



Linguistic Resources and Transformer-based Models for the Machine Translations between Luri and Yazdi Dialects versus Standard Persian

Bahmani, Zahra¹

Mirbeygi, Mohaddeseh²

Hashemi Dijujin, Negin³

Nouri, Marzieh⁴

Amani, Mahsa⁵

Asgari, Ehsaneddin⁶ 

Soleymani Baghshah, Mahdiah⁷

Beigy, Hamid⁸

Movaghar, Ali⁹

Moghimi, Afzal¹⁰

PhD. Student, Department of Computer Engineering, Sharif University of Technology, AI Group, Tehran, Iran
PhD. Student, Department of Computer Engineering, Sharif University of Technology, AI Group, Tehran, Iran
M.Sc. Student, Department of Computer Engineering, Sharif University of Technology, AI Group, Tehran, Iran
Research Assistant, Language Processing and Digital Humanities Lab., Sharif University of Technology, Tehran, Iran
B.Sc. Student, Department of Computer Engineering, Sharif University of Technology, AI Group, Tehran, Iran
Scientist at Qatar Computing Research Institute, Doha, Qatar
Associate Professor, Department of Computer Engineering, Sharif University of Technology, AI Group, Tehran, Iran
Associate Professor, Department of Computer Engineering, Sharif University of Technology, AI Group, Tehran, Iran
Professor, Department of Computer Engineering, Sharif University of Technology, AI Group, Tehran, Iran
Other

Abstract

Despite recent advances in developing language technologies for the standard Persian dialect, the official Iranian language, a large number of Iranian language variations remained computationally unexplored. Iranian languages, e.g., Kurdi, Azeri, and many Persian dialects are examples of low-resource language distinctions lacking significant linguistic resources such as machine-readable lexicons or part-of-speech (POS) taggers. Efforts in developing language technologies for such languages can significantly contribute to language survival in the digital era and promote cultural diversity. To the best of our knowledge, for the first time, we created linguistic resources for the Luri and the Yazdi dialects by introducing the first parallel corpora between these language variations and the modern Persian language. In this study, we train neural encoder-decoders (1) recurrent sequence-to-sequence and (2) transformer-based machine translation models and evaluate the trained model using BLEU score on an unseen test dataset.


Keywords: Natural language processing, Luri dialects, Yazdi dialect, Translations of Persian languages, Deep neural translation.

1. zahra.bahmani2009@gmail.com 2. m.mirbeygi@sharif.edu 3. eihashemi@ce.sharif.edu
4. marzieh.nouri@parsi.ai 5. mahsaama@ce.sharif.edu 6. easgari@hbku.edu.qa
(Corresponding Author) 7. soleymani@sharif.edu 8. beigy@sharif.edu 9.
movaghar@sharif.edu 10. afzalmoghime@gmail.com

How to cite: Bahmani, Z., Mirbeygi, M., Hashemi Dijujin, N., Nouri, M., Amani, M., Asgari, E., Soleymani Baghshah, M., Beigy, H., Movaghar, A., & Moghimi, A. (2024). Linguistic Resources and Transformer-based Models for the Machine Translations between Luri and Yazdi Dialects versus Standard Persian. *Language and Linguistics*, 19(37), 153 - 172. doi: 10.30465/lsi.2024.43456.1640



ترجمه ماشینی مبتنی بر مدل مبدل برای گویش‌های لری بویراحمدی و یزدی به فارسی معیار و گسترش منابع زبانی رایانشی مرتبط^۱

دانشجوی دکتری، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، گروه هوش مصنوعی، تهران، ایران	بهمنی، زهرا
دانشجوی دکتری، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، گروه هوش مصنوعی، تهران، ایران	میریگی، محدثه
دانشجوی کارشناسی ارشد، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، گروه هوش مصنوعی، تهران، ایران	هاشمی دیجوجین، نگین
پژوهشگر، دانشگاه صنعتی شریف، آزمایشگاه پردازش هوشمند متن و زبان، گروه هوش مصنوعی، تهران، ایران	نوری، مرضیه
دانشجوی کارشناسی، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، گروه هوش مصنوعی، تهران، ایران	امانی، مهسا
پژوهشگر علوم انسانی محاسباتی. موسسه تحقیقات محاسباتی قطر، دوحه، قطر	عسگری، احسان‌الدین 
دانشیار، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، گروه هوش مصنوعی، تهران، ایران	سلیمانی باغشاهی، مهدیه
دانشیار، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، گروه هوش مصنوعی، تهران، ایران	بیگی، حمید
استاد، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر، گروه هوش مصنوعی، تهران، ایران	موقر، علی
سایر	مقیمی، افضل

چکیده

با وجود تلاش‌های گسترده رایانشی بر روی گویش معیار فارسی، سایر لهجه‌ها و گویش‌ها و زبان‌های ایرانی کمتر مورد توجه محققین حوزه زبان‌شناسی رایانشی قرار گرفته‌اند. یکی از مهم‌ترین چالش‌های کار رایانشی بر روی این تمایزهای زبانی، نبود یک مجموعه داده دیجیتال و استاندارد است. در این پژوهش اولین مجموعه داده تک‌زبانه و نیز داده موازی بر روی گویش‌های لری و یزدی که گویش‌هایی با منابع محدود محسوب می‌شوند، در مقابل فارسی معیار ارائه شده است. در ادامه مدل‌های یادگیری ژرف ترجمه ماشینی کدگذار-کدگشا در دو نوع مدل شبکه عصبی بازگشتی و مدل ژرف مبدل برای این گویش‌ها به فارسی معیار توسعه یافته و ارزیابی گردیده است. در این پژوهش که اولین تلاش محاسباتی روی این دو گویش محسوب می‌شود، برای ترجمه لری به فارسی و فارسی به لری به امتیاز BLEU به ترتیب ۷/۳۹ و ۳/۲۹ رسیدیم. این امتیاز برای گویش یزدی به فارسی معیار و برعکس به ۰/۷۳ و ۰/۷۷ رسید. که نشان می‌دهند پژوهش‌های فراتری برای توسعه منابع زبانی برای این زبان نیاز است.

کلیدواژه: پردازش زبان طبیعی، گویش لری، گویش یزدی، ترجمه تمایزهای زبانی ایرانی، مدل یادگیری ژرف ترجمه ماشینی

