$$Corr(V_1...V_m) = \left[\frac{\sum_{i=1}^{m} \frac{(V_i - \mu V_i)(V_{i+1} - \mu V_{i+1})}{\sigma V_i \sigma V_{i+1}}}{m} \right] \quad (۳-۱۱)$$

محاسبه مقادیر V_i و V_{i+1} در عبارت (۳-۱۲) نشان داده شده است.

$$V = \{v_1, v_2, ..., v_{n-5}\}$$
$$v_i = D_{i,j} \quad (۳-۱۲)$$
$$\max\_coherence(p) = \max\_corr(v_1..v_m)$$

در این فرمول‌ها V_i سطرهای ماتریس پاراگراف، n تعداد جملات پاراگراف، D_{ij} فاصله گذر جمله i به جمله j در پنجره شش جمله‌ای، *max_corr(v1..vm)* ماکزیمم ضریب همبستگی بردارهای تولید شده و *max_coherence(p)* ماگزیمم انسجام محلی پاراگراف مورد مطالعه است. در ادامه الگوریتم ارزیابی انسجام درون پاراگرافی قابل مشاهده است.

۱. شروع
۲. تعیین پنجره شش جمله‌ای برای اندازه‌گیری فاصله گذر در جملات متوالی در ابتدای پاراگراف
    ۲,۱. محاسبه فاصله گذر جمله اول را با جمله دوم
    ۲,۲. محاسبه فاصله گذر جمله اول را با جمله سوم
    ۲,۳. محاسبه فاصله گذر جمله اول را با جمله چهارم
    ۲,۴. محاسبه فاصله گذر جمله اول را با جمله پنجم
    ۲,۵. محاسبه فاصله گذر جمله اول را با جمله ششم
    ۲,۶. تشکیل سطر اول ماتریس گذر پاراگراف
۳. حرکت پنجره به سمت جلو (جمله دوم تا هفتم)
    ۳,۱. تکرار عملیات مرحله ۲ برای پنجره جدید
    ۳,۲. تشکیل سطر بعدی ماتریس گذر پاراگراف
۴. تکرار مرحله سه به تعداد n-5 بار (nتعداد جملات موجود در پاراگراف است)
۵. ارزیابی انسجام پاراگراف با محاسبه ضریب همبستگی سطرهای ماتریس گذر
۶. پایان

۳-۴-۳ ارزیابی انسجام عمومی

برای ارزیابی انسجام عمومی (انسجام پاراگراف‌های متوالی) به صورت زیر عمل می‌شود. ابتدا یک پاراگراف مجازی تولید می‌شود. برای ایجاد این پاراگراف ابتدا عنوان متن در جایگاه نخستین جمله آن قرار گرفته و سپس اولین جمله از هر پاراگراف (جمله موضوعی آن) به صورت متوالی در جایگاه جملات بعدی پاراگراف مجازی قرار می‌گیرد. در نهایت روش ارائه شده برای ارزیابی انسجام محلی پاراگرافی بر روی این پاراگراف مجازی نیز اعمال شده که نتیجه آن میزان انسجام عمومی متن است. در ادامه الگوریتم ارزیابی انسجام درون پاراگرافی قابل مشاهده است.

۱. شروع
۲. ایجاد پاراگراف مجازی کل متن (انتخاب عنوان متن در جایگاه جمله موضوعی پاراگراف، انتخاب جملات موضوعی هر پاراگراف به صورت متوالی در جایگاه جملات بعدی پاراگراف مجازی)
۳. انجام عملیات پیش‌پردازش بر روی پاراگراف مجازی و تولید لیست‌های جملات
۴. اعمال الگوریتم ارزیابی انسجام درون پاراگرافی بر روی پاراگراف مجازی ایجاد شده
۵. ارزیابی انسجام عمومی متن با محاسبه ضریب همبستگی سطرهای ماتریس گذر پاراگراف مجازی
۶. پایان

۴-۴-۳ نتیجه گیری

مسئله مورد بررسی در این فصل، معرفی روشی برای ارزیابی همزمان انسجام عمومی و محلی با استفاده از ویژگی‌های آماری موجود در متن و بدون نیاز به مفاهیم معنایی واژگان است. با توجه به اینکه تمرکز روش معرفی شده بر روی متن‌های داستانی، روایی و نمایشی است از پیش‌پردازش‌های ویژه‌ای برای پردازش اولیه و آماده سازی متن ورودی استفاده شده است. این شکل پیش‌پردازش برای متون داستانی با حجم بزرگ، تعداد جملات زیاد و حجم زیاد واژگان محاوره‌ای و غیر معمول مناسب است. در رویکرد معرفی شده پس از پیش‌پردازش‌های معرفی شده جملات به لیست‌هایی از واژگان مهم موجود در متن و به صورت استاندارد تبدیل شده و سپس با بکارگیری الگوریتم word2vec و تعبیه کلمه واژگان موجود در لیست‌ها تبدیل به بردارهای عددی شده‌اند.

سپس بر خلاف رویکردهای پیشین که برای ارزیابی وابستگی مفهومی جملات از شباهت بردارهای واژگان بین جملات استفاده می‌کرده‌اند، این رساله از ماتریس‌های گذر واژگان از یک جمله به جمله بعدی (واژگان لیست مبدا و لیست مقصد) استفاده شده که در جای خود مهمترین نوآوری روش پیشنهادی است. این فاصله گذر با بکارگیری فاصله هر واژه از لیست مبدا با تمامی واژه‌های لیست مقصد بر اساس سه معیار استاندارد فاصله اقلیدسی، معکوس شباهت کسینوسی و فاصله هلینگر فاصله گذر دو جمله مورد بررسی را محاسبه کرده است. فاصله کمتر جملات متوالی و غیر متوالی با فاصله بیشتر

از هم نشانه انسجام بیشتر بخش‌های متفاوت متن است. رویکردهای قبلی برای ارزیابی انسجام کل متن گراف آن را تشکیل داده که رئوس آن جملات و وزن یال‌های آن مقدار شباهت و وابستگی آنان به هم بوده است. این عمل در متن‌های بزرگ موجب افزایش بیش از حد گراف، رابطه‌های بین رئوس آن و در نتیجه بسیار بالای‌شد. اما در ریکرد پیشنهادی با ایجاد یک پنجره شش جمله‌ای و حرکت آن بر روی کل متن موجب ایجاد یک ماتریس ۵ در m شده که در آن m=n-5 بوده و n تعداد جملات متن است. با توجه به محدودیت تعداد جملات هر متن و پردازش خطی (مرتبه زمانب n) روش پیشنهادی از سرعت و دقت بسیار بالاتری در مقایسه با گراف تولید شده با ارتباطات زیاد رئوس آن (مرتبه زمانی 2n) شده است. در نهایت با ارزیابی وابستگی سطرهای ماتریس گذر ایجاد شده به ترتیب حضور در متن انسجام محلی و عمومی به طور همزمان ارزیابی می‌شود.

فصل ۴ ارزیابی و نتایج

۴-۱ مقدمه

در این فصل به معرفی سیستم مورد استفاده جهت آزمایش، پایگاه‌های داده استفاده شده، تحلیل، تفسیر و نتایج روش معرفی شده بر روی آنها پرداخته می‌شود.

۴-۲ سیستم مورد استفاده جهت آزمایش

ارزیابی آزمایشات بر روی سیستمی با مشخصات سخت افزاری مطابق با جدول (۴-۱) اجرا شده و جهت پیاده‌سازی الگوریتم‌ها از زبان برنامه‌نویسی پایتون استفاده شده است.

جدول ۴-۱: مشخصات سیستم مورد آزمایش

سخت افزار و سیستم عامل	مدل
پردازنده	Intel Core 2 Duo T6570 @2.1 GHz
حافظه اصلی	GB ۴,۰
حافظه جانبی	GB 1000
سیستم عامل	Microsoft Windows 10

۴-۳ معرفی پایگاه داده

پایگاه داده‌های بکار گرفته شده در این رساله عبارت از دو پایگاه داده استاندارد بکار گرفته شده در رویکردهای پیشین و یک پایگاه داده ایجاد شده جدید متشکل از متن‌هایی استاندارد و منسجم است. در ادامه توضیحاتی در مورد این دو مجموعه داده ارائه می‌گردد.

