Journal of Electronical & Cyber Defence
Vol. 9, No. 4, 2022, Serial No. 36

Automatic Persian Text Generation Using Rule-Based Models and Word Embedding

O. Hajipoor[1], S. Sadat Sadidpour[2*]

* Malek-Ashtar University of Technology, Tehran, IR Iran

(Received: 27/01/2021, Accepted: 04/04/2021)

ABSTRACT

Natural language generation comes from natural language processing. Natural language is generated from a machine system such as a knowledge base. Although NLG systems have been around for a long time, the commercial applications of this technology have recently increased. In NLG, the system needs to decide how to put a concept into words. The ability to create meaningful text plays a key role in many natural language processing applications such as machine translation, speech and image-to-text conversions. The aim of this paper is to provide a method for generating text using artificial intelligence methods with the correct structure and starting point for generating Persian (Farsi) texts. In other words, the method presented in this article can produce various long Persian texts, maintaining the intended meaning and the Persian language structure. In order to advance the generation of text, an attempt has been made to use a combination of machine learning methods with probabilistic models. In the proposed model, probabilistic models are used to extract the rules and Word2vec is used to embed the text, and then in the generation phase, a combination of the two and a cosine distance are used. The results indicate the presentation of a model whose generation text has the appropriate structure, concept and variety. This model is also optimal in terms of ergonomics and complexity .

Keywords: Natural language generation, automatic text generation, language model, rule-based method, Word Embedding

* Corresponding Author Email: sadidpour@mut.ac.ir

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Publisher: Imam Hussein University © Authors

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۴، زمستان ۱۴۰۰، ص ۵۴- ۴۳

علمی- پژوهشی

تولید خودکار متن فارسی با استفاده مدل‌های مبتنی بر قاعده و تعبیه واژگان

امید حاجی‌پور[۱]، سعیده سادات سدیدپور[۲*]

۱- دانشجوی دکتری هوش‌مصنوعی، دانشگاه صنعتی امیرکبیر ۲- استادیار، دانشگاه صنعتی مالک‌اشتر،تهران، ایران

(دریافت: ۱۳۹۹/۱۱/۰۸، پذیرش: ۱۴۰۰/۰۱/۱۵)

چکیده

تولید زبان طبیعی از پردازش زبان طبیعی حاصل می‌شود. زبان طبیعی از یک سیستم ارائه ماشینی مانند پایگاه دانش تولید می‌شود. سیستم‌های NLG از مدت‌ها پیش وجود داشته اما فنّاوری آن به صورت ابزار تجاری اخیراً به‌صورت گسترده به وجود آمده است. در NLG، سیستم نیاز به تصمیم‌گیری در مورد چگونگی قرار دادن یک مفهوم در کلمات دارد. توانایی ایجاد متن معنی‌دار نقش کلیدی در بسیاری از کاربردهای پردازش زبان طبیعی مانند ترجمه ماشین، گفتار و تبدیل عکس به متن دارد. هدف این پروژه ارائه روشی برای تولید متن با استفاده از روش‌های هوش مصنوعی و با ساختار درست و آغازی برای تولید متن فارسی است. به عبارت دیگر در این مقاله روشی ارائه شده که قادر به تولید متن طولانی متنوع علاوه بر حفظ معنا و ساختار در زبان فارسی می‌باشد. جهت پیشبرد تولید متن سعی شده از ترکیب روش‌های یادگیری ماشین با مدل‌های احتمالاتی، استفاده شود. در مدل پیشنهادی از مدل‌های احتمالاتی برای استخراج قوانین و از Word2vec برای برداری‌سازی متن استفاده شده و سپس در فاز تولید از ترکیب این دو و فاصله کسینوسی استفاده می‌شود. نتایج نشان‌دهنده ارائه مدلی بوده که متن تولیدی آن دارای ساختار، مفهوم و تنوع مناسب می‌باشد. همچنین این مدل از نظر انسانی و پیچیدگی نیز بهینه می‌باشد.

کلید واژه‌ها: تولید زبان طبیعی، تولید خودکار متن، مدل زبانی، روش مبتنی بر قاعده، تعبیه کلمات

۱- مقدمه

در سال‌های اخیر تلاش‌هایی برای تولید متن[۱] در زبان‌های انگلیسی و چینی شده است اما محققین هنوز در ابتدای راه تولید متن با استفاده از روش‌های یادگیری عمیق هستند. چالش‌های متن فراتر از تصویر است، در متن مفاهیمی وجود دارد که نوشته نشده است و حتی برای خود انسان نیز دارای ابهام است.

درصورت وجود یک سیستم تولیدکننده برای زبان طبیعی و متن مصنوعی، می‌توان یک فضای وسیع از فرصت‌های باز را متصور شد. به‌عنوان‌مثال؛ ربات چت کننده می‌تواند با یک انسان در یک روش غیر رباتیک صحبت کند که بسیار نزدیک‌تر به شیوه‌ای است که انسان با دیگران صحبت می‌کند. در این راستا که روش‌های جدید در تولید متن بتوانند مانند تجزیه و تحلیل تصاویر مؤثر باشند، همچنان یک سؤال باز است.

از تولید متن برای کاربردهای زیادی استفاده می‌شود. به‌عنوان نمونه در ترجمه‌ی ماشینی یعنی ترجمه از زبانی به زبان دیگر که نمونه‌ای از کاربرد متن به متن[۲] است، کاربرد دارد. به‌عنوان نمونه‌ای دیگر خلاصه‌سازی[۳] از دیگر کاربردهای متن به متن در فرایند تولید متن می‌باشد که در آن متنی به هر اندازه به‌عنوان ورودی به سیستم داده می‌شود و متنی با اندازه‌ی محدود به‌عنوان خلاصه ارائه می‌گردد. همچنین سیستم‌های پرسش و پاسخ[۴] که پرسشی را به‌عنوان ورودی گرفته و پاسخی متناسب با آن ارائه می‌کنند از دیگر کاربردهای متن به متن در تولید متن می‌باشد. اما تولید متن به کاربردهای متن به متن محدود نمی‌شود. به‌عنوان نمونه می‌توان به ایجاد عنوانی مناسب برای یک تصویر[۵] اشاره کرد. این فرایند نمونه‌ای از کاربردهای تصویر به متن[۶] تولید متن می‌باشد.

تقاضای متون زبان طبیعی که هر نوع اطلاعاتی را ارائه می‌دهند، در حال افزایش است؛ بنابراین احتمال دارد که NLG در آینده فناوری کلیدی باشد (شاخص خوبی از این تعداد قابل‌توجهی از شرکت‌های NLG است که در سال‌های اخیر ظهور کرده‌اند). به‌عنوان یک نتیجه، بسیاری از سیستم‌های NLG یک کاربرد عملی پیدا کرده‌اند، درحالی‌که تقاضای برنامه‌های کاربردی زندگی واقعی، تأثیر روبه‌رشد در رویکردها و سؤالات مورد نظر در زمینه NLG دارد.

*رایانامه نویسنده مسئول: sadidpour@mut.ac.ir

[1] text generation
[2] text to text
[3] summarization
[4] question answering
[5] image captioning
[6] image to text

* این مقاله یک مقاله با دسترسی آزاد است که تحت شرایط و ضوابط مجوز Creative Commons Attribution (CC BY) توزیع شده است.
© نویسندگان
ناشر: دانشگاه جامع امام حسین (ع)

به‌طورکلی هدف تولید خودکار متن تعامل انسان با ماشین بوده و در این وظیفه از پردازش زبان طبیعی، سیستم می‌خواهد به صورت کاملاً خودکار متنی را تولید کند که متن تولید شده توسط انسان نزدیک بوده و دارای ۴ ویژگی مهم باشد: ۱- ساختار مناسب، ۲- گرامر مناسب، ۳- مفهوم مناسب و ۴- نوآوری و تنوع. به طور معمول در اکثر روش‌ها سه مورد اول با مورد تنوع همواره در تعارض بوده و هرجا متن تولید شده دارای ساختار، گرامر و معنای خوب باشد، از تنوع و نوآوری پایین برخوردار است و برعکس. به‌عنوان‌مثال در روش‌های مبتنی بر الگو که در بخش ۲ توضیح داده می‌شود، استفاده از یک الگوی خاص که تنها کلمات در آن جایگزین می‌شوند دارای گرامر و معنای درست می‌باشد، اما متن تولیدی با متنی که الگو از روی آن ساخته شده تفاوت چندانی نداشته و به‌عبارت‌دیگر تنوع آن پایین است.

