

Fajik: A Neural Encoder-Decoder Model along with Required Language Resources for an Accurate Tajiki-Persian Transliteration

Sadraei Javaheri, Mohammad Ali¹ 

Master Student at Sharif University, Artificial Intelligence Group, Department of Computer Engineering Data Science & Machine Learning Lab, Tehran. Iran.

Asgari, Ehsan² 

Applied Science and Technology, University of California Berkeley Language Processing and Digital Humanities Lab, AI Group, SUT, Tehran. Iran.

Rabiee, Hamid Reza³ 

Professor at Sharif University, Artificial Intelligence Group, Department of Computer Engineering Data Science & Machine Learning Lab, Tehran. Iran

Abstract

The Tajik language, also known as Tajiki Persian, is a variation of Persian (Farsi) spoken in Tajikistan. One of the main distinctions between Iranian Persian and Tajiki Persian is the writing script. Since the early 1900s, Tajiki has been written in the Cyrillic script. Although the difference between spoken Tajiki and spoken Persian is not significant, the script difference caused a cultural break between these two nations making many cultural resources (e.g., poems, stories, etc.) unavailable for the newer generations. An automatic and accurate transliterator model can again fill the gap between Persians and Tajiks and facilitate transfer learning between these two language variations. The efforts on Persian-Tajiki transliteration have primarily been rule-based and context-independent, causing many errors. Deep learning methods, particularly neural language models, revolutionized computational linguistics in the last decade by providing language understanding through contextualized representations. In this work, we create a Persian-Tajiki transliteration dataset for training and evaluation purposes. We also train an accurate neural sequence-to-sequence model transliterating between Tajiki-Persian and Persian-Tajiki.

Keywords: Deep learning, Encoder-decoder model, Transliteration, Tajiki Persian


1. m.sadraei@sharif.edu


2. ehsan.asgari@sharif.edu


3. rabiee@sharif.edu

How to Cite: Sadraei Javaheri, M. A., Asgari, E., & Rabiee, H. R. (2023). Fajik: A Neural Encoder-Decoder Model along with Required Language Resources for an Accurate Tajiki-Persian Transliteration. *Language and Linguistics*, 18(35), 87-103. doi: 10.30465/lsi.2023.8500

فاجیک: مدل شبکه عصبی ژرف رشته به رشته و منابع زبانی مرتبط برای نویسه‌گردانی میان تاجیکی و فارسی

صدراپی جواهری، محمدعلی  دانشجوی کارشناسی ارشد هوش مصنوعی دانشکده مهندسی کامپیوتر شریف، تهران، ایران

عسگری، احسان‌الدین  استادیار گروه هوش مصنوعی دانشکده مهندسی کامپیوتر شریف، تهران، ایران

ربیعی، حمیدرضا  استاد گروه هوش مصنوعی دانشکده مهندسی کامپیوتر شریف، تهران، ایران

چکیده: زبان تاجیکی، یا به شکل درست‌تر گویش تاجیکی زبان فارسی، از گونه‌های زنده زبان فارسی است که در کشور تاجیکستان رواج دارد. مهم‌ترین تفاوت فارسی رایج در ایران با فارسی تاجیکی در سیستم نوشتار است که این تفاوت سبب ایجاد گسست فرهنگی میان دو ملتی که گفتار هم را متوجه می‌شوند، شده است. ساخت سامانه‌ای برای تبدیل این دو نوشتار می‌تواند به نزدیکتر کردن این دو ملت کمک شایان‌ذکری کند. گذشته تلاش‌هایی برای ساخت نویسه‌گردان رایانشی بین این دو گونه نوشتار صورت گرفته است. این سامانه‌ها با استفاده از روش‌های سنتی و قانون‌محور^۱ این کار را انجام می‌دادند و برای همین خطاهای قابل توجهی در خروجی خود دارند. در این پژوهش تلاش شده است که با کمک روش‌های یادگیری ژرف این نویسه‌گردانی انجام شود. در این پژوهش ابتدا پیکره موازی میان فارسی و تاجیکی جمع‌آوری شده است. سپس با کمک

مدل‌های رشته به رشته یک سامانه نویسه‌گردانی با عملکرد بهتر نسبت به سامانه‌های گذشته ایجاد شده است.

کلیدواژه‌ها: یادگیری ژرف، مدل رشته به رشته، نویسه‌گردانی، فارسی تاجیکی.

