Razi University

Research in Western Iranian Languages and Dialects

ISSN Print: 2345-2579 ISSN Online: 2676-573X

Linguistics Society of Iran

Capabilities and Limitations of Persian Stemming in Natural Language Processing

Maryam Assadi[1], **Vida Shaghaghi**[2✉], and **Mohsen Kahani**[3]

1. Ph.D. Student in Linguistics, Department of Linguistics, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran. E-mail: assadi.marya@gmail.com
2. Corresponding Author, Professor, Department of Linguistics, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran. E-mail: vshaghaghi@hotmail.com
3. Professor, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran. E-mail: kahani@um.ac.ir

Article Info

Article type:
Research Article

Article history
Received: 30 Jan 2024
Received in revised form: 13 Mar 2024
Accepted: 16 Mar 2024
Published online: 21 Mar 2025

Keywords:
morphology,
morphological analysis,
Persian language,
stemming,
Natural Language Processing (NLP),
pre-processing,
sequence to sequence model.

ABSTRACT

This article presents a review of stemming techniques for the Persian language, encompassing structural methods, statistical approaches, and lookup tables. In addition, we explore the potential improvement of Persian stemming by drawing insights from theoretical research and experimental results on languages sharing common challenges with Persian. Through a meticulous analysis, we propose the incorporation of Byte Pair Encoding (BPE) and Sequence-to-Sequence (Seq2Seq) models into the Persian stemming framework. This recommendation is rooted in the unique strengths of these methods, tailored to address Persian's intricate morphology, extensive loanword integration, and script diversity. BPE excels in capturing prevalent morphemes and managing out-of-vocabulary terms, while Seq2Seq models show promise in decoding implicit morphological rules and accommodating linguistic idiosyncrasies. In light of Persian's status as a low-resource language in need of advanced technological resources, we put forward a novel enhancement for Persian stemming. This enhancement leverages both BPE and Seq2Seq models within a unified NLP pipeline, signifying a promising path for further research in Persian language processing. By harnessing linguistic insights, this approach has the potential to contribute significantly to bridging the digital language divide for Persian.

Cite this article: Assadi, M., Shaghaghi, V., & Kahani, M. (2025). Capabilities and limitations of Persian stemming in natural language processing. *Research in Western Iranian Languages and Dialects, 13*(1), 1-17. http//doi.org/10.22126/jlw.2024.10234.1748 (in Persian).

DOI: https://doi.org/10.22126/jlw.2024.10234.1748

Publisher: Razi University

Introduction

The realm of Natural Language Processing (NLP) stands at the frontier of innovation, continually propelled by the evolving need to unravel the complexities inherent in language. Within this dynamic landscape, the Persian language emerges as a challenging domain, demanding specialized attention due to its rich morphology, intricate script, diverse word formation processes, and extensive usage of loanwords. A pivotal facet of linguistic processing in this context is stemming, a technique aimed at reducing words to their root or base form, providing the foundation for a spectrum of NLP tasks. Stemming has garnered substantial recognition for enhancing computational efficiency, text analysis, and information retrieval. This study embarks on a meticulous exploration of stemming techniques tailored to the idiosyncrasies of Persian. By examining established structural and statistical methods, as well as dictionary-based approaches, this investigation evaluates their effectiveness and identifies their limitations in the Persian language context. Moreover, the study advocates for an innovative integration of advanced machine learning methods, specifically Sequence-to-Sequence (Seq2Seq) models and Byte Pair Encoding (BPE), to unlock the potential for significantly enhanced Persian stemming. In presenting a roadmap towards bridging the linguistic digital divide, this study aims to invigorate the future landscape of Persian language processing and contribute to the broader advancement of NLP across diverse linguistic domains.

Material and Methods

The study comprehensively evaluates various stemming methodologies for the Persian language, focusing on their applicability and efficacy. The material utilized for this research primarily includes Persian language datasets, linguistic resources, and NLP tools. The datasets encompass a diverse collection of Persian texts, covering different domains and styles of writing to ensure a broad representation of language usage. Linguistic resources consist of lexicons, grammatical rules, and linguistic databases essential for understanding Persian morphology. In addition, various NLP tools are employed, ranging from structural and statistical stemmers to advanced machine learning models like Byte Pair Encoding (BPE) and Sequence-to-Sequence (Seq2Seq) models. These tools are utilized to analyze and process the datasets, generating insights into the strengths and weaknesses of each stemming technique. The methods involve a systematic evaluation of these tools, considering factors such as accuracy, computational efficiency, and adaptability to the complexities of the Persian language. The study also proposes an innovative approach that integrates BPE and Seq2Seq models, demonstrating a potential enhancement in Persian stemming. The material and methods employed in this research aim to provide a comprehensive understanding of stemming techniques for the Persian language and pave the way for advancements in Persian language processing.