در این مقاله یک سیستم مبتنی بر قاعده (قانون) و با تکیه بر درک مفاهیم توسط Word2vec [۱] که برای تعبیه‌سازی[1] می‌باشد، ارائه شده که این دو را با هم ترکیب کرده و متن‌هایی دارای ساختار، گرامر و معنای مناسب تولید کرده و تنوع را نیز در متن‌های تولیدی لحاظ می‌کند. همچنین این سیستم قادر به تولید هر نوع جمله (طولانی یا کوتاه) می‌باشد.

تولید متن یک نوع ترجمه ماشینی است، به این صورت که با توجه به پیکره متنی منبع S، متن ترجمه‌شده T در زبان مقصد، T_{best} به عنوان مناسب‌ترین ترجمه انتخاب می‌شود، اگر که احتمال P(T|S) را حداکثر کند [۲].

$$T_{best} = \arg\max_T P(T|S) = \arg\max_T P(S|T)P(T) \quad (۱)$$

که در آن P(S|T) نشان‌دهنده مدل استفاده شده برای جایگزینی کلمات یا عبارات در یک زبان منبع با زبان مقصد است. این مدل، یک مدل ترجمه می‌باشد. P (T) نشان‌دهنده یک مدل زبان است که برای مرتب‌سازی کلمات یا عبارات ترجمه‌شده در زبان مقصد استفاده می‌شود. ورودی مدل زبان "سبد واژگان[2] " است و هدف مدل اساساً این است که کلمات را دوباره مرتب کند. در این مرحله، فرضیه‌ای وجود دارد که می‌گوید احکام طبیعی را می‌توان صرفاً با مرتب‌سازی دوباره کلمات با استفاده از یک مدل ترجمه ایجاد کرد. اگر مجموعه‌ای اصلی از کلمات زبان مقصد به صورت k بیان شود، معادله ۱ به‌صورت ۲ خواهد شد.

$$P(T|S) = P(k|S)P(T|k) \quad (۲)$$

P(k|S) در این معادله، نشان‌دهنده یک مدل است که یک مجموعه از کلمات کلیدی در زبان مقصد را بر اساس متن زبان اصلی ارائه می‌دهد. P(T|k) نشان‌دهنده یک مدل بوده که متن S را با توجه به مجموعه‌ای از کلمات اصلی k ارائه می‌دهد.

تولید محتوای مناسب در هر حوزه یک چالش بزرگ و نگران‌کننده است و حتی در بعضی زمینه‌ها تهدید و فرصت جهت پیشبرد اهداف اجتماعی، اقتصادی و سیاسی است. برای مثال تهیه اسناد گزارشی جهت پیشبرد و شناسایی بازارهای هدف بر اساس داده‌های تولیدشده در بستر فضای مجازی به‌صورت انسانی کاری هزینه‌بر است که نیاز به روش‌هایی برای تبدیل داده‌ها به متن در این زمینه مشهود است. برای رسیدن به این هدف، ابتدا باید بتوان در تولید متن خودکار به نتیجه رسید تا این هدف محقق شود.

موضوع تولید خودکار زبان طبیعی یکی از موضوعات مطرح روز بوده که به خصوص، با توجه به همه گیر شدن شبکه‌های اجتماعی مورد توجه بیشتری قرار گرفته است. یکی از اصلی‌ترین کاربردهای این موضوع تولید خبر و تولید شایعه است که یکی از موارد مهم در پدافند به شمار می‌رود. این موضوع در زمینه‌های دیگر از جمله سیاست نیز کاربرد بسیاری دارد. به عنوان مثال در انتخابات اخیر و خرداد ۱۴۰۰، منافقین از تولید خبر در شبکه‌های اجتماعی برای دور کردن مردم از فضای انتخابات استفاده می‌کردند. این نمونه‌ای از بار منفی این وظیفه بوده و برعکس این قضیه نیز جزو مسائل مهم می‌باشد. لذا وجود سیستمی برای تولید خودکار متن و بدون دخالت انسانی ضروری است.

به‌طورکلی نوآوری‌های این مقاله به شرح زیر است:

۱- ارائه روشی برای تولید خودکار متن در زبان فارسی

۲- توانایی مدل در تولید متون طولانی و کوتاه

۳- ایجاد تنوع در متون تولید شده در حین حفظ کیفیت

۴- عدم وجود مشکلات گرامری در متون تولیدی

۵- توانایی تنظیم مدل در زمینه‌های مختلف

۲- مروری بر تولید خودکار متن

یکی از روش‌ها در تولید زبان سیستم‌های مبتنی بر قاعده[3] هستند که این سیستم‌ها بر اساس قوانینی که هدف کلی سیستم‌ها را تعیین می‌کنند، ساخته می‌شوند. متون تولید شده با این روش حتماً باید طوری تولید شوند که قوانین از پیش تعیین شده را ارضا کنند. چندین سیستم تولید مبتنی بر قاعده کلی عمومی توسعه یافته است که بعضی از آن‌ها به صورت عمومی در دسترس هستند.

[1] Embedding
[2] Bag Of Word
[3] Rule Based

در [۳] یک چارچوب مستقل از دامنه برای خلاصه‌سازی خودکار متن پیشنهاد شده است. همان‌طور که اشاره شد خلاصه‌سازی بخش کوچک‌تری از تولید متن می‌باشد. فرآیند ابتدا متن منبع را دسته‌بندی کرده و سپس مجموعه‌ای بهینه و مطلوب از مجموعه‌های قوانین یا وزن و روش‌های متداول را روی آن اعمال می‌کند. در این روش از طبقه‌بندی متن برای حل مشکل وابستگی به دامنه استفاده شده است.

همچنین یکی از مشکلات سیستم‌های NLG تفاوت در گرامر و ساختارهای زبانی متفاوت است که این موضوع را می‌توان با روش‌های مبتنی بر قاعده رفع کرد [۴]. در این مقاله سیستمی نرم‌افزاری بر اساس روش‌های مبتنی بر قواعد ارائه شده که تا به امروز میلیون‌ها متن را تولید کرده است.

مقاله‌ی [۵] یک روش مبتنی بر الگوی مبتنی بر برچسب[1]، به نسل زبانی طبیعی است. این ایده لغوی سازی را به کل عبارات گسترش می‌دهد که مشابه با اصالت اصطلاحات در گرامر TAG است. علاوه بر این، نوع دیگری از الگوها وجود دارد که مشتقات جزئی نام دارد. در این روش‌ها معمولاً در فاز اول از درخت اشتقاق استفاده می‌شود. در مقاله‌ای دیگر که بر این عقیده استوار است که رویکردهای مبتنی بر الگو[2] به قابلیت نگهداری و کیفیت خروجی وابسته هستند. برخی از سیستم‌های اخیر NLG که خود را "مبتنی بر الگو" می‌نامند این ادعاها را نشان می‌دهند [۶]. در [۷] نیز از ۶۰ هزار الگو برای تولید متن استفاده شده و ۳ مدل عمده برای تولید متن معرفی کرده است.

مدل مخفی مارکوف برای نسخه‌ی متن مناسب است. مقاله [۸] کاربرد مدل‌های پنهان مارکوف را برای تولید متن در زبان لهستانی ارائه می‌دهد. در این روش یک برنامه تولید متن با استفاده از مدل پیشنهادی مدل مخفی مارکوف توسعه یافت. این برنامه از یک متن مرجع برای یادگیری توالی‌های ممکن نامه استفاده می‌کند. نتایج پردازش متن نیز مورد بحث قرار گرفته است. رویکرد ارائه شده همچنین می‌تواند در روند تشخیص گفتار مفید باشد. دامنه سنتز و شناسایی گفتار به طور قابل ملاحظه‌ای طی ۳۰ سال گذشته به دلیل توسعه تلفن همراه، که در آن به طور گسترده استفاده می‌شود، تکامل یافته است. در روش‌های محبوب سنتز و تجزیه و تحلیل سخنرانی‌ها، مدل‌های پنهان مارکوف (HMM) مورد استفاده قرار می‌گیرند [۹, ۱۰].

از دیگر روش‌های جدید تولید خودکار متن، می‌توان به یادگیری تقویتی [۱۱]، خودکدگذارهای متغیر[3] (VAE) [۱۲] و شبکه‌های مولد تقابلی[4] (GAN) [۱۳] اشاره کرد.