۱ مقدمه

فارسی تاجیکی از گویش‌های زبان فارسی است که اکثر متکلمان آن در کشور تاجیکستان زندگی می‌کنند. بخش قابل توجهی از مردم ازبکستان به خصوص در شهرهای سمرقند و بخارا، از گفتار و نوشتار تاجیکی استفاده می‌کنند. برخی زبان‌شناسان تاجیکی را یک زبان جدا و برخی دیگر آن را گویشی از زبان فارسی در نظر می‌گیرند. با وجود این که در کلمات، نوشتار و دستور زبان این دو دارای تفاوت‌هایی نسبت به یکدیگر هستند، فهم متقابل^۱ بین متکلمین فارسی ایرانی و فارسی تاجیکی وجود دارد. به همین سبب در این مقاله تاجیکی به عنوان یک گویش در نظر گرفته می‌شود.

تاجیکستان کشوری در شمال شرق افغانستان است که مرز مشترکی با ایران ندارد. مساحت این کشور ۱۴۳ هزار کیلومتر مربع است و نزدیک ۱۰ میلیون نفر جمعیت دارد. همانگونه که در نقشه شکل مشاهده می‌شود فارسی زبانان از جنوب غرب ایران تا منطقه تاجیکستان و اطراف آن حضور دارند. از منظر جغرافیایی تاجیکستان دورترین منطقه فارسی زبان نسبت به ایران است که همین موضوع سبب شده است تفاوت‌های بین این دو گونه زبان فارسی در گذر زمان به وجود بیاید.



شکل ۱- مناطقی با بیشینه متکلم فارسی (پلاتون^۲، ۲۰۱۲)

1. mutual intelligibility

2. A. Platon

در سطوح مختلف زبان، تفاوت‌هایی بین فارسی تاجیکی و فارسی ایرانی وجود دارد. برای مثال در فارسی تاجیکی کمتر از کلمات عربی و فرانسوی استفاده می‌شود، در عوض تعداد زیادی از کلمات روسی در زبان حضور دارند. تفاوت این دو فقط در سطح کلمات نیست و حتی در دستور زبان تفاوت‌های اندکی وجود دارد. اما بزرگترین تفاوتی که باعث فاصله افتادن بین مردم تاجیکستان و ایران شده است، تفاوت در سیستم نوشتار است.

مردم تاجیکستان قبل از سال ۱۳۰۷ هجری شمسی مانند سایر فارسی‌زبانان با الفبای فارسی - عربی نوشتار می‌کردند. اما از سال ۱۳۰۷ تا سال ۱۳۱۹ تحت فشار حکومت شوروی الفبای مردم تاجیکستان دو مرتبه تغییر پیدا کرد. چند سالی در این منطقه استفاده از الفبای لاتین اجباری شد و سپس الفبای سیریلیک جایگزین الفبای لاتین شد (کلر^۱، ۲۰۰۱). تغییر خط در تاجیکستان سبب شد که تاجیکیان نسبت به همزبانان ایرانی و افغانستانی خود تا حدودی بیگانه شوند و ارتباطات مناسبی میان آنان صورت نگیرد. همچنین این تغییر سبب شده است که طیف زیادی از منابع فارسی برایشان غیر قابل استفاده باشد. کمبود منابعی که با فارسی تاجیکی نوشته شده‌اند، باعث شده است که مردم تاجیکستان به زبان روسی گرایش بیشتری داشته باشند. (خودکولوا^۲، ۲۰۱۵) چرا که نوشتارهای ادبی و علمی زیادی به زبان روسی وجود دارد.

متأسفانه مردم کمی در تاجیکستان با الفبای فارسی آشنا هستند. با استفاده از نویسه‌گردانی^۳ میان این دو زبان می‌توان فاصله میان این دو قوم را کم کرد. نویسه‌گردانی با کمک نیروی انسانی، فقط می‌تواند تعداد محدودی از آثار را به نوشتار دیگر برگرداند. همچنین منبع بزرگی مانند وب فارسی که روز به روز به حجم آن افزوده می‌شود، نمی‌تواند به گونه‌ای غیر خودکار نویسه‌گردانی کرد. برای همین لازم است که این کار به شکل رایانشی و خودکار انجام گیرد.

نویسه‌گردان‌های رایانشی متنوعی در سال‌های اخیر عرضه شده‌اند. برای نمونه می‌توان به نویسه‌گردان میان انگلیسی و پنجابی (وانگ^۴ و همکاران، ۲۰۱۲)، انگلیسی و چینی (شاو^۵ و نیوره^۶، ۲۰۱۶) یا حتی انگلیسی و فارسی (محصولی و صفابنخس، ۲۰۱۷) اشاره کرد. در گذشته تلاش‌هایی برای انجام نویسه‌گردانی رایانشی میان فارسی و تاجیکی صورت گرفته است. سایت پرشین تاجیک^۷ و سامانه به‌روزیان^۸ دو وب‌گاه رایگان برای انجام این کار

