Language and Linguistics
Spring & Summer 2023, 18 (35)
DOI: 10.30465/LSI.2023.8498

Original Research
Received: 2022-10-26
Accepted: 2023 - 5 – 2
ISSN: 23223847

Using Transformer-based Neural Models for Converting Informal to Formal Text in Persian

Momtazi, Saeedeh[1]
Associate Professor of Computer Engineering Department, Amirkabir University of Technology

Adibian, Majid[2]
Master student of Computer Engineering Department, Amirkabir University of Technology

Abstract

Nowadays, there is a significant growth in various data types, including textual data, which is produced through various methods, especially on social networks. These user-generated texts, however, are often informal and contain many errors, which make them unusable for many natural language processing tasks. In this research, we address the problem of informal texts and propose a model to convert informal text to formal text in Persian language. To this aim, two state-of-the-art sequence-to-sequence models, namely the encoder-decoder and the transformer-based models, are used. In addition to neural network models, we present a set of rules for converting informal text to formal text, and examine the impact of using these rules alongside each of the two models. Evaluation of the proposed models shows that the best performance has reached an accuracy of 70.7% in the SacreBLEU metric by using the combination of the transformer-based model and the set of rules.

Keywords: Natural Language Processing, Conversion of Informal Text to Formal Text, Encoder-Decoder Model, Transformer-based Model.

1. momtazi@aut.ac.ir 2. majid77@aut.ac.ir
How to cite: Momtazi, S. & Adibian, M. (2023). Using Transformer-based Neural Models for Converting Informal to Formal Text in Persian. *Language and Linguistics*, *18*(35), 47-69. doi: 10. 30465/lsi.2023.8498

زبان و زبان‌شناسی
دوره ۱۸، شماره ۳۵، بهار و تابستان، سال ۱۴۰۱، ۴۷- ۶۹
DOI: 10.30465/LSI.2023.8498

مقاله پژوهشی
تاریخ دریافت: ۱۴۰۱/۰۸/۰۴؛ تاریخ پذیرش: ۱۴۰۲/۰۲/۱۲؛ تاریخ انتشار: بهار ۱۴۰۲
ISSN: 23223847

تبدیل متن محاوره‌ای به رسمی فارسی با استفاده از شبکه‌های عصبی مبتنی بر مبدل

ممتازی، سعیده دانشیار دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران
ادیبیان، مجید کارشناسی ارشد دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده: در دنیای امروز شاهد رشد تولید داده‌های مختلف از جمله داده‌های متنی هستیم و همواره حجم زیادی از داده‌های متنی به روش‌های مختلف به خصوص در شبکه‌های اجتماعی تولید می‌شود. ولی این متن‌ها غالباً غیررسمی و دارای خطاهای بسیاری هستند که باعث می‌شود امکان استفاده از آن‌ها در بسیاری از پردازش‌های زبان طبیعی وجود نداشته باشد. در این مقاله به تبدیل متن محاوره‌ای به متن رسمی در زبان فارسی پرداخته شده است. برای این منظور دو مدل کدگذار-کدگشا و مبدل که از به‌روزترین مدل‌های دنباله-به-دنباله هستند پیاده‌سازی شده‌اند. در کنار استفاده از مدل‌های شبکهٔ عصبی، مجموعه‌ای از قواعد در تبدیل متن محاوره‌ای به رسمی فراهم شده‌اند و تأثیر استفاده از این قواعد در کنار هریک از دو مدل بررسی شده است. در انتها نتایج مدل‌های گفته‌شده مقایسه شده‌اند که در بهترین حالت این نتایج به‌دست‌آمده به دقت ۷۰.۷ درصد در معیار بلویِ ارتقایافته رسیده‌اند.

کلیدواژه‌ها: پردازش زبان طبیعی، تبدیل متن محاوره‌ای به رسمی، مدل کدگذار-کدگشا، مدل مبدل.

۱ مقدمه

دنیای امروز دنیای استفاده از هوش مصنوعی در بسیاری از جنبه‌های زندگی است. رشد

روزافزون هوش مصنوعی در زمینه‌های مختلف تحقیقاتی و علمی تا کاربردهـای تجـاری و صنعتی بر همگان روشن است. تنوع کاربردهای هوش مصنوعی تا جایی است کـه حتـی در زندگی روزمره مکرراً با این کاربردها روبه‌رو می‌شویم. در زمینهٔ هوش مصنوعی روش‌هـا و الگوریتم‌های مختلفی وجود دارند که از جمله این روش‌ها که در سال‌هـای اخیـر پرکـاربرد شده و توانسته باعث بهبود نتایج بسیاری از فعالیت‌ها شـود، اسـتفاده از شـبکه‌هـای عصـبی عمیق است. شبکه‌های عصـبی عمیـق خـود نیـز بـه انـواع مختلفـی تقسـیم می‌شـوند و در کاربردهای مختلف باید نوع مناسب با آن کاربرد را مورد استفاده قرار داد. هرچند اسـتفاده از شبکه‌های عصبی عمیق توانسته کمـک زیـادی بـه بهبـود نتـایج در بسـیاری از فعالیت‌هـای هوش مصنوعی داشته باشد ولی دارای معایبی هم هست که از جمله آن نیاز این مـدل‌ها بـه حجم بسیار زیاد داده مناسب برای آن فعالیت جهت آموزش مدل است.

یکی از شاخه‌های هوش مصنوعی، پردازش زبان طبیعی اسـت کـه شـامل پردازش‌هـای مختلف بر روی متن است. در این حوزه برای آموزش شبکه‌های عصبی عمیق حجم بسـیار بالایی از داده‌های متنی مـورد نیـاز اسـت. امـروزه بـا گسـترش شـبکه‌های اجتمـاعی و در دسترس قرار گرفتن تکنولوژی‌های مختلف، روزانه متن‌هـای بسـیاری توسـط افـراد نوشـته می‌شود و در نتیجهٔ آن حجم بسیار زیادی از داده‌های متنی تولید می‌شود. ولی با ایـن حـال، همچنان مشکل فراهم کردن داده‌های متنیِ مناسب وجود دارد که دلیل آن غیررسـمی بـودن بسیاری از داده‌های متنی است که به شکل‌های مختلف تولید می‌شود؛ چراکه بسیاری از ایـن متن‌ها به‌صورت محاوره‌ای است. علاوه بر این، خود زبان نیز دارای ویژگی‌های ذاتی اسـت کـه باعـث تغییـرات گسـترده و متنـوعی در آن می‌شـود. از جملـهٔ ایـن ویژگی‌هـا وجـود گویش‌های محلی متنوع، مختصرنویسی‌ها در نوشتار روزمره، رایج شدن اصطلاحات متنـوع و متغیر در زمان و به وجود آمدن کلمات جدید در اثر اتفاقات در دنیای واقعی هسـتند. ایـن ویژگی زبان، مشکل پویایی متون نوشته شده و افزایش غیراستاندارد بـودن آن‌هـا را تشـدید می‌کند.

از جمله اقداماتی که در این زمینه انجام می‌شـود پردازش‌هـای پایـه اسـت کـه بـر روی متن‌ها صورت می‌گیرد تا متن ظاهر استانداردتری پیدا کند و بتـوان از شـکل مناسـب‌تری از آن برای فعالیت‌های مختلف استفاده کرد. یک نمونه از این پردازش‌ها تبدیل متن محـاوره‌ای به رسمی است.

در شرایطی که حجم بسیار بالای داده‌هـای متنـیِ تولیـد شـده از شـبکه‌هـای اجتمـاعی معمولاً به‌صورت محاوره‌ای هستند، اگر بتوانیم مشکلات این محتـواهـا را برطـرف کنـیم و

آن‌ها را به شکل رسمی تبدیل کنیم به حجم بسیار بالایی از داده‌های متنیِ استاندارد دست می‌یابیم. این داده‌های متنیِ استاندارد می‌توانند در بسیاری از پردازش‌های متنی و فعالیت‌هایی که بر روی متن‌ها انجام می‌شود استفاده شود و جهش قابل توجهی در بهبود نتایج این فعالیت‌ها ایجاد کند.

در زبان فارسی نیز همچون سایر زبان‌ها موضوعِ گفته‌شده مطرح است و با وجود داده‌های متنی فراوان فارسی به‌دلیل غیراستانداردبودن این داده‌ها به‌خصوص در شبکه‌های اجتماعی، برای اجرای فعالیت‌های مختلف بر روی متن‌های فارسی داده‌های استاندارد با کمبود روبه‌رو است. با تبدیل این داده‌های غیررسمی به داده‌های رسمی به‌کمک روش‌های تبدیل متن محاوره‌ای به رسمی می‌توان شاهد فراهم شدن داده‌های مناسب بیشتری برای بسیاری از فعالیت‌های متنی و در نتیجه بهبود نتایج آن‌ها بود.