ایده یادگیری تقویتی که در مدل‌های مولد عمیق استفاده می‌شود، نتایج امیدوارکننده‌ای را نشان داده است. یادگیری تقویتی، ناحیه‌ای از یادگیری عمیق بوده که در آن مدل با تعامل با محیط اطراف یاد گرفته و برای انجام کارها به آن‌ها پاداش می‌دهد [۱۱]. الگوریتم‌های یادگیری تقویتی را می‌توان با استفاده از مفاهیم عامل، محیط، حالت، عمل و پاداش درک کرد [۱۴]. مدل‌سازی تولید متن به عنوان مسئله یادگیری تقویتی یک ایده‌ی کلیدی است. این مفهوم برای اولین بار توسط باخمن و همکاران بررسی شده که مشکلات تولید دنباله ممکن است به عنوان مسئله تصمیم‌گیری پی‌درپی فرموله شود [۱۵]. با الهام از این مفهوم و یادگیری تقویتی، روشی جایگزین برای آموزش مدل تولیدی ارائه شد [۱۶].

VAE ها به عنوان یکی از روش‌های معروف برای یادگیری بدون نظارت بر توزیع‌های پیچیده توسعه‌یافته‌اند. کاربردهای VAE برای تولید داده‌های گسسته (متن) محدود است. مسئله اصلی استفاده از VAE برای تولید متن ، فروپاشی KL[5] است (بدین معنی که وقتی کدگشا از هدف آموزش قدرتمندتر می‌شود، می‌توان با استراتژی غلط آن را حل کرد)، یعنی کدگشا بدون در نظر گرفتن فضای نهفته، خروجی تولید می‌کند. اگر ترم KL صفر باشد، احتمال پسین[6] مستقل از داده‌های ورودی است [۱۷].

تلاش اخیر مدل مولد متن در VAE توسط بوومن و همکاران پیشنهاد شده که از شبکه‌های عصبی بازگشتی برای گرفتن ویژگی‌های عمومی جملات (به‌عنوان‌مثال موضوع، سبک) در متغیرهای کلی[7] استفاده می‌کند [۱۸]. در یکی دیگر از جدیدترین این روش‌ها، به جای استفاده از توزیع نرمال، تولید متن را در یک فضای نهفته هذلولی بررسی کرده تا نمایش‌های سلسله‌مراتبی پیوسته را بیاموزد. این مدل که Apo-VAE نامیده شده، با اتخاذ فرمول اولیه از واگرایی KL، یک روش یادگیری تقابلی را برای توانمندسازی آموزش مدل قوی معرفی کرده است [۱۹].

یادگیری عمیق نحوه کار، محاسبه، تجزیه‌وتحلیل و سهولت زندگی ما را تغییر داده است. ما به سیستم‌ها آموخته‌ایم که

[1] Tag
[2] Pattern Based

[3] Variational Auto-Encoders
[4] Generate Adversarial Networks
[5] kullback labier collapse
[6] Posterior
[7] Universal variable

چیزها را برای خود رقم بزنند، بسیاری از معماری‌های یادگیری عمیق را می‌توان به‌خاطر موفقیت خلاقانه آن اعتبار داد. با وجود این، هیچ موفقیت عمده‌ای توسط مدل‌های مولد عمیق ایجاد نشده که به دلیل عدم توانایی آن‌ها در محاسبات احتمالی غیرقابل‌حل است. اما راه‌حلی ارائه شده که می‌تواند مشکلات مدل‌های تولیدی را حل کند که شبکه‌های مولد تقابلی (GAN) نامیده می‌شوند [۱۳].

GAN یک الگوریتم محبوب یادگیری عمیق بوده که متفاوت از شبکه عصبی مرسوم، رویکردی تقابلی دارد. GAN حاوی دو مدل است که به روش تقابلی آموزش می‌بینند. اول، مولد نمونه‌های داده را تولید کرده و سپس متمایزکننده این نمونه‌های داده را به‌عنوان داده واقعی (داده‌های آموزشی) یا جعلی (تولید شده توسط تولیدکننده) طبقه‌بندی می‌کند. هدف تولیدکننده تولید نمونه‌هایی است که بسیار نزدیک به داده‌های واقعی باشد، به‌گونه‌ای که بتواند متمایزکننده را گمراه کند و هدف متمایزکننده طبقه‌بندی دقیق این دو نوع نمونه داده است.

از اولین کارها در زمینه تولید متن با شبکه‌های تقابلی توسط یو و همکاران تحت عنوان SeqGAN در سال ۲۰۱۷ پیشنهاد شد که از امتیاز پیش‌بینی متمایزکننده برای هدایت مولد استفاده می‌کند [۲۰]. در این مدل تولیدکننده G توالی s را تولید کرده و متمایزکننده آن را به عنوان دنباله واقعی (پاداش بالا) یا جعلی (پاداش کم) پیش‌بینی می‌کند. هدف از مدل مولد (سیاست) تولید توالی از حالت اولیه بوده تا پاداش نهایی مورد انتظار به حداکثر برسد.

هنگامی که خروجی در شبکه‌های تقابلی گسسته باشد، بازگرداندن گرادیان در شبکه دشوار است. برای پرداختن به این موضوع، MaliGAN پیشنهاد شد [۲۱]. همچنین RankGAN توسط لین و همکاران برای تولید توصیف متن با کیفیت بالا ارائه شده است [۲۲].

چالش اصلی تولید توالی متن طولانی، پراکنده بودن سیگنال هدایت شده باینری بوده که فقط در صورت تولید کل نمونه ارائه می‌شود. برای کاهش این مشکل LeakGAN پیشنهاد شده که ترکیبی از تطبیق ویژگی‌ها و یادگیری تقویتی سلسله مراتبی است [۲۳]. این مدل شامل ۲ ماژول مدیر[1] و کارگر[2] در تولیدکننده سلسله مراتبی G بوده که مدیر (LSTM بوده که به عنوان واسطه استفاده شده و بازنمایی ویژگی را از متمایز کننده D دریافت می‌کند) و کارگر (از این ویژگی‌های دریافت شده از متمایزکننده به عنوان سیگنال راهنما برای تولیدکننده استفاده می‌کند). اطلاعات دریافتی از متمایزکننده به عنوان اطلاعات نشت شده شناخته می‌شود.

اما همان‌طور که بیان شد یکی از مشکلات اصلی تولید متن حفظ تعادل بین کیفیت و تنوع متن تولیدی است که برای این مشکل در [۲۴]، یک چارچوب تولید متن دسته‌ای جدید، GAN با آگاهی از دسته (CatGAN) ، پیشنهادشده است. CatGAN یک مدل آگاه از طبقه‌بندی را برای تولید متن دسته‌ای (طبقه‌ای) و یک الگوریتم یادگیری تکاملی سلسله مراتبی برای آموزش مدل و به دست آوردن تعادل بین کیفیت نمونه و تنوع فراهم می‌کند.

همچنین در [۲۵]، یک رویکرد جدید انجمنی پیشنهادشده که هدف آن بهبود عملکرد تولید متن تقابلی از طریق کاهش سرعت فروپاشی حالت[1] آموزش است. همچنین در [۲۶] یک روش تقابلی دیگر ارائه شده که احساسات را نیز، در متون تولید شده در نظر می‌گیرد.

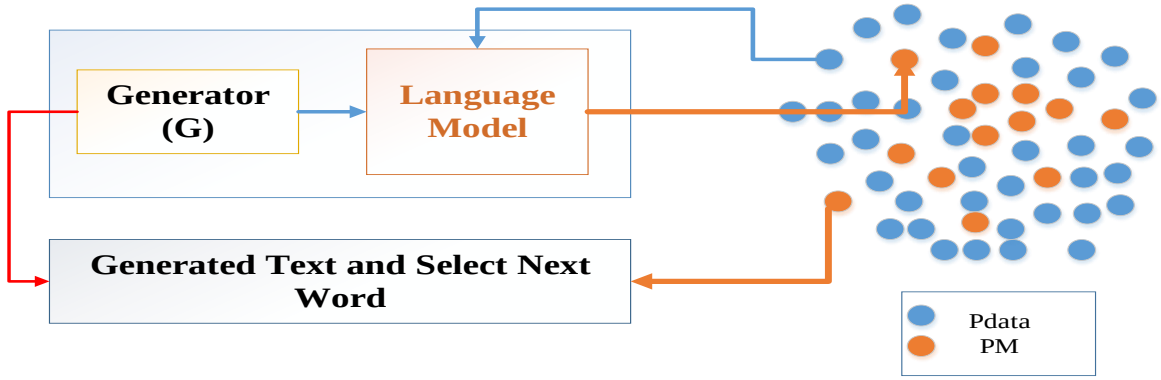
بسیاری از روش‌های تقابلی دیگر موجود در مقالات برای بهبود تولید متن به کار برده شده است. به‌عنوان‌مثال؛ روش‌های MaliGAN [۲۱]، TextGAN [۲۷]، GSGAN [۲۸] و MaskGAN [۲۹] از جمله این بهبودها می‌باشند.