1. S. Keller

4. P. Wang

7. www.persian-tajik.ir

2. N. Khudoikulova

5. Y. Shao

8. pertoj.com/payvand/index.php

3. transliteration

6. Y. Nivre

هستند. همچنین با استفاده از روش‌های آماری تلاشی برای انجام این نویسه‌گردانی صورت گرفته است. (دیویس^۱، ۲۰۱۲) این سامانه‌ها با استفاده از روش‌های سستی زبان‌شناسی رایانشی، نویسه‌گردانی را انجام می‌دهند و در خروجی‌هایشان اشتباهات زیادی وجود دارد. طراحی سامانه‌ای با خطای کمتر و انعطاف‌پذیری بیشتر هدفی است که در این پژوهش روی آن کار شده است. در دهه گذشته، روش‌های مبتنی بر یادگیری ژرف^۲ تحولات گسترده‌ای در زبان‌شناسی رایانشی ایجاد کردند و نشان دادند که عملکرد بهتری نسبت به روش‌های سستی دارند. مدل‌های یادگیری ژرف در زبان‌شناسی رایانشی انواع متفاوتی دارند. دسته‌ای از مدل‌ها که با نام مدل‌های رشته‌به‌رشته^۳ شناخته می‌شوند برای انجام کارهایی از قبیل ترجمه، خلاصه‌سازی و نویسه‌گردانی استفاده می‌شوند. یک مدل رشته به رشته نوعی مدل زبانی شرطی است که با دانستن یک رشته در زبان مبدا، محتمل‌ترین رشته زبان مقصد را تولید می‌کند. سرعت بالای پیشرفت مدل‌های یادگیری سبب شد که در دهه گذشته مدل‌های رشته‌به‌رشته متنوعی معرفی شوند. ابتدا مدل‌های مبتنی بر شبکه عصبی بازگشتی^۴ رواج داشتند. (سوتسکور^۵ و همکاران، ۲۰۱۴) این مدل‌ها هنگام تولید رشته مقصد دچار فراموشی اطلاعات رشته مبدا می‌شدند. در اینجا محققان با معرفی کردن فرایند توجه^۶ تلاش کردند که این مشکل را حل کنند. (باهدانو^۷ و همکاران، ۲۰۱۵) اما پیشرفت مدل‌ها همینجا متوقف نشد. در سال ۲۰۱۷ محققین توانستند با معرفی مبدل‌ها^۸ بهترین مدل رشته به رشته حال حاضر را بسازند (وسوانی^۹ و همکاران، ۲۰۱۷). طبق اطلاعات در دسترس ما تاکنون تلاشی برای استفاده از مدل‌های رشته‌به‌رشته در نویسه‌گردانی میان تاجیکی و فارسی صورت نگرفته است. در این پژوهش سعی شده است که یک سامانه رایانشی دقیق برای نویسه‌گردانی میان این دو نوشتار طراحی شود. این سامانه با رویکرد پیکره‌محور^{۱۰} به کمک شبکه‌های عصبی ژرف^{۱۱} پیاده‌سازی شده است.

۲ چالش‌های موجود

نوشتار فارسی و نوشتار تاجیکی چند تفاوت عمده دارند که باعث می‌شود نویسه‌گردانی کار به نسبت دشواری باشد. در این بخش، تعدادی از این چالش‌ها را بررسی می‌کنیم.

- | | | |
|-----------------------------------|--------------------------|-------------------------|
| 1. C. I. Davis | 2. deep learning | 3. sequence to sequence |
| 4. recurrent neural network (RNN) | 5. I. Sutskever | 6. attention mechanism |
| 7. B. Bahdanau | 8. transformers | 9. A. Vaswani |
| 10. corpus-driven | 11. deep neural networks | |

۲-۱ هم‌خوان‌های متفاوت در الفبای فارسی

در الفبای فارسی هم‌خوان‌های^۱ تکراری زیادی وجود دارد. این موضوع در تبدیل تاجیکی به فارسی سبب می‌شود که املائی درست یک کلمه برای رایانه مشخص نباشد. همچنین کلمات هم‌آوا^۲ ممکن است در فارسی املائی متفاوتی داشته باشند اما در نوشتار تاجیکی به یک شکل نوشته شوند. برای مثل دو کلمه «حیات» و «حیاط» هر دو در تاجیکی به یک شکل نوشته می‌شوند. (مگردومیان^۳ و پرواز^۴، ۲۰۰۸)

مثال (۱)

xaɛt
/hæjɔ:t/
حیات

مثال (۲)

xaɛt
/hæjɔ:t/
حیاط

۲-۲ عدم نوشتن واژه‌های کوتاه و تنوین در نوشتار فارسی

در نوشتار رایج زبان فارسی در نقطه مقابل نوشتار تاجیکی، واژه‌های کوتاه و تنوین نوشته نمی‌شوند. به همین دلیل، در تبدیل فارسی به تاجیکی، رایانه باید بتواند خوانش درست کلمه را حدس زده و آن را به شکل درست به تاجیکی برگرداند. این مشکل زمانی چالش‌برانگیزتر می‌شود که برای آن کلمه چند شکل مختلف خوانش وجود داشته باشد (برجیان ۱۳۷۸). به مثال‌های زیر توجه کنید:

مثال (۳)

мулк
/mulk/
مُلک

مثال (۴)

малак
/mælæk/
مَلک

1. consonant
3. K. Megerdoomian

2. homophone
4. D. Parvaz

مثال (۵)

малик
/mælik/
ملک

۲-۳ عدم نوشتن کسره اضافه در نوشتار فارسی

کسره اضافه یکی از عناصر مهم در زبان فارسی است که کاربرد آن پیوند دادن کلمات برای ایجاد ترکیب‌های مختلف است. کسره اضافه مانند سایر واژه‌های کوتاه، در نوشتار فارسی نوشته نمی‌شود اما در نوشتار تاجیکی نوشته می‌شود. برای همین اگر قرار باشد نویسه‌گردانی از فارسی به تاجیکی صورت گیرد مدل باید بتواند مکان‌هایی در جمله که دارای کسره اضافه است را تشخیص دهد. در مثال زیر حرف آخر کلمه «کتاب» همان کسره اضافه است.

مثال (۶)

Китоби ман
/kitɔːbi mæn/
کتاب من

۲-۴ اتصال نشانه مفعولی به کلمه قبل خود در الفبای تاجیکی

در نوشتار تاجیکی نشانه مفعولی را به عنوان پسوند، به کلمه قبلی خود متصل می‌کنند. در نویسه‌گردانی از تاجیکی به فارسی مدل باید بتواند متوجه شود که «را»ی موجود در آخر کلمه نشانه مفعولی است یا متعلق به خود کلمه است. در مثال ۱ را متعلق به خود کلمه «صحرا» است ولی در مثال ۲ «را»ی مفعولی است که به کلمه «کتاب» متصل شده است.

مثال (۷)

Ман ба сахро рафтам.
/mæn bæ səhrɔː ræftəm/
من به صحرا رفتم.

مثال (۸)

Китобро хондам.
/kitɔːbrɔː xɔːndæm/
کتاب را خواندم.

۲-۵ اختلاف تلفظ بین دو کلمه یکسان بر اثر گذر زمان

استفاده از الفبای عربی برای نوشتار فارسی قدمت بسیار زیادی دارد. برای همین با این که

تلفظ کلمات در گذر زمان دچار تغییر شده‌اند نوشتار آن‌ها شکل قدیمی خود را حفظ کرده است. اما نوشتار تاجیکی قدمتی کمتر از صد سال دارد و برای همین شکل نوشتاری کلمات بسیار به تلفظ فعلیشان در منطقه تاجیکستان شباهت دارد. حتی برخی کلمات که در مناطق مختلف تاجیکستان به شکل‌های متفاوتی تلفظ می‌شوند، دارای چند املائی صحیح هستند. در نویسه‌گردانی موضوع اختلاف تلفظ بین فارسی و تاجیکی مشکل‌ساز است. برای مثال کلمه «تاریخ» در زبان تاجیکی دارای یک «ع» ساکن در تلفظ خود است که در نوشتار نیز نوشته می‌شود:

مثال (۹)

таърих
/tæʔriχ/

تاریخ

۳ روش‌شناسی

در این پژوهش از رویکرد پیکره‌محور استفاده شده است. در این رویکرد نیاز است که پیکره موازی^۱ بین مبدا و مقصد در دسترس باشد. طبق بررسی‌های ما هیچ پیکره موازی آماده‌ای یافت نشد. برای همین در این پژوهش برای اولین بار با خزش^۲ و بگانه‌های مختلف، یک پیکره موازی جمع‌آوری شده است.

۳-۱ دادگان

حجم نوشتار تاجیکی موجود در وب بسیار اندک است. البته تلاش‌هایی در گذشته برای جمع‌آورده پیکره تک‌زبانه تاجیکی صورت گرفته است اما حجم آن‌ها کم است و هیچ کدام از این پیکره‌ها یک پیکره موازی بین فارسی و تاجیکی نیست. (دوودف^۳ و غیره ۲۰۱۱) برای همین برای گام اول این پژوهش، یک پیکره موازی از کلمات و غزلیات جمع‌آوری شده است.

منبع اول فرهنگ لغت واژه‌جو^۴ است. وبگاه واژه‌جو یک فرهنگ لغت تاجیکی است که معادل نوشتار فارسی تعداد زیادی از کلمات در آن وجود دارد. نزدیک ۷۰ هزار کلمه از این

1. parallel corpus
3. G. Dovudov

2. crawl
4. vazhaju.tj

وبگاه استخراج شده است. متأسفانه اندکی از این کلمات دارای غلط نوشتاری در نوشتار فارسی خود هستند. این موضوع سبب ایجاد نوفه^۱ در داده جمع‌آوری شده می‌گردد و باید با افزایش حجم داده نسبت به این نوفه مقاوم شد.