برای اجرای این تبدیل از گذشته روش‌هایی پیشنهاد شده است و همواره سعی شده عملکرد این روش‌ها بهبود یابد. با بررسی روش‌های جدید و مطرح در این زمینه و همچنین آزمودن راه‌کارهای جدید برای آن و بهبود آن‌ها برای زبان فارسی می‌توان به‌روشی بهینه برای تبدیل متن محاوره‌ای به رسمی در زبان فارسی رسید.

در پژوهش حاضر برای این منظور از شبکه‌های عصبی عمیق استفاده شده است. برای این هدف ابتدا دو مدل مختلف مبتنی بر شبکه‌های عصبی یعنی مدل مبدل[1] و مدل کدگذار-کدگشا[2] با استفاده از شبکه‌های بازگشتی که مدل‌های به‌روز در این حوزه هستند پیشنهاد شده‌اند و بهترین حالت ممکن در هریک به‌دست آمده است. همچنین بررسی‌هایی بر روی الگوهای تبدیل متن محاوره‌ای به رسمی انجام شده و با استفاده از این الگوها، قواعدی ساخته شده است که سعی شده با استفاده از آن‌ها در کنار استفاده از شبکه‌های عصبی عمیق نتایج بهبود یابد.

در ادامه ابتدا کارهای انجام شده در این زمینه در زبان فارسی و سایر زبان‌ها بررسی شده‌است و سپس مدل پیشنهادی و نحوه آموزش آن توضیح داده شده است. در انتها روش ارزیابی نتایج آورده شده و نتایج حاصل از مدل پیشنهادی مورد بررسی قرار گرفته است.

۲ پیشینه کارهای انجام‌شده در تبدیل متن محاوره‌ای به رسمی

در تبدیل متن محاوره‌ای به رسمی در زبان فارسی پژوهش‌های محدودی انجام شده است.

1. transformer
2. encoder-decoder

به همین دلیل در این قسمت پس از بررسی کارهای انجام‌شده در فارسی بـه مقـالات اخیـر مربوط به سایر زبان‌ها هم می‌پردازیم.

۱-۲ در زبان فارسی

در مقالهٔ ارائه‌شده توسط آرمین و شمس‌فـرد (۲۰۱۱: ۱۷۳٤–۱۷۲٤) در سـال ۲۰۱۱، روشـی مبتنی بر قاعده و با استفاده از مدل‌سازی آماری برای تبدیل متن محاوره‌ای بـه رسـمی ارائـه شده است. در ابتدا، روش ساخت پیکره متن مورد نیاز توضیح داده شده است. این پیکره بـا استفاده از کتاب‌هایی که متن محـاوره‌ای دارنـد و زیرنـویس فارسـی چنـدین فـیلم و ده‌هـا وبلاگ استخراج شده است که در نهایتـاً پیکـره‌ای بـا ۴۴ هـزار کلمـه فـراهم شـده اسـت. الگوریتم پیشنهادی در این مقاله به‌طور کلـی شـامل واحدسـازی[1] مـتن محـاوره‌ای و یـافتن کلمات پیشنهادی برای شکل رسمی آن‌ها با اسـتفاده از مجموعـه‌ای از قـوانین و در نهایـت رتبه‌بندی مجموعهٔ پیشنهادی بـا اسـتفاده از روش احتمـالاتی اسـت. در ایـن مقالـه بـیش از بیست قانون برای این تبدیل ساخته شده و در کنار آن نیز پایگـاه داده‌ای بـرای تبـدیل مـتن محاوره‌ای به رسمی برای کلماتی که قانون خاصی ندارند فراهم شده است. همچنـین بـرای یافتن احتمال هریک از کلمات کاندید از بایگرم[2]ها استفاده شده است و احتمـال وقـوع هـر کلمه وابسته به کلمهٔ قبلی آن از پیکره «بیجن خان» بدست آمده است.

در مقاله رسولی و همکاران (۲۰۲۰) روشی برای تبدیل مـتن محـاوره‌ای بـه رسـمی در زبان فارسی بررسی شده است. روش ارائـه شـده در ایـن مقالـه اسـتفاده از استانداردسـازی دنباله–به–دنباله[3] در ترجمهٔ ماشـینی اسـت بـه‌طوری کـه ورودی مـدل دنبالـه‌ای از کلمـاتِ محاوره‌ای و خروجـی آن دنبالـه‌ای از کلمـات اسـتاندارد اسـت و مـدل استفاده‌شـده مـدل ترجمه‌ای با استفاده از شبکه مبدل با شش لایه بر پایه برت[4] است. در این مقاله بـه‌دلیل نبـود داده‌های محاوره‌ای به رسمی از مجموعه قواعدی برای تبـدیل مـتن رسـمی بـه محـاوره‌ای استفاده شده است تـا دادهٔ مناسـب بـرای آمـوزش مـدل سـاخته شـود و از مـتن صـفحات ویکی‌پدیا و یک میلیون جمله از پیکره «میزان» به‌عنوان متن اسـتاندارد اسـتفاده شـده اسـت. همچنین برای داده‌های تست از نظرات موجود در سایت‌هـای خبـری و ترجمـه فـیلم‌هـا و دیالوگ‌های فیلم‌ها استفاده شده است. برای محاسبهٔ دقت پایه از ابزار «هضم» اسـتفاده شـده است که از قواعد برای این تبدیل استفاده می‌کند.

1. tokenization
2. bigram
3. sequence-to-sequence
4. Bert

معصومی و همکاران (۲۰۲۰) نیز در مقالـه‌ای روشـی بـرای جمـع‌آوری و اعتبارسـنجی متون محاوره‌ای ارائه کرده‌انـد. عمـده تمرکـز ایـن مقالـه بـر روی روش جمـع‌آوری متـون محاوره‌ای و شکل رسمی آن‌ها است. روشی که ارائـه شـده اسـتفاده از پیـام‌رسـان تلگـرام[1] به‌عنوان یک پیام‌رسان محبوب در ایران برای فراهم‌کردن شکل رسـمی جمـلات محـاوره‌ای است. به این صورت که ابتدا مجموعهٔ بزرگی از داده‌های متنی محاوره‌ای از توییتر و تلگـرام و سایت دیجی‌کالا فراهم شده است و سپس در پیام‌رسان تلگرام سیستمی طراحی شـده کـه افراد می‌توانند جملات محاوره‌ای را دریافت کنند و سپس شکل رسمی آن‌هـا را بنویسـند و این نوشته‌ها به‌عنوان جملات کاندید برای شکل رسمی آن جملهٔ محاوره‌ای در نظـر گرفتـه می‌شود. همچنین افراد می‌توانند در این سیستم به جملات کاندید برای هر جمله محـاوره‌ای رای دهند و در نتیجهٔ این کار با توجه به تعداد افرادی که هـر جملـهٔ کاندیـد را قابـل قبـول دانسته‌اند می‌توان جملهٔ رسـمی را بـرای آن جملـهٔ محـاوره‌ای تشـخیص داد. ایـن سیسـتم طراحی شده «TeleCrowd» نامیده شده است و توانسته در زمان کوتاه‌تر و با هزینـهٔ کمتـری نسبت به حالت عادی ۲۷۰۰ جمله کاندید و ۲۱۰۰۰ رای را فراهم کند. همچنین در انتهـا بـا ارزیابی نتایج به‌دست آمده با استفاده از معیار بلو[2] به امتیـاز ۰.۵۴ رسـیده اسـت. ایـن معیـار یک روش رایج در ترجمهٔ ماشینی است و در قسمت‌های بعدی بـه‌طور کامـل توضـیح داده خواهد شد.