همان‌طور که در بخش مقدمه اشاره شد، در این مقاله روشی ارائه شده که جزو اولین کارها در زمینه تولید متن فارسی بوده و فاقد مشکلات روش‌های مطرح شده از جمله فروپاشی حالت، تعیین پاداش و یا عدم توانایی در تولید متون طولانی می‌باشد.

۳- روش پیشنهادی

در این بخش روش پیشنهادی، مطرح و مورد بررسی قرار می‌گیرد. در ابتدای بخش مراحل پیش‌پردازش و آماده‌سازی متون شرح داده می‌شود. سپس روش اصلی مورد بررسی قرار می‌گیرد. ایده کلی از شبکه‌های مولد تقابلی گرفته شده است. در شبکه‌های تقابلی از متمایزکننده برای جریمه کردن تولیدکننده استفاده می‌شود تا تولیدکننده متون واقعی‌تر تولید کند. اما ایده این مقاله بر این است که حتی در دنیای واقعی، همیشه جریمه کردن راه‌حل مناسب نبوده و گاهی راهنمایی کردن بهتر از جریمه کردن جواب می‌دهد؛ لذا در مدل پیشنهادی متمایزکننده از مدل حذف شده و به‌جای آن، از یک مدل زبانی که شامل قوانین و بردارهای استخراج شده توسط Word2vec هست، استفاده می‌شود. به‌طورکلی روش پیشنهادی در شکل ۱ نشان داده شده است.

[1] Manager
[2] Worker



شکل (۱): دیاگرام مدل پیشنهادی (P_{data} توزیع کلمات ورودی به مدل و P_M توزیع کلمات محدود شده توسط مدل زبانی و مرحله استخراج زبانی است)

به‌طورکلی این مدل شامل دو بخش اصلی، یعنی استخراج قوانین با کمک مدل زبانی N- تایی[۲] و سپس تولید متن می‌باشد. در بخش استخراج قوانین که در مدل زبانی تعبیه‌شده، بر اساس تعداد N، N- تایی‌های متن ورودی یا یک پیکره بزرگ به عنوان قاعده استخراج می‌شود. سپس در فاز تولید از این قوانین برای تولید استفاده می‌گردد.

۳-۱- پیش‌پردازش

پیش‌پردازش متن، مرحله آماده‌سازی متن برای ورود به سامانه‌های پردازشی متون می‌باشد. سامانه‌های پردازشی متون، در صورتی عملکرد مناسبی ارائه می‌دهند که ورودی مناسب برای آن‌ها تأمین شود. هرچه پیش‌پردازش متن، با کیفیت بالاتری انجام شود، کارایی سامانه بیشتر خواهد شد.

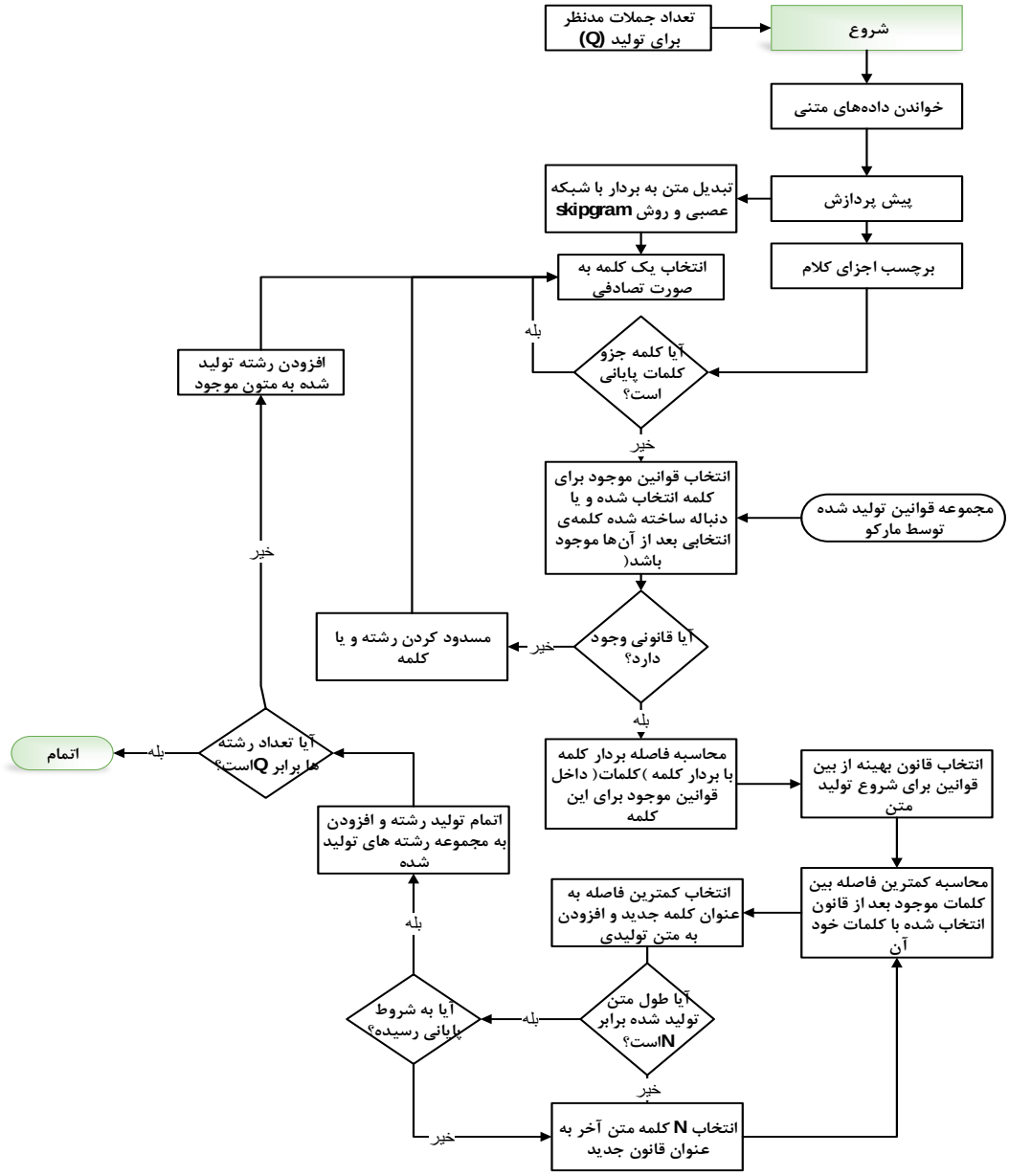
پردازش زبان فارسی از جهاتی با پردازش زبان انگلیسی تفاوت دارد. در زبان انگلیسی تمامی حروف و تمامی کلمات جدا از هم و با قاعده‌ای مشخص نوشته می‌شوند و این در حالی است که در زبان فارسی بعضی از حروف به هم چسبیده هستند، برخی از حروف جدا از هم نوشته می‌شوند، بعضی از کلمات یکپارچه‌اند، بعضی از کلمات با فاصله یا نیم‌فاصله به دو یا چند بخش تقسیم می‌شوند.

تمامی حوزه‌های مرتبط با پردازش زبان طبیعی به نحوی با متون واقعی سروکار دارند. صورت‌های غیراستاندارد نویسه‌ها و کلمات، به فراوانی در این نوع متون نوشته دیده می‌شوند. قبل از این که بتوان از این متون به منظور استفاده در سیستم‌های تبدیل متن به گفتار، ترجمه ماشینی، بازشناسی حروف فارسی، خلاصه‌ساز فارسی، جستجو در متون فارسی و غیره استفاده کرد و یا در پایگاه داده ذخیره نمود، باید ابتدا پیش‌پردازشی روی آن‌ها انجام گیرد تا صورت‌های غیراستاندارد به شکل استاندارد تبدیل گردند.