Results and Discussion

The study yields substantial insights into stemming techniques for the Persian language. Structural stemmers, relying on predefined linguistic rules, demonstrate simplicity and effectiveness but face challenges in handling exceptions, especially within verb conjugations, and in coping with the nuances of the Persian script. Statistical stemmers, driven by probabilistic models and machine learning, display adaptability to linguistic variations and prove efficient in managing irregular forms and homographic affixes inherent in Persian. Despite the need for significant training data, they outperform rule-based counterparts in managing complexities. Lookup tables, leveraging dictionaries, offer an alternative approach, accounting for irregularities and variations in word forms. However, they demand a comprehensive and precise dictionary, presenting a challenge given Persian's extensive vocabulary and the complexity introduced by loanwords. Furthermore, the rich and intricate morphology of Persian, characterized by diverse word formation processes, compounds the challenges of stemming. Several linguistic factors,

including compound verbs, diverse plural suffixes, Persian script intricacies, loanwords, and code-switching pose substantial hurdles in morphological analysis. These complexities underline the critical need for advanced approaches such as deep learning techniques, particularly Sequence-to-Sequence (Seq2Seq) models and Byte Pair Encoding (BPE), to enhance stemming accuracy and adaptability. The proposed integration of Seq2Seq and BPE models within a unified NLP pipeline presents a promising avenue for addressing the intricacies of Persian morphology and advancing the field of Persian language processing. Overall, the study underscores the importance of considering language-specific properties and employing advanced methodologies to bridge the digital language divide effectively.

Conclusion

This comprehensive exploration of stemming techniques for the Persian language sheds light on critical aspects of natural language processing in a linguistically rich and diverse context. The study underscores the significance of stemming as a fundamental pre-processing step, vital for enhancing various NLP applications such as information retrieval, text mining, sentiment analysis, and machine translation. The assessment of traditional structural and statistical stemming methods, alongside dictionary-based approaches, illuminates their strengths and limitations in addressing the intricacies of Persian morphology, script, and vocabulary. These insights guide the proposal of a synergistic integration of advanced methods, particularly Sequence-to-Sequence (Seq2Seq) models and Byte Pair Encoding (BPE), as a potential pathway to significantly enhance Persian stemming. By emphasizing the need for language-specific considerations and advanced computational methodologies, this study contributes to the ongoing efforts to bridge the digital language divide, paving the way for future research in Persian language processing and, by extension, to enriching NLP advancements in diverse linguistic landscapes.

Ethical Considerations
Not applicable

Funding
Not applicable

Conflict of interest
The authors declare no conflict of interest.

مطالعات زبان‌ها و گویش‌های غرب ایران
شاپا چاپی: ۲۵۷۹ – ۲۳۴۵ شاپا الکترونیکی: ۵۷۳x – ۲۶۷۶
دانشگاه رازی
انجمن زبان‌شناسی ایران

امکانات و کاستی‌های ستاک‌یابی زبان فارسی در پردازش زبان طبیعی

مریم اسدی[۱] | ویدا شقاقی[۲]✉ | محسن کاهانی[۳]

۱. دانشجوی دکتری زبان‌شناسی، گروه زبان‌شناسی، دانشکدهٔ ادبیات و زبان‌های خارجی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: assadi.marya@gmail.com
۲. نویسندهٔ مسئول، استاد، گروه زبان‌شناسی، دانشکدهٔ ادبیات و زبان‌های خارجی، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: vshaghaghi@hotmail.com
۳. استاد، گروه مهندسی کامپیوتر، دانشکدهٔ فنی و مهندسی، دانشگاه فردوسی، مشهد، ایران. رایانامه: kahani@um.ac.ir

اطلاعات مقاله

نوع مقاله: پژوهشی

تاریخ دریافت: ۱۴۰۲/۱۱/۱۰
تاریخ بازنگری: ۱۴۰۲/۱۲/۲۳
تاریخ پذیرش: ۱۴۰۲/۱۲/۲۶
تاریخ انتشار: ۱۴۰۴/۱/۱

کلیدواژه‌ها:
صرف،
تحلیل صرفی،
زبان فارسی،
ستاک‌یابی،
پردازش زبان طبیعی،
پیش‌پردازش،
الگوی دنباله‌به‌دنباله.