۴-۳-۱ پایگاه داده‌های استاندارد بکار گرفته شده در رویکردهای پیشین

پایگاه داده‌های پیشین بکار گرفته شده در این رساله دو پایگاه داده Earthquakes و Accidents است. پایگاه داده‌های ذکر شده حاوی اخبار موجود در روزنامه‌ها و اخبار ارائه شده توسط مقامات دولتی است (شکل ۴-۱). این پایگاه‌های داده شامل دو مجموعه داده بوده و از طریق وب در دسترس است[۱]. مجموعه نخست حاوی متون ارائه شده در خبرگزاری آوشیتدپرس از زلزله‌های اتفاق افتاده در آمریکای شمالی

[۱] http://people.csail.mit.edu/regina/coherence

بوده که میانگین اندازه جملات آن ۱۰٫۴ واژه است. مجموعه دوم حاوی اخبار و اطلاعات ارائه شده توسط انجمن ملی حمل و نقل ایمن از تلفات و تصادفات هوایی است. میانگین تعداد واژه موجود در هر جمله در این مجموعه ۱۱٫۵ واژه است. در هر بخش از این پایگاه‌داده یک متن اصلی و متن‌های ایجاد شده جدید از جایگشت‌های متفاوت جمله‌های آن وجود دارد. این پایگاه‌داده تا به حال توسط بسیاری از روش‌های معرفی شده تشخیص انسجام متن بکار گرفته شده است [۲۰] [۳۶] [۴۵]. فرض اصل تمامی روش‌های استفاده کننده بر این بوده است که متن اصلی از انسجام قابل قبولی برخوردار بوده و متون دیگری که از جایگشت‌های جملات آن ایجاد شده‌اند باید دارای امتیاز انسجام کمتری از آن باشند. به عنوان مثال در روش بکار گرفته شده توسط بارزیلای [۲۲] به ازای هر محاسبه انسجام یک بردار شامل دو مقدار میزان انسجام متن اصلی و متن با ایجاد جایگشت جدید جملات ایجاد شده است. در نهایت به ازای k متن با n جایگشت موجود، تعداد k*n بردار دو مقداری تولید شده است. مقدار فاصله دو مقدار هر بردار مشخص‌کننده دقت سیستم طراحی شده است.

This is preliminary information subject to change and may contain errors.
Any errors report will be corrected when the final report has been completed.
On January 13 1994 about 1230 hours Greenwich Mean Time a beech be-90 n46wa registered to Charles kuykendall Wilmer Texas.
The ditching was precipitated by an in-flight fire during cruise flight.
The German national commercial pilot and the sole occupant received minor injuries.
... ...

شکل ۴-۱: بخشی از پایگاه داده Accidents

۴-۳-۲ پایگاه داده ایجاد شده

پایگاه داده مورد استفاده یک مجموعه متن انتخابی بیست داستان از سری داستان‌های هانس کریستین آندرسن بوده که انسجام پیش فرض آنان به دلیل اینکه توسط یک نویسنده ماهر تولید شده است پذیرفته شده است. در این مجموعه برای هر داستان ده نمونه متن غیر منسجم (۴-۱) تولید شده است:

$$d_i : i \in [1,...,10] \qquad (۱-۴)$$

این ده نمونه عبارت از پنج نمونه متن غیر منسجم با جابجایی اتفاقی جملات آن به میزان ۱۰ درصد، ۲۰ درصد، ۳۰ درصد، ۴۰ درصد و ۵۰ درصد و پنج نمونه متن غیر منسجم با حذف اتفاقی جملات (ایجاد متن کوتاه‌تر) هستند. پنج نمونه خلاصه ایجاد شده شامل دو متن با حذف ۱۰ درصد جملات، دو متن با حذف ۲۰ درصد جملات و یک متن با حذف ۳۰ درصد جملات هستند. متن‌های خلاصه بدون اعمال

هیچگونه الگوی خلاصه‌سازی بوده و کاملا اتفاقی ایجاد شده تا انسجام و پیوستگی موضوعی آنان کاهش پیدا کند. در نتیجه در پایگاه داده مورد مطالعه ۲۲۰ متن شامل ۲۰ متن دارای انسجام کامل و ۲۰۰ متن با انسجام کاهش یافته با درصدهای ذکر شده به دو صورت حذف اتفاقی بخش‌های مختلف و جابجایی اتفاقی جملات است. در نتیجه متن‌های ایجاد شده با جابجایی یا کاهش جملات دارای کاهش انسجامی متناسب با میزان جابجایی یا کاهش جملات هستند که در شکل (۴-۲) نمونه‌ای از این متن‌های ایجاد شده را نشان داده شده است.

There was once a woman who wished very much to have a little child, but she could not obtain her wish.

At last she went to a fairy, and said, I should so very much like to have a little child; can you tell me where I can find one?

Here is a barleycorn of a different kind to those which grow in the farmers' fields, and which the chickens eat; put it into a flower-pot, and see what will happen.

Thank you, said the woman, and she gave the fairy twelve shillings, which was the price of the barleycorn.

Then she went home and planted it, and immediately there grew up a large handsome flower, something like a tulip in appearance, but with its leaves tightly closed as if it were still a bud.

شکل ۴-۲: بخشی از پایگاه داده ایجاد شده

۴-۴ ارزیابی مدل پیشنهادی

در این پژوهش مقایسه‌ای بین رویکردهای مبتنی بر شبکه موجودیت و رویکردهای استفاده کننده از بردارهای عددی word2vec انجام شده است. این مقایسه مشخص می‌کند روش‌های استفاده کننده از بردارهای عددی علاوه بر اینکه از الگوریتمی ساده‌تر برخوردار بوده دقت خوبی در مقایسه با روش‌های معنایی دارند. مدل معرفی شده در این رساله بر روی تمام جفت متن‌های موجود و ده نمونه غیر منسجم تولید شده (d_o, d_i) هر متن اعمال شده و کاهش انسجام در هر جفت با درصد جابجایی جملات یا خلاصه‌سازی مقایسه می‌شود. هر چه این دو مقدار به هم نزدیک‌تر باشند دقت مدل طراحی شده بالاتر است. جهت ارزیابی اولیه مدل ده متن از متن‌های موجود در پایگاه داده ایجاد شده به همراه تمامی نمونه‌های غیر منسجم تولیدی از هر کدام انتخاب شده است. برای ارزیابی تاثیر طول متن بر کارایی روش سعی شده است که متن‌های انتخابی دارای طول متفاوت (تعداد جملات کم و تعداد جملات زیاد) باشند. جدول (۴-۲) متن‌های انتخابی به همراه نمونه‌های غیر منسجم هر کدام را نشان می‌دهد. در این جدول تعداد جملات متن‌های انتخابی در هر چهار حالت (متن اصلی، ده درصد کاهش یافته، بیست

درصد کاهش یافته و سی درصد کاهش یافته) مشخص شده است.