اگر حروف، نشانه‌های نگارشی و کلمات فارسی به شکل یکسانی نوشته نشوند، متون مورد استفاده قابل تحلیل توسط سامانه‌های رایانه‌ای نخواهند بود. به‌عنوان‌مثال اگر هنجارسازی روی دو داده "رئیس‌جمهور" و "رییس‌جمهور" اعمال نشود، سیستم این دو را دو عبارت جدا در نظر می‌گیرد که روی نتایج تأثیر زیادی دارد. طی فرایند هنجارسازی، علائم نگارشی، حروف، فاصله‌های بین کلمات، اختصارات و غیره بدون ایجاد تغییرات معنایی در متن به شکل استاندارد تبدیل می‌گردند؛ بنابراین، بایستی از یک استاندارد مشترک برای پیش‌پردازش و پردازش متون استفاده کرد [۳۰].

یکی از مشکلات زبان فارسی وجود چند نمونه مختلف از یک نویسه و حرف است که کار جستجو در متون فارسی را مشکل می‌کند. در این مرحله کاراکترهای غیراستاندارد با کاراکترهای استاندارد جایگزین می‌شوند و کاراکترهای اضافی نیز بسته به نوع پردازش از بین می‌روند تا واژه‌های یکسان در تمامی متن به یک صورت نوشته‌شده باشند. به‌عنوان‌مثال، برای دو کلمه «مسئله» و «مسأله»، کل متن هنجار شده و یکی از این دو و یا یک کلمه جایگزین مانند «مسئله» به عنوان کلمه مرجع انتخاب می‌شود. از این قبیل کلمات می‌توان «رییس» و «رئیس»، کلماتی که دارای حروف "ی" و "ک" عربی می‌شوند و ... را نام برد. همچنین همه‌ی فاصله‌ها نیز هنجار و به یک حالت تبدیل شده است. به‌عنوان‌مثال "میرفت"، "می‌رفت" و "می-رفت" هر سه دارای یک معنا و مفهوم هستند، اما اگر پردازش روی آن‌ها صورت نگیرد، دارای نتایج متفاوتی خواهند بود.

[1] Mode collapse
[2] n-gram



شکل (۲): دیاگرام فاز تولید در مدل پیشنهادی (N طول دنباله مدنظر و Q تعداد جملات مدنظر برای تولید است.)

در زبان‌شناسی پیکره‌ای، برچسب‌گذاری اجزای کلام[1]، در واقع عمل انتساب برچسب به کلمات تشکیل‌دهنده یک متن یا یک پیکره است. این برچسب‌گذاری براساس نقش آن کلمه در متن، مانند اسم، فعل، قید، صفت، و غیره صورت می‌گیرد. بعضی کلمات ممکن است یک یا چند برچسب داشته باشند. اگر یک کلمه بیش از یک برچسب داشته باشد، نیاز به ابهام‌زدایی دارد.

۳-۲- استخراج قوانین

روش پیشنهادی تا حدود زیادی مانند روش مارکوف می‌باشد. درواقع استخراج قوانین در روش پیشنهادی، مطابق مدل مارکوف صورت می‌گیرد، اما به جای محاسبه احتمال، از مدل Word2vec و همچنین معیار فاصله کسینوسی استفاده شده است.

برای استخراج قوانین، در ابتدا تمام تک کلمه‌ای‌های ممکن استخراج می‌شود. سپس، قوانین به قوانین ۲ تایی بسط داده شده و ۱ تایی‌هایی که قابل بسط نباشند، فیلتر خواهند شد. این عمل برای ۳ تایی‌ها نیز ادامه پیدا می‌کند. در واقع این عمل تا جایی ادامه خواهد داشت که در n تایی‌ها، n برابر با طول مجاز برای قوانین شود که این طول را کاربر وارد خواهد کرد (n). اگر کاربر طول خاصی را وارد نکند، بر اساس آزمایشات طول ۳ که بهترین طول است، به طور خودکار توسط سیستم در نظر گرفته می‌شود.اگر کلمه‌ای زودتر به n برسد، قوانین ۱ تایی، ۲ تایی و ... تا n تایی برای این کلمه به مجموعه قوانین افزوده شده و اگر دیگر نتوان قانونی تولید کرد، مجموعه قوانین توسط این بخش

[1] Part of speech tag

برگردانده می‌شود. در واقع مجموعه قوانین نهایی شامل دنباله‌های ۱، ۲، ...، n تایی می‌باشد. این قوانین می‌توانند روی یک پیکره بزرگ و از قبل استخراج شوند. به‌عنوان مثال اگر فرض شود دو جمله زیر موجود است:

مثال ۱:
جمله ۱- علی به مدرسه می‌رود و درس می‌خواند.
جمله ۲- علی به مدرسه آمد و درس خود را نخوانده بود.

برای این دو جمله اگر فرض شود n برابر ۲ باشد، مجموعه قوانین تشکیل شده به‌صورت جدول (۱) خواهد بود:

جدول (۱): مثال مجموعه قوانین تشکیل شده

دنباله کلمات	قوانین
علی به	مدرسه
به مدرسه	می‌رود، آمد
مدرسه می‌رود	و
می‌رود و	درس
و درس	می‌خواند، خود
آمد و	درس
درس خود	را
خود را	نخوانده
را نخوانده	بود

که این مجموعه قوانین برای متون آموزشی زیاد بسیار بیشتر خواهد بود و این شانس تولید متن جدید را افزایش خواهد داد. همین امر باعث ایجاد تنوع در متن جدید و درعین‌حال حفظ ساختار جمله می‌شود.

۳-۳- تولید متن

روند کلی فاز تولید در شکل (۲) نشان داده شده است. پس از استخراج قوانین، در این بخش کلمات موجود در متن آموزشی برای بهبود نتایج خروجی برچسب اجزای کلام می‌خورند. برای شروع هر دنباله‌ی متنی که قرار است تولید شود، در ابتدا یک کلمه به صورت تصادفی انتخاب شده، سپس بررسی می‌شود که این کلمه داخل مجموعه‌ی قوانین موجود هست یا خیر. همچنین برای بررسی عدم انتخاب کلمات پایانی، از برچسب اجزای کلام کمک گرفته می‌شود.

در واقع از بین قوانین موجود که در مجموعه‌ی آموزشی وجود دارد، برای تولید متن جدید کمک گرفته می‌شود و اگر کلمه‌ای شامل قانون نباشد (در شروع کار فقط این احتمال برای کلمات پایان هر پاراگراف و یا متن وجود دارد، اما در ادامه احتمال این امر برای دنباله‌های ۲تایی به بعد خیلی افزایش میابد)، روند به بن‌بست رسیده و کلمه جدیدی برای شروع انتخاب خواهد شد. این حالت بیانگر نیمه نظارتی بودن مدل بوده، زیرا علت رد کردن این دنباله عدم وجود آن تا دنباله n تایی می‌باشد (n حداکثر طول قوانین موجود در مجموعه قوانین است).

حال اگر مجموعه قوانین و دنباله‌های موجود در آن شامل این کلمه باشند، تمام دنباله‌های موجود (۲ تایی، ۳ تایی و... و nتایی) مرتبط با این کلمه استخراج شده و با توجه به مدل word2vec برای هر دنباله از قوانین بردار استخراج می‌شود. سپس بین این کلمه و کلمات تمامی قوانین از فاصله کسینوسی استفاده کرده و قانونی که دارای کمترین فاصله بین کلماتش باشد به عنوان قاعده اصلی انتخاب خواهد شد که این امر دارای بهترین نتایج در بین معیارهای فاصله‌ی موجود است. در واقع محاسبه‌ی فاصله برای کلمات با توجه به شبکه‌ی عصبی، پیوستگی بین کلمات در تمام حالات در نظر گرفته می‌شود.

بعد از مشخص شدن قانون، بخش اصلی تولید متن شروع شده و ازآنجایی‌که مجموعه قوانین شامل کلمات ممکن بعد از دنباله یا کلمه‌ی فعلی است، پیوستگی در متن حفظ می‌شود و متن تولید شده دارای ساختار و مفهوم خواهد بود. سپس بین کلمات موجود در قوانین و کلمه یا کلمات دنباله مجدداً فاصله‌ی کسینوسی حساب شده و کلمه‌ای که کمترین فاصله را داشته باشد به‌عنوان کلمه جدید در نظر می‌گیرد.