چکیده

برای سرعت‌بخشیدن و آسانی انتقال و گسترش دانش، فرایندهای ذخیره و مبادلهٔ اطلاعات خودکارسازی می‌شوند. پردازش زبان طبیعی از محورهای این خودکارسازی است. زبان‌شناسان نظری می‌توانند در پیشبرد مطالعات پردازش زبان طبیعی نقش تأثیرگذاری ایفا کنند. آن‌ها باتکیه‌بر دستاوردهای مطالعات زبان‌شناختی می‌توانند با شناسایی شباهت‌های زبان‌ها به یکدیگر ابزاری را که متخصصان پردازش زبان طبیعی برای زبانی مشخص طراحی کرده‌اند، براساس شباهت برای زبان دیگری پیشنهاد دهند. به‌عبارتی، زبان‌شناسان نظری می‌توانند به تعمیم نتایج پژوهش‌های پردازش زبان طبیعی کمک کنند. در این مقاله، رویکردهای ستاک‌یابی زبان فارسی ازمنظر زبان‌شناسی نظری مطالعه و تحلیل شده‌اند. تحلیل صرفی از مراحل پردازش زبان طبیعی است که به صورت کلمه می‌پردازد. ستاک‌یابی نیز از مراحل اصلی تحلیل صرفی است که بر کاهش صورت واژهٔ تصریف‌شده یا واژهٔ مشتق تا رسیدن به ریشه یا ستاک تمرکز دارد. ازنظر زبانی، غنای صرفی، مسائل خط فارسی و منابع محدود باعث شده‌اند ستاک‌یابی در زبان فارسی به پژوهشی دشوار تبدیل شود. پیمودن این مراحل دشوار در گرو طراحی روش‌هایی کارآمد برای مؤلفه‌های خاص زبان فارسی است. پس از تحلیل رویکردهای مختلف ستاک‌یابی همچون رویکردهای ساختاری، آماری و یادگیری عمیق برای زبان‌هایی با مسائل مشابه مسائل زبان فارسی، ستاک‌یابی با استفاده از الگوی دنباله‌به‌دنباله برای زبان فارسی پیشنهاد می‌شود.

استناد: اسدی، مریم؛ شقاقی، ویدا؛ کاهانی، محسن (۱۴۰۴). امکانات و کاستی‌های ستاک‌یابی زبان فارسی در پردازش زبان طبیعی. *مطالعات زبان‌ها و گویش‌های غرب ایران*، ۱۳(۱)، ۱-۱۷. http//doi.org/10.22126/jlw.2024.10234.1748

DOI: http//doi.org/10.22126/jlw.2024.10234.1748
ناشر: دانشگاه رازی

۱– مقدمه

انسان برای ارتقای سطح زندگی‌اش پیوسته مشغول نوآوری و تولید دانش بوده است. اکنون که دانش با سرعتی بیش از همیشه درحالِ تولید است برای سرعت‌بخشیدن و آسانی در انتقال و گسترش دانش، فرایندهای ذخیره و انتقال اطلاعات خودکارسازی[1] می‌شوند. پردازش زبان طبیعی[2] از محورهای اصلی این خودکارسازی است.

این مقاله با مطالعه و بررسی رویکردهای گوناگون ستاک‌یابی[3] بر اهمیت تحلیل صرفی[4] در کاربردهای گوناگون پردازش زبان فارسی تأکید می‌کند. تحلیل صرفی یکی از مراحل پردازش زبان طبیعی است که به صورت کلمه[5] می‌پردازد. ستاک‌یابی نیز از مراحل اصلی تحلیل صرفی است که بر کاهش صورت کلمهٔ تصریف‌شده[6] یا مشتق[7] تا رسیدن به ریشه یا ستاک تمرکز دارد.

تحلیل صرفی از مراحل نخستین و اصلی بسیاری از فرایندهای پردازش زبان طبیعی است و برای مقاصد فراوان و گوناگونی مانند هنجارسازی متن[8]، برچسب‌گذاری اجزای کلام[9] و درک معنایی[10] استفاده می‌شود. در کاربرد، ستاک‌یابی آثار مثبت گسترده و متنوعی دارد؛ به‌این‌دلیل، ارتقای فرایندها و ابزارهای ستاک‌یابی ضرورت می‌یابند. در پژوهش حاضر، چالش‌های ستاک‌یابی در پردازش رایانشی زبان فارسی به‌طور مشخص مطالعه شده‌اند؛ این چالش‌های زبان‌ویژه در سطوح مختلف ساخت‌واژه، معنی‌شناسی، نحو و رسم‌الخط وجود دارند. در پایان مقاله راهکاری رایانشی برای بهبود امکانات ستاک‌یابی موجود ارائه می‌شود. علاوه‌براین، به اهمیت نقش زبان‌شناسان نظری در مطالعات پردازش زبان طبیعی نیز اشاره می‌کنیم.