۲-۲ در سایر زبان‌ها

در مقالـهٔ ارائـه شـده توسـط منسـفیلد[3] و همکـاران (۲۰۱۹: ۱۹۶–۱۹۰) بـرای تبـدیل مـتن محاوره‌ای به متن رسمی در زبان انگلیسی به‌عنوان یک قدم مهم در تبدیل مـتن بـه گفتـار از روش‌های ترجمهٔ ماشینی استفاده شده اسـت. در ایـن مقالـه از روش دنبالـه–بـه–دنبالـه در ترجمهٔ ماشینی استفاده شده است و ورودی و خروجی دنبالـه‌ای از اجـزای کلمـات اسـت. یعنی جملهٔ مورد نظر به‌صورت دنباله‌ای از زیر کلمات داده می‌شود و دنبالـه‌ای از ایـن زیـر کلمات دریافت می‌شود و این زیر کلمات در واقع اجزای سازندهٔ کلمه هسـتند کـه ترکیـب آن‌ها کلمه را می‌سازد. این مورد برای آن اسـت کـه نسـبت بـه حالـت اسـتفاده از حـروف وابستگی‌های با فاصله زیاد را بهتر در نظر بگیرد. نتایج نشان می‌دهد با استفاده از ایـن روش

1. Telegram 2. BLEU 3. C. Mansifeld

و البته استفاده از حجم بالای داده‌های متنی آماده در این زمینه در زبان انگلیسی خروجی‌ها به امتیاز بلوی بالاتر از ۹۸ درصد رسیده‌اند.

در مقالهٔ آرفین[1] و تیون[2] (۲۰۲۰) روشی برای تبدیل متن محاوره‌ای به رسمی در زبان مالایی بیان شده است. در این مقاله ابتدا انواع تغییراتی که در متن‌هایی که در شبکه‌های اجتماعی ایجاد می‌شود گفته شده که باعث می‌شود متن‌های شبکه‌های اجتماعی برای استفاده در بسیاری از پردازش‌های زبان طبیعی مناسب نباشند. روش ارائه شده در این مقاله استفاده از مجموعه قواعدی برای تبدیل متن محاوره‌ای به رسمی است. روش ارزیابی در این مقاله بررسی دقت برچسب‌گذاری اجزای سخن[3] در استفاده از متن رسمی شده به وسیله این روش است. درواقع هرچه رسمی‌سازی در این روش بهتر باشد، دقت برچسب‌گذاری اجزای سخن نیز بیشتر خواهد بود. نتایج نیز نشان می‌دهد که استفاده از این قواعد باعث بهبود محسوس دقت برچسب‌گذاری شده است که نشان‌دهنده آن است که رسمی‌سازی انجام‌شده روشی مناسب برای تبدیل متن محاوره‌ای به رسمی در متون شبکه‌های اجتماعی است و همچنین این روش می‌تواند باعث بهبود دقت سایر پردازش‌های زبان طبیعی شود.

در مقالهٔ کوژیربایف[4] و یسنبایف[5] (۲۰۲۰: ۱۱۵–۱۲۲)، روشی برای استانداردسازی در زبان قزاقستانی ارائه شده است. در این روش از راه‌کارهای ترجمهٔ ماشینی استفاده شده است تا متن غیراستاندارد به حالت استاندارد تبدیل شود و در سایر کارهای پردازش زبان طبیعی بتوان از این متن استفاده کرد. در این مقاله از دو روش آماری و شبکهٔ عصبی استفاده شده است. از روش آماری به‌عنوان یک روش پایه استفاده شده و روش آن مبتنی بر عبارات سه کلمه‌ای است و دقت به‌دست‌آمده از آن به‌عنوان دقت پایه محاسبه شده است. در روش شبکهٔ عصبی از مدل دنباله–به–دنباله مبتنی بر کلمات در ترجمهٔ ماشینی استفاده شده است و شبکهٔ عصبی استفاده‌شده شبکه کدگذار–کدگشا با استفاده از حافظهٔ کوتاه‌مدت بلند[6] است. در نهایت، برای ارزیابی نتایج به‌دست‌آمده از این مدل‌ها از معیار بلو استفاده شده است. نتایج به‌دست‌آمده نشان می‌دهد در روش آماری به امتیاز ۲۱.۶۷ و در روش شبکهٔ عصبی به امتیاز ۲۹.۷۴ در معیار بلو رسیده است.

در مقالهٔ لیو[7] و همکاران (۲۰۲۰: ۱۸۷۹–۱۸۶۸)، روشی برای تبدیل اصطلاحات

1. S. N. A. N. Ariffin
2. S. Tiun
3. Part-of-speech tagging
4. Z. Kozhirbayev
5. Z. Yessenbayev
6. LSTM
7. Y. Liu

محاوره‌ای که پزشکان در پرونده‌های پزشکی ثبت کرده‌اند به اصطلاحات استاندارد در زبان چینی ارائه شده است. در این مقاله در ابتدا روشی از ترکیب مدل برت[1] و الگوریتم شباهت متن ارائه شده است که با استفاده از چند کلمه‌ای‌ها، کلمه‌های کاندید برای جایگزینی کلمهٔ محاوره‌ای به دست می‌آید و سپس با استفاده از روش دسته‌بندی در برت کلمهٔ مناسب از بین کلمات کاندید جدا می‌شود. در ادامه برای بهبود روش پیشنهادشده سعی شده است اندازهٔ مجموعه کلمات کاندید جایگزینی افزایش یابد تا امکان یافتن کلمهٔ درست بیشتر شده و دقت بهبود پیدا کند و در عین حال از ساختار آن برای حذف نویزها از این مجموعه استفاده شده است. بر اساس نتایج ارائه‌شده دقت با استفاده از روش ابتدایی ۸۳.۵ درصد بوده که پس از اعمال راه‌کار گفته شده دقت تا ۰.٦ درصد بهبود پیدا کرده است و در عین حال حجم محاسبات ۲٦.۷ درصد کمتر شده است.

در مقاله‌ای از آرورا[2] و کنسال[3] (۲۰۱۹: ۱–۱٤)، به محاوره‌ای و غیررسمی بودن پیام‌ها در شبکهٔ اجتماعی توییتر و در زبان‌های مختلف پرداخته شده است که باعث می‌شود برای بسیاری از پردازش‌های زبان طبیعی مانند تحلیل احساس متن، این متن‌ها مناسب نباشند. در این مقاله روشی برای نرمال‌سازی ارائه شده و سپس به تحلیل احساس در پیام‌های توییتری پرداخته شده است. در روش پیشنهادی برای نرمال‌سازی پیام‌های غیررسمی توییتر سه قدم شامل واحدسازی و تشخیص و جایگزینی کلمات خارج از لغات زبان و ریشه‌یابی کلمات مطرح شده است. سپس برای تحلیل احساس از شبکهٔ عصبی پیچشی[4] برپایهٔ حروف استفاده شده است تا با استفاده از دو مرحلهٔ گفته‌شده بتوان تحلیل احساس برای پیام‌های غیررسمی توییتری را انجام داد. در نتیجهٔ آن سه مزیت برای آن مطرح شده که شامل امکان تحلیل احساس برای داده‌های غیررسمی، مدیریت حافظه در اثر استفاده از حروف به جای کلمات و بهبود دقت در تحلیل احساس متن‌های غیررسمی هستند.

۳ تبدیل متن محاوره‌ای به رسمی

زبان یک وسیله ارتباطی است که براساس تعریف سوسور[5] (۱۹۱٦) دارای صورت و معنا است. صورت زبان به‌صورت خط نوشتاری یا آواهای زبانی تجلی می‌کند. ویژگی مهم خط این است که قابل پردازش است. خط غیررسمی زیرگونه‌ای از خط فارسی است که در

1. Bert
2. M. Arora
3. V. Kansal
4. Convolutional Neural Network
5. F. De Saussure

فضای مجازی به‌دلیل سهولت و سرعت در تایپ حروف مورد استفاده قرار می‌گیرد. امید طبیب‌زاده (۱۳۹۸) فارسی غیررسمی را گونهٔ گفتاریِ نوشتار فارسی معرفی کرده است. اگرچه این خط توسط انسان قابل درک است اما مدل‌های رایانشی قابلیت پردازش دقیق آنها را ندارند چراکه مدل‌های زبانی مورد استفاده در ابزارهای رایانشی براساس داده‌های معیار آموزش دیده‌اند. بر همین اساس برای پردازش داده‌های غیررسمی نیازمند تبدیل این ساختار خط به یک ساختار معیار هستیم. سؤال اصلیِ پژوهش حاضر این است که آیا می‌توان با استفاده از رویکردهای نوین در شبکه‌های عصبی و یادگیری عمیق خط غیررسمی را به ساختار معیار آن تبدیل کرد.

همان‌طور که در قسمت‌های قبل گفته شد، در بسیاری از پردازش‌های زبان طبیعی نیاز است که متن مورد استفاده به شکل رسمی باشد و یادگیری مدل‌های مختلف در این زمینه بر روی متن رسمی انجام می‌شود و استفاده از متن‌هایی که به‌شکل غیررسمی هستند نتیجهٔ مناسبی نخواهد داشت. ازطرفی، حجم زیادی از متن‌های موجود متن‌های شبکه‌های اجتماعی و مکالمات سریال‌ها و فیلم‌ها است که عموماً به‌صورت محاوره‌ای هستند.