به‌عنوان مثال اگر متن تولیدی تا به اینجا شامل کلمات $x_1, x_2, \dots, x_n$ باشد (n برابر حداکثر طول n تایی‌ها در مجموعه قوانین است)، برای تعیین کلمه جدید x_{n+1} از رابطه $distance(x_{n+1}, [x_1, x_2, \dots, x_n])$ استفاده می‌شود که هر xi خود یک بردار m تایی (تعداد ابعاد در نظر گرفته شده برای تعبیه‌سازی[1] در Word2vec) می‌باشد. تابع distance نیز می‌تواند توابع فاصله‌ای باشد که در واقع فاصله بین بردار تعبیه شده x_{n+1} را با هر یک از x_iها محاسبه کرده و میانگین آن را به‌عنوان فاصله نهایی در نظر می‌گیرد. برای انتخاب کلمه هدف بر اساس n کلمه آخر از رابطه (۳) استفاده می‌گردد:

$$Distance(w_i, \{w_1, w_2, \dots, w_n\}) = \frac{dist(w_i, w_1) + dist(w_i, w_2) + \dots + dist(w_i, w_n)}{n} \quad (۳)$$

که در آن wi بردار ویژگی کلمه i ام موجود در مجموعه قوانین، {$w_1, w_2, \dots, w_n$} بردار ویژگی n کلمه آخر دنباله تولید شده توسط Word2vec و dist تابعی برای محاسبه فاصله بین بردارهاست که در این تحقیق از فاصله کسینوسی استفاده شده است. با اینکه فاصله کسینوسی و اقلیدسی تقریباً دارای نتایج یکسانی هستند ولی معمولاً فاصله کسینوسی دارای نتایج بهتری نسبت به سایر روش‌ها می‌باشد [۳۰].

پس از این‌که کلمه جدید مشخص گردید، این کلمه به دنباله

موجود افزوده می‌شود و در فاز بعدی شرط خاتمه بررسی می‌شود که در این بخش شرط خاتمه رسیدن به فعل و یا علامت‌های پایانی جمله با توجه به برچسب اجزای کلام متن می‌باشد. اگر شرط خاتمه ارضا نشده باشند، n کلمه آخر متن تا به اینجا به‌عنوان دنباله جدید انتخاب می‌گردد. سپس مجدداً تمام کلمات بعد این دنباله از مجموعه قوانین استخراج شده و بین این کلمات و کلمه‌های موجود در دنباله، فاصله محاسبه گردیده و کمترین فاصله به عنوان کلمه جدید انتخاب می‌شود و این روند تا زمانی ادامه میابد که شرط پایانی اولیه اتمام یابد. در این بخش در واقع یک جمله آزمون تولید گردیده است.

در واقع در این بخش چون از قوانین استفاده می‌شود، متن تولید شده دارای ساختار مناسب بوده و از طرفی چون از Word2vec برای تعیین کلمه بعدی استفاده می‌شود، از نظر معنایی نیز متن تولیدی دارای معنا می‌باشد. همچنین از هرچه قوانین موجود افزایش یابد، تنوع در این سیستم بالا می‌رود. پس به‌طورکلی هم کیفیت و هم تنوع در متون تولیدی در این مدل در نظر گرفته می‌شود.

به‌عنوان‌مثال بر اساس مثال ۱، با فرض وجود دو جمله ۳ و ۴:

جمله ۳: محمد به مدرسه وارد شد.
جمله ۴: حسین به مدرسه وارد شده است.

مجموعه قوانین با فرض ساخته شدن توسط یک مدل زبانی جامع، قاعدتاً تکمیل‌تر شده و لذا قوانین ۲تایی "به مدرسه" علاوه بر "می‌رود" و "آمد"، شامل کلمه‌ی "وارد" هم خواهد بود. حال با فرض این که سیستم تا به اینجا جمله‌ی "علی به مدرسه" را تولید کرده، در فاز تولید بین بردار کلمات این جمله با سه کلمه موجود در مجموعه قوانین، یعنی "می‌رود"، "آمد" و "وارد" فاصله کسینوسی محاسبه‌شده و کلمه‌ی دارای کمترین فاصله، به عنوان کلمه هدف انتخاب می‌شود. با فرض وجود تنها ۴ جمله بیان شده و با توجه به خاصیت Word2vec و تکرار کلمه "وارد" بعد از ۲ تایی "به مدرسه"، این کلمه هدف بوده و به دنباله قبلی افزوده می‌شود. لذا دنباله جدید "علی به مدرسه وارد" می‌باشد که در هیچ‌کدام از ۴ جمله قبل موجود نیست. به‌عبارت‌دیگر بیان کننده جدید بودن و تنوع متن تولیدی می‌باشد. این امر برای پیش‌بینی سایر کلمات نیز ادامه می‌یابد.

۴- نتایج

در این بخش تمامی نتایج از جنبه‌های مختلف بررسی خواهند شد. در کلیه مراحل برای ارزیابی ۱۰۰۰ جمله توسط مدل تولید و سپس ارزیابی صورت گرفته است.

مجموعه داده: برای آموزش و ارزیابی مدل از مجموعه خبرهای جمع‌آوری شده توسط دانشگاه مالک‌اشتر استفاده شده و بخشی از آن که دارای ۱۹۴۵۰۴ جمله بوده، انتخاب شده است.

معیارهای ارزیابی: برای ارزیابی نیز از دو معیار Bleu و ROUGE-L استفاده شده است. معیار Bleu به این صورت است که جمله تولید شده را با جملات آموزش مقایسه کرده و نزدیک‌ترین جمله به آن را یافته، سپس Bleu_n بر اساس nتایی‌ها جمله مرجع را با جمله تولید شده مقایسه کرده و امتیاز به دست می‌آید. بالا بودن این معیار به معنی خوب بودن مدل است اما از آنجایی‌که در دو مجموعه جدید داده تولید شده، جملات کوتاه هستند، این معیار به تنهایی نمی‌تواند معیار خوبی باشد و از معیار ROUGE_L استفاده می‌شود.

البته در معیار Bleu هم هر چه n بیشتر شود و دقت بالاتر باشد، نشان‌دهنده مفهوم و تنوع مدل است، اما بالا بودن بیش از حد این معیار و پایین بودن معیار ROUGE_L نشان‌دهنده عدم تناسب و بیش برازش می‌باشد. ROUGE-L معیاری است که برای اولین بار برای خلاصه‌سازی ارائه شده است. ROUGE-L یک اندازه‌گیری F است که بر اساس طولانی‌ترین زیرمجموعه مشترک (LCS) بین متن تولید شده و متن مرجع می‌باشد. ازآنجایی‌که معیار خاصی برای تولید متن وجود نداشته و در اکثر مقالات از این معیارها استفاده می‌شود، ما هم در این گزارش به‌صورت ترکیبی این معیارها را تحلیل خواهیم کرد. به‌طور کلی مدلی بهتر خواهد بود که علاوه بر تناسب بین معیارهای Bleu، بین این معیار و ROUGE_L نیز تناسب برقرار باشد.

مدل پایه: همچنین برای ارزیابی بهتر نتایج، از مدل Bi-LSTM برای ارزیابی استفاده شده است. در این مدل دنباله کلمات به صورت سری زمانی در نظر گرفته شده و مدلBi-LSTM سعی در پیش‌بینی کلمه جدید دارد. این مدل به علت استفاده در اکثر مدل‌های مولد از جمله شبکه‌های GAN به‌عنوان مدل پایه انتخاب شده است.

۴-۱- مقایسه نتایج و توانایی مدل

در این روش از یک مدل Bi-LSTM دو لایه برای تولید متن جدید از یک شبکه یادگیری عمیق چند لایه (۲ تا ۵ لایه) با تعداد نرون مختلف استفاده شده و در ابتدا پس از این‌که به هر کلمه یک اندیس اختصاص داده شود و به‌صورت واژه‌نامه تبدیل شوند، این کلمات به صورت سری زمانی درآمده و پس از آموزش مدل، برای تعیین کلمات و

تولید متن جدید، از خروجی لایه softmax استفاده شده و پس از این که کلمات واژه‌نامه بین صفر و یک مقدار گرفتند، از احتمالات استفاده شده و به عدد هر کلمه تنوع اعمال شده و کلماتی که احتمال انتخاب آن‌ها به عنوان کلمه‌ی جدید بیشتر باشد، به‌عنوان خروجی انتخاب می‌شوند. مقایسه این نتایج در جدول (۱) نشان داده شده است. برای مقایسه روش‌ها، برای هر کدام ۱۰۰۰ جمله به صورت کوتاه (کمتر از ۱۵ کلمه) و ۱۰۰۰ جمله طولانی (بیشتر از ۲۰ کلمه) تولید شده و ارزیابی انجام شده است.