دانش نظری و امکانات عملیاتی در هر دو زمینهٔ ستاک‌یابی و مراحل دیگر تحلیل صرفی، هنوز پاسخ‌گوی چالش‌ها و محدودیت‌های موجود نیستند. شناسایی انواع واژه‌های بسیط (تجزیه‌ناپذیر) و غیربسیط (تجزیه‌پذیر) در تحلیل صرفی به‌رغم اهمیت فراوانی که در پردازش زبان طبیعی دارد، ساده انگاشته می‌شود. انتظار می‌رود که حذف وندهای اشتقاقی و تصریفی به شناسایی واحد واژگانی موسوم به ریشه/ستاک منجر شود. در تحلیل واژهٔ نویسندگی به اجزائی چون [نویس*nevis*]فعل، [ـَنده*ande*]پسوندساسم‌ساز، [–گ*g*–]پسوندمیانجی، [–ی–*i*]پسوندساسم‌ساز دست می‌یابیم؛ اما در تحلیل واژهٔ دیگری چون «پلکانی» [پله]اسم، [–ک]پسوندمیانجی، [–ان]پسوندجمع، [–ی]پسوند صفت‌ساز مشخص می‌شود که پسوند اشتقاقی پس از پسوند تصریفی به پایه افزوده شده است. مطالعهٔ دقیق‌تر گواه آن است که در این ساخت، پسوند –ان فقط پسوند تصریفی جمع نیست و در این واژه احتمالاً نقش دیگری ایفا می‌کند. چنین تحلیل‌هایی به شناخت بهتر و دقیق‌تر اجزای ساختاری واژه‌ها می‌انجامد و در ستاک‌یابی دقیق و بهینه ضرورت دارند.

همچنین برای مطالعهٔ جامع ستاک‌یابی، بررسی پیشینهٔ پژوهش‌ها اهمیت دارد؛ ازاین‌رو، به‌اختصار به نخستین گام‌های این مسیر اشاره می‌شود. نخستین اثر پژوهشی دربارهٔ ستاک‌یابی به لاوینز[11] (۱۹۶۸) تعلق دارد. در الگوریتم لاوینز که برای زبان انگلیسی طراحی شده بود به پیچیدگی‌های موجود در زبان نظیر صورت‌های بی‌قاعده و وندهای گوناگون پرداخته شده بود (لاوینز، ۱۹۶۸). مقالهٔ مهم و تأثیرگذار بعدی به سالتون و لِسک[12] (۱۹۶۵) مربوط می‌شود.

اگرچه پیشرفت‌های چشمگیر در روش‌های ستاک‌یابی در سال ۱۹۸۰ حاصل شد که همگی محصول انتشار مقاله‌ای از پورتر[13] (۱۹۸۰) بود، این الگوریتم چنان ساده، کارآمد و قابل‌اعتماد است که امروزه هم استفاده می‌شود. الگوریتم پورتر پایانه‌های تصریفی و وندهای کلمه‌ها را حذف می‌کند و کلمه را به ریشه می‌رساند. طی سال‌های بعد، این الگوریتم‌ها اصلاح و به‌روزرسانی شدند و برای هریک نسخه‌های جدیدتری منتشر شد تا پاسخ‌گوی مسائل زبان‌ویژه و پروژه‌های خاص و جدید باشند.

---

1. automation
2. Natural Language Processing (NLP)
3. stemming
4. morphological analysis
5. word form
6. inflected
7. derived
8. text normalization
9. part of speech tagging
10. semantic understanding
11. J. B. Lovins
12. G. Salton & M. E. Lesk
13. M. F. Porter

ریشه‌یاب اِسنوبال[۱]، ریشه‌یاب لَنکَستِر[۲] و ریشه‌یاب کروِتز[۳] چند نمونه از ریشه‌یاب‌های جدید با امکانات بیشتر هستند (شارما[۴]، ۲۰۱۲). پس از ظهور یادگیری ماشینی و یادگیری عمیق، روش‌های پیچیده‌تری برای ستاک‌یابی آزموده شدند؛ نظیر استفاده از شبکه‌های عصبی برای درک ساختار صرفی کلمات. ستاک‌یابی همچنان از سرفصل‌های اصلی پژوهش، سرمایه‌گذاری‌ها در پردازش زبان طبیعی و بسیاری از کاربردهای[۵] پردازش زبان طبیعی مانند ترجمهٔ ماشینی، بازیابی اطلاعات و استخراج متن است (چای[۶]، ۲۰۲۳).