برای افزایش حجم داده به عنوان منبع دوم از غزلیات سعدی و حافظ استفاده شده است. استفاده از غزلیات این مزیت را دارد که در سطح مصراع‌ها می‌توان پیکره موازی ایجاد کرد. با کمک وبگاه گنجور^۲ نسخه فارسی غزلیات و به کمک وبگاه روشن فکر^۳ نسخه تاجیکی آن به دست آمده است. چالشی که در جمع‌آوری این داده وجود داشت این بود که تعدادی از مصراع‌ها با هم اختلافات ریزی داشتند که عامل این موضوع اختلاف نسخ مختلف یک غزل است. در جدول زیر می‌توانید مشخصات آماری این داده را مشاهده کنید:

جدول ۱- مشخصات آماری پیکره موازی جمع‌آوری شده

نام مجموعه داده	تعداد سطرها
واژه‌جو	۷۳۰۵۶ کلمه یا عبارت
غزلیات حافظ	۷۴۹۶ مصرع
غزلیات سعدی	۱۱۶۷۸ مصرع

۲-۳ مدل استفاده شده

در این پژوهش از معماری یادگیری ژرف، شبکه عصبی بازگشتی همراه با توجه^۴ (باهداناو و همکاران، ۲۰۱۵) استفاده شده است. در دهه اخیر این معماری در زبان‌شناسی رایانشی تحول زیادی ایجاد کرده است. فرایند توجه عملکرد معماری‌های شبکه عصبی بازگشتی را برای ترجمه و نویسه‌گردانی بهبود می‌دهد. به طور دقیق‌تر معماری مورد استفاده قرار گرفته این پژوهش «واحد بازگشتی دروازه»^۵ (چو^۶ و همکاران، ۲۰۱۴) همراه با توجه است.

شبکه عصبی بازگشتی استفاده شده در این پژوهش از نوع رشته به رشته^۷ است که دارای دو بخش رمزگذار^۸ و رمزگشا^۹ است. بخش رمزگذار نوشتار ورودی را به یک بردار^{۱۰} تبدیل می‌کند و بخش رمزگشا تلاش می‌کند که از بردار خروجی مذکور، نوشتار

1. noise

4. with attention

7. Seq2Seq

9. decoder

2. ganjoo.net

5. gated recurrent unit (GRU)

8. encoder

10. vector

3. ravshanfikt.tj

6. K. Cho

مقصد را حرف به حرف تولید کند. بخش رمزگذار مدل این پژوهش دو طرفه^۱ و بخش رمزگشای آن یک طرفه است. فرایند توجه در هر مرحله رمزگشایی به بردارهای مراحل میانی رمزگذار مراجعه می‌کند که این موضوع سبب می‌شود که مدل بتواند با دقت بیشتری رمزگشایی کند. به این مراجعه کردن فرایند توجه گفته می‌شود.

واحد بازگشتی دروازه نوعی شبکه عصبی بازگشتی است که در هر مرحله متغیری برای تعیین مقدار فراموشی حالت قبلی و پذیرش حالت جدید دارد که از منظر ریاضی به شکل زیر تعریف می‌شود.

$$\begin{aligned} z_t &= \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \\ \hat{h}_t &= \phi_h(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \\ h_t &= z_t \odot \hat{h}_t + (1 - z_t) \odot h_{t-1} \end{aligned}$$

فرمول ۱ ریاضیات پشت واحد بازگشتی دروازه

فرایند توجه این پژوهش بر اساس شباهت فاصله کسینوسی با استفاده از بیشینه نرم^۲ تعریف شده است. اگر بردار فعلی رمزگشا را s_t ، تمامی حالت‌های میانی رمزگذار را h_i و بردار خروجی توجه را c_t بنامیم به زبان ریاضی خواهیم داشت:

$$\begin{aligned} a_{ti} &= \text{softmax}(s_t, h_i) = \frac{\exp(\langle s_t, h_i \rangle)}{\sum_{j=0}^T \exp(\langle s_t, h_j \rangle)} \\ c_t &= \mathbb{E}_{a_t}[h] = \sum_{j=0}^T \langle a_{tj}, h_j \rangle \end{aligned}$$