از نتایج در جدول (۱) مشخص است که با پیچیده‌تر شدن متن، جملات تولید شده توسط Bi-LSTM با توجه به معیار Rouge از دقت کمتری در ارتباط بین جملات تولید شده برخوردار خواهند بود ولی از نظر ساختاری دارای نتایج خوبی در معیار Bleu هستند.

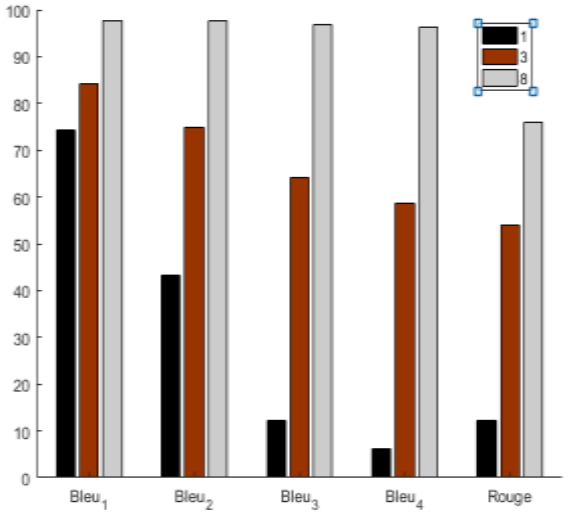
جدول (۱): مقایسه توانایی و نتایج مدل پایه و روش پیشنهادی

مدل	Bleu_1	Bleu_2	Bleu_3	Bleu_4	Rouge
متن طولانی و روش پیشنهادی	۸۴/۳۳	۷۴/۸۲	۶۴/۲۴	۵۸/۵۹	۵۴/۱۰
متن طولانی و روش Bi-LSTM	۹۰/۸۲	۷۵/۱۸	۶۳/۴۱	۵۶/۱۳	۲۱/۲۲
متن کوتاه و روش پیشنهادی	۸۹/۴۷	۸۶/۳۴	۷۸/۴۳	۷۲/۱۷	۴۳/۶۷
متن کوتاه و روش Bi-LSTM	۹۷/۲	۷۷/۳۱	۳۹/۱۳	۱۷/۰۶	۳۹/۳۶

مدل پیشنهادی نسبت به مدل پایه نتایج بهتری داشته و متون تولید شده، تا حدودی معیارهای تولید متن را محقق می‌سازند. مخصوصاً در بخش خبرهای پیچیده مشخص است که معیار Bleu نسبت به داده‌ی ساده کمتر شده که این طبیعی بوده و قابل قبول می‌باشد، اما نکته قابل توجه افزایش معیار Rouge است که نشان از درک ارتباط بین جملات تولید شده توسط مدل می‌باشد. به‌طورکلی باتوجه‌به این نتایج مشخص است که مدل پیشنهادی چه در تولید متن کوتاه و چه تولید متن طولانی موفق عمل کرده است.

۴-۲- بررسی تأثیر پارامتر n

در روش پیشنهادی سعی بر این است که با استفاده از Word2vec از متن ویژگی استخراج شده و سپس برای تولید و انتخاب کلمه‌ی بعد از ترکیب این ویژگی‌ها و قوانین استفاده شود. در این روش از مدل word2vec و به‌صورت skip-gram کلمات به بردار تبدیل می‌شود. اما همان‌طور که بیان شد، یکی از پارامترهایی که در تعیین کلمه‌ی بعد تأثیر دارد، n بوده که در واقع هم برای دنباله‌ی قوانین (n تایی) و هم تعداد کلمات دنباله‌ی قبل برای تعیین کلمه جدید استفاده می‌شود. در این بخش تأثیر این پارامتر برای n های برابر با ۱، ۳ و ۸ بررسی شده که نتایج آن در شکل (۳) نشان داده شده است.



شکل (۳): بررسی تأثیر n های مختلف در تولید متن توسط روش پیشنهادی

بر اساس نتایج به‌دست‌آمده در شکل (۲) و جملات تولید شده، مشخص است که در نظر گرفتن تعداد nهای زیاد عملاً باعث بیش‌برازش مدل شده و ازآنجایی‌که Rouge یک معیار بر اساس LCS است مشخص می‌کند که تعداد حالات ۸ بسیار شبیه به متن آموزشی می‌باشد. از طرفی افزایش بیش از حد معیارهای Bleu_4 و Blue_3 صرفاً به معنای خوب بودن مدل نیست و با

توجه به متون تولید شده مشخص است که مدل برای ۸ بیش‌برازش شده است. در واقع طبق معیار انسانی هم خود انسان نمی‌تواند با این دقت متون خبری طولانی تولید کند.

از طرفی برای تعداد حالات ۱ با توجه به این معیارها می‌توان دریافت که مدل دچار کم‌برازش شده و از یک جایی به بعد داده‌های پرت تولید می‌کند. این موضوع با کاهش شدید Bleu_4 و Blue_3 قابل مشاهده است که در Bleu_4 این امر حتی به صفر می‌رود. با توجه به نتایج به دست آمده بهترین طول برای مدل‌های پیشنهادی ۳ می‌باشد که این امر علاوه بر مفهومی شدن جملات، به تنوع بیشتر متون نیز کمک می‌کند.

۴-۳- ارزیابی انسانی

در این بخش مدل پیشنهادی، ازنظر گرامر، معنا، نوآوری و تنوع توسط انسان ارزیابی‌شده است. برای این منظور هر یک از این پارامترها اعدادی بین ۱ تا ۵ را اختیار کرده که ۵ بهترین و ۱ بدترین می‌باشد. نتایج در جدول (۲) نشان داده شده است. برای این منظور ۱۰۰ جمله تولیدشده و ارزیابی‌شده است.

جدول (۲): ارزیابی انسانی مدل پیشنهادی

معیار	گرامر	معنا	نوآوری	تنوع	مجموع
امتیاز	۴/۶۴	۴/۳۹	۳/۹۷	۳/۸۵	۴/۲۱

با توجه به نتایج مشخص است که روش پیشنهادی ازآنجایی‌که مبتنی بر قانون و ترکیب آن با Word2vec می‌باشد، ازنظر گرامر امتیاز بسیار بالایی دارد. همچنین این امر باعث افزایش دقت در معنای جملات تولیدشده می‌شود. اما چون این مدل بر اساس قوانین عمل کرده و از قوانین تبعیت می‌کند، نوآوری پایین‌تری دارد. علت آن هم این است که اکثر کلمات پیش‌بینی‌شده بر اساس حالات قبل بوده و باعث نزدیک شدن داده تولیدشده به داده آموزش می‌شود. از طرفی چون این مدل بر اساس فاصله بوده، در حفظ ارتباط معنایی بین جملات در خبرهای چندجمله‌ای ضعیف‌تر عمل می‌کند.

۴-۴- پیچیدگی مدل

از آنجایی‌که بخش اصلی مدل پیشنهادی مربوط به استخراج قوانین بوده و برای برداری‌سازی می‌توان از شبکه‌های پیش آموزش‌دیده هم استفاده کرد، در این بخش پیچیدگی زمانی مدل برای استخراج این قوانین روی بخش‌های مختلف مجموعه داده ارزیابی شده است. نتایج به‌دست‌آمده در جدول ۳ نشان داده شده است.

جدول (۳): پیچیدگی زمانی مدل پیشنهادی در استخراج قوانین برحسب ثانیه

تعداد جملات	تعداد قوانین	زمان
۱۲۸۸۶	۳۷۴,۶۴۰	۴۱/۳۲
۴۵۴۹۸	۱,۳۵۳,۹۹۲	۱۵۰/۲۵
۸۰۲۱۷	۲,۹۹۴,۲۵۷	۳۷۴/۹۴
۱۹۴۵۰۶	۶,۱۴۶,۲۰۱	۱۵۲۲/۸۶

بر اساس جدول (۳) مشخص است که بین تعداد جملات و زمان استفاده شده برای استخراج قوانین رابطه مستقیم وجود دارد، در واقع اگر تعداد جملات k برابر شود، میزان زمان مصرفی هم k برابر می‌شود.