۲- رویکردهای ستاک‌یابی

ستاک‌یابی پیشینه‌ای نسبتاً طولانی دارد و پیشرفت رویکردهای ستاک‌یابی حاصل کاربردهایی مانند بازیابی اطلاعات، تحلیل متن، ترجمهٔ ماشینی و شیوه‌های یادگیری ماشین بوده است. در این مقاله تکامل رویکردهای ستاک‌یابی از روش‌های ساده‌تر قاعده‌بنیاد[۷] تا الگوریتم‌های پیچیدهٔ آماری و یادگیری ماشینی بررسی می‌شوند.

۲-۱ ستاک‌یابی قاعده‌بنیاد

نخستین الگوریتم‌های ستاک‌یابی از نوع قاعده‌بنیاد بودند. طراحی این الگوریتم‌ها به این ترتیب بود که وندها تا دستیابی به ریشه حذف شوند. این الگوریتم‌ها ساده و سرراست[۸] بودند و براساس مجموعه‌قواعدی دست‌نویس برای زبانی مشخص عمل می‌کردند. اولین الگوریتم ستاک‌یابی قاعده‌بنیاد که به‌طور گسترده به آن توجه شد، ستاک‌یاب لاوینز (۱۹۶۸) بود. این ستاک‌یاب برای زبان انگلیسی طراحی شده بود و ۲۹۴ قاعدهٔ حذف پایانه و ۳۵ قاعدهٔ بازنویسی را دربرداشت. کارایی ریشه‌یاب لاوینز در مدیریت مسائل صرفی زبان انگلیسی مطلوب بود (لاوینز، ۱۹۶۸).

الگوریتم ستاک‌یابی پورتر (۱۹۸۰) که برای زبان انگلیسی طراحی شده است، در چند مرحله قواعدی بافت ـآزاد[۹] را اعمال می‌کند. این الگوریتم باوجودِ سادگی از دیگر الگوریتم‌های ستاک‌یابی قاعده‌بنیاد اولیه کارایی مطلوب‌تری داشت و به‌دلیل همین عملکردِ پذیرفتنی همچنان استفاده می‌شود (پورتر، ۱۹۸۰).

۲-۲ جدول جست‌وجو[۱۰]

هم‌زمان با ظهور روش‌های رایانشی با توانایی ذخیرهٔ اطلاعات و جست‌وجوی سریع در جدول‌های بزرگ، ستاک‌یابی براساس جدول جست‌وجو محبوبیت پیدا کرد. جدول جست‌وجو یا ستاک‌یاب واژه‌نامه‌بنیاد[۱۱] از فهرستی ازپیش‌تهیه‌شده، شامل کلمه‌ها و ریشهٔ آن‌ها استفاده می‌کند. جدول جست‌وجو به‌جای استفاده از قواعد زبانی، هر کلمه را صرفاً در فهرستی جست‌وجو و آن را با ریشه/ستاک جایگزین می‌کند. مدیریت صورت‌های بی‌قاعده‌ای که ستاک‌یاب‌های قاعده‌بنیاد را گمراه می‌کنند، برای جدول جست‌وجو آسان است. جدول جست‌وجو در ترکیب با دیگر رویکردها، ابزاری ارزشمند و کارآمد است (شریفلو و شمس‌فرد، ۲۰۰۸).

۲-۳ ستاک‌یابی آماری[۱۲]

ستاک‌یاب آماری در پیکره‌ای از واژه‌ها برای تعیین ریشه یا ستاک به تحلیل آماری می‌پردازد تا طی این فرایند الگوهای واژه‌سازی نظیر احتمال اضافه‌شدن وند مشخصی به ریشه‌ای مشخص را شناسایی کند. در ستاک‌یابی آماری مرحلهٔ یادگیری[۱۳] وجود ندارد. درعوض، براساس فراوانی[۱۴] و توزیع[۱۵] صورت‌کلمه‌های مختلف در پیکره، احتمال بروز هر ریشه محاسبهٔ آماری می‌شود (شارما، ۲۰۱۲).