۱ مقدمه

بی‌شک گسترش شتابان فضای مجازی و ارائه‌ی خدمات گوناگون در بستر آن یکی از چشم‌گیرترین اتفاقات دهه‌های اخیر محسوب می‌شود. در این میان پردازش زبان طبیعی و درک مفاهیمی که کاربران با نقش‌های گوناگون در این فضا رد و بدل می‌کنند، نقشی اساسی در هوشمندسازی این خدمات دارد. ابزارهایی از جمله مترجم برخط گوگل^۱، دستیارهای شخصی هوشمند مانند الکسا^۲ و سیری^۳، و نیز برنامه‌هایی مانند گرامرلی^۴ توانسته‌اند خدمات مفیدی به سخنوران زبان‌های تحت پوشش خود ارائه دهند. این ابزارها بیشتر برای زبان‌های رسمی کشورها که منابع گسترده‌ی صوتی یا مکتوب دارند، ارائه می‌شوند؛ اما در عمده‌ی کشورها درصد قابل توجهی از بومیان هر منطقه از زبان یا گویش خاص خود در ارتباطات نوشتاری و گفتاری بهره می‌گیرند. به ویژه، با رواج نوشتار غیررسمی در فضای مجازی بخش قابل توجهی از نوشته‌های کاربران در قالب گویش بیان می‌شوند و از این رو، درک ماشینی از متون گویش‌های مختلف در مسائل مختلفی مانند تحلیل احساسات، تشخیص موضوع، تبلیغات شخصی‌سازی‌شده، استخراج عبارات کلیدی و قیود زمانی و مکانی، و سیستم‌های ترجمه‌ی ماشینی اهمیت پیدا می‌کند. این در حالی است که صنعت امروز کمتر به بحث گویش‌ها می‌پردازد. البته با توجه به ناکافی بودن دادگان برای آموزش مدل‌های هوشمند نمی‌توان انتظار فراتری داشت (هرات و همکاران، ۲۰۱۹). عدم توسعه‌ی فناوری، این زبان‌ها را با خطر انقراض مواجه کرده است. حفظ تنوع زبانی ایران‌زمین از رسالت‌های حوزه زبان‌شناسی رایانشی با توسعه‌ی فناوری برای زبان‌ها در کشور است.

توسعه‌ی ترجمه ماشینی، یکی از حوزه‌های اصلی مورد نیاز زبان‌هایی با منابع محدود است که می‌تواند کاربردهای گسترده‌ای در زمینه‌های فرهنگی، اقتصادی، امدادسانی و غیره داشته باشد. ایجاد دادگان موازی بین گویش‌ها و زبان‌های استاندارد از ملزومات اصلی ایجاد یک سامانه ترجمه ماشینی است. در سال‌های اخیر پژوهش‌های گوناگونی در زمینه‌ی ایجاد این منابع و استفاده از آن‌ها در تشخیص و ترجمه‌ی گویش انجام شده است. برای نمونه، (کاستا-یوسا، ۲۰۱۸) بین دو نسخه‌ی اروپایی و برزیلی زبان پرتغالی ترجمه ماشینی انجام دادند. گامان و همکاران در (گمان و همکاران، ۲۰۲۰) با استفاده از دادگان MOROCO که برگرفته از اخبار زبان رومانیایی است، وظایف تشخیص گویش و شناسایی موضوع بین گویش‌های مختلف را روی دو گویش رومانیایی بررسی کرده‌اند. همچنین یک مجموعه دادگان جدید

1. Google
2. Alexa
3. Siri
4. Grammarly

شامل ۵۰۰۰ توییت بر روی این دو گویش ارائه داده‌اند. در (هرات و همکاران، ۲۰۱۹) بحث ترجمه‌ی ماشینی روی گویش‌های عربی به تفصیل بررسی شده است. این مقاله پس از بررسی چالش‌های پردازش زبان‌های طبیعی روی گویش‌های عربی و معرفی دادگان موجود در مطالعات پیشین، به مقوله‌ی ترجمه‌ی ماشینی بین گویش‌های عربی و عربی استاندارد می‌پردازد؛ مثلاً^۱ گویش‌های مصری^۱ (محمد و همکاران، ۲۰۱۲؛ سالوم و همکاران، ۲۰۱۲)، شامی^۲ (سالوم و همکاران، ۲۰۱۲؛ مفتوح و همکاران، ۲۰۱۵)، تونس^۳ (مفتوح و همکاران، ۲۰۱۵)، و عراقی (سالوم و همکاران، ۲۰۱۲). طبق این مقاله، یکی از مزایای ترجمه‌ی ماشینی از گویش به زبان استاندارد، ایجاد یک پل برای ترجمه‌ی گویش به زبان‌های دیگر است (سواف، ۲۰۱۰) که در گویش‌های فارسی نیز می‌تواند صادق باشد؛ مثلاً^۴ ترجمه ابتدا از گویش به زبان استاندارد، و پس از آن به زبان انگلیسی صورت بگیرد. همچنین (بانیاتا و همکاران، ۲۰۲۱) برای نخستین بار، یک روش مبتنی بر مبدل برای ترجمه گویش‌های بومی عربی (عربی شامی، مغربی^۴، و عراقی) به عربی استاندارد ارائه کردند و طبق آزمایش‌های آنان، مدل پیشنهادی توانسته است ترجمه‌ی به نسبت با کیفیتی را ارائه دهد؛ و حتی با توجه به انجام ترجمه به کمک زیرکلمات، بر مشکل لغات ناشناخته نیز تا حدی غلبه کند. تلاش‌های دیگری نیز در زمینه‌ی زبان‌ها یا گویش‌هایی با منابع کم انجام شده‌اند و سعی شده است تا بین چنین زبان‌هایی با نزدیک‌ترین زبان دارای منابع گسترده ارتباط ایجاد کنند؛ برای نمونه (Haddow, B., 2013) بین آلمانی اتریشی^۵ و گویش وینی^۶، (ناکوی، ۲۰۱۲) بین زبان اندونزیایی^۷ و مالایی انگلیسی^۸ و (اسکانل، ۲۰۰۶) بین زبان ایرلندی^۹ و زبان بومی گالیک اسکاتلندی^{۱۰} مطالعاتی انجام داده‌اند.

گویش‌ها و زبان‌های محلی، منابعی بسیار غنی برای تحقیقات در زمینه جامعه‌شناسی، مردم‌شناسی، شناخت تاریخ و ادبیات یک سرزمین هستند. هر یک از گویش‌های متنوع ایرانی همچون لری، کردی، بلوچی، گیلکی، بخشی از هویت مردمانش را نشان می‌دهد و بستری برای انتقال ارزش‌های فرهنگی از یک نسل به نسل دیگر است (داوری و جانی، ۱۳۹۷). علی‌رغم تلاش‌هایی که در زبان‌ها و فرهنگ‌های گوناگون بر روی گویش‌ها صورت گرفته،

-
1. Egyptian
 2. Levantine
 3. Tunisian
 4. Maghrebi
 5. Austrian German
 6. Viennese
 7. Indonesian
 8. English using Malay
 9. Irish
 10. Scottish Gaelic

تاکنون دادگان منسجمی از گویش‌های مختلف فارسی و سایر زبان‌های ایرانی برای کاربردهای رایانشی ارائه نشده است.

کارهای انجام‌شده در مطالعه‌ی حاضر: در این مطالعه، گویش لری بویراحمدی و گویش یزدی مورد بررسی قرار گرفته‌اند و دادگان موازی به تفکیک برای هر کدام از دو گویش جمع‌آوری شده‌اند. هم‌چنین از دو مدل یادگیری ژرف برای ترجمه‌ی ماشینی استفاده شده تا اولین تلاش و پایه‌ای برای پژوهش‌های آتی روی این زبان‌ها باشد. در ادامه، پس از معرفی این دو گویش و بررسی چالش‌های موجود، به معرفی دادگان و مدل ترجمه‌ی استفاده شده برای ترجمه‌ی زبان لری و گویش یزدی به فارسی معیار و برعکس پرداخته و نتایج و کارهای آینده ارائه خواهند شد.