جدول ۴-۲: تعداد جملات متن‌های انتخابی به همراه نمونه‌های غیر منسجم کاهش یافته جملات

الف	ب	ج	د	ه
ST_45	۴۵	۴۱	۳۷	۳۳
ST_68	۶۸	۶۲	۵۶	۵۰
ST_70	۷۰	۶۳	۵۸	۴۹
ST_82	۸۲	۷۴	۶۶	۵۸
ST_86	۸۶	۷۸	۷۰	۶۲
ST_101	۱۰۱	۹۱	۸۱	۷۱
ST_111	۱۱۱	۱۰۰	۸۹	۷۸
ST_169	۱۶۹	۱۵۳	۱۳۷	۱۲۲
ST_192	۱۹۲	۱۷۳	۱۵۴	۱۳۵
ST_260	۲۶۰	۲۳۴	۲۰۸	۱۸۲

این جدول ده متن انتخابی را نشان داده که منتخبی از متن‌های کوتاه، متوسط و بلند از داستان‌های هانس کریستین آندرسن هستند. در این ده داستان انتخابی متون بین ۴۵ جمله تا ۷۰ جمله جزو داستان‌های کوتاه، متون بین ۸۲ جمله تا ۱۱۱ جمله در زمره داستان‌های با طول متوسط و متون بین ۱۶۹ تا ۲۶۰ از متون بلند هستند. در جدول فوق ستون (ج) تعداد جملات پس از کاهش انسجام به ده درصد، ستون (د) تعداد جملات پس از کاهش انسجام به بیست درصد حذف اتفاقی جملات و ستون (ه) تعداد جملات پس از کاهش انسجام به سی درصد حذف اتفاقی جملات است. با توجه به اینکه حذف جملات از متون داستانی عملی مرسوم بوده و همان خلاصه‌سازی استخراجی است، اما حذف اتفاقی جملات موجب کاهش شدید انسجام شده و حذف جملات با درصدهای متفاوت و درصدهای مختلف موجب کاهش‌های انسجامی متفاوتی در متن خروجی خواهند شد. به همین دلیل تولید پایگاه داده متن با جملات کاهش یافته در سایر پایگاه داده‌های قبلی ارزیابی متن مانند Earthquakes و Accidents مرسوم نبوده ولی در این پایگاه داده و در این پژوهش معرفی شده است.

بخش دیگری از پایگاه داده تولید شده از جابجایی جملات موجود در متن ایجاد شده است. این جابجایی‌ها با میزان ۱۰ درصد، ۲۰ درصد، ۳۰ درصد، ۴۰ درصد و ۵۰ درصد تولید شده‌اند. جابجایی جملات برای تولید متون غیر منسجم استاندارد در تمامی پایگاه داده‌ها مرسوم بوده و دو پایگاه داده استاندارد Earthquakes و Accidents نیز دارای نمونه‌های غیر منسجم متن اصلی با درصد جابجایی جملات هستند. اما آنچه مسلم است جابجایی جملات در متون داستانی بسیار با اهمیت بوده، زیرا جابجایی جملات تاثیر بسیار زیادی بر انسجام متن حاصل داشته مقدار کاهش انسجام بیش از درصد جابجایی جملات خواهد بود. در حالی که در پایگاه‌داده‌های متون خبری علمی جابجایی جملات تاثیری کمتر از میزان درصد جابجایی در کاهش انسجام داشته است. دلیل اصلی تولید پایگاه داده تولید شده

این مهم بوده، زیرا به هیچ عنوان نمی‌توان انسجام متن‌های داستانی را توسط سیستمی که با متون غیر داستانی آموزش داد.

برای ارزیابی اولیه ابتدا یک متن به همراه ده نمونه متن کاهش انسجام یافته آن را انتخاب و مدل پیشنهادی را بر روی آن اعمال کرده‌ایم. متن انتخابی شامل ۱۹۲ جمله و ۴۵۰۶ واژه است. مدل پیشنهادی بر روی مجموعه‌ای ده متنی شامل پنج نمونه متن با درصد جابجایی جملات ۱۰ تا ۵۰ درصد، دو نمونه متن خلاصه شده (حذف اتفاقی جملات) ۱۰ درصدی، دو نمونه متن خلاصه شده (حذف اتفاقی جملات) ۲۰ درصدی و یک نمونه متن خلاصه شده (حذف اتفاقی جملات) ۳۰ درصدی انجام شده است. نتایج حاصل از مقایسه انسجام در نمونه‌های کاهش یافته از نطر تئوری، دقت تشخیص انسجام توسط مدل پیشنهادی و دقت تشخیص مدل پیشنهادی در مقایسه با مقدار واقعی تئوری در جدول (۴-۳) آمده است. منظور از دقت تشخیص مدل پیشنهادی در مقایسه با مقدار تئوری خروجی کسر این دو مقدار بر هم بوده و مشخص می‌کند مدل ارائه شده تا چه میزان توانسته است انسجام واقعی را ارزیابی بکند.

جدول ۴-۳: نتایج حاصل از اعمال مدل پیشنهادی و مقایسه نتیجه حاصل با دقت انسجام از نظر تئوری بر روی ده نمونه کاهش انسجام یافته از یک متن بلند

نوع کاهش انسجام	الف	ب	ج
جابجایی جملات	۹۰٪	۸۲٪	۹۱,۱۱٪
	۸۰٪	۶۹٪	۸۶,۲۵٪
	۷۰٪	۶۰٪	۸۵,۷۱٪
	۶۰٪	۵۸٪	۹۶,۶۷٪
	۵۰٪	۳۹٪	۷۸٪
کاهش اتفاقی تعداد جملات	۹۰٪ (نمونه نخست)	۷۷٪	۸۵,۵۶٪
	۹۰٪ (نمونه دوم)	۸۱٪	۹۰٪
	۸۰٪ (نمونه نخست)	۶۶٪	۸۲,۵٪
	۸۰٪ (نمونه دوم)	۶۸٪	۸۵٪
	۷۰٪	۵۲٪	۷۴,۲۹٪
دقت مدل			**۸۵,۵۱**

در این جدول متن بلند داستانی (۱۹۲ جمله‌ای) و نمونه‌های غیر منسجم آن مورد ارزیابی قرار گرفته است. در ستون (الف) متن غیر منسجم با درصد کاهش انسجام نشان داده شده که انتظار تئوری برای ارزیابی را مشخص می‌کند. به عنوان مثال در سطر اول متن موجود با ده درصد جابجایی جملات بوده که از نظر تئوری میزان انسجام آن باید نود درصد باشد. آنچه در ستون (ب) نشان داده می‌شود دقت تشخیص انسجام توسط مدل پیشنهادی بوده که از میزان دقت تئوری پایین‌تر است. این تفاوت دقت طبیعی بوده و دلیل آن همان کاهش شدید انسجام متون داستانی با جابجایی جملات و یا حذف اتفاقی

آنان در مقایسه با سایر متون بوده که در بخش‌های قبلی ذکر شد. ستون (ج) دقت تشخیص مدل پیشنهادی در مقایسه با مقدار واقعی تئوری را نشان داده که از تقسیم نسبت ستون (ب) بر ستون (الف) حاصل شده است. این مقدار فاصله مقدار ارزیابی انسجام واقعی با انسجام تئوری را نشان داده که در مقایسه با سایر رویکردهای پیشین، حالات ویژه متون داستانی و حجم نسبتا بالای متن مورد پردازش مقداربسیار بهینه‌ای است. میانگین دقت محاسبه شده (۸۵،۵۱ درصد) برای تمام حالات نشان دهنده این مهم است.

تا به حال مدل‌های مبتنی بر شبکه موجودیت و روش‌های مبتنی بر تئوری گراف‌ها دو نمونه از روش‌های مهم و پر استفاده ارزیابی انسجام متن بوده‌اند. این دو مدل به تنهایی و یا با ترکیب با سایر الگوریتم‌ها نتایج متفاوتی را ارائه داده‌اند. تا به حال کارایی بالای روش‌های مبتنی بر شبکه موجودیت در بسیاری از رویکردها بکار گرفته شده و ثابت گردیده است، اما بیشترین قدرت و تاکید آنان بر ارزیابی انسجام محلی بوده است. پس از معرفی روش‌های ترکیبی و ترکیب مدل شبکه موجودیت و گراف‌ها، رویکردهای معرفی شده توانایی ارزیابی انسجام عمومی را تا حدودی یافته و انسجام عمومی را در بخش بزرگ‌تری از متن (جملات با فاصله بیشتر) ارزیابی کرده‌اند. اما چالش روش‌های ترکیبی با بزرگ‌تر شدن محدوده ارزیابی و افزایش اندازه گراف مشخص‌تر شده و دقت حاصل کاهش می‌یابد. دلیل اصلی کاهش دقت ذکر شده، تصویر کردن گراف دو قسمتی در یک گراف بدون جهت و کاهش اطلاعات مربوط به وابستگی جملات به هم است. رویکرد پیشنهادی لیوما و تریسان با ارائه پارامترهایی قابل استخراج از خود گراف دوقسمتی و عدم نیاز به تصویر کردن آن به گراف معمولی غیر جهت دار این چالش را تا حدودی کاهش داده است [۲۰]. این روش به دلیل ترکیب دو رویکرد شبکه موجودیت و تئوری گراف، بهبود چالش رویکردهای قبلی (کاهش اطلاعات گراف دو قسمتی با تصویر کردن به یک گراف یک قسمتی بدون جهت) و همچنین ارزیابی انسجام عمومی در سطحی وسیع‌تر نسبت به رویکردهای قبلی، گزینه مناسبی جهت مقایسه با روش پیشنهادی ما و ارزیابی آن است. جهت ارزیابی مدل پیشنهادی این رساله (ECEM)[۱]، مدل مربوطه را با روش ارائه شده (BGSEG)[۲] [۲۰] توسط لیوما و تریسان مقایسه کرده‌ایم.