1. Snowball stemmer
2. Lancaster stemmer
3. Krovetz stemmer
4. D. Sharma
5. task
6. C. Chai
7. Rule-based stemming
8. straightforward
9. context free
10. Look up table
11. Dictionary-based stemmer
12. Statistical stemming
13. learning phase
14. frequency
15. distribution

۲-۴ ستاک‌یابی مبتنی‌بر یادگیری ماشینی[1]

ستاک‌یاب‌های مبتنی‌بر یادگیری ماشینی از الگوریتم‌هایی استفاده می‌کنند که بر پیکره‌های نمونه‌های برچسب‌گذاری‌شده آموزش داده شده‌اند. این نمونه‌ها معمولاً جفت‌های صورت کلمه و ریشه هستند و الگوریتم یادگیری ماشینی تغییر(های) لازم برای کاهش صورت کلمه به ریشه را از داده‌های آموزشی می‌آموزد. ستاک‌یاب‌های مبتنی‌بر یادگیری ماشینی به‌دلیل توانایی مدل‌سازی برای الگوهای پیچیدهٔ زبانی نسبت‌به ستاک‌یاب‌های آماری دقت و انعطاف بیشتری دارند؛ اما محدودیت این ستاک‌یاب‌ها در وابستگی به مرحلهٔ یادگیری و تأمین داده‌های آموزشی برچسب‌گذاری‌شدهٔ مطلوب است (مصطفی و دیگران، ۲۰۱۷).

بااینکه ستاک‌یابی آماری و ستاک‌یابی مبتنی‌بر یادگیری ماشینی دو رویکرد مجزا و متفاوت برای شناسایی ریشهٔ کلمه هستند، گاهی با یکدیگر هم‌پوشانی پیدا می‌کنند. باید تأکید شود که اساس کار ستاک‌یابی آماری محاسبهٔ فراوانی و توزیع صورت کلمه‌هاست؛ درحالی‌که ستاک‌یابی مبتنی‌بر یادگیری ماشینی از نمونه‌های برچسب‌گذاری‌شده و الگوریتم‌های یادگیری ماشینی استفاده می‌کند.

۳- روش‌شناسی

بخشی از مطالعات صرف، تنوع‌های نظام‌مند صورت و معنی کلمه‌ها را بررسی می‌کند. دستاوردهای مطالعات صرف امکان درک چگونگی ساخت و دگرگونی کلمه‌ها در هر زبان را فراهم می‌کنند. دو زمینهٔ پژوهشی ممکن در صرف، «صرف نظری» و «صرف رایانشی» هستند. صرف نظری به مطالعه، توصیف و پیش‌بینی الگوهای صورت کلمه‌های ممکن زبان می‌پردازد. صرف رایانشی نیز کاربرد عملی این نظریه‌ها را برای پردازش زبان‌های طبیعی در روش‌های رایانشی مطالعه می‌کند.

در پژوهش‌های صرفی الگوهای حاکم بر ساختمان کلمه در زبان‌های گوناگون شناسایی می‌شوند، این مسیر پژوهشی استفاده از نظریه‌های زبان‌شناسی با رویکرد تحلیلی را دربرمی‌گیرد. پژوهشگر صرف پرسش‌هایی دربارهٔ تصریف‌ها و اشتقاق‌های اسم‌ها، فعل‌ها و قیدها و چگونگی پیوند[2] یا تغییر تکواژها در هر زبان را مطالعه می‌کند. به‌علاوه، تعامل قاعده‌های صرفی و واجی یا همان پدیده‌های صرفی-واجی[3] را نیز بررسی می‌کند.

ازسوی دیگر، صرف رایانشی به اجرای اصول و نظریه‌های صرف در روش‌های رایانشی می‌پردازد. به این معنی که برنامه‌هایی رایانشی طراحی می‌شوند که قادرند کلمه‌های زبان انسان را درک، تولید و تحلیل کنند؛ می‌توان به نمونه‌های غلط‌یاب املا[4]، برنامهٔ تبدیل متن به گفتار[5]، بازیابی اطلاعات[6] و ترجمهٔ ماشینی اشاره کرد.

صرف رایانشی همچنین تولید تحلیلگر و تولیدکنندهٔ[7] صرفی را در دستور کار خود دارد. تحلیلگر صرفی صورت تصریف‌شدهٔ کلمه را دریافت می‌کند و ریشه[8] یا بن‌واژه[9] همراه مشخصه‌های[10] صرفی نظیر زمان دستوری، حالت و جنسیت را به دست می‌دهد. تولیدکنندهٔ صرفی[11] ریشه و مجموعهٔ مشخصهٔ صرفی را دریافت می‌کند و صورت تصریف‌شده را به دست می‌دهد (کوردی، ۲۰۱۶).