فرمول ۲ ریاضیات نحوه محاسبات توجه

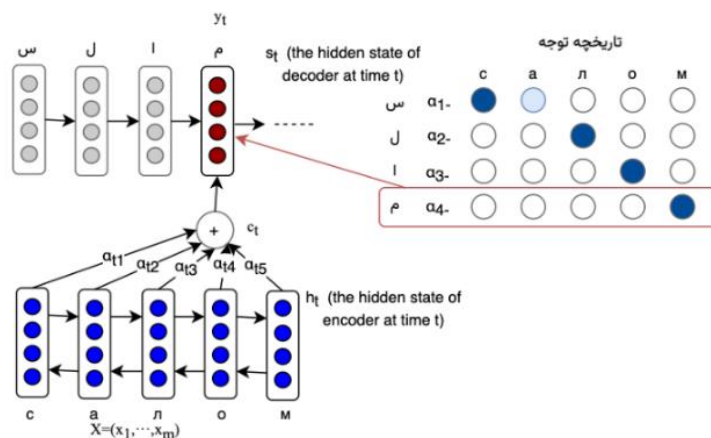
در رابطه بالا به کمک بیشینه نرم به یک توزیع جرمی احتمالی^۳ بر روی بردارهای میانی رمزگذار می‌رسیم. به شکل شهودی این توزیع مقدار توجه به هر بخش ورودی را مشخص

1. bidirectional

2. soft max

3. probability mass function (PMF)

می‌کند. پس از آن امید ریاضی بردارهای میانی ورودی بر اساس توزیع مذکور محاسبه می‌شود که حاصل آن بردار خروجی توجه است. بردار خروجی توجه همراه با سایر بردارها به رمزگشا داده می‌شود. دادن بردار خروجی توجه به رمزگشا تاثیر شایانی روی کیفیت خروجی مدل می‌گذارد. در شکل این مراحل نشان داده شده است. هر ردیف تاریخچه توجه، نشان‌دهنده توزیع جرمی احتمال محاسبه شده است. خانه‌های پررنگ‌تر دارای احتمال بیشتری هستند.



شکل ۲- نمایی از فرایند توجه برای شبکه عصبی بازگشتی با کدگذار تک لایه دو طرفه و کدگشای تک لایه یک طرفه برای وروی «caiom» و خروجی «سلام» همراه با نمایشی از ماتریس توجه هر حرف خروجی نسبت به حرفهای ورودی

۴ ارزیابی

در این پژوهش از ۸۰ درصد داده‌ها برای آموزش، ۱۰ درصد برای اعتبارسنجی و ۱۰ درصد برای آزمایش استفاده شده است.

برای ارزیابی مدل از کمینه فاصله ویرایشی^۱ استفاده شده است. در واقع مدلی بهتر است که خروجی‌های آن کمینه فاصله ویرایشی کمتری نسبت به مدل دیگر داشته باشد. در ازای تک تک داده‌ها نمی‌توان کمینه فاصله ویرایشی را مقایسه کرد به همین خاطر سه معیار متفاوت روی کمینه فاصله ویرایشی تعریف شده است.

1. minimum edit distance (MED)

معیار اول میانگین کمینه فاصله ویرایشی‌ها است. با این معیار می‌توان بدون مقایسه تک‌تک کلمات، به طور متوسط عملکرد مدل را مورد ارزیابی قرار داد. معیار دوم درصد داده‌هایی است که دقیقاً درست پیش‌بینی شده‌اند. به نوعی این معیار آن کسر از داده‌هایی را حساب می‌کند که کمینه فاصله ویرایشی برابر با صفر دارند. معیار سوم درصد داده‌هایی است که کمینه فاصله ویرایش کوچک‌تر یا مساوی یک دارند. طبیعتاً معیار سوم همیشه بزرگ‌تر یا مساوی معیار دوم است. وجود معیار سوم به ما کمک می‌کند که بدانیم اگر در خروجی مدل کمی اغماض کنیم چه کسری از اوقات خروجی درست را برگردانده است.

۵ نتایج

فراسنجه‌های مدل یادگرفته شده در جدول آمده‌اند.

جدول ۲- مشخصات مدل

معماری	تعداد لایه‌ها	اندازه جاسازی ^۲	اندازه مخفی ^۳
GRU	۲	۱۲۸	۱۰۲۴

۵-۱ نتیجه ارزیابی مدل

برای سنجیدن عملکرد مدل در مقایسه با دیگر سامانه‌های موجود، سه معیار گفته شده روی داده‌های آزمایش واژه‌جو و غزلیات سعدی محاسبه شده است.