در این مرحله انسجام متن ۱۹۲ جمله‌ای انتخاب شده قبل به همراه هر ده نمونه کاهش یافته انسجام را با هر دو روش ECEM و BGSEG ارزیابی کرده‌ایم. با توجه به اینکه این متن جزو متن‌های بلند محسوب شده، نتیجه حاصل برتری روش ECEM را مشخص می‌کند. جدول (۴-۴) و شکل‌های (۴-۳) و (۴-۴) نتایج حاصل از اعمال هر دو روش را بر روی متن انتخابی و نمونه‌های غیر منسجم آنان نشان می‌دهد. در این جدول برای هر متن ده نمونه با انسجام کاهش یافته تولید شده که پنج نمونه اول با جابجایی جملات (ده، بیست، سی، چهل و پنجاه درصد) و پنج نمونه دوم با حذف اتفاقی جملات (دو متن با ده درصد کاهش، دو متن با بیست درصد کاهش و یک متن با سی درصد کاهش) ایجاد گردیده‌اند.

---

[۱] Embedding-based coherence evaluation model

[۲] Bipartite graph structure of entity grids

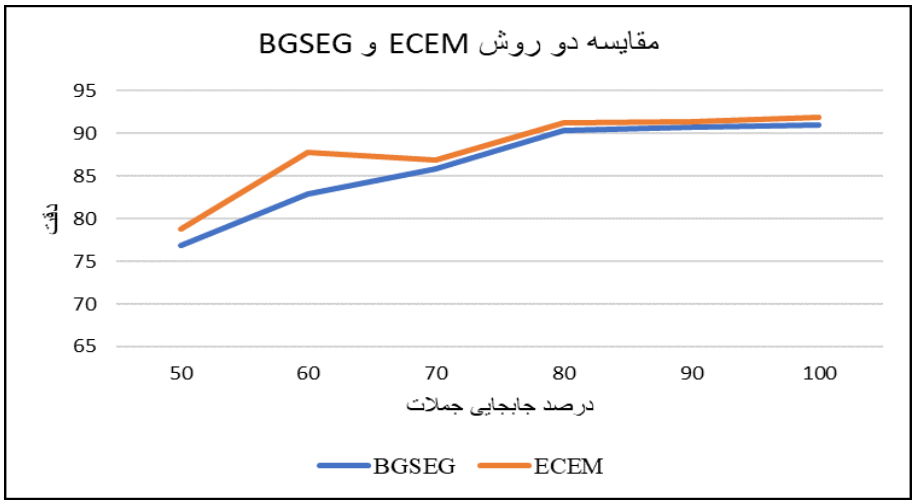
ستون "انسجام تئوری" وابستگی موضوعی جملات موجود در متن‌های تولیدی را از نظر تئوری مشخص می‌کند. دو ستون "دقت مدل BGSEG" و " دقت مدل ECEM" ارزیابی انسجام بدست آمده هر نمونه را توسط دو روش فوق نشان می‌دهد. ستون‌های "دقت واقعی مدل BGSEG" و " دقت واقعی مدل ECEM" انسجام واقعی دو روش را محاسبه کرده‌اند که این مقدار از تقسیم دقت هر مدل بر انسجام تئوری بدست آمده است. همانطور که در نتایج حاصل مشخص است مدل ECEM در نمونه‌های تولید شده با جابجایی جملات ۱.٤٦ درصد بهینه‌سازی و در نمونه‌های حذف اتفاقی جملات ۸٦ صدم درصد بهینه‌سازی دارد. در مجموع بهینه‌سازی مدل معرفی شده در هر دو شکل کاهش جملات ۱.۱٦ درصد بوده است.

جدول ۴-۴: نتایج حاصل از اعمال مدل پیشنهادی (ECEM) و مدل لیوما و تریسان(BGSEG) بر روی یک متن به همراه ده نمونه متن کاهش انسجام یافته

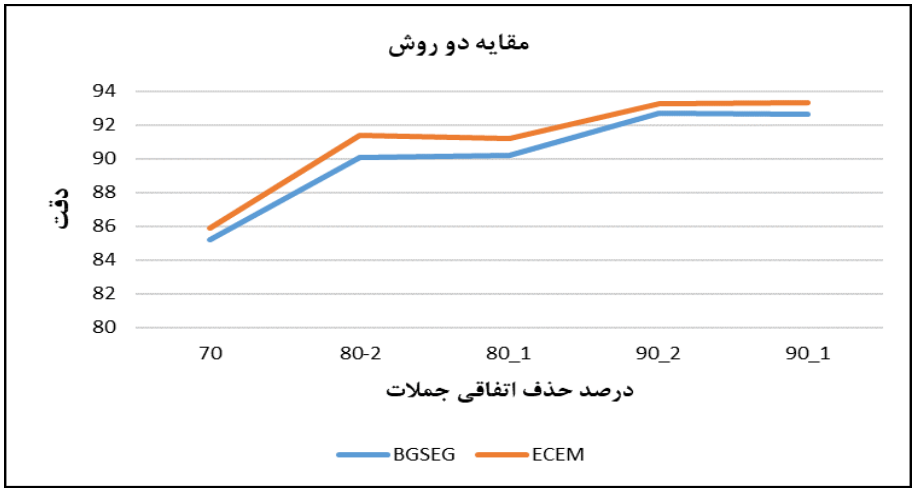
نوع کاهش انسجام	انسجام تئوری	دقت مدل BGSEG	دقت مدل ECEM	دقت واقعی مدل BGSEG	دقت واقعی مدل ECEM
متن اصلی	۱۰۰	۸۹,۱۲	۸۹,۸۷	۹۰,۰۲	۹۱,۸۷
جابجایی جملات	۹۰	۸۱,۶۸	۸۲,۲۲	۹۰,۷۶	۹۱,۳۶
	۸۰	۷۲,۲۳	۷۲,۹۸	۹۰,۲۹	۹۱,۲۳
	۷۰	۵۹,۸۷	۶۰,۱۱	۸۶,۸۷	۸۵,۸۷
	۶۰	۴۹,۷۵	۵۲,۶۶	۸۲,۹۲	۸۷,۷۷
	۵۰	۳۸,۴۵	۳۹,۲۵	۷۶,۹۰	۷۸,۸۰
میانگین جابجایی				**۸۵ .۵۵**	**۸۷ .۰۰**
کاهش اتفاقی تعداد جملات (خلاصه‌سازی اتفاقی)	۹۰(۱)	۸۳,۳۵	۸۴,۱۵	۹۲,۶۱	۹۳,۳۲
	۹۰(۲)	۸۳,۶۸	۸۳,۹۵	۹۲,۶۸	۹۳,۲۸
	۸۰(۱)	۷۲,۱۷	۷۲,۹۵	۹۰,۲۱	۹۱,۱۹
	۸۰(۲)	۷۲,۰۵	۷۳,۱۱	۹۰,۰۶	۹۱,۳۹
	۷۰	۵۹,۶۵	۶۰,۱۳	۸۵,۲۱	۸۵,۹۰
میانگین خلاصه‌سازی				۹۰,۱۵	۹۱,۰۲
میانگین کل				**۸۸ .۲۲**	**۸۹.۵**