۲. معرفی گویش‌های لری بویراحمدی و یزدی

زبان‌های ایرانی از نظر زبان‌شناسی به سه دسته ایرانی باستانی، ایرانی میانه و ایرانی نو تقسیم‌بندی می‌شوند. زبان‌های ایرانی نو به دو دسته‌ی زبان‌های شرقی (پشتو، آسی، ارموری، پراچی و پامیری) و غربی تقسیم می‌شوند (محسنی، ۱۳۹۲). زبان‌های غربی شامل سه دسته‌ی زبان‌های شمال غربی (کردی، آذری، کرمانجی و لکی)، مرکزی و جنوب غربی (تاتی، قفقازی، لاری و لری) هستند. زبان‌های مرکزی به چهار دسته‌ی شمال غربی، جنوب غربی، شمال شرقی و جنوب شرقی تقسیم می‌شوند. زبان دری زرتشتی زیرمجموعه‌ی زبان‌های جنوب شرقیست که شامل گویش یزدی نیز می‌شود (محسنی، ۱۳۹۲). با توجه به هدف مطالعه‌ی حاضر، در ادامه به معرفی زبان لری و گویش یزدی می‌پردازیم و برخی از چالش‌های آن‌ها را بررسی می‌کنیم.

۱.۲. معرفی زبان لری و چالش‌های آن

زبان لری یکی از زبان‌های ایرانی غربی است که بیش از چهار میلیون نفر متکلم دارد. بخش اعظم این افراد در استان‌های کهگیلویه و بویراحمد، همدان، لرستان، چهارمحال و بختیاری، و خوزستان زندگی می‌کنند. درباره ریشه این زبان نظریات متعددی وجود دارد؛ برخی آن را یک گویش می‌دانند، در حالی که برخی دیگر اعتقاد دارند لری یک زبان مستقل است و خود دارای گویش‌های متنوعی است (مجیدی، حق بی، ۱۳۹۷).^۱ لری بویراحمدی به عنوان یکی از

۱- در پژوهش مدرسی (۱۳۶۸) گویش بدین شکل تعریف شده است: «هرگاه دو گونه‌ی زبانی بدون آموزش آگاهانه، در حد ایجاد ارتباط معمول برای گویندگان با یکدیگر، قابل فهم باشد، آن دو گونه، دو گویش متفاوت از یک واحد محسوب می‌شوند و در غیر این صورت باید آن‌ها را دو زبان جداگانه دانست».

گویش‌های اصلی لری، در منطقه شرق و شمال استان کهگیلویه و بویراحمد که به نام بویراحمد شناخته می‌شود، رواج دارد.

سایر گونه‌های گویش لری، یا به این گویش نزدیک هستند، و یا به دلیل مجاورت به گویش بختیاری نزدیک شده‌اند. گویش بویراحمدی بازمانده‌ی فارسی میانه است که خود از بازماندگان فارسی باستان به شمار می‌رود (طاهری، ۱۳۹۱). از این گویش به گویش ممسنی نیز یاد می‌شود.

همانگونه که گفته شد، در زمینه‌ی پردازش رایانشی، بر روی گویش‌های محلی ایرانی، کارهای تحقیقاتی چندانی انجام نشده است. گویش لری بویراحمدی نیز مستثنی نبوده، و طبق بررسی‌های ما، کار تحقیقاتی قابل توجهی روی آن انجام نگرفته است. یکی از مهم‌ترین دلایل این مسئله می‌تواند نبود یک منبع دیجیتال از متون لهجه‌ی لری بویراحمدی و معانی آن باشد. حتی تعداد منابعی که شامل متون گویش به همراه ترجمه‌ی فارسی استاندارد باشند بسیار محدود هستند.

در میان همین منابع محدود هم مهم‌ترین چالش جمع‌آوری داده، نبودن ساختار نوشتار استاندارد در گویش‌های محلی است. از آنجایی که معمولاً متون برخط یکی از مهم‌ترین منابع در جمع‌آوری مجموعه‌ی استاندارد زبانی به شمار می‌روند، نبودن یک ساختار استاندارد در فرم نوشتاری گویش‌ها موجب می‌شود که کلمات با ساختارهای نوشتاری متفاوتی در متون ظاهر شوند، و این تفاوت‌ها فرآیند یادگیری ماشین را با چالش مواجه می‌کند. این مسئله حتی در متون کتاب‌های منتشرشده نیز مشاهده می‌شود و معمولاً ساختار نوشتاری کلمات از یک استاندارد واحد پیروی نمی‌کنند. در این رابطه مثال ۱ و ۲ از زبان لری در جداول ۱ و ۲ آمده است.

همان طور که در مثالها قابل مشاهده است، یک کلمه واحد «کن»، در دو متن لری با دو نوشتار متفاوت آمده است. اختلاف لهجه در میان متکلمان گویش نیز باعث تشدید اختلاف در ساختار نوشتاری گویش می‌شود.

جدول ۱: مثال ۱، جمله لری، حاوی فعل «کن»، که در زبان لری معادل با کلمه «کُ ko» است

سیلم کُ که ای باره و منزل برسونم								جمله لری
برسونم	منزل	و	باره	ای	که	کُ	سیلم	کلمات
berasunom	manzel	va	bāra	i	ke	ko	seylom	IPA
برسانم	مقصد	به	بار را	این	که	کن	نگاهم	ترجمه کلمه
نگاهم کن تا این بار را به مقصد برسانم								ترجمه جمله

جدول ۲: مثال ۲، جمله لری، حاوی فعل «کن»، که در اینجا با نوشتار «ko» آمده است

و هموره تو بیو چپ کو و آبادی یار									جمله لری	
یار	آبادی	و	کو	چپ	بیو	تو	ره	همو	و	کلمات
yār	ābādi	va	ko	čap	beyow	te	ra:	hamu	va	IPA
یار	روستا	از	کن	چپ	بیا	تو	راه	همان	از	ترجمه کلمه
تو از همان مسیر بیا و راحت را به سمت روستای یار تغییر بده									ترجمه جمله	

علاوه بر چالش‌هایی که در گردآوری داده وجود دارد، پژوهش روی متون گویش لری بویراحمدی با چالش‌هایی روبه‌رو است که در نظر گرفتن آن‌ها در پژوهش‌های آتی ضروری به نظر می‌رسد. برخی از این چالش‌ها عبارتند از:

- مختصرگویی و کوتاه شدن طول کلمات: در گویش‌های محلی، مانند شکل گفتاری عامیانه‌ی زبان‌ها، تمایل به مختصرگویی وجود دارد. در گویش بویراحمدی نیز این مسئله باعث شده که کلمات طول نسبتاً کوتاه‌تری داشته باشند؛ به طوری که تعداد زیادی فعل با طول یک حرف دیده می‌شود. در مثال ۳ و ۴ از مختصرگویی در جداول ۳ و ۴ آمده است. همان طور که در این مثالها قابل مشاهده است، افعال «است» و «هستی» از زبان فارسی فقط با یک حرف در زبان لری بیان شده‌اند.