با بررسی پیشینهٔ کار در تبدیل متن محاوره‌ای به رسمی مشخص می‌شود که این کار در زبان فارسی کمتر مورد توجه بوده است و مقالات کمی در این مورد وجود دارد. علاوه بر آن راه‌کارهای موجود عموماً استفاده از قواعدی در تبدیل متن محاوره‌ای به رسمی است که این روش‌ها هرچند در مواردی نتایج خوبی دارند ولی ممکن است با تغییرات اشتباه خود باعث به‌وجودآمدن خطاهای جدید شوند. همچنین در سال‌های اخیر راه‌کارهایی در استفاده از شبکه‌های عصبی در این زمینه ارائه شده است که نگاه این روش‌ها برمبنای کلمات است و برای پوشش دادن انواع کلمات محاوره‌ای با استفاده از مجموعه‌ای از قواعد، کلمات محاوره‌ای ساخته شده و استفاده می‌شوند. این روش‌ها به‌دلیل استفاده از شبکه‌های عصبی به دقت‌های خوبی رسیده‌اند ولی به‌دلیل استفاده از واحدهای کلمه‌ای نمی‌توانند تمامی حالات برای کلمات محاوره‌ای را پوشش دهند و در موارد خاص دچار مشکل می‌شوند.

در این مقاله سعی شده است با ارائه راه‌کاری جدید در زمینهٔ تبدیل محاره به رسمی به بهبود این مورد در زبان فارسی بپردازیم. در این راستا تغییر نگاهی به مسئله داشته‌ایم و از واحدهای حروف به‌جای کلمات به‌عنوان واحدهای سازندهٔ متن استفاده شده است. همچنین از شبکه‌های عصبی جدید برای آموزش مدل و بهبود دقت خروجی‌ها استفاده شده است. در کنار موارد گفته‌شده مجموعه قواعدی برای تبدیل متن محاوره‌ای به رسمی در زبان فارسی ساخته شده است که تأثیر استفاده از این قواعد بررسی خواهد شد. در ادامه مراحل استفاده شده در مدل پیشنهادی ارائه می‌شود.

۱-۳ آماده‌سازی داده‌ها

برای تبدیل متن محاوره‌ای به رسمی نیاز است که مجموعه‌ای بزرگ از جملات و عبارات محاوره‌ای و معادل رسمی آن‌ها را داشته باشیم. برای این منظور از مجموعه داده‌ای استفاده شده است که حاصل کار مقالهٔ ارائه‌شده از معصومی و همکاران (۲۰۲۰) است که در بخش پیشینه به آن پرداختیم. این مجموعه داده شامل ۴۵۰۱ جملهٔ محاوره‌ای و شکل رسمی آن‌ها است که به‌وسیله نیروهای انسانی تهیه و ارزیابی شده‌اند و دارای کیفیت خوبی است. در جدول ۱ نمونه‌هایی از این داده‌ها را می‌بینیم.

جدول ۱- چند نمونه از داده‌های محاوره‌ای و شکل رسمی آن‌ها در داده‌ها

جملهٔ محاوره‌ای	جملهٔ رسمی
خونه آخرین پناهه	خانه آخرین پناه است
هوا سرد شده بچه هامو از بالکن آوردم تو خونه یخ نزنن	هوا سرد شده بچه هایم را از بالکن آوردم توی خانه که یخ نزنند
کاش همین الانه میشد برم دریا تا حالم خوب شه	کاش همین الان میشد بروم دریا تا حالم خوب شود

برای فراهم کردن دادهٔ مناسب در گام اول لازم است ابتدا مجموعه پیش‌پردازش‌هایی بر روی داده‌های متنی خام انجام شود تا به شکل مناسب تبدیل شوند. این مجموعه پاک‌سازی‌ها مجموعه‌ای از توابع ساده هستند که بر روی متن ورودی اجرا می‌شوند و متن پاک‌سازی شده را تولید می‌کنند. این توابع شامل جایگزینی حروف عربی، جایگزینی حروف انگلیسی، حذف علائم نگارشی، جایگزینی اعداد و پاک‌سازی نهایی متن هستند.

در تبدیل متن محاوره‌ای به رسمی به‌طور کلی دنباله‌ای از کلمات محاوره‌ای به دنباله‌ای از کلمات رسمی تبدیل می‌شود. در این گونه از مسائل استفاده از شبکه‌های عصبی دنباله-به-دنباله بسیار مناسب هستند. برای آموزش این مدل‌ها نیاز است که مجموعه‌ای از دنباله کلمات محاوره‌ای و مجموعه‌ای از دنباله کلمات رسمی متناظر با آن‌ها را داشته باشیم.

در این پژوهش با بررسی‌های بسیار و ارزیابی‌های متعدد برای اندازهٔ دنبالهٔ کلمات مورد استفاده در آموزش مدل و تعداد قدم‌ها در ساخت این دنباله‌ها، مشخص شد که حالت بهینه استفاده از دنباله‌های ۸ کلمه‌ای با گام‌های ۲ کلمه‌ای از جملات است. به عنوان مثال در یک جملهٔ محاوره‌ای و شکل رسمی آن در داده‌های آموزشی، ۸ کلمهٔ ابتدایی این جملات را به‌عنوان دنبالهٔ اول استفاده می‌کنیم و سپس دو کلمه به جلو می‌رویم و ۸ کلمهٔ بعدی (که ۶ کلمهٔ آن با دنباله قبلی مشترک است) را استفاده می‌کنیم. و به همین شکل ادامه می‌دهیم.

در ادامه از ۴۴۰۰ جمله از جملات موجود برای ساخت این دنباله‌ها استفاده شده و باقی جملات به‌عنوان داده‌های آزمون در نظر گرفته شده‌اند. در نتیجهٔ پردازش گفته‌شده، ۳۰۸۲۱ دنباله کلمه به دست آمده است.

۲-۳ مدل پیشنهادی

پس از انجام فعالیت‌های گفته شده در زمینه فراهم کردن داده مجموعه‌ای از زوج داده‌ها خواهیم داشت که هر زوج شامل دنباله کلمات محاوره‌ای و دنباله کلمات رسمی آن‌ها خواهد بود.

همان‌طور که اشاره شد نوع نگاه ما به مسئله نسبت به کارهای مشابهی که در این زمینه انجام شده متفاوت است و نگاه ما به جای دنباله‌ای از کلمات به‌صورت دنباله‌ای از حروف است. در واقع در متن محاوره‌ای انواع کلمات با شکل‌های مختلف ممکن است ظاهر شود و در نگاه کلمه‌ای باید به هر کلمه و انواع محاوره‌ای آن عددی نسبت داد که این امر با توجه به انواع کلمات محاوره‌ای و تغییرات مدام آن‌ها غیرممکن است و تنها در صورتی که انواع خاص و محدودی از کلمات محاوره‌ای را در نظر بگیریم ممکن خواهد بود که این خود یک محدودیت و ضعف محسوب می‌شود. به همین دلیل در این پژوهش به جای استفاده از کلمات به‌عنوان اجزاء سازنده متن از حروف استفاده می‌کنیم که با آموزش مدل با استفاده از این دنباله حروف مشکل گفته شده نیز برطرف خواهد شد و این ضعف وجود نخواهد داشت. همچنین استفاده از واحد حروف به جای کلمات این امکان را فراهم می‌آورد که مدل شبکهٔ عصبی به‌طور ضمنی قواعد تبدیل متن محاوره‌ای به رسمی را حتی در داخل ساختار کلمات یاد بگیرد.

با توجه به مطالب گفته شده دنباله‌های کلماتی که در مرحله قبل به دست آورده‌ایم را به دنباله‌هایی از حروف تبدیل می‌کنیم تا مناسب روش ارائه شده در این پژوهش برای تبدیل متن محاوره‌ای به رسمی شود. حال به هریک از حروف شناسه‌ای را نسبت می‌دهیم و دنباله حروف به دست آمده از مرحله قبل را به دنباله‌ای از اعداد تبدیل می‌کنیم و هر دنباله را در یک بردار عددی ذخیره می‌کنیم. در انتهای فرایند گفته شده مجموعه‌ای از زوج بردارهای عددی داریم که در هر زوج بردار، اولی مربوط به عبارت محاوره‌ای و دومی مربوط به عبارت رسمی آن است.