طراحی، توسعه[12] و به‌کارگیری الگوهای زبان‌شناختی پیشرفته در الگوریتم‌های هوش مصنوعی و یادگیری ماشینی به برهم‌کنش[13] صرف نظری و صرف رایانشی وابسته است. صرف نظری اصول و ساختارهای صوری را فراهم می‌کند و صرف رایانشی برای کاربرد عملی این اصول، روش‌شناسی[14] ارائه می‌دهد. الگوریتم‌هایی که پژوهشگران صرف رایانشی تهیه می‌کنند می‌توانند شکاف‌ها و ابهام‌های الگوهای نظری را در کاربرد منعکس کنند و به‌این‌ترتیب، امکان اصلاح و ارتقای پژوهش‌ها فراهم شود.

---

1. Machine learning stemming
2. concatenation
3. morphophonology
4. spell checking
5. text to speech
6. information retreival
7. generator
8. root
9. lemma
10. feature
11. Morphological generator
12. develop
13. interaction
14. methodology

پردازش رایانشی بعضی از ویژگی‌های خاص نظام صرفی زبان فارسی دشوار است. نخست نظام تصریفی فعل زبان فارسی خودنمایی می‌کند. فعل‌های فارسی اطلاعات دستوری فراوانی همچون زمان دستوری، وجه، نمود و شخص را در صورت‌های تصریف‌شده بیان می‌کنند. هریک از این صورت‌های تصریف‌شده اطلاعات معنایی خاصی را همراه دارند؛ این فراوانی صورت‌ها از دشواری‌های ستاک‌یابی است. تحلیلگر صرفی باید چنان دقیق و جامع باشد که بتواند هر صورت فعل را به‌درستی تحلیل کند و بر نظام فعل فارسی نظارت داشته باشد (اَمتراپ[1] و دیگران، ۲۰۰۰).

چگونگی وندافزایی در زبان فارسی نیز بر مسائل تحلیل صرفی می‌افزاید. تحلیل کلمه‌های طولانی که از چند تکواژ و وند ساخته شده‌اند، دشوار است؛ زیرا هر وند تغییر معنایی و مقولهٔ دستوری منحصربه‌فرد خود را ایجاد می‌کند.

موانع ستاک‌یابی در زبان فارسی به پیچیدگی‌های زبانی محدود نمی‌شوند. محدودیت‌های کاربردی مانند کمبود منابع زبانی برچسب‌گذاری‌شدهٔ معتبر برای فارسی نیز به‌شدت بر عملکرد روش‌های ستاک‌یابی اثر می‌گذارد.

افزون بر آنچه گفته شد خط فارسی نیز بر مسائل این مسیر می‌افزاید. خط فارسی بسیاری از واکه‌ها را نمایش نمی‌دهد. به‌این‌دلیل، در زبان فارسی هم‌نگاره‌های[2] زیادی وجود دارد؛ مانند صورت فعل‌های تصریف‌شده. برای نمونه، در جملهٔ «آرام‌تر! بچه را کُشتی»، کلمهٔ «کُشتی» فعل است؛ اما در جملهٔ «هر روز می‌روم باشگاهِ کُشتی»، «کُشتی» اسم است. بنابراین، بافت و رابطهٔ کلمه‌ها با یکدیگر در جمله برای رفع ابهام مهم است.

اختلافِ‌نظر دربارهٔ املای واموا‌ژه‌های عربی مسئلهٔ دیگری است که باید به آن توجه شود. ازاین‌رو، می‌توان نتیجه گرفت خط فارسی از علل ایجاد ابهام در تحلیل صرفی است (اسلامی، ۱۳۸۱).

از دیگر دشواری‌های ترجمهٔ ماشینی، به‌ویژه از زبان فارسی، مسئلهٔ رمزگردانی[3] گویشوران است. رمزگردانی به معنی استفاده از دو یا چند زبان یا گویش در گفت‌وگویی مشخص است (متیوز[4]، ۲۰۰۷). رمزگردانی در کلام فارسی‌زبانان به‌خصوص در جامعهٔ ایرانیِ دورازمیهن[5] بسامد زیادی دارد. ازاین‌رو، در پردازش متن‌هایی که در فضای مجازی و شبکه‌های اجتماعی منتشر می‌شوند بایست به احتمال حضور کلمه یا ساختی از زبانی دیگر توجه کرد. همچنین در گفتار فارسی‌زبانان تحصیل‌کرده استفاده از واژه‌های انگلیسی یا فرانسه مشاهده می‌شود.