جدول ۳: مثال ۳، جمله لری، حاوی حرف «ه» به معنای «است»

سنگینه اگر درد مو درمون مو ایبه							جمله لری
ایبه	مو	درمون	مو	درد	اگر	سنگینه	کلمات
iyē	mo	darmuna	mo	dard	agar	sangine	IPA
میدهد	من	درمان به	من	درد	اگر	سنگین است	ترجمه فارسی
مسئله‌ی سنگینی است، که درد من مرا درمان میکند							ترجمه جمله

جدول ۴: مثال ۴، جمله لری، حاوی حرف «سی» به معنای «هستی»

مث سایه و دیندا و نهامی و تونَ اینخوم									جمله لری
ایخوم	تونَ	و	نهامی	و	دیندا	و	سایه	مث	کلمات
ixom	tuna	vo	nahāmi	vo	dīndā	va	sāya	mesə	IPA
میخواهم	تورا	و	جلویم هستی	و	دنبال	از	سایه	مثل	ترجمه کلمه
مثل سایه پشت سر و جلوی من هستی (همیشه با منی) و تورا می‌خواهم									ترجمه جمله

کلمات هم‌نگاره^۱: در گویش لری بویراحمادی مانند خیلی از گویش‌ها و زبان‌های دیگر، برخی از کلمات هم‌نگاره با نوشتار یکسان و معانی مختلف مشاهده می‌شوند. همچنین تمایل به مختصرگویی گاهی باعث تولید کلماتی با ساختار یکسان شده است که برخی اوقات، هم در گفتار و هم در نوشتار یکی هستند، و در سایر موارد فقط در نوشتار با هم معادل هستند. این مسئله در فرآیند یادگیری کلمات تأثیر منفی خواهد داشت. مثال ۵ و ۶ از اشعار لری که شامل کلمات هم‌نگاره است در جداول ۵ و ۶ آمده است. همانطور که در نمونه‌های ذکر شده مشخص است کلمه «سیل [seyl]»، با نوشتار و گفتار یکسان ولی معانی متفاوت در دو جمله مختلف آمده است. همچنین نمونه‌های از نوشتار یکسان و گفتار متفاوت در جداول ۷ و ۸ آمده است.

جدول ۵: مثال ۵، جمله لری، حاوی کلمه «سیل [seyl]» به معنای «نگاه»

بالی مال اومم برم سیل کردم و من کپر									جمله لری
کپر	من	و	کردم	سیل	برم	اومم	مال	بالی	کلمات
kapar	men	va	kerdom	seyl	berram	umam	māl	bāley	IPA
چادر	داخل	از	کردم	نگاه	بروم	آدمم	روستا	بالای	ترجمه کلمه
از بالای روستا رد می‌شدم، داخل چادر را نگاه کردم									ترجمه جمله

1 Homograph

جدول ۶: مثال ۶، جمله لری، حاوی کلمه «سیل seyl» به معنای «سیل»

نه بی سیل و طیفون، نه بی تش بلا								جمله لری	
بلا	تش	بی	نه	طیفون	و	سیل	بی	نه	کلمات
balā	taš	bey	na	tifun	o	seyl	bey	na	IPA
بلا	آتش	با	نه	طوفان	و	سیل	با	نه	ترجمه کلمه
نه با سیل و طوفان، نه با آتش و قضا و بلا								ترجمه جمله	

جدول ۷: مثال ۷، جمله لری، حاوی کلمه «سی si» به معنای «برای»

سی خاطر ای باغه که باد ورتکناشه							جمله لری
ورتکناشه	باد	که	باغه	ای	خاطر	سی	کلمات
vartakanāše	bāḍ	ke	bāye	i	xātere	si	IPA
پاره اش کرده است	باد	که	باغ است	این	خاطر	برای	ترجمه فارسی
برای خاطر این باغ است، که باد آن را تکه پاره کرده است							ترجمه جمله

جدول ۸: مثال ۸، جمله لری، حاوی کلمه «سی sey» به معنای «نگاه»

دلم ایخو یه غزل وت بنویسم سی کُ							جمله لری	
کُ	سی	بنویسم	وت	غزل	یه	ایخو	دلم	کلمات
ko	sey	benvisom	vat	γazal	ya	ixo	delom	IPA
کن	نگاه	بنویسم	از تو	غزل	یک	میخواهد	دلم	ترجمه فارسی
دلم میخواهد از تو غزلی بنویسم، نگاه کن							ترجمه جمله	

۲.۲. معرفی گویش یزدی و چالش‌های آن

ریشه‌ی گویش یزدی زبان دری زرتشتی است که تفاوت‌های آوایی و بعضاً معنایی با هم دارند. ریشه‌ی بعضی واژگان را در زبان‌های کهن مانند هندی باستان، سانسکریت، اوستایی، و زبان‌های دوره‌ی میانه می‌توان جست‌وجو کرد. برای مثال، واژه «پسر pusaer» به معنای فرزند ذکور در

زبان‌های سانسکریت، پارسی باستان، پهلوی و هند باستان به ترتیب به صورت 'pu:sra'، 'pu:tra' و 'pu:sær' تلفظ می‌شد. واژه‌ی «خش» [xæʃ] به معنای خوب، لذیذ، خوشمزه نیز در پهلوی و دوره‌ی باستان و میانه به ترتیب به شکل [xvæʃ]، [xwæʃ] و [xæʃ] تلفظ می‌شد و تلفظ فارسی معیار کنونی [xuʃ] است، اما تلفظ در گویش یزدی با دوره‌ی میانه یکسان است. واژه «پیر» [pi:yær] به معنای پدر نیز در پهلوی [pi:dær] بوده است (رمضان‌خانی، ۹۱).

در واژگان گویش یزدی فرآیندهای آوایی مانند ابدال، حذف، اضافه و قلب صورت می‌گیرد. به عنوان مثالی از فرآیند ابدال، واژه‌ی «آستر» به صورت [assær] تلفظ می‌شود. نمونه‌ای از فرآیند حذف هم کلمه «گفت» است که در گویش یزدی به صورت [guf] بیان می‌شود. فرآیند قلب نیز در کلمه «کتف» از فارسی معیار صورت می‌گیرد و در گویش یزدی [keft] گفته می‌شود (صادق زاده، رمضان‌خانی، ۱۳۹۸). مواردی از این قبیل بین فارسی استاندارد و گویش یزدی اختلاف ایجاد می‌کنند و انتظار می‌رود یک مدل ترجمه‌ی ماشینی موفق بتواند چنین تبدیلی‌هایی را درک کند.

در گویش یزدی نیز مانند گویش بویراحمدی محدودیت منابع از جمله مهم‌ترین چالش‌ها است. تعداد منابع دیجیتالی که در نوشتار آن‌ها از گویش یزدی استفاده شده بسیار محدود هستند. همچنین، مسئله استفاده از زبان محاوره‌ای در نوشتار گویش یزدی، کار پردازش این گویش را با مشکلاتی روبه‌رو می‌کند.

۳. روش تحقیق

هدف مطالعه‌ی حاضر، معرفی دادگان موازی برای انجام ترجمه‌ی ماشینی برای گویش لری و گویش یزدی به فارسی معیار و برعکس است. در ابتدا به معرفی دادگان موجود می‌پردازیم و در دو بخش بعدی مدل‌های ترجمه روی دادگان پیشنهادی را شرح می‌دهیم. سپس نحوه‌ی آموزش و شیوه‌ی ارزیابی را بیان می‌کنیم.