جدول ۳- عملکرد از تاجیکی به فارسی بر روی داده آزمایش واژه‌جو

میانگین MED	درصد MED=0	درصد MED=1	
0.13	89%	97%	مدل این پژوهش
0.23	82%	96%	بهروزیان
0.46	72%	89%	پرشین تاجیک

1. hyperparameter
2. embedding size
3. hidden size

جدول ۴- عملکرد از تاجیکی به فارسی بر روی داده آزمایش غزلیات سعدی

درصد MED=1	درصد MED=0	میانگین MED	
84%	68%	0.67	مدل این پژوهش
76%	51%	0.88	بهروزیان
55%	30%	1.84	پرشین تاجیک

جدول ۵- عملکرد از فارسی به تاجیکی بر روی داده آزمایش وازه‌جو

درصد MED=1	درصد MED=0	میانگین MED	
94%	81%	0.27	مدل این پژوهش
80%	47%	0.82	بهروزیان
51%	25%	1.62	پرشین تاجیک

جدول ۶- عملکرد از فارسی به تاجیکی بر روی داده آزمایش غزلیات سعدی

درصد MED=1	درصد MED=0	میانگین MED	
58%	33%	1.8	مدل این پژوهش
11%	2%	3.97	بهروزیان
7%	0%	4.54	پرشین تاجیک

۵-۲ محدودیت‌ها و خطاها

هر چند روش مورد استفاده این پژوهش عملکرد بهتری نسبت به روش‌های سنتی دارد اما متأسفانه بی‌نقص نیست که مهم‌ترین عامل آن حجم کم پیکره‌های استفاده شده است. در این بخش دو خطا را به عنوان نمونه نگاه می‌کنیم.

۵-۲-۱ خطای املائی هم‌خوان‌ها برای تبدیل فارسی به تاجیکی

مدل‌های مبتنی بر یادگیری این پژوهش برای این که املائی درست یک کلمه را یاد بگیرد

لازم است که آن را در مثال‌های متنوع در داده آموزش مشاهده کرده باشد. در مثال بخش ۲-۱ در مورد کلمه «حیات» صحبت شد اما متاسفانه در داده جمع‌آوری شده توسط ما هیچ مثالی برای کاربرد کلمه «حیات» در جمله آورده نشده است برای همین مدل همیشه این کلمه را به املائی «حیات» می‌نویسد.

(ورودی)

хаёти хона
/hæjɔ:ti χɔ:næ/

(خروجی سامانه)

حیات خانه

(خروجی درست)

حیات خانه

۲-۵-۲ عدم تشخیص واژه‌های کوتاه مناسب برای خواندن کلمه

کلمات زیادی از فارسی که در سده اخیر به زبان اضافه شدند در تاجیکی حضور ندارند. به سبب همین موضوع در دادگان جمع‌آوری شده نیز نیستند. در اینجا مدل باید تلفظ کلمه را حدس بزند که کار ساده‌ای نیست و با خطاهای زیادی همراه است. مانند کلمه «تمبر» در مثال زیر:

(ورودی)

تمبر

(خروجی سامانه)

тамаббур
/tæmæbbur/

(خروجی درست)

тамбр
/tæmbr/

۳-۵ اعتبارسنجی نتایج

در این پژوهش برای بررسی پایداری یادگیری، از اعتبارسنجی متقابل ۱۰ لایه^۱ استفاده شده

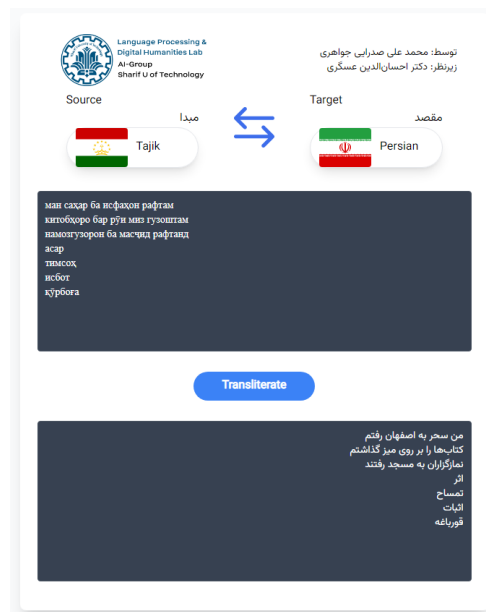
است. ترکیب داده‌های واژه‌جو و غزلیات به کمک ویژگی‌های tf-idf به ۱۰ خوشه تقسیم شده‌اند و بر روی این خوشه‌ها اعتبارسنجی متقابل صورت گرفته است. نتیجه نهایی در جدول قابل مشاهده است.

جدول ۷- نتایج اعتبارسنجی متقابل ۱۰ لایه بر روی کل پیکره

درصد MED=1	درصد MED=0	میانگین MED	
90% ± 3	75% ± 5	0.47 ± 0.17	تاجیکی به فارسی
77% ± 5	58% ± 7	1.00 ± 0.28	فارسی به تاجیکی

۵-۴ سامانه تحت وب

مدل طراحی شده در یک سامانه تحت وب از طریق آدرس fajik.parsi.ai در دسترس است. برای گرفتن خروجی بهتر پیشنهاد می‌شود که ورودی سامانه را به شکل چند رشته کوتاه در چند خط بدهید. در شکل تصویری از سامانه طراحی شده با مثال‌هایی برای تبدیل تاجیکی به فارسی قابل مشاهده است.