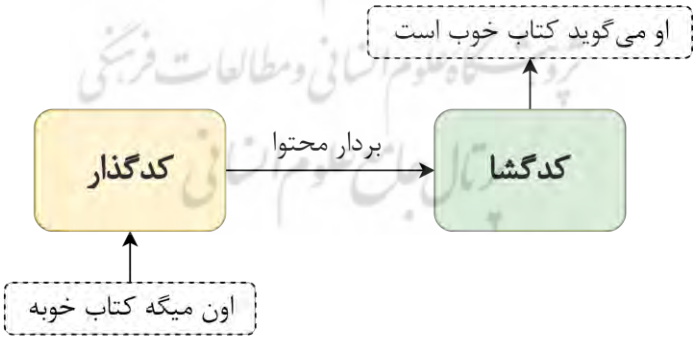
پس از تبدیل داده‌های متنی به فضای برداری، جهت استفاده از این داده‌ها دو رویکرد مختلف مبتنی بر شبکه‌های عصبی ارائه می‌شود که عبارتند از شبکه‌های کدگذار-کدگشا و

مبدل. در هر دوی این مدل‌ها آموزش بر روی داده‌های آموزشی انجام می‌شود و سپس دقت بر روی داده‌های ارزیابی به دست می‌آید و پارامترهای این مدل‌ها تغییر داده می‌شوند تا بهترین دقت بر روی داده‌های ارزیابی به دست آید.

۱-۲-۳ مدل کدگذار-گدگشا

در پردازش داده‌های متنی از آنجا که ورودی‌ها به‌صورت دنباله‌ای از داده‌ها هستند شبکه‌های عصبی مکرر بسیار مناسب هستند. مسائل دنباله-به-دنباله نوع خاصی از این گونه مسائل هستند که در آن‌ها هم ورودی و هم خروجی دنباله‌ای از داده‌ها است. مدل کدگذار-کدگشا برای حل این مسائل به وجود آمده‌اند. مثال معروف این گونه مسائل، مسائل ترجمهٔ ماشینی هستند که در آن‌ها دنباله‌ای از کلمات زبان مبدا به‌عنوان ورودی داده می‌شود و دنباله‌ای از کلمات زبان مقصد (که ترجمه دنباله کلمات ورودی هستند) به‌عنوان خروجی به دست می‌آید.

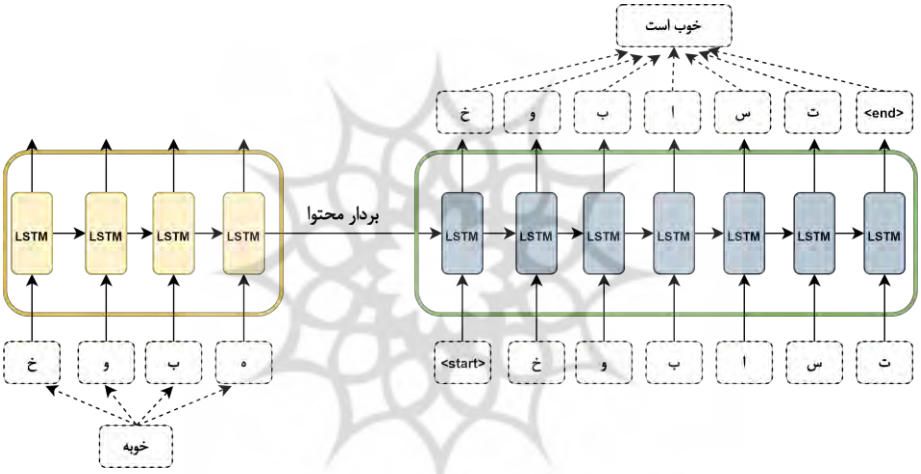
در نگاه سطح بالا از معماری این مدل که در شکل ۱ ارائه شده است می‌توان این مدل را به دو قسمت کدگذار و کدگشا تقسیم کرد که به وسیله برداری که از یکی به دیگری انتقال می‌یابد، به هم متصل هستند. در قسمت کدگذار هر واحد از دنبالهٔ ورودی پردازش می‌شود و تمام اطلاعات مربوط به دنبالهٔ ورودی را در برداری با طول ثابت جمع می‌کند. در قسمت کدگشا برداری که در قسمت قبل ساخته شده خوانده می‌شود و سعی می‌شود توالی هدف به‌صورت واحد پیش‌بینی شود.



شکل ۱- معماری کلی مدل کدگذار-کدگشا

با نگاه دقیق‌تر به هر قسمت از دو قسمت گفته شده، هریک می‌تواند یک مدل حافظهٔ

کوتاه‌مدت طولانی باشد که به یک‌دیگر متصل شده‌اند. در شکل ۲ این مدل در یک نمونه از تبدیل متن محاوره‌ای به رسمی با نگاه دقیق‌تر نشان داده شده است. همان‌طور که در این تصویر دیده می‌شود خروجی‌های قسمت کدگذار در نظر گرفته نمی‌شود و تنها خروجی که به مرحله بعد می‌رود، مورد استفاده است. این خروجی در نهایت پس از یک دنباله از ورودی‌ها به قسمت کدگشا می‌رود. ورودی اولیه در این قسمت علامت <start> است که به‌معنی شروع دنباله خروجی است و با این ورودی و برداری که از قسمت کدگذار آمده کلمه اول به دست می‌آید. سپس این خروجی به‌عنوان ورودی مرحله بعدی استفاده می‌شود و این فرایند تا رسیدن به علامت <end> ادامه می‌یابد.



شکل ۲- مدل کدگذار-کدگشا با استفاده از مدل حافظهٔ کوتاه‌مدت بلند

در مدل پیشنهادی مبتنی بر کدگذار-کدگشا در پژوهش حاضر در هر دو قسمت کدگذار و کدگشا از لایه‌های حافظهٔ کوتاه‌مدت بلند دوطرفه استفاده شده است تا با توجه به محتوای قبل و بعد از هر کلمه آموزش صورت گیرد. در انتها و قبل از خروجی نیز یک لایهٔ توجه[1] ساده اضافه شده است تا دقت بهبود یابد. در این مدل از تابع آدام[2] به‌عنوان بهینه‌ساز و از بیشینهٔ نرم[3] به‌عنوان تابع فعال‌سازی استفاده شده است. داده‌هایی که برای آموزش به مدل داده شده سه دسته هستند. ورودی قسمت کدگذار دنباله متن محاوره‌ای

1. attention layer 2. Adam 3. softmax

است و ورودی قسمت کدگشا دنبالهٔ رسمی است که با کد مربوط به <start> شروع می‌شوند و خروجی مدل هم دنبالهٔ رسمی است که با کد مربوط به <end> پایان می‌یابند. در هر دو مدل گفته‌شده پس از چند مرحله تکرار آموزش بر روی کل داده‌های آموزشی به دقت قابل قبولی بر روی این داده‌ها می‌رسیم.

۲-۲-۳ مدل مبدل

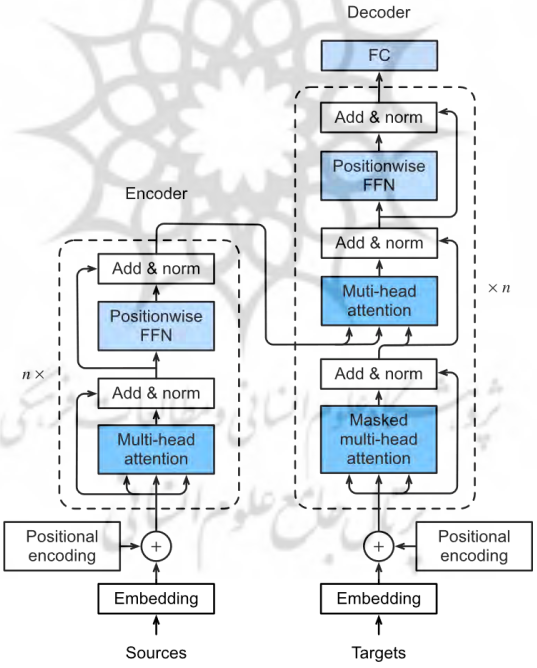
مدل مبدل نامی است که به معماری کدگذار-کدگشا بر اساس لایه‌های خودتوجه[1] داده شده است. تفاوت اصلی این مدل با مدل کدگذار-کدگشا که در قسمت قبل گفتیم آن است که دنبالهٔ ورودی به‌صورت متوالی پردازش نمی‌شوند بلکه می‌توانند به‌صورت موازی پردازش شوند و انتقال یابند که این امر باعث بهینه شدن استفاده از واحد پردازش گرافیکی[2] و افزایش سرعت آموزش می‌شود. به‌عبارتی در حالت استفاده از روش قبل برای دریافت جمله رسمی از جمله ورودی محاوره‌ای باید مرحله به مرحله کلمات ورودی را وارد می‌کردیم ولی در مدل مبدل تمام متن جمله را هم‌زمان انتقال می‌دهیم و خروجی را دریافت می‌کنیم.