در این جدول در ستون انسجام تئوری اندازه تغییرات ایجاد شده در متن به صورت جابجایی جملات و یا حذف آنان مشخص شده است. به عنوان مثال در بخش جابجایی جملات عدد ۹۰ مشخص کننده این است که ده درصد جملات با هم جابجا شده که از نظر تئوری انسجام متن ده درصد کاهش یافته و به میزان نود درصد رسیده است. اما مدل BGSEG این انسجام را ۸۱,۶۸ محاسبه کرده که در ستون دقت مدل BGSEG مشخص شده است. با توجه به اینکه انسجام واقعی نود درصد بوده و انسجام محاسبه شده توسط مدل نامبرده ۸۱,۶۸ بوده دقت واقعی مدل BGSEG از تقسیم نسبت دو مقدار تئوری و واقعی

بدست آمده و مقدار ۹۰,۷۶ بوده و در ستون دقت واقعی مدل BGSEG مشخص شده است. به همین صورت مقدار انسجام همان متن (با دقا انسجامی تئوری ۹۰ درصد) توسط مدل پیشنهادی این رساله ۸۲,۲۲ ارزیابی شده و در ستون دقت مدل ECEM نشان داده شده است. به همین صورت دقت واقعی مدل پیشنهادی ما نیز از تقسیم انسجام ارزیابی شده بر انسجام واقعی محاسبه شده که مقدار ۹۱,۲۳ بوده و در ستون دقت واقعی مدل ECEM مشخص گردیده است. در این جدول تمامی نمونه متن‌های غیر منسجم با جابجایی جملات (۱۰، ۲۰، ۳۰، ۴۰، ۵۰ درصد) و حذف اتفاقی جملات (۱۰، ۲۰، ۳۰ درصد) نیز محاسبه شده و دقت واقعی دو مدل در دو ستون دقت واقعی مدل BGSEG و دقت واقعی مدل ECEM نمایش داده شده است. همانطور که مشاهده می‌شود با توجه به ادعای قبلی جابجایی جملات در متون داستانی و روایی تاثیر بسیار بالاتری از کاهش اتفاقی جملات در کاهش انسجام در مقایسه با متون خبری و علمی دارند. زیرا کاهش جملات به نوعی همان عمل خلاصه سازی بوده که در متون داستانی امری طبیعی است. اما جابجایی جملات در متون داستانی دنباله مفاهیم روایی داستان را بر هم زده، موجب کاهش شدید انسجام آن شده که در متون خبری و علمی شدت تاثیر آن به این حد نیست.



شکل ۴-۳: مقایسه دو مدل با متن‌های غیر منسجم با جملات جابجا شده



شکل ۴-۴: مقایسه دو مدل با متن‌های غیر منسجم با جملات اتفاقی حذف شده

در مرحله بعد ارزیابی و مقایسه دو روش را بر روی پایگاه داده ایجاد شده انجام داده‌ایم. بدین منظور ده متن با اندازه‌های متفاوت به همراه ده نمونه کاهش یافته انسجام هر کدام را انتخاب کرده و هر دو روش را بر روی آنان اعمال کردیم. با انجام این آزمایش برتری روش پیشنهاد شده بر روی متن‌های بزرگ‌تر مشخص شده و کارایی و دقت بالاتر آن در متن‌های با تعداد جمله بالا نشان داده شده است. از ده متن انتخابی چهار متن کوتاه، سه متن متوسط و سه متن بلند و با تعداد جملاتی زیادتر است. با توجه به آزمایش انجام شده دقت برای متن‌های کوتاه و با تعداد جملات کمتر مدل BGSEG در مقایسه با مدل پیشنهادی (ECEM) از دقت بیشتری برخوردار بوده و در متن‌های متوسط تفاوت چندانی ندارد. اما برتری مدل پیشنهادی در متن‌های بلند و با تعداد جملات بیشتر نمایان شده و برتری ۱,۹۵ درصدی را نشان می‌دهد. میانگین دقت مدل معرفی شده در کل متن‌های مورد ارزیابی نیز بیشتر بوده و بر روی گروه متن‌های کوتاه و بلند نیز ۰,۴۱ درصد بهبود را نمایش می‌دهد. این مقادیر درجدول و نمودار ۴-۵ نمایش داده شده است.

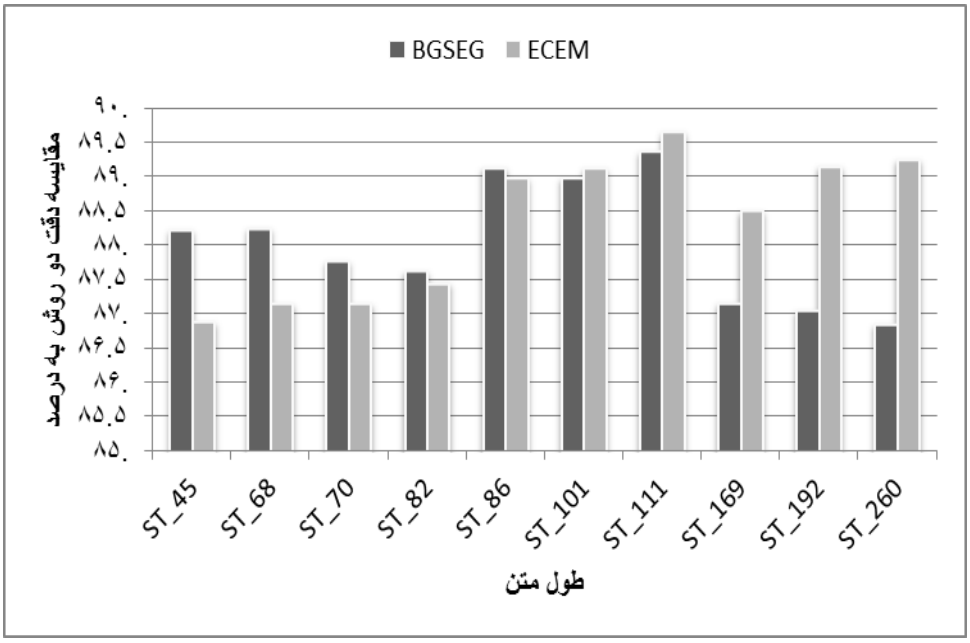
جدول ۴-۵: مقایسه دو مدل BGSEG و ECEM

تعداد جملات متن	دقت مدل BGSEG	دقت مدل ECEM
۴۵	۸۸,۲۲	۸۶,۸۸
۶۸	۸۸,۲۵	۸۷,۱۵
۷۰	۸۷,۷۷	۸۷,۱۵
۸۲	۸۷,۶۲	۸۷,۴۵
میانگین	**۸۷,۹۷**	**۸۷,۱۶**
۸۶	۸۹,۱۳	۸۸,۹۸
۱۰۱	۸۸,۹۸	۸۹,۱۲
۱۱۱	۸۹,۳۶	۸۹,۶۵

میانگین	**۸۹,۱۶**	**۸۹,۲۵**
۱۶۹	۸۷,۱۶	۸۸,۵۰
۱۹۲	۸۷,۰۵	۸۹,۱۵
۲۶۰	۸۶,۸۵	۸۹,۲۵
میانگین	**۸۷,۰۲**	**۸۸,۹۷**
میانگین کل	**۸۸,۰۵**	**۸۸,۴۶**