برای مجموعه داده اصلی این زمان حدود ۲۵ دقیقه می‌باشد در نتیجه با توجه به رابطه‌ای که بین تعداد جملات و زمان وجود دارد، این زمان برای یک پیکره بزرگ که حدود ۳ میلیون جمله داشته باشد، تقریباً ۷ ساعت زمان می‌برد که در دنیای امروز زمان بسیار کمی می‌باشد. لازم به ذکر است که این نتایج با Cpu گرفته شده است.

۴-۵- نمونه‌های تولید شده توسط سیستم پیشنهادی

در این بخش نمونه‌هایی از جملات تولید شده در مقایسه با جملات اصلی موجود در پیکره آموزشی، در جدول (۴) نشان داده شده است. همان‌طور که در این جدول مشاهده می‌شود، نمونه‌های تولید شده در مقایسه با نمونه‌های اصلی ساختار را حفظ کرده‌اند. همچنین در این مقایسه مشاهده می‌شود که این نمونه‌ها، علاوه بر این که دارای معنا بوده، نسبت به متن اصلی از تنوع لازم و نوآوری نیز برخوردار می‌باشند.

۵- نتیجه‌گیری

در این مقاله، مدلی خودکار برای تولید متن در زبان فارسی مبتنی بر قاعده و Word2vec ارائه شده که مشکلات روش‌های خودکار از جمله تولید متن طولانی و حفظ تنوع در حین حفظ کیفیت را ندارد. در این مدل در ابتدا قوانین ساخته شده و سپس بر اساس این قوانین و بردارهای تعبیه شده کلمات، کلمه جدید پیش‌بینی می‌شود. این مدل قادر به تولید متون طولانی و کوتاه با حفظ ساختار، گرامر و تنوع می‌باشد. همچنین باتوجه‌به متون نشان داده شده در جدول (۴) مشخص است که متون تولید شده مفهومی و پایدار هستند.

جدول (۴): نمونه متون طولانی و کوتاه تولید شده توسط مدل

نوع متن	متن
اصلی	جوانب مختلف از جمله شرایط خاص اقتصادی کشور، رکود چند ساله در اقتصاد و افزایش ناچیز حقوق و دستمزدها احساس ما این است تغییر قیمت حامل‌های انرژی منجر به ایجاد شوک قیمتی و تورمی در اقتصاد خواهد شد
تولید شده	جوانب مختلف از جمله شرایط خاص اقتصادی کشور، رکود چند ساله در اقتصاد و سیاست خواهد_داشت.
اصلی	می‌تواند ارتباط چهره به چهره نباشد و مردم کارهای اداری خود را از طریق اینترنت و تلفن حل و فصل کنند.
تولید شده	می‌تواند ارتباط چهره به چهره نباشد و مردم کارهای اداری خود را از مردم محروم کشور بشنوند و با دولت در ارتباط نباشند.
اصلی	مصاحبه‌های اخیر با شخصیت‌های جنجالی از جمله وکیل بابک زنجانی مسائلی را مورد بحث قرار داده که در رسانه ملی آنگونه مورد توجه قرار نگرفته.
تولید شده	شخصیت‌های جنجالی از جمله وکیل بابک زنجانی چیست ؟ توصیه من به وکیل او نیست، بلکه یک اعلامیه است.
اصلی	اتفاقا فردی نزدیک به دولت و مواضع وزارت ارتباطات بود، این بار به مذاق مالکان مخابراتی که در تصرف دانشکده مخابرات خود را بازنده تعامل با دولت می‌دیدند، خوش نیامد و وی را کنار گذاشتند.
تولید شده	آمار اشتغال صفر بود و همه این بار به مذاق مالکان مخابراتی که در تصرف دانشکده مخابرات خود را بازنده تعامل با دولت میدیدند، خوش نیامد و وی را کنار گذاشتند.
اصلی	بعد تلویزیون، بعد ویدئو، بعد ماهواره و بعد فضای مجازی که با آن مبارزه کردند.
تولید شده	ماهواره و بعد فضای مجازی که با آن روبه رو بود، از مهمترین دلایل این مسأله می‌دانستند.

همچنین این مدل از نظر انسانی امتیاز قابل قبولی را کسب کرده و باتوجه به پیچیدگی مدل، با این که مبتنی بر قانون بوده و روش‌های مبتنی بر قانون ذاتاً دارای سربار زمانی بالایی هستند، از آنجایی‌که قوانین در بخش تولید توسط فاصله کسینوسی محدود می‌شوند این سربار کاهش داشته و خود بخش استخراج قوانین نیز نسبت به مدل‌های مبتنی بر قانون موجود زمان بسیار کمی را مصرف می‌کند.

۶- مراجع

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[2] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, et al., "A statistical approach to machine translation," Computational linguistics, vol. 16, 1990.

[3] Y. K. Meena and D. Gopalani, "Domain independent framework for automatic text summarization," Procedia Computer Science, vol. 48, pp. 722-727, 2015.

[4] A. Bauer, N. Hoedoro, and A. Schneider, "Rule-based Approach to Text Generation in Natural Language-Automated Text Markup Language (ATML3)," in Challenge+ DC@ RuleML, 2015.

[5] T. Becker, "Practical, template–based natural language generation with tag," in Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+ 6), 2002, pp. 80-83.

[6] K. V. Deemter, M. Theune, and E. Krahmer, "Real versus template-based natural language generation: A false opposition?," Computational Linguistics, vol. 31, pp. 15-24, 2005.

[7] A. Ratnaparkhi, "Trainable methods for surface natural language generation," in Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, 2000, pp. 194-201.

[8] G. Szymanski and Z. Ciota, "Hidden Markov models suitable for text generation," in WSEAS International Conference on Signal, Speech and Image Processing (WSEAS ICOSSIP 2002), pp. 3081-3084.

[9] S. R. Eddy, G. Mitchison, and R. Durbin, "Maximum discrimination hidden Markov models of sequence consensus," Journal of Computational Biology, vol. 2, pp. 9-23, 1995.

[10] S. R. Eddy, "Multiple alignment using hidden Markov models," in Ismb, 1995, pp. 114-120.

[11] A. Skymind, "Beginner's Guide to Deep Reinforcement Learning," ed, 2019.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672-2680.

[14] T. Iqbal and S. Qureshi, "The Survey: Text Generation Models in Deep Learning," Journal of King Saud University-Computer and Information Sciences, 2020.

[15] P. Bachman and D. Precup, "Data generation as sequential decision making," in Advances in Neural Information Processing Systems, 2015, pp. 3249-3257.

[16] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, et al., "An actor-critic algorithm for sequence prediction," arXiv preprint arXiv:1607.07086, 2016.

[17] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Understanding posterior collapse in generative latent variable models," 2019.

[18] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," arXiv preprint arXiv:1511.06349, 2015.

[19] S. Dai, Z. Gan, Y. Cheng, C. Tao, L. Carin, and J. Liu, "APo-VAE: Text Generation in Hyperbolic Space," arXiv preprint arXiv:2005.00054, 2020.

[20] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient," in AAAI, 2017, pp. 2852-2858.

[21] T. Che, Y. Li, R. Zhang, R. D. Hjelm, W. Li, Y. Song, et al., "Maximum-likelihood augmented discrete generative adversarial networks," 2017.

[22] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," in Advances in Neural Information Processing Systems, 2017, pp. 3155-3165.

[23] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. J. a. p. a. Wang, "Long text generation via adversarial training with leaked information," 2017.

[24] Z. Liu, J. Wang, and Z. Liang, "CatGAN: Category-Aware Generative Adversarial Networks with Hierarchical Evolutionary Learning for Category Text Generation," in AAAI, 2020, pp. 8425-8432.

[25] H. Yin, D. Li, X. Li, and P. Li, "Meta-CoTGAN: A Meta Cooperative Training Paradigm for Improving Adversarial Text Generation," in AAAI, 2020, pp. 9466-9473.

[26] K. Wang and X. Wan, "SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks," in IJCAI, 2018, pp. 4446-4452.

[27] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, et al., "Adversarial feature matching for text generation," 2017.[28] M. J. Kusner and J. M. J. a. p. a. Hernández-Lobato, "Gans for sequences of discrete elements with the gumbel-softmax distribution," 2016.

[29] W. Fedus, I. Goodfellow, and A. M. J. a. p. a. Dai, "Maskgan: Better text generation via filling in the _," 2018.

[۳۰] ا. حاجی‌پور و س. س. سدیدپور، "استخراج خودکار کلمات کلیدی متون کوتاه فارسی با استفاده از word2vec"، پدافند الکترونیکی و سایبری، vol. 8, pp. 105-114, 2020.