غنای صرفی، مسائل خط فارسی، چندزبانگی و منابع محدود باعث شده‌اند ستاک‌یابی در زبان فارسی به مسئله‌ای دشوار تبدل شود. پشت‌سرگذاشتن این مراحل دشوار در گروی طراحی روش‌هایی کارآمد برای مؤلفه‌های زبان‌ویژهٔ زبان فارسی است.

چالش‌های صرفی زبان فارسی در ستاک‌یابی

زبان فارسی، به‌ویژه در نظام فعلی، دارای تصریف غنی است؛ به این معنی که کلمه با تغییر نقش یا معنی ممکن است صورت‌های مختلفی داشته باشد. همچنین زبان فارسی فرایندهای واژه‌سازی متنوعی مانند اشتقاق[6]، ترکیب[7]، تکرار[8]، تبدیل یا صفر[9]، اختصارسازی[10]، کوتاه‌سازی[11]، سرواژه‌سازی[12]، آمیزش[13] و گسترش استعاری[14] را دارد (شقاقی، ۱۳۹۲). مسائل مربوط به این گوناگونی در اینجا بررسی می‌شوند.

در بحث ستاک‌یابی مرور تعریف دقیق چهار اصطلاح زبان‌شناختی بن‌واژه، ستاک، ریشه و پایه ضروری است؛ زیرا این چند اصطلاح مهم در پژوهش‌های زبان‌شناسی نظری و زبان‌شناسی رایانشی گاهی به‌جای یکدیگر به کار رفته‌اند و از دقت و شفافیت پیشینهٔ

1. J. W. Amtrup
2. homograph
3. code switch
4. P. H. Matthews
5. diaspora
6. derivation
7. compounding
8. reduplication
9. conversion
10. abbreviation
11. shortening
12. initialism
13. blending
14. overextension

پژوهشی این بحث کاسته‌اند.

بن‌واژه[1]: «واژه‌ای که مبنای تعدادی واژه و صورت کلمه است و در فرهنگ به‌عنوان مدخل درج می‌شود» (شقاقی، ۱۳۹۴: ۳۲).

ستاک[2]: پایهٔ صورت کلمهٔ تصریف‌شده یا صورتی از واژه که پس از حذف وند تصریفی باقی می‌ماند و به‌عنوان پایه برای پردازش صرفی عمل می‌کند. ستاک‌های «کتاب، خونگرم، خور، گو» پس از حذف وندهای تصریفی از صورت کلمه‌های «کتاب‌ها، خونگرم‌تر، نخوردیم، می‌گوید» باقی می‌ماند. (شقاقی، ۱۳۹۴: ۹۳)

ریشه[3]: بخشی از واژه که پس از حذف همهٔ وندها باقی می‌ماند. صورت مبنا برای تولید دیگر صورت‌ها ازطریق ایجاد تغییر در پایه یا وندافزایی (تصریفی یا اشتقاقی). پایه‌ای غیرقابل‌تجزیه که فقط از یک تکواژ تشکیل شده است. ریشه ممکن است پایهٔ آزاد یا وابسته باشد. (شقاقی، ۱۳۹۴: ۸۵)

پایه[4]: ریشه یا ستاکی که وند به آن متصل می‌شود. با اعمال فرایند صرفی بر تکواژ یا بخشی از واژه که پایه محسوب می‌شود می‌توان واژه‌های جدید (مشتق، مرکب[5] یا مکرر[6]) یا صورت کلمه‌های متفاوت را ساخت. قواعد صرفی بر پایه (واژه بسیط یا غیربسیط) عمل می‌کنند. (شقاقی، ۱۳۹۴: ۳۶)

باتوجه‌به آنچه تاکنون دربارهٔ ستاک‌یابی در پردازش زبان طبیعی مطرح شده است و تعریف‌های زبان‌شناختی که در بالا شرح داده شد، همنامی[7] دردسرسازی در زبان انگلیسی نمایان می‌شود. معادل انگلیسی ریشه[8]، به معنی صورت کاهش‌یافتهٔ کلمه در فرایند ستاک‌یابی پردازش زبان طبیعی، *stem* است؛ اما در زبان‌شناسی *stem* معادل ستاک است و بخشی از کلمه که پس از حذف همهٔ وندها باقی می‌ماند *root* معادل «ریشه» است. این مسئله ممکن است استفادهٔ میان‌رشته‌ای از منابع زبان‌شناسی و پردازش زبان طبیعی را دشوار کند.

پس از مروری بر تعریف چهار اصطلاح زبان‌شناختی بن‌واژه، ستاک، ریشه و پایه، به چالش‌های صرفی زبان فارسی در ستاک‌یابی پرداخته می‌