شکل ۳- سامانه نهایی آماده شده برای نویسه‌گردانی

۶ نتیجه‌گیری

در این پژوهش، به مساله مهم تبدیل نویسه فارسی میان گونه تاجیکی و فارسی رایج در ایران پرداختیم که مساله پراهمیت در پیوند دو ملت همسایه با پیشینه و زبان مشترک است. ما برای اولین بار از روشهای یادگیری ژرف برای حل دقیق‌تر و مبتنی بر سیاق این مساله بهره بردیم. نشان داده شد که «واحد بازگشتی دروازه» عملکرد مناسبی برای نویسه‌گردانی نسبت به روش‌های سنتی دارد. روش ما بر روی هر دوی داده‌های واژه‌جو و غزلیات سعدی که در این پژوهش بررسی شد، دقت بالاتری نیست به روشهای بهروزیان و پرشین تاجیک دارد. این بهبود در غزلیات سعدی، به فاصله ویرایشی‌ای در حدود نصف فاصله ویرایشی سایر مدلها می‌رسد.

انتظار می‌رود، گسترش دادگان به عملکرد بهتر این مدل نیز کمک کند. یکی دیگر از راهکارهای بهبود عملکرد استفاده از معماری‌های یادگیری ژرف ترنسفورمرهاست که در سالهای اخیر تحولات زیادی در زبانشناسی رایانشی و به طور کلی یادگیری ژرف ایجاد کرده‌اند که در این پژوهش از آنها استفاده نشده است. علت استفاده نشدن این معماری در این پژوهش، کم بودن حجم داده برای آموزش این معماری بوده است. برای کسب عملکرد بهتر به وسیله مدل‌ها با حجم داده کم، نیاز است که از ترندهای متنوعی استفاده شود که گسترش توامان آنها با داده‌های یادگیری گام بعدی این پژوهش خواهد بود.

منابع

برجیان، حبیب (۱۳۷۸). «ساختمان خط تاجیکی». *نامه فرهنگستان*. ۱۰۳-۱۱۶.

- Bahdanau, D., & K. Cho, & Y. Bengio (2015). "Neural machine translation by jointly learning to align and translate". *The Hilton San Diego: 3rd International Conference on Learning Representations*.
- Cho, K., et al. (2014). "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar.
- Davis, C. I. (2012). "Tajik-Farsi Persian Transliteration Using Statistical Machine Translation". *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey.
- Dovudov, G. et al. (2011). "Building a 50M Corpus of Tajik Language". *Recent Advances in Slavonic Natural Language Processing (RASLAN)*. 89-95.
- Keller, S. (2001). *To Moscow, Not Mecca: The Soviet Campaign Against Islam in Central Asia*. 1917-1941 .
- Khudoikulova, N. (2015). "Linguistic situation in Tajikistan: language use in public space". *Russian Journal of Communication*. 7(2): 164-178.

- Mahsuli, M. M., & R. Safabakhsh (2017). "English to Persian transliteration using attention-based approach in deep learning". *Iranian Conference on Electrical Engineering (ICEE)*. Tehran, Iran .
- Megerdoomian, K., & D. Parvaz (2008). "Low-Density Language Bootstrapping: the Case of Tajiki Persian". *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.
- Platon, A. (2012). *Persian Language Location Map*. 28 11. https://commons.wikimedia.org/wiki/File:Persian_Language_Location_Map.svg.
- Shao, Y., & J. Nivre (2016). "Applying Neural Networks to English-Chinese Named Entity Transliteration". *Proceedings of the Sixth Named Entity Workshop* .
- Sutskever, I, & O. Vinyals, & Q. V. Le (2014). "Sequence to Sequence Learning with Neural Networks". *Advances in neural information processing systems 27*. Montreal, Canada.
- Vaswani, A. et al. (2017). "Attention is all you need". *Advances in neural information processing systems 30*. Long Beach, California.
- Wang, P., & P. Nakov, & H. Tou Ng (2012). "Source Language Adaptation for Resource-Poor Machine Translation". *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 286–296.

استناد به این مقاله: صدرایی، محمدعلی؛ عسگری، احسان‌الدین و ربیعی، حمیدرضا (۱۴۰۱). فاجیک: مدل شبکه عصبی ژرف کدگذار-کدگشای و منابع زبانی مرتبط برای نویسه‌گردانی میان تاجیکی و فارسی. *زبان و زبان‌شناسی* ۱۸ (۳۵)، ۸۷-۱۰۳. doi: 10.30465/lsi.2023.850013:26