۴-۵ نتیجه گیری

در این فصل، به ارزیابی و مقایسه مدل معرفی شده با یکی از مدل‌های معروف پرداخته شد. مدل قابل مقایسه (BGSEG) یکی از روش‌های مورد تایید درحوزه ارزیابی انسجام محلی و عمومی به طور همزمان بوده که از ترکیب دو رویکرد مبتنی بر موجودیت و مبتنی بر گراف استفاده کرده است. رویکرد مبتنی بر موجودیت از روش‌های پایه‌ای و اصلی ارزیابی انسجام بوده که قدرت آن در ارزیابی انسجام محلی اثبات شده است. رویکردهای مبتنی بر گراف نیز در ترکیب با روش‌های مبتنی بر موجودیت برای ارزیابی انسجام عمومی بکار گرفته شده که به دلیل کاهش حجم گراف در تصویر کردن گراف دو قسمتی حاصل از دقت پایینی برخوردار بودند. اما روش BGSEG با عدم تصویر سازی گراف دوقسمتی دقت مدل خود را بالا برده و یکی از بهترین رویکردهای پیشنهادی تا به امروز در حوزه ارزیابی انسجام عمومی و محلی در متن‌های خبری و کوتاه بوده است. اما به دلیل این عدم کاهش حجم گراف در متون بزرگ با مشکل محاسبات سنگین و پیچیدگی زمانی و حافظه‌ای بالا درگیر می‌شد. روش معرفی شده در این رساله با عدم نیاز به موارد ذکر شده از نتیجه بهتری در متن‌های با طول بیشتر و داستانی داشته است. به صورتی که در متن‌های بلند و با بیش از دویست جمله از بهینه سازی نتیجه‌ای درخور توجه داشته است. نتیجه یکی از آزمایشات بر روی متن دویست و شصت جمله‌ای با دو روش پیشنهادی و BGSEG بهینه‌سازی ۲,۴ درصدی را نشان می‌دهد.



شکل ۴-۵: مقایسه دو مدل BGSEG و ECEM

پیوست

در این بخش به ارائه یک مثال در مورد الگوی طراحی شده می‌پردازیم. متن زیر بخشی از یک داستان از نمونه داستان‌های هانس کریستین آندرسن موجود در پایگاه داده تولید شده است (۲۶۰ جمله). این متن با ده درصد جابجایی (جابجایی اتفاقی ۲۶ جمله در کل متن) تبدیل به یک متن با ده درصد کاهش انسجامی تبدیل شده است:

There was once a woman who wished very much to have a little child. She went to a fairy and said: “I should so very much like to have a little child. Can you tell me where I can find one?” “Oh, that’s no problem,” said the fairy. “Here is a seed. It is not like most seeds – plant it into a flowerpot and see what will happen.” “Thank you,” said the woman and she paid the fairy. Then she went home and planted it, and in no time up grew a lovely large flower, like a tulip but with its leaves tightly closed, as if it were still a bud. “It is a beautiful flower,” said the woman, and she kissed the red and golden-colored petals. As she did so the flower opened and she could see that it was a real tulip. But within the flower, sat a very delicate and graceful little maiden. She was hardly half as long as a thumb so they gave her the name of Little Thumb, or Thumbelina, because she was so small. A polished walnut shell was used for her cot, her bed was made of the leaves of violets and her bedcover was a rose-leaf. Here she slept at night, but during the day she played on a table, where the peasant wife had placed a plate full of water.

……..

سپس پیش‌پردازش‌های لازم بر روی ان اعمال شده، هرجمله به یک لیست از واژگان با امتیاز بالا جهت مقایسه و اعمال فاصله گذر بین جملات تبدیل شده است. در ادامه چهار لیست اول مشخص شده است:

['there', 'woman', 'wish', 'much', 'little', 'child']

['she', go, 'fairy', say, 'I', 'much', 'like', 'little', 'child']

['tell', 'find', 'one']

['oh', 'easily', 'manage', say, 'fairy']

……..

قطعه کد زیر برنامه پیش‌پردازش کننده وتولید کننده این لیست‌ها به زبان پایتون را نشان می‌دهد:

```
import nltk
import numpy as np
import pandas as pd
nltk.download("stopwords")
from nltk.tokenize import sent_tokenize
```