۳.۱. دادگان پژوهش

به طور کلی تعداد منابعی که بتوان از آن‌ها جهت تهیه‌ی مجموعه داده‌ی استاندارد برای گویش بویراحمدی استفاده کرد بسیار محدود هستند. یکی از غنی‌ترین منابع موجود (مقیمی و همکاران، ۱۴۰۰) است. این کتاب شامل هزاران کلمه، عبارت، ضرب‌المثل، و جمله با گویش لری بویراحمدی به همراه تلفظ و ترجمه‌ی آن‌ها به فارسی معیار است. جهت ایجاد دادگان پیشنهادی، ۳۰۰۰ جمله با گویش بویراحمدی همراه با معادل فارسی آن‌ها به صورت دستی استخراج شده است. از مهم‌ترین ویژگی‌های این مجموعه داده یک‌دستی آن است؛ زیرا همه‌ی

جملات از یک منبع تهیه شده‌اند و از یک ساختار نوشتاری استاندارد و لهجی واحدی پیروی می‌کنند.

برای گویش یزدی، از دیالوگ‌های کتاب یزدی «ننه زهرا و پسرش» (عسکری، ۱۴۰۰) و بلاگ خوش‌ترین شهر کویر^۱ استفاده شده است. گزارشی از مجموعه دادگان در جدول ۹ آمده است.

جدول ۹: گزارش دادگان پژوهش

نوع داده	منبع	تعداد زوج داده		تعداد واحد زبانی	تعداد واحد یکتا
لری به فارسی	(مقیم و همکاران، ۱۴۰۰)	۸۰۲۵		۱۰۹۳۰	۴۰۱۴
یزدی به فارسی	(عسکری، ۱۴۰۰)	۵۶۱	۸۲۴	۷۲۴۶	۳۶۱۱
	بلاگ خوش‌ترین شهر کویر	۲۶۳			

۲.۳. مدل‌های رشته‌به‌رشته‌ی ترجمه بر روی دادگان پیشنهادی

در این مقاله، ترجمه‌ی ماشینی از گویش‌های یزدی و لری بویراحمدی به فارسی استاندارد، به عنوان یکی از کاربردهای دادگان ارائه‌شده انتخاب شده است. برای این منظور ابتدا جملات زبان مبدأ و زبان مقصد به رشته‌ای از کلمه‌ها تبدیل می‌شوند، سپس از دو مدل رشته‌به‌رشته با ساختار کدگذار-کدگشا^۲ برای ترجمه استفاده می‌شود. این معماری از دو بخش کدگذار و کدگشا تشکیل شده است. بخش کدگذار متن ورودی را به یک بردار تبدیل کرده و بخش کدگشا از این بردار برای تولید ترجمه‌ی متن ورودی به صورت کلمه به کلمه استفاده می‌کند.

مدل اول معماری بازگشتی^۳ با سازوکار توجه^۴ بین کدگذار و کدگشا است و مدل دوم، ساختاری مبتنی بر مبدل دارد. در ادامه به شرح مختصر نحوه‌ی استخراج واحدهای زبانی ترجمه و ساختار مدل پرداخته می‌شود.

واحدهای زبانی ترجمه

یکی از عوامل مؤثر در ترجمه‌ی ماشینی بهینه برای ترجمه با منابع محدود، شکستن لغات به واژه‌های معنی‌دار است که از طریق کم کردن مؤلفه‌ها به مدلی بهینه کمک می‌کند. یک روش

1. <http://mrb123.blogfa.com/category/9>

2. Encoder-decoder

3. Recurrent

4. Attention

بدون نظارت برای استخراج این زیرکلمات، استفاده از الگوریتم^۱ Byte Pair Encoding یا BPE است که کلمات را به زیرکلمات پربسامد می‌شکند. در این پژوهش نیز از همین روش برای تبدیل جملات به واحدهای زبانی استفاده شده است. استفاده از BPE باعث می‌شود تعمیم‌پذیری به لغات خارج از دامنه‌ی آموزش نیز بهبود پیدا کند. به همین منظور برای هر دو گویش، جملات مبدأ و مقصد را به شکل جداگانه با استفاده از BPE به کلمه تبدیل کرده‌ایم. از طرفی اندازه‌ی دامنه‌ی لغات را که به عنوان یک مؤلفه در مدل BPE مطرح است، عدد کوچک‌تری در نظر گرفته‌ایم؛ چرا که آزمایش‌ها نشان داده‌اند برای زبان‌هایی با منابع اندک، این کار می‌تواند تأثیر مثبتی در بهبود کیفیت ترجمه ایجاد کند. گفتنی است در این کار از پیش‌پردازش‌هایی مانند حذف حروف اضافه یا حرکات استفاده نشده است؛ زیرا با توجه به لحن گفتاری جملات گویش، تشخیص حروف اضافه از کلمات در مواردی دشوار است و بخش‌های رایج در پیش‌پردازش متون باید برای هر گویشی به طور اختصاصی بازنویسی شوند.

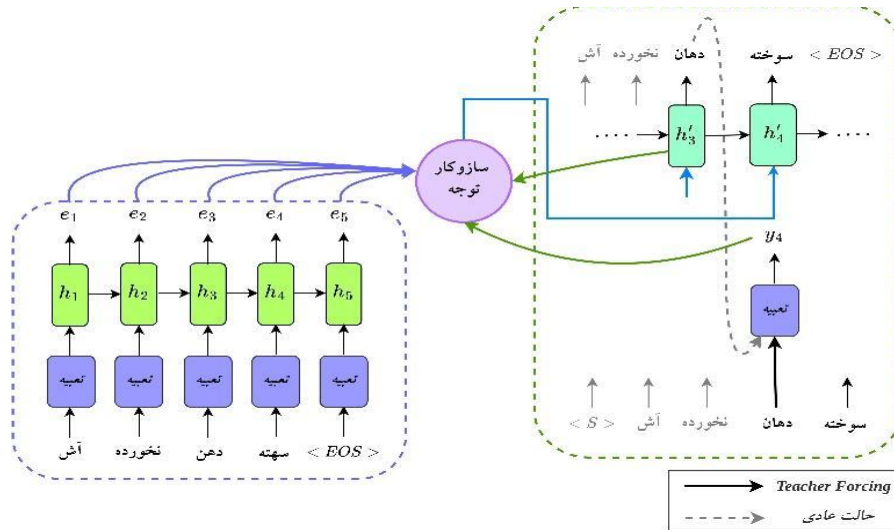
شرح مدل بازگشتی

کدگذار این مدل متشکل از یک لایه‌ی حافظه کوتاه‌مدت طولانی (LSTM)^۲ بعد از یک لایه‌ی تعبیه‌سازی^۳ واحدهای زبانی است. کدگشا نیز معماری مشابهی دارد، اما بین دو بخش مدل، با استفاده از سازوکار توجه بین خروجی‌های کدگذار، ورودی، و حالت نهان کدگشا ارتباط برقرار می‌شود. تصویری کلی از مدل در شکل ۱ آمده است.

همان‌طور که در شکل نشان داده شده است، کلمه‌های جمله‌ی لری «آش نخورده دهن سهته» و جمله‌ی فارسی متناظر آن، «آش نخورده دهان سوخته»، پس از تعبیه، به واحدهای بازگشتی داده می‌شوند. منظور از h_i یا h'_i حالت نهان واحد بازگشتی و منظور از e_i خروجی کدگذار در زمان i است. هم‌چنین کلمه‌های $\langle S \rangle$ و $\langle EOS \rangle$ به ترتیب شروع و پایان رشته را نشان می‌دهند. در حین ترجمه، کدگشا تا زمانی که کلمه $\langle EOS \rangle$ در خروجی تولید نشده باشد به کار خود ادامه می‌دهد. هم‌چنین اگر در زمان آموزش از روش کمک معلم^۴ استفاده شود، ورودی کدگشا در هر قدم به طور مستقیم از رشته‌ی هدف برداشته می‌شود، اما در حالت عادی و نیز در زمان ارزیابی، ورودی کدگشا همان خروجی گام قبلی است.