به‌طور کلی معماریِ این مدل شامل دو بخش کدگذار و کدگشا است. در هر دو قسمت ابتدا یک لایهٔ تعبیه‌سازیِ[3] وجود دارد که هر کلمه را به فضای برداری می‌برد و برای آن که محل کلمه هم در نظر گرفته شود از تعبیه‌سازی مکانی[4] نیز استفاده شده است. در لایه‌های بعدی در قسمت کدگذار با یک لایهٔ توجه مقدار ارتباط هر کلمه به دیگر کلمه‌ها به دست می‌آید و سپس با یک لایهٔ تغذیه رو به جلو[5] آن را به شکل مناسب برای قسمت کدگشا تبدیل می‌کند که در انتهای این قسمت، بردارهایی برای هر کلمه به‌صورت موازی به دست می‌آید. در قسمت کدگشا در ابتدا توجهِ پوشانده‌شده[6] وجود دارد که تأثیر قسمت‌های مختلف ماتریس را در محاسبه و پیش‌بینی کلمات نتیجه مشخص می‌کند و قسمت‌های بی‌تأثیر را صفر قرار می‌دهد. دو لایهٔ بعدی این قسمت مانند قبل است که البته از خروجی قسمت کدگذار استفاده می‌کند (وسوانی[7] و همکاران، ۲۰۱۷). شکل ۳ معماری یک شبکه مبدل را نمایش می‌دهد.

در مدل مبدل پیشنهادی در پژوهش حاضر دو قسمت کدگذار و کدگشا پیاده‌سازی

1. self-attention
2. graphics processing unit (GPU)
3. embedding
4. positional embedding
5. feed forward
6. masked attention
7. A. Vaswani

شده است. در قسمت کدگذار ابتدا داده‌ها ورودی در بردارهای عددی تعبیه می‌شوند و سپس تعبیه مکانی بر روی این داده‌ها انجام می‌شود و در انتها از دو لایهٔ کدگذار استفاده می‌شود که در هریک از این لایه‌ها از لایه‌های توجه استفاده شده است. قسمت کدگشا هم مانند قسمت کدگذار است و در انتهای آن دو لایهٔ گدگشا وجود دارد که در آن از لایه‌های توجه استفاده شده است. همچنین برای تابع بهینه‌ساز[1] از آدام و برای تابع فعال‌سازی[2] بیشینهٔ نرم به کار گرفته شده است. مشابه مدل قبل داده‌های مورد استفاده در این مدل سه قسمت هستند. قسمت اول داده‌ها، ورودی بخش کدگذار است که دنباله محاوره‌ای است، قسمت دوم داده‌ها ورودی بخش کدگشا است که دنبالهٔ رسمی است و با کد مربوط به <start> شروع می‌شوند و قسمت سوم داده‌های استفاده شده در این مدل داده‌های خروجی قسمت کدگشا هستند که دنبالهٔ رسمی است و فقط با کد مربوط به <end> پایان می‌یابد. سپس این داده‌ها برای آموزش به مدل داده می‌شود.



شکل ۳- معماری مدل مبدل (واسوانی و همکاران،۲۰۱۷)

1. optimaizer function

2. activation function

۳-۳ استفاده از قواعد در تبدیل متن محاوره‌ای به رسمی

در تبدیل متن محاوره‌ای به رسمی در کنار استفاده از مدل‌های شبکه‌های عصبی می‌توان از قواعدی در این تبدیل استفاده کرد به‌طوری که حتی برخی از کارهای انجام شده در این زمینه تنها با استفاده از این قواعد، متن محاوره‌ای را به رسمی تبدیل می‌کنند.

در این قسمت مجموعه‌ای از قواعد تبدیل متن محاوره‌ای به رسمی در فارسی را بیان می‌کنیم که با پیاده‌سازی این قواعد و استفاده از آن‌ها در کنار مدل شبکهٔ عصبی تغییرات دقت در استفاده از این قواعد بررسی می‌شود (آرمین و شمس‌فرد، ۲۰۱۱: ۱۷۳٤-۱۷۲٤):

- قاعده اول: تبدیل «ون» به «ان» و تبدیل «وم» به «ام» در صورتی که کلمه جدید در مجموعه لغات فارسی موجود باشد؛ مانند تبدیل «خونه» به «خانه» و «کدوم» به «کدام»
- قاعده دوم: تبدیل «ا» در انتهای کلمه به «ها» در صورتی که باقی کلمه به جز «ا» انتهایی آن در مجموعه لغات فارسی موجود باشد؛ مانند «درختا» به «درخت‌ها»
- قاعده سوم: تبدیل «رو» به «را» و «و» در انتهای کلمه به «را» به‌صورت یک کلمه مجزا در صورتی که باقی کلمه به جز «و» انتهایی آن در مجموعه لغات فارسی موجود باشد؛ مانند «کتابو» به «کتاب را»
- قاعده چهارم: تبدیل «ا» در انتهای کلمه به «اه» در صورتی که باقی کلمه به جز «ا» انتهایی آن در مجموعه لغات فارسی موجود باشد؛ مانند «سیا» به «سیاه»
- قاعده پنجم: تبدیل ضمایر محاوره‌ای به رسمی که به‌صورت زیر است:

o مون: مان	o اون: آن
o تون: تان	o ایشون: ایشان
o شون: شان	o اونها، اونا: آنها

- قاعده ششم: تبدیل ضمایر بعد از حروف اضافه از حالت محاوره‌ای به رسمی که بعضی از آن‌ها را در زیر می‌بینیم:

o	ازم: از من	o	ازت: از تو		و ...
o	بهم: به من	o	بهت: به تو		و ...
o	برام: برای من	o	برات: برای تو		و ...
o	باهام: با من	o	باهات: با تو		و...

- قاعده هفتم: تبدیل «ه» در انتهای کلمه به «است» در صورتی که باقی کلمه به جز «ه» انتهایی آن در مجموعه لغات فارسی موجود باشد؛ مانند «خوبه» به «خوب است»

- قاعده هشتم: در این قسمت برخی از کلمات پرتکراری که شکل محاوره‌ای آن‌ها قاعده خاصی ندارد را به‌صورت موردی به شکل رسمی آن‌ها تبدیل می‌کنیم؛ مانند «چار» به «چهار»، «کوچیک» به «کوچک»، «شیش» به «شش»، «دیگه» به «دیگر» و....

مشخص است که استفاده از این قواعد به تنهایی دقت کمتری از شبکه‌های عصبی خواهد داشت چراکه این قواعد به‌صورت توابعی هستند که بر روی هریک از کلمات اجرا می‌شوند و هرچند بسیاری از کلمات محاوره‌ای را به شکل رسمی آن‌ها تبدیل می‌کنند ولی کلمات زیادی را هم به اشتباه تغییر می‌دهند. به‌عنوان مثال اجرای قاعده هفتم بر روی کلمهٔ «نامه» آن را به کلمهٔ «نام است» تبدیل می‌کند که یک خطا در این فرایند است. همچنین این روش نمی‌تواند از محتوای متن و کلمات مجاور برای تشخیص شکل رسمی کلمه مورد نظر استفاده کند و از این نظر نیز دقت کمی خواهد داشت. برای بهبود این شرایط به‌ازای هریک از قوانین گفته‌شده تعدادی نمونهٔ پرکاربردتر را فراهم کرده‌ایم و در مجموع حدود صد نمونه از همهٔ قواعد به دست آمده است. حال با جایگزین کردن نمونه‌های فراهم‌شده در متن مورد نظر، این قواعد را اثر می‌دهیم. همچنین اثر این قواعد را قبل و بعد از شبکه‌های عصبی مقایسه و بررسی خواهیم کرد.

۴ ارزیابی

ارزیابی نتایج حاصل از تبدیل متن محاوره‌ای به رسمی آسان نیست. در بهترین حالت آن بهتر است یک شخص تک‌تک کلمات در هریک از جملات نتیجه را بررسی کند و تعداد کلمات محاوره‌ای که به‌درستی تغییر کرده‌اند و تعداد خطاهای ناخواسته‌ای که ایجاد شده‌اند را بشمارد و سپس بر اساس این اعداد ارزیابی و مقایسه انجام شود. ولی این کار علاوه بر این که نیاز به وقت و هزینهٔ بسیار زیادی دارد در ارزیابی داده‌های بزرگ غیرممکن است چراکه معمولاً ارزیابی‌های متعددی برای یک کار پژوهشی انجام می‌شود. به همین دلیل از روش‌های ارزیابی خودکار استفاده می‌شود که در ادامه یکی از معروف‌ترین آن‌ها در این حوزه توضیح داده شده است.