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
my_Lemmatizer=WordNetLemmatizer()
#ِDefine variables and data types----------
sentence_List=[]
allWords_List=[]
nonRepeadedWordsList=[]
#Open files----------
file=open("a5.txt")
text=file.read()
text=text.lower()
#Unwanted Characters Deleting----------
unwanted_alphabet=["'", ",", "", '!', ';', '“', '”','-']
for alpha in unwanted_alphabet:
    text=text.replace(alpha, "")
#Text segmentation & Stopword Removing----------
sentences=sent_tokenize(text)
for x in range(len(sentences)):
    sentences[x]=word_tokenize(sentences[x])
    for wsent in range(len(sentences[x])):
        sentences[x][wsent]=my_Lemmatizer.lemmatize(sentences[x][wsent], pos='v')
        allWords_List.append(sentences[x][wsent])
    sentences[x]=[w for w in sentences[x] if w not in stopwords.words('english')]
    sentence_List.append(sentences[x])
#Print preprocessed and normalized types in text--------
allWords_List=[w for w in allWords_List if w not in stopwords.words('english')]
nonRepeadedWordsList=list(set(allWords_List))
print(nonRepeadedWordsList)
```

مراجع

[۱] خراسانی ا. و علی‌نژاد ح. (۱۳۹۴) "بررسی عناصر انسجام متن در نفثه‌المصدور بر اساس نظریّهٔ هالیدی و حسن" *متن پژوهی ادبی*، شماره ۱۹، دوره ۶: ص ۷-۳۱.

[۴۰] شریفی ع. مهدوی م ا. (۱۳۹۷) "رویکردی با ناظر در استخراج واژگان کلیدی اسناد فارسی با استفاده از زنجیره‌های لغوی" *فصل نامه پردازش علائم و داده‌ها*، شماره ۴، دوره ۳۸.

[۲۷] قاسم زاده، م. (۱۳۹۶) روش پژوهش نگارش و ارائه در علوم مهندسی کامپیوتر. چاپ اول. انتشارات دانشگاه یزد، ص ۹۲.

[2] M. A. K. Halliday M.A. K, Hasan R (1976) "Cohesion in English," ed: London: Longman.

[3] Crossley S. A, Kyle K and McNamara D. S. (2016) "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion," Behavior research methods. 48, 4, pp 1227-1237.

[4] Mikolov. T, Sutskever. I, Chen. K, Corrado. G, and Dean. J., (2013) "Distributed representations of words and phrases and their compositionality", in Advances in neural information processing systems, pp. 1-9.

[5] Yannakoudakis. H and Briscoe. T., (2012) "Modeling coherence in ESOL learner texts," in Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 33-43: Association for Computational Linguistics.

[6] Burstein. J, Tetreault. J, and Andreyev. S., (2010) "Using entity-based features to model coherence in student essays," in Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics, pp. 681-684: Association for Computational Linguistics.

[7] Higgins. D, Burstein. J, Marcu. D, and Gentile. C., (2004) "Evaluating multiple aspects of coherence in student essays," in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004.

[8] Lin. Z, Liu. C, Ng. H. T, and Kan. M. Y., (2012) "Combining coherence models and machine translation evaluation metrics for summarization evaluation," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 1006-1014: Association for Computational Linguistics.

[9] Xiong. D, Ding. Y, Zhang. M, and Tan. C. L., (2013) "Lexical chain based cohesion models for document-level statistical machine translation," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1563-1573.

[10] Fox. H. J., (2002) "Phrasal cohesion and statistical machine translation, " in Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 304-3111: Association for Computational Linguistics.

[11] Celikyilmaz. A and Hakkani-Tür. D., (2011) "Discovery of topically coherent sentences for extractive summarization," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 491-499: Association for Computational Linguistics.

[12] Parveen. D, and Strube. M., (2015) "Integrating Importance, Non-Redundancy and Coherence in Graph-Based Extractive Summarization," Twenty-Fourth International Joint Conference on Artificial Intelligence, pp. 1298-1304.

[13] Zhang. R., (2011) "Sentence ordering driven by local and global coherence for summary generation," in Proceedings of the ACL 2011 Student Session, pp. 6-11: Association for Computational Linguistics.
[14] Severyn. A, and Moschitti. A., (2016) "Modeling relational information in question-answer pairs with convolutional neural networks", arXiv preprint arXiv :1604.01178.
[١٥] Putra. J. W. G and Tokunaga. T., (2017) "Evaluating text coherence based on semantic similarity graph," in Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, pp. 76-85.
[١٦] Lee. G. H, and Lee. K. J., (2017) "Automatic Text Summarization Using Reinforcement Learning with Embedding Features," in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, pp. 193-197.
[١٧] Jacobmeyer. H., (2004) "Graham Swift, Ever After: A Study in Intertextuality," webdoc. sub. gwdg. de.
[١٨] Oddo J. (2013) "Precontextualization and the rhetoric of futurity: Foretelling Colin Powell's UN address on NBC News." Discourse & Communication, 7, 1, pp 25-53.
[١٩] Kusner. M, Sun. Y, Kolkin. N, and Weinberger. K., (2015) "From word embeddings to document distances," in International Conference on Machine Learning, pp. 957-966.
[٢٠] Lioma. C, Tarissan. F, Simonsen. J. G, Petersen. C, and Larsen. B, (2016) "Exploiting the bipartite structure of entity grids for document coherence and retrieval," in Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, pp. 11-20: ACM.
[٢١] Severyn. A, and Moschitti. A., (2015) "Learning to rank short text pairs with convolutional deep neural networks", in Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, ACM.
[٢٢] Barzilay R and Lapata M. (2008) "Modeling local coherence: An entity-based approach ",Computational Linguistics, 34, 1, pp 1-34.
[٢٣] https://en.wikipedia.org/wiki/Paragraph/.
[٢٤] Kalchbrenner. N, Grefenstette. E, and Blunsom. P., (2014) "A convolutional neural network for modelling sentences," arXiv preprint arXiv:1404.2188.
[٢٥] Karuna. P, Purohit. H, Uzuner. O, Jajodia. S, and Ganesan. R., (2018) "Enhancing Cohesion and Coherence of Fake Text to Improve Believability for Deceiving Cyber Attackers", Proceedings of the First International Workshop on Language Cognition and Computational Models, pp. 31–40.
[٢٦] Cardenas. R, Bello. K, Coronado. A, and Villota. E., (2018) "Improving Topic Coherence Using Entity Extraction Denoising," The Prague Bulletin of Mathematical Linguistics, 110, 1, pp 85-101.
[٢٨] Rosenfeld. R., (1996) "A maximum entropy approach to adaptive statistical language modeling," Computer Speech & Language, 10, 3, pp 187-228.
[٢٩] Luhn H. P. (1958) "A business intelligence system" IBM Journal of Research and Development. 2, 4, pp 314-319.
[٣٠] Foltz P. W, Kintsch W and Landauer. T. K. (1998) "The measurement of textual coherence with latent semantic analysis," Discourse processes. 25, 2-3, pp 285-307.
[٣١] Barzilay. R and Lapata. M., (2005) "Modeling local coherence: An entity-based approach", In Proceedings of the ACL, pp. 141–148.
[٣٢] Zhang. M, Feng. V. W, Qin. B, Hirst. G, Liu. T, and Huang. J, (2015), "Encoding world knowledge in the evaluation of local coherence," in Proceedings of the 2015

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1087-1096.
[٣٣] Mohiuddin. T, Joty. S, and Nguyen. D. T., (2018) "Coherence Modeling of Asynchronous Conversations: A Neural Entity Grid Approach", arXiv:1805.02275v1 [cs.CL.
[٣٤] Nguyen. D. T and Joty. S, (2017) "A neural local coherence model," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1320-1330.
[٣٥] Xu. F, Du. S, Li. M, and Wang. M, (2017) "An entity-driven recursive neural network model for chinese discourse coherence modeling," arXiv preprint arXiv:1704.04336.
[٣٦] Blanco. R and Lioma. C. (2012) "Graph-based term weighting for information retrieval," Information retrieval, 15, 1, pp. 54-92.
[٣٧] Guinaudeau. C and Strube. M., (2013) "Graph-based local coherence modeling," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 93-103.
[٣٨] Petersen. C, Lioma. C, Simonsen. J. G, and Larsen. B., (2015) "Entropy and graph based modelling of document coherence using discourse entities: An application to IR," in Proceedings of the 2015 International Conference on The Theory of Information Retrieval, pp. 191-200: ACM.
[٣٩] Mesgar. M, and Strube. M., (2015) "Graph-based coherence modeling for assessing readability," in Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, pp. 309-318.
[٤١] Somasundaran. S, Burstein. J, and Chodorow. M., (2014) "Lexical chaining for measuring discourse coherence quality in test-taker essays," in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 950-961.
[٤٢] Htet Myet. L, Chang. C, and Kim. P., (2018) "An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms", Soft Computing, 22, 12, pp 4013- 4023.
[٤٣] Li. J, and Jurafsky. D., (2016) "Neural net models for open-domain discourse coherence," arXiv preprint arXiv:1606.01545.
[٤٤] Logeswaran. L, Lee. H, and Radev. D., (2016) "Sentence ordering using recurrent neural networks," arXiv preprint arXiv:1611.02654.
[٤٥] Logeswaran. L, Lee. H, and Radev. D., (2018) "Sentence Ordering and Coherence Modeling using Recurrent Neural Networks," Thirty-Second AAAI Conference on Artificial Intelligence.
[٤٦] Kiddon. C, Zettlemoyer. L, and Choi. Y., (2016) "Globally coherent text generation with neural checklist models," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 329-339.
[٤٧] Kim. Y., (2014) "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882.
[٤٨] Zhang. X, and LeCun. Y., (2015) "Text understanding from scratch," arXiv preprint arXiv:1502.01710.

[٤٩] Firth. R., (1957) "2. A Note on Descent Groups in Polynesia," Man, 57, pp 4-8.
[٥٠] Johnson. R, and Zhang. T., (2014) "Effective use of word order for text categorization with convolutional neural networks," arXiv preprint arXiv:1412.1058.
[٥١] Johnson. R, and Zhang. T., (2015) "Semi-supervised convolutional neural networks for text categorization via region embedding," in Advances in neural information processing systems, pp. 919-927.
[٥٢] Nguyen. T. H, and Grishman. R., (2015) "Relation extraction: Perspective from convolutional neural networks," in Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pp. 39-48.
[٥٣] Bakarov. A., (2018) "A Survey of Word Embeddings Evaluation Methods," arXiv preprint arXiv:1801.09536.
[٥٤] Zhang. Y, and Wallace. B., (2015) "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification ", arXiv preprint arXiv:1510.03820.
[٥٥] Pennington. J, Socher. R, and Manning. C., (2014) "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543.
[56] Christensen. J, Soderland. S, and Etzioni. O., (2013) "Towards coherent multi-document summarization," in Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 1163-1173.