1 رمزگذاری زوج بایت

2. Long Short-Term Memory
3. Embedding
4. Teacher Forcing



شکل ۱. تصویر کلی مدل بازگشتی

شرح مدل مبتنی بر مبدل

بخش کدگذار، از دو لایه‌ی کدگذار که روی هم پشته^۱ شده‌اند، تشکیل شده است. در مورد بخش کدگشا نیز همین شرایط برقرار است. در هر قسمت، خروجی لایه‌ی قبلی، به عنوان ورودی وارد لایه‌ی بعد می‌شود.

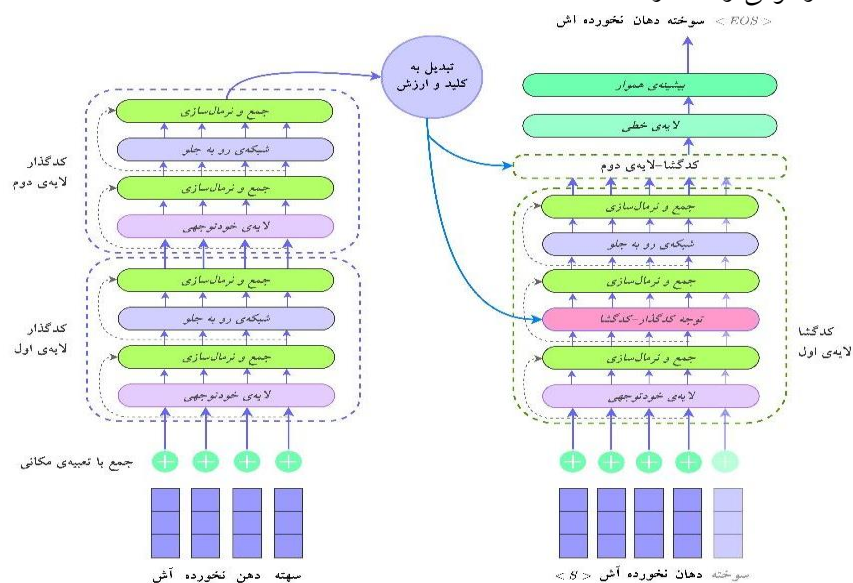
هر کدگذار از دو زیرلایه‌ی خودتوجهی^۲ و شبکه‌ی عصبی روبه‌جلو تشکیل شده است. همچنین از یک اتصال باقی‌مانده^۳ در اطراف هر یک از زیرلایه‌ها استفاده می‌شود و یک لایه نرمال‌سازی پس از آن قرار می‌گیرد. هر کدگشا، علاوه بر زیرلایه‌های ذکر شده برای کدگذار، یک زیرلایه‌ی دیگر نیز به نام توجه کدگذار-کدگشا^۴ دارد. این لایه سازوکار توجه چندسراسری^۵ را روی خروجی کدگذار اعمال می‌کند. مشابه کدگذار، در بخش کدگشا نیز یک اتصال باقی‌مانده در اطراف هر زیرلایه وجود دارد و به دنبال آن، یک لایه نرمال‌سازی انجام می‌شود.

همانطور که اشاره شد، ورودی هر یک از کدگذارها، خروجی کدگذار قبلی است. به این ترتیب نیاز است در ابتدا کلمات ورودی با استفاده از یک الگوریتم تعبیه‌سازی به بردارهایی نظیر شوند. پس از این مرحله نیاز است موقعیت کلمات در جمله نیز به نحوی در بردارهای نهایی گنجانده شود؛ چرا که در داده‌های زبانی، ترتیب اهمیت بالایی دارد. بعنوان مثال، دو عبارت «کار

1. Stack
2. Self-attention
3. Residual Connection
4. Encoder-Decoder attention
5. Multi-head attention

برای زندگی» و «زندگی برای کار» با آنکه متشکل از تعدادی کلمه‌ی کاملاً مشترک هستند، به دلیل تفاوت در موقعیت کلمات، معانی کاملاً متفاوتی منتقل می‌کنند. برای لحاظ کردن این مورد در معماری مبتنی بر مبدل از تعبیه‌ی مکانی^۱ استفاده می‌شود. به این معنا که جایگاه هر کلمه در مرحله‌ی تعبیه‌سازی به همراه خود کلمه در بردار نهایی آن واژه تعبیه می‌شود، و هر بردار علاوه بر محتوای کلمه، اطلاعاتی از جایگاه کلمه در جمله را نیز در خود دارد. بدین ترتیب بردار دیگری که به کمک تابعی سینوسی برای هر کلمه محاسبه می‌شود، به بردار مربوط به کلمه اضافه می‌شود.

پس از اعمال موارد گفته شده، خروجی وارد کدگذار اول می‌شود. خروجی آخرین کدگشا در هر مرحله یک بردار خواهد بود، اما آنچه ما انتظار داریم، یک کلمه در زبان مقصد است. لذا برای استخراج کلمات از بردارهای تولید شده، ابتدا به یک لایه شبکه عصبی خطی نیاز داریم که بردار فعلی را به برداری با اندازه‌ی بزرگتر نظیر کند. برای مثال، اگر تعداد لغاتی که از پیش آموخته شده است، ۲۰۰۰ باشد، اندازه‌ی برداری که پس از اعمال لایه‌ی خطی خواهیم داشت، برابر ۲۰۰۰ خواهد بود. در این صورت هر درایه از بردار را می‌توان متناظر با یک کلمه از کل لغات در نظر گرفت. نهایتاً با اعمال تابع بیشینه‌ی هموار^۲ روی این بردار، بالاترین احتمال، به عنوان کلمه‌ی خروجی در نظر گرفته می‌شود. تصویری کلی از معماری یاد شده در شکل ۲ آمده است (وسوانی و همکاران، ۲۰۱۷).



شکل ۲. تصور کلی مدل مبتنی بر مبدل

1. Positional Encoding
2. Softmax

۳.۳. نحوه ارزیابی

از مهم‌ترین معیارهای سنجش ترجمه ماشینی معیار BLEU می‌باشد که در این کار نیز مورد استفاده قرار گرفته است. معیار BLEU سعی می‌کند شباهت رشته‌ی ترجمه‌ی تولیدشده و رشته‌ی مرجع را با انطباق n-gramها بسنجد. رابطه‌ی محاسبه این معیار به صورت زیر است:

$$BLEU\ SCORE = BP \times \exp\left(\frac{1}{N} \sum_{n=1}^4 P_n\right) \quad (1)$$

$$P_n = \frac{\text{تعداد } n\text{-gram رشته تولید شده}}{\text{تعداد } n\text{-gram رشته مرجع}} \quad (2)$$

$$BP = \exp\left(\min\left(1 - \frac{\text{طول مرجع}}{\text{طول ترجمه تولید شده}}, 0\right)\right) \quad (3)$$

جریمه‌ی کوتاهی^۱ که در رابطه بالا با BP نشان داده شده است، از اختصاص نابجای مقادیر بالای BLEU به جملات کوتاه جلوگیری می‌کند. عبارت P_n نیز دقت براساس n-gram مربوطه است که n در بازه یک تا چهار قرار دارد.