۴-۱ معیار ارزیابی

در ارزیابی نتایج و مقایسه با سایر روش‌ها در ابتدا باید معیاری برای ارزیابی نتایج تبدیل

متن محاوره‌ای به رسمی انتخاب کنیم. در مسائل ترجمهٔ ماشینی معمولاً از معیار بلو استفاده می‌شود که در مقاله‌ای از پاپینینی[1] و همکاران (۲۰۰۲: ۳۱۸–۳۱۱) ارائه شد. ایدهٔ اصلی این روش آن است که هرچه نتیجهٔ ترجمهٔ ماشینی به ترجمهٔ یک انسان نزدیک‌تر بود ترجمهٔ بهتری است و باید امتیاز بیشتری بگیرد. در واقع این روش میزان نزدیکی خروجی ترجمهٔ ماشینی به ترجمهٔ انسان را به دست می‌آورد. در این روش معمولاً جملات یا عبارت‌هایی در نظر گرفته می‌شوند و این معیار برای هریک از آن‌ها محاسبه می‌شود و در نهایت برای کل جملات، میانگین مقادیر اندازه گرفته شده به دست می‌آید.

در این روش یک خروجی برای ترجمهٔ ماشینی و چند جملهٔ مرجع مناسب که ترجمهٔ خوبی هستند در نظر گرفته می‌شود و این معیار یک عدد بین صفر و یک را به‌عنوان میزان شباهت خروجی ترجمهٔ ماشینی به جملات مرجع مشخص می‌کند و هرچه عدد بزرگتر باشد، شباهت بیشتر و دقت ترجمه نیز بیشتر خواهد بود.

روش گفته شده در محاسبه دقت ترجمهٔ ماشینی روشی معروف است ولی مشکلات زیادی هم دارد. این روش یک معیار واحد نیست بلکه به مجموعه‌ای از پارامترها نیاز دارد. پیش‌پردازش‌ها تأثیر زیادی بر روی این معیار دارند و استفاده از جملات مرجع مختلفِ پردازش‌شده نتایج متفاوتی را به دنبال خواهد داشت.

در مقالهٔ ارائه‌شده از پُست[2] (۲۰۱۸) برای حل این مشکل راه‌حل ساده‌ای پیشنهاد شد که طبق آن باید گروه‌های تحقیقاتی معیار بلویی که گزارش می‌کنند با استفاده از یک طرح پیش‌پردازشی داخلی باشد و پارامترهایی که استفاده می‌کنند را صریحا بیان کنند. بر این اساس، معیار بلوی ارتقایافته[3] معرفی شد که همان معیار بلو را ارائه می‌دهد و با کارهایی همچون بارگیری و ذخیره مجموعه‌های ارزیابی «WMT»[4] و «IWSLT 2017»[5] و حذف نیاز به کاربر برای رسیدگی به جملات مرجع سعی کرده مشکلات گفته‌شده در این معیار را برطرف کند. در نتیجه این روش، یک معیار استاندارد در گزارش نتایج و مقایسه با مقالات و سایر روش‌ها است.

با توجه به مطالب گفته شده در این پژوهش از معیار بلوی ارتقاءیافته برای ارزیابی نتایج تبدیل متن محاوره‌ای به رسمی و همچنین مقایسه با ابزارهای موجود استفاده می‌کنیم.

1. K. Papineni 2. M. Post 3. SacreBLEU
4. Workshop on Machine Translation
5. International Workshop on Spoken Language Translation

۲-۴ نتایج

در ارزیابی نتایج به‌دست‌آمده از این پژوهش، مجموعه‌ای شامل ۱۰۰ جملهٔ محاوره‌ای و معادل رسمی آن را از داده‌هایی که در قسمت آموزش مدل توضیح داده شد جدا کرده‌ایم و بررسی شده است که این داده‌ها در مجموعه داده‌های آموزشی حضور نداشته باشند تا نتایج به‌دست‌آمده از آن قابل اطمینان باشد.

خروجی مورد نظر در این قسمت از پژوهش در دو حالت استفاده از مدل کدگذار-کدگشا و استفاده از مدل مبدل به دست آمده است و همچنین در هریک از این دو مورد تأثیر استفاده از قواعد توضیح داده شده در قسمت‌های قبل در کنار استفاده از این مدل‌ها بررسی شده است. نتایج حاصل از این آزمایش‌ها در جدول ۲ آورده شده است.

همان‌طور که در قسمت قبل گفته شد، برای ارزیابی نتایج از معیار بلوی ارتقاءیافته استفاده می‌کنیم که ابزار آمادهٔ آن در زبان پایتون موجود است و در استفاده از آن، جملات رسمی موجود را به‌عنوان جملات مرجع و خروجی مدل ارائه شده یا ابزار مورد بررسی را به‌عنوان جمله کاندید در نظر می‌گیریم تا امتیاز در آن خروجی به دست آید. سپس با میانگین‌گیری از امتیازات تمام خروجی‌ها دقت آن مدل یا ابزار در تبدیل متن محاوره‌ای به رسمی با استفاده از این معیار به دست خواهد آمد.

پس از به دست آوردن نتایج این پژوهش برای بررسی میزان دقتی که این روش داشته است، نتایج را با نتایج در مقالات دیگر و همچنین ابزارهای موجود مقایسه می‌کنیم. در مقالهٔ ارائه‌شده از آرمین و شمس‌فرد (۲۰۱۱: ۱۷۳٤-۱۷۲٤) که در بخش پیشینهٔ کار توضیح داده شد تبدیل متن محاوره‌ای به رسمی در فارسی با استفاده از مدل برت انجام شده است که البته بررسی در سطح کلمات بوده است که جزئیات آن در قسمت مربوط گفته شده است. دقت به‌دست‌آمده در این مقاله با استفاده از همین معیار ارزیابی حدود ۶۲ درصد است که البته داده‌های ارزیابی در این مقاله در دسترس ما قرار ندارد. در مقالهٔ معصومی و همکاران (۲۰۲۰) که باز هم در بخش پیشینهٔ کار توضیح داده شد، روشی برای فراهم کردن دادهٔ مناسب برای تبدیل متن محاوره‌ای به رسمی در زبان فارسی ارائه شده است و داده‌های آموزشی استفاده‌شده در این پژوهش نیز از خروجی همین مقاله است. در این مقاله در انتها از معیار بلو در ارزیابی نتایج داده‌های فراهم شده استفاده شده است که با استفاده از این معیار به دقت ۵۴ درصد رسیده است. این مدل در جدول ۲ با نام «TeleCrowd» آمده است. شایان ذکر است اگرچه دادگان پژوهش حاضر با دادگان معصومی و همکاران یکسان است، با توجه به آن که نوع معیار استفاده شده در این مقاله با معیار ارزیابی در پژوهش حاضر دقیقا یکی نیست نمی‌توان مقایسه دقیقی با نتایج آن انجام داد.

همچنین در مجموعه ابزارهایی که متن‌کاوی فارسی‌یار[1] در زبان فارسی فراهم کرده است تبدیل متن محاوره‌ای به رسمی نیز وجود دارد. به همین دلیل این ابزار را در ارزیابی قرار می‌دهیم و نتایج تبدیل متن محاوره‌ای به رسمی در این ابزار را هم به‌دست می‌آوریم و دقت به‌دست آمده از آن را با دقت خروجی پژوهش انجام شده مقایسه و بررسی خواهیم کرد.

نتایج به‌دست‌آمده از ابزار گفته‌شده و در کنار خروجی این پژوهش در حالات مختلفی که توضیح داده شد در جدول ۲ آورده شده است. همان طور که از نتایج مشخص است خروجیِ به‌دست‌آمده از این پژوهش دقت بهتری از ابزار گفته‌شده دارد و در میان روش‌های پیاده‌سازی شده، روش استفاده از مدل مبدل در حالت تأثیر دادن قواعد قبل از شبکهٔ عصبی دقت بهتری کسب کرده است.