[57] Zhang. R, Li. W, Liu.N and Gao. D (2016) "Coherent narrative summarization with a cognitive model," Computer Speech & Language, 35, pp 134-160.
[58] Gambhir. M and Gupta. V (2017) "Recent automatic text summarization techniques: a survey," Artificial Intelligence Review, 47, 1, pp 1-66.
[59] Barzilay. R, and Yen Kan. M., (2001) "SIMFINDER: A Flexible Clustering Tool for Summarization", In Proceedings of the NAACL Workshop on Automatic Summarization, pp. 41-49.
[60] Alonso. L, and Fuentes. M., (2003) "Cohesion and Coherence for Automatic Summarization", Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, pp. 1-8.
[61] Anjaneyulu. M, Sarma. S. S. V. N, Vijaya Pal Reddy. P, Prem Chander. K, and Nagaprasad. S., (2018) "Sentence Similarity Using Syntactic and Semantic Features for Multi-document Summarization", International Conference on Innovative Computing and Communications.
[62] Liu. P, Saleh. M, Pot. E, Goodrich. B, Sepassi. R, Kaiser. L, and Shazeer. N., (2018) "Generating wikipedia by summarizing long sequences," arXiv preprint arXiv:1801.10198.
[63] Han. A. L. F, and Wong. D. F., (2016) "Machine translation evaluation: A survey," arXiv preprint arXiv:1605.04515.
[64] Sim Smith. K. M., (2017) "Coherence in Machine Translation," Department of Computer Science, University of Sheffield.
[65] Xiong. D, and Zhang. M., (2013) "A Topic-Based Coherence Model for Statistical Machine Translation," Twenty-Seventh AAAI Conference on Artificial Intelligence.
[66] Smith. K. S, Aziz. W, and Specia. L., (2015) "A proposal for a coherence corpus in machine translation," in Proceedings of the Second Workshop on Discourse in Machine Translation, pp. 52-58.

[67] Xiong. D, Zhang. M, and Wang. X., (2015) "Topic-based coherence modeling for statistical machine translation," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 23, 3, pp 483-493.
[68] Wong. B, and Kit. C., (2012) "Extending machine translation evaluation metrics with lexical cohesion to document level," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1060-1068: Association for Computational Linguistics.
[69] Smith. K. S, Aziz. W, and Specia. L., (2016) "The Trouble with Machine Translation Coherence," in Proceedings of the 19th Annual Conference of the European Association for Machine Translation, pp. 178-189.
[70] Xiong. H, He. Z, Wu. H, and Wang. H., (2019) "Modeling Coherence for Discourse Neural Machine Translation", The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), pp. 7338- 7345.
[71] Knott. A, Dale. R., (2005) "Choosing a set of coherence relations for text generation: a Data Driven- approach", European Workshop on Trends in Natural Language Generation, pp. 47-67.
[72] Ferreira. T. C, Krahmer. E, and Wubben. S., (2016) "Towards more variation in text generation: Developing and evaluating variation models for choice of referential form," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 568-577.
[73] Zhang. Y, Gan. Z, Fan. K, Chen. Z, Henao. R, Shen. D, and Carin. L., (2017) "Adversarial feature matching for text generation," Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 4006-4015.
[74] Cho. W, Zhang. P, Zhang. Y, Li. X, Galley. M, Brockett. C, Wang. M, and Gao. J., (2019) "Towards coherent and cohesive long-form text generation", Proceedings of the First Workshop on Narrative Understanding, pp. 1–11, Association for Computational Linguistic.
[75] Siddharthan. A (2014) "A survey of research on text simplification," ITL-International Journal of Applied Linguistics,165, 2, pp 259-298.
[٧٦] Ma. S, and Sun. X (2017) "A semantic relevance based neural network for text summarization and text simplification," arXiv preprint arXiv:1710.02318.
[٧٧] Sulem. E, Abend. O, and Rappoport. A., (2018) "Semantic structural evaluation for text simplification," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), vol. 1, pp. 685-696.
[٧٨] Chandrasekar. R, and Srinivas. B., (1996) "Automatic Induction of Rules for Text Simplification", Knowledge-Based Systems, 10, 3, pp 183–190.
[٧٩] Siddharthan. A (2006) "Syntactic simplification and text cohesion," Research on Language and Computation, 4, 1, pp 77-109.
[٨٠] Leroy. G and Kauchak. D (2013) "A user study measuring the effect of lexical simplification and coherence enhancement on perceived and actual text difficulty", International Journal of Medical Informatics, pp 717-730.
[٨١] Van den Bercken. L, Sips. R. J, and Lofi. C., (2019) "Evaluating Neural Text Simplification in the Medical Domain", In The World Wide Web Conference, pp. 3286-3292, ACM.
[٨٢] Song. W, Fu. R, Liu. L, and Liu. T., (2015) "Discourse element identification in student essays based on global and local cohesion," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2255-2261.

[٨٣] Huang. G, Tan. M, Huang. S, Mo. R, and Zhou. Y., (2017) "A discourse coherence model for analyzing Chinese students' essay," in Progress in Informatics and Computing (PIC), 2017 International Conference on, pp. 430-434: IEEE.
[٨٤] Louis. A, and Nenkova. A., (2012) "A coherence model based on syntactic patterns," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1157-1168: Association for Computational Linguistics.
[85] Vijayarani. S, Ilamathi. M. J and Nithya. M., (2015) "Preprocessing techniques for text mining-an overview," International Journal of Computer Science & Communication Networks, 5, 1, pp 7-16.
[86] Denny. M. J and Spirling. A (2018) "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it," Political Analysis, 26, 2, pp 168-189.
[87] Simard. M (1998) "Automatic insertion of accents in French text," in Proceedings of the Third Conference on Empirical Methods for Natural Language Processing, pp. 27-35.
[88] Iyyer. M, Manjunatha. V, Boyd-Graber. J and Daume. H (2015) “Deep unordered composition rivals syntactic methods for text classification”, In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1681-1691.
[89] Wieting. J, Bansal. M, Gimpel. K and Livescu. K (2016) “Embedding words and sentences via character n-grams”, In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, pp. 1504–1515.
[90] Ermakova. L, Mothe. J, and Firsov. A., (2017) "A Metric for Sentence Ordering Assessment Based on Topic-Comment Structure," in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1061-1064: ACM.
[91] Arunsirot. S (2013) "An Analysis of Textual Metafunction in Thai EFL Students' Writing," Novitas-ROYAL (Research on Youth and Language), 7, 2, pp 160-174.

Abstract

Discourse coherence modeling evaluation is an important subfield in Natural Language Processing. In recent years, there has been an increasing interest but challenging task in text coherence evaluation. Document coherence evaluation methods are divided into two main categories of local and global coherence evaluation. Using automatic methods for evaluating or increasing the quality of coherence is considered the most important goal of all text processing systems such as document summarization, text generation, text simplification, statistical machine translation, mode detection, question answering, student essay scoring, produced documents by unskilled people and combined topic texts by unskilled persons. Therefore, all of the machine-driven NLP tasks tend to measure the coherence in order to improve their processing algorithm.

In recent years, there have been several investigations into text coherence evaluation. It is also high-quality systems are designed with the ability to produce very close texts to human written. However, most of proposed models are engaging with semantic and linguistic concepts of text. The most important challenge of them is limiting to a particular area, lack of applicability and expansion into other languages, complex algorithms and inaccuracies. Most of the previous approaches require strong assumptions and specific features to evaluate the coherence. Discovering the different text sections relationship and features selection are often has been done by users. Most proposed methods often assess local coherence limited to only a few adjacent sentences. Their accuracy in evaluating global coherence, especially in long documents, is not acceptable and low accuracy. Previous important and existing approaches, such as entity and graph-based models, are much involved with semantic and linguistic concepts. By limiting themselves to available word co-occurrence information in sequential sentences within a short part of a text, these methods have engaged with inaccuracy in public coherence evaluation. It is also there is few offered approaches that evaluated local and global coherence simultaneously. Methods which evaluate local and global coherence concurrently, only had an acceptable local coherence accuracy and do not have global coherence precision. One of their greatest challenges is their limitation on long text coherence evaluation and suitable for low number sentences documents.

In this thesis, we attempt to assess the coherence and sentences dependency in whole text using statistical approaches and text hidden knowledge. Using Google's word2vec algorithm, the proposed approach converts words into numeric vectors and sentences into numeric matrices. Applying statistical approaches based on recent results in word embeddings, presented method introduces a simple and efficient model called "ECEM" and studies how to incorporate the external word correlation knowledge to assess both local and global coherence simultaneously. It is also assessing the local topic integrity of text at the paragraph level regardless of word meaning and handcrafted rules. The global coherence in proposed method is evaluated by sequence paragraph dependency. The most important feature of the proposed model is the ability to simultaneously assess high precision local and global coherence in large and high-number sentences. The combined presented local and global coherence evaluation method does not depend on subject matter and words concept, and has the ability to extend and apply to other languages

Keywords: Text coherence, local coherence, global coherence, word vector space, language models.

Shahrood University of

Technology

Faculty of Computer Engineering

Ph.D. Thesis in Artificial Intelligence Engineering.

A new model for text coherence evaluation using statistical characteristics

By: Mohamad Abdolahi

Supervisor:
Dr. Morteza Zahedi

Advisor:
Dr. Hoda Mashayekhi

January, 2020