تنظیمات آزمایش و نحوه‌ی آموزش

برای آموزش هر دو نوع معماری، ۸۰٪ دادگان هر گویش برای آموزش مدل، ۱۰٪ برای راستی‌آزمایی و تنظیم فرامولفه‌ها، و ۱۰٪ برای آزمون به کار رفته است. داده‌های آزمون و ارزیابی از نمونه‌های بلندتر انتخاب شده و از کلمات و عبارات کوتاه، تنها در روند آموزش بهره گرفته شده است. در هر دو مدل، نرخ یادگیری به صورت تطبیقی^۲ تنظیم شده و در طول آموزش بسته به روند آن، تغییر می‌کند. از آنتروپی متقاطع^۳ به عنوان تابع هزینه‌ی هر دو مدل استفاده شده و برای هر دو گویش لری و یزدی، مدل‌ها ۵۰۰ دور آموزش دیده‌اند.

۴. نتایج آزمایش‌ها

در این بخش به بررسی نتایج حاصل از آزمایش‌های صورت گرفته و تحلیل خروجی‌های آن‌ها می‌پردازیم.

1. Brevity Penalty
2. Adaptive
3. Cross Entropy

۱.۴. مقایسه‌ی امتیاز BLEU

نتیجه‌ی امتیاز BLEU در هر آزمایش برای دادگان ارزیابی و آزمون در جدول 10 آورده شده است. در مجموع با وجود تعداد نسبتاً کم دادگان حاضر، تأثیر بسزای استفاده از مدل مبتنی بر مبدل بر نتایج هر دو گویش، به خصوص لری، مشهود است.

جدول ۱۰. نتایج دادگان ارزیابی و آزمون

مدل	گویش	جهت ترجمه	ارزیابی	آزمون
مدل کدگذار-کدگشا بازگشتی	لری	به فارسی	0.19 %	0.14 %
		از فارسی	0.10 %	0.14 %
	یزدی	به فارسی	0.15 %	0.09 %
		از فارسی	0.10 %	0.12 %
مدل کدگذار-کدگشا بر پایه‌ی مبدل	لری	به فارسی	7.48 %	7.39 %
		از فارسی	4.30 %	3.29 %
	یزدی	به فارسی	0.64 %	0.73 %
		از فارسی	0.49 %	0.77 %

۲.۴. تحلیل و بررسی خروجی‌های مدل مبدل

همانطور که در بخش قبل قابل مشاهده است، مدل مبدل عملکرد بهتری بر روی هر دو گویش داشته است. در این قسمت، به تحلیل چند خروجی این مدل و مدل بازگشتی برای هر دو گویش می‌پردازیم. برخی از نمونه‌های خروجی ترجمه‌ی ماشینی هر دو مدل روی داده لری در جدول 11 آورده شده است. همانطور که مشاهده می‌شود، خروجی مدل مبدل کاملاً نزدیک به متن مرجع، و در مواردی دقیقاً همان است. در حالی که خروجی مدل بازگشتی همانطور که از نتایج قابل پیشبینی است، از ترجمه درست، بسیار فاصله دارد، و گاهاً، خروجی کاملاً بیربطی را ارائه داده است.

در جدول 12 نمونه‌هایی از ترجمه‌ی ماشینی مدل مبتنی بر مبدل و مدل بازگشتی، گویش یزدی آورده شده است که شامل ترجمه از یزدی به فارسی و برعکس می‌باشد.

همانطور که مشاهده می‌شود، به دلیل کمبود منابع زبانی، این مدل قادر به ترجمه‌ی صحیح جملات نیست؛ با این حال پیداست که مدل مبتنی بر مبدل، توانسته الگوهایی را برای ترجمه یاد بگیرد. برای نمونه، در مثال اول بخش‌هایی از ترجمه به مرجع نزدیک هستند و یا در مثال دوم، با اینکه ترجمه به درستی انجام نشده است، اما در حالت یزدی به فارسی، مدل توانسته است منفی بودن جمله را درک کند. این مورد در مثال سوم نیز دیده می‌شود که مدل توانسته مثبت بودن فعل را در هر دو حالت ترجمه تشخیص دهد. در اینجا نیز همانطور که انتظار می‌رود، نتایج حاصل از مدل بازگشتی مطلوب نیست.

جدول ۱۱: برخی از نمونه‌های خروجی مدل مبتنی بر مبدل و مدل بازگشتی، مربوط به مجموعه داده‌ی لری

متون ترجمه ماشینی			متون اصلی		
فارسی-لری بازگشتی	فارسی-لری مبدل	لری-فارسی بازگشتی	لری- فارسی مبدل	فارسی	لری
و ش	یک ناگهی اومه	به اندازه	با حرف آمد	به حرف آمد	و حرف اومه
یه دهسه	آدم کم دهسی یه	آدم دست	آدم کم دستی است	آدم کم دستی است	آدم کم دهسی یه
وشش ایکنه	یالا ایکاره بکُ	آدم یک اندازه-	یا الله این کار را بکن	یا الله این کار را بکن	یالا ای کاره بکُ

نتایج به دست آمده بر به وضوح نشان می‌دهد که مدل مبدل به مراتب از مدل بازگشتی جهت ترجمه زبانهای محلی بهتر عمل میکند. خطاهای حاصل در خروجی بیش از هر مسئلهی دیگری تحت تأثیر حجم پایین دادگان است. بیشک افزایش دادگان باعث بهبود نتایج خواهد شد. همچنین وجود ساختارهایی که در بخش ۱.۲ به آن به طور مفصل پرداخته شده است، کار ترجمه را برای مدل‌های مورد استفاده دشوار کرده است.

جدول ۱۲: برخی از نمونه‌های خروجی مدل مبتنی بر مبدل و مدل بازگشتی، مربوط به مجموعه داده یزدی

متون ترجمه شده ماشینی			متون اصلی		
فارسی-یزدی بازگشتی	فارسی-یزدی مبدل	یزدی-فارسی بازگشتی	یزدی-فارسی مبدل	فارسی	یزدی
حالا و و	معلومه! کسی گیرونه.	ی	معلوم است خوب دیگر کسی نیست.	معلوم است دیگر کسی به فکر گرسنگان نیست.	معلومه دیه کسی یادشه گُشنا نیس.
حالا	گف: دلتون دردو میتونم.	ی	گفت خیلی خوب نیست.	گفت دلتان درد نمی گرفت.	گف: دلتون درد نمگرف.
حالا	جُمُلِجِیمه بود	ی	حناسب هم داشت	حرف مان هم درست بود	حرفونن درس بود

۵. نتیجه‌گیری و کارهای آینده

در این پژوهش برای نخستین بار دو مجموعه داده‌ی موازی برای گسترش پژوهش روی گویش‌های لری و یزدی معرفی شدند و مدل‌های رشته‌به‌رشته‌ی نسبتاً کارآمدی روی آن‌ها

آموزش دیدند. با توجه به محدودیت منابع دیجیتال مناسب، چالش‌های زیادی بر سر تهیه داده‌های قابل استفاده وجود داشته است.