جدول۲- مقایسهٔ عملکرد پژوهش با ابزارهای موجود در تبدیل متن محاوره‌ای به رسمی در فارسی

ابزارها و روش‌های تبدیل متن محاوره‌ای به رسمی در فارسی		دقت بر اساس معیار بلوی ارتقاء یافته
مدل‌های پیشنهادی	مدل کدگذار-گدگشا	۶۱.۳۳
	مدل کدگذار-گدگشا + اثر دادن قواعد در ابتدا	۶۱.۸۸
	مدل کدگذار-گدگشا + اثر دادن قواعد در انتها	۶۱.۳۳
	مدل مبدل	۶۹.۹۶
	مدل مبدل + اثر دادن قواعد در ابتدا	۷۰.۷
	مدل مبدل + اثر دادن قواعد در انتها	۶۹.۹۶
مدل‌های پایه	فارسی‌یار	۴۳.۰۹
	TeleCrowd	٥٤ (معیار بلو)

با بررسی نتایج به دست آمده می‌توان موارد زیر را برداشت کرد:

- مدل‌های پیاده‌سازی شده توانسته‌اند به کارایی خوبی در تبدیل متن محاوره‌ای به رسمی برسند.

1. https://text-mining.ir/

- با مقایسهٔ مدل مبدل و مدل کدگذار-کدگشا دیده می‌شود که مدل مبدل توانسته عملکرد بهتری داشته باشد که دلیل این موضوع می‌تواند قوی‌تر بودن مدل مبدل به‌دلیل استفاده از لایه‌های توجه باشد.
- استفاده از قواعد تبدیل متن محاوره‌ای به رسمی قبل از مدل‌های شبکهٔ عصبی توانسته است باعث بهبود اندک نتایج شود. در حالی که استفاده از این قواعد بعد از شبکه‌های عصبی تغییری در نتایج ایجاد نکرده است. این موضوع به‌دلیل آن است که مدل‌های شبکهٔ عصبی توانسته‌اند تغییراتی که قواعد اعمال می‌کننده را خود بیاموزند و در نتیجه استفاده از این قواعد بعد از مدل‌های شبکهٔ عصبی تأثیری نداشته است.

در جدول ۳ نمونه‌هایی از خروجی تبدیل متن محاوره‌ای به رسمی در این پژوهش نشان داده شده‌اند. همان‌طور که دیده می‌شود، مدل به‌خوبی توانسته بسیاری از حالت‌های محاوره‌ای را به شکل رسمی آن‌ها تبدیل کند ولی در مواردی هم دچار مشکل شده است. از جمله مواردی که باعث شده نتایج به دقت بهتری نرسند آن است که بعضی از کلمات محاوره‌ای بدون تغییر مانده‌اند. دلیل اصلی این امر می‌تواند تعداد کم داده‌ها باشد که باعث شده مدل موارد مشابه را در زمان آموزش تجربه نکند و در زمان آزمون نتواند این موارد را اصلاح کند. از دیگر مشکلاتی که ممکن است در خروجی‌ها دیده شود تغییر اشتباه یک کلمه به کلمه‌ای بی‌معنی است که از معایب استفاده از حروف به جای کلمات در آموزش مدل است. البته مدل سعی می‌کند با آموزش مناسب از این تغییرات اشتباه جلوگیری کند و این خطاها به ندرت اتفاق خواهند افتاد.

جدول ۳- نمونه‌هایی از خروجی پژوهش در تبدیل متن محاوره‌ای به رسمی

جمله محاوره‌ای	جمله رسمی	خروجی مدل
ولی واقعیت خوب نیس	ولی واقعیت خوب نیست	ولی واقعیت خوب نیست
کاش میشد آمارشونو در بیارم	کاش می‌شد آمارشان را در بیاورم	کاش می‌شد آمارشان را در بیاورم
بزرگترین مشکل بیکاری اینه که تعطیلی نداره	بزرگترین مشکل بیکاری این است که تعطیلی ندارد	بزرگترین مشکل بیکاری این است که تعطیلی ندارد
طبیعیه سرزنش نکن خودتو	طبیعی است سرزنش نکن خودت را	طبیعی است سرزنش نکن خودت را
پس فعلا در حد شایعست	پس فعلا در حد شایعه است	پس فعلا در حد شایعست
اینجا دم خونه ما بیشتر از بیستا سگ هست	اینجا دم خانه ما بیشتر از بیست تا سگ هست	اینجا دم خانه ما بیشتر از بیستا سگ هست

در این پژوهش از مجموعه‌ای ۴۴۰۰ تایی از جملات محاوره‌ای و شکل رسمی آن‌ها استفاده شده است که در صورتی که از داده‌های آموزشی بیشتری استفاده شود مدل مورد نظر حالت‌های متنوع‌تری از شکل‌های محاوره‌ای را آموزش می‌بینید و در نتیجه دقت خروجی بهتر خواهد شد.

۵ جمع‌بندی و نتیجه‌گیری

این پژوهش بر روی تبدیل متن محاوره‌ای به رسمی در زبان فارسی تمرکز داشت که در آن ابتدا به بررسی مقالات و کارهای انجام شده در این زمینه‌ها به خصوص در زبان فارسی پرداخته شد و در ادامه مدل‌های استفاده شده توضیح داده شده و دو مدل کدگذار-کدگشا و مبدل پیاده‌سازی شده است.

برای ساخت داده آموزشی در تبدیل متن محاوره‌ای به رسمی از داده‌های آموزشی ارائه شده توسط معصومی و همکاران (معصومی،۲۰۲۰) استفاده شد. همچنین در کنار مدل ساخته شده از مجموعه قواعدی برای این تبدیل استفاده شده است.

در بخش ارزیابی‌ها در تبدیل متن محاوره‌ای به رسمی معیار ارزیابی بلوی ارتقاء یافته استفاده شده است که معمولاً در ترجمهٔ ماشینی استفاده می‌شود. نتایج به دست آمده در این قسمت نسبت به ابزار و روش‌های موجود بهبود قابل توجهی را نشان می‌دهد و بسیاری از کلمات محاوره‌ای در این جملات را به شکل رسمی آن تبدیل کرده است. همچنین در این قسمت خروجی مدل مبدل توانست به دقت بالاتری از مدل کدگذار-کدگشا دست یابد.

یکی از محدودیت‌های موجود در این پژوهش حجم کم داده‌های محاوره‌ای و شکل رسمی آن‌ها بود که تأثیر زیادی در آموزش شبکه‌های عصبی می‌گذارد. انتظار می‌رود در صورت فراهم شدن داده‌های بیشتر و همچنین ارزیابی مدل‌های متنوع‌تر بتوان به نتایج بهتری در تبدیل متن محاوره‌ای به رسمی در زبان فارسی رسید.

منابع

طبیب‌زاده، امید (۱۳۹۸). *مبانی و دستور خط فارسی شکسته براساس صد سال آثار داستانی و نمایشی*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

Ariffin, S. N. A. N., & S. Tiun (2020). "Rule-based text normalization for Malay social media texts". *International Journal of Advanced Computer Science and Applications*. 11(10), 156-162

Armin, N., & M. Shamsfard (2011). "converting Persian colloquium text to formal by n-grams". *Computer Society of Iran. for statistical machine translation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724-1734.

Arora, M., & V. Kansal (2019). "Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis". *Social Network Analysis and Mining*, 9(1), 1-14.

De Saussure, F. (1916). *Course de linguistique générale*. Lausanne, Paris: Payot

Kozhirbayev, Z., & Z. Yessenbayev (2020). "Kazakh text normalization using machine translation approaches". CEUR Workshop Proceedings, Vol. 2780, CEUR-WS, 115-122.

Liu, Y., et al. (2020). "An Advanced ICD-9 Terminology Standardization Method Based on BERT and Text Similarity". *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*. Springer, Cham, 1868-1879.

Mansfield, C., et al. (2019). "Neural text normalization with subword units". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol2 (Industry Papers), 190-196.

Masoumi, V., Salehi, M., Veisi, H., Haddadian, G., Ranjbar, V., & Sahebdel, M. (2020). *TeleCrowd: A Crowdsourcing Approach to Create Informal to Formal Text Corpora*. arXiv preprint arXiv:2004.11771.

Papineni, K., et al. (2002). "Bleu: a method for automatic evaluation of machine translation". *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311-318).

Post, M. (2018). *A call for clarity in reporting BLEU scores*. arXiv preprint arXiv:1804.08771.

Rasooli, M. S., et al. (2020). *Automatic Standardization of Colloquial Persian*. arXiv preprint arXiv:2012.05879.

Vaswani, A., et al. (2017). "Attention is all you need". *Advances in neural information processing systems*. 30, 6000–6010.

استناد به این مقاله: ممتازی، سعیده و ادیبیان، مجید (۱۴۰۱). تبدیل متن محاوره به رسمی فارسی با استفاده از شبکه‌های عصبی مبتنی بر مبدل زبان و زبان‌شناسی ۱۸(۳۵)، ۴۷_۶۹. doi: 10.30465/lsi.2023.8498