بی‌شک برای استمرار پژوهش‌های آتی، نیاز به تهیه داده‌های بیشتر و گسترش مجموعه داده‌ی فعلی پابرجا خواهد بود. دادگان بالقوه‌ی فراوانی در بستر وب و فضای مجازی اجتماعی وجود دارد که می‌تواند منبع ارزشمندی برای گسترش پژوهش و حتی آموزش مدل‌های دنیای واقعی روی گویش‌ها به شمار رود. البته استفاده از این نوع داده تنها به شرط پیش‌پردازش مناسب ممکن است. فراتر از تنوع‌های زبانی، الگوی نوشتاری در فضای غیررسمی گاهی به سلیقه‌های شخصی و غیراستانداردی می‌رسد که چالش بزرگی برای همگون‌سازی آن‌ها در یک کار رایانشی ایجاد خواهد کرد. معرفی رابط‌های کاربری برای جمع‌آوری دادگان کنترل‌شده از افراد مسلط به گویش می‌تواند یک جایگزین کوچک، اما مناسب باشد. هم‌چنین پیش‌پردازش مختص به گویش گاهی مهم در هر کدام از وظایف محتمل برای این دادگان است که نیازمند بررسی بیشتر و حتی توسعه‌ی کتابخانه‌هایی جهت تسهیل سیر پردازش در مطالعات مرتبط است.

از دیگر جهات پژوهشی می‌توان به بررسی هم‌بستگی بین گویش‌های مختلف و کمک دادگان گویش‌های مجزا به یکدیگر برای آموزش مدل‌های مشترک اشاره کرد. بی‌شک کار بر روی گویش‌های فارسی در ابتدای مسیر خود است و هم‌چنان جای کار بر روی معماری مدل‌ها و روش آموزش وجود دارد. دادگان جمع‌آوری شده در این پژوهش می‌تواند نقطه‌ی شروعی برای کارهای آتی در رابطه با گویش‌های مذکور باشد.

تقدیر و تشکر

بدین‌وسیله از جناب آقای افضل مقیمی و همکارانشان در تألیف لغت‌نامه گری بویراحمدی، که حاصل کار ارزشمند خود را در اختیار این پژوهش قرار دادند، صمیمانه تشکر می‌کنیم. هم‌چنین از آقایان سید یاسین موسوی، امیرعلی ابراهیم‌زاده، و محمدجواد هزاره جهت همکاری در استخراج خودکار دادگان از لغت‌نامه تقدیر و تشکر به عمل می‌آوریم.

منابع

- داوری و جانی، پریسا و ابراهیم (۱۳۹۷)، «بررسی و توصیف زبان‌شناختی انواع ضمیر در گویش لری کامفیروز»، فصلنامه ادبیات و زبان‌های محلی ایران زمین، دوره ۸، شماره ۳، ص ۴۷-۶۲
- رمضان‌خانی، صدیقه (۱۳۹۱)، «بررسی برخی واژگان یزدی و مقایسه آن‌ها با زبان‌های باستانی»، ششمین همایش پژوهش‌های ادبی، تهران
- صادق زاده و رمضان‌خانی، محمود و صدیقه (۱۳۹۸)، «بررسی تطبیقی - موضوعی ساختار واژگان در گویش یزدی»، فصلنامه علمی پژوهشی زبان و ادب فارسی، شماره ۴۰.

طاهری، اسفندیار (۱۳۹۱)، «ریشه شناسی چند واژه از لری بویراحمدی»، ادب پژوهی، دوره ۶، شماره ۲۰، ص ۷۵-۸۸

عسکری کامران، محمد تقی (۱۴۰۰)، ننه زهرا و پسرش، یزد یادداشت نو.
 مجیدی و حق بی، لایلا و فریده (۱۳۹۷)، «نمود فعل در زبان لری و گونه‌های آن، فصلنامه مطالعات زبان
 ها و گویش های غرب ایران»، دانشکده ادبیات و علوم انسانی، دانشگاه رازی کرمانشاه، سال
 ششم، شماره ۲۲، ص ۹۳-۱۱۰
 مدرسی، یحیی (۱۳۶۸)، «درآمدی بر جامعه شناسی زبان، تهران»، مؤسسه مطالعات و تحقیقات فرهنگی.
 محسنی، محمدرضا (۱۳۹۲)، «پان ترکیسم ایران و آذربایجان»، چاپ دوم، نشر سمرقند، ص ۱۳۱-۱۳۰
 مقیمی و نظری و خالقی زاده و مقیمی، افضل و جلیل و محمد هادی و جبار (۱۴۰۰)، «فرهنگ واژه‌های
 لری بویراحمدی»، تهران: زیتون سبز.

- Anastasopoulos, A, et al. "Endangered languages meet Modern NLP." *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*. 2020.
- Asgari, E and Schütze, H. 2017. Past, Present, Future: "A Computational Investigation of the Typology of Tense in 1000 Languages". In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113-124, Copenhagen, Denmark. Association for Computational Linguistics.
- Baniata, L. H., Ampomah, I., & Park, S. (2021). "A Transformer-Based Neural Machine Translation Model for Arabic Dialects That Utilizes Subword Units". *Sensors*, 21(19), 6509.
- Găman, M., & Ionescu, R. T. (2020). "The unreasonable effectiveness of machine learning in Moldavian versus Romanian dialect identification". *International Journal of Intelligent Systems*.
- Harrat, S., Meftouh, K., & Smaili, K. (2019). "Machine translation for Arabic dialects (survey)". *Information Processing & Management*, 56(2), 262-273.
- King, B. P. "Practical Natural Language Processing for Low-Resource Languages" (Doctoral dissertation). *University of Michigan*. (2015).
- Haddow, B., Hernández, A., Neubarth, F., & Trost, H. (2013, September). "Corpus development for machine translation between standard and dialectal varieties". In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants* (pp. 7-14).
- Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., & Smaili, K. (2015, October). "Machine translation experiments on PADIC: A parallel Arabic dialect corpus". In *Proceedings of the 29th Pacific Asia conference on language, information and computation* (pp. 26-34).
- Mohamed, E., Mohit, B., & Oflazer, K. (2012, July). "Transforming standard Arabic to colloquial Arabic". In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Volume 2: Short Papers, 176-180.
- Mutton, A., Dras, M., Wan, S., & Dale, R. (2007, June). GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 344-351).
- Nakov, P., & Ng, H. T. (2012). "Improving statistical machine translation for a resource-poor language using related resource-rich languages". *Journal of Artificial Intelligence Research*, 44, 179-222.
- Ruiz Costa-Jussà, M., Zampieri, M., & Pal, S. (2018). "A neural approach to language variety translation". In *COLING 2018: The 27th International Conference on*

- Computational Linguistics: Proceedings of the Conference: August 20-26, 2018 Santa Fe, New Mexico, USA*. Association for Computational Linguistics.
- Salloum, W., & Habash, N. (2012, December). Elissa: "A dialectal to standard Arabic machine translation system". In *Proceedings of COLING 2012: Demonstration Papers* (pp. 385-392).
- Sawaf, H. (2010). "Arabic dialect handling in hybrid machine translation". In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*.
- Scannell, K. P. (2006). "Machine translation for closely related language pairs". In *Proceedings of the Workshop Strategies for developing machine translation for minority languages* (pp. 103-109).
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). "Sequence to sequence learning with neural networks". *Advances in neural information processing systems*, 27.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. (2017). "Attention is all you need". *Advances in neural information processing systems*, 30.
- Wolk, K., & Koržinek, D. (2016). Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking. *arXiv preprint arXiv:1601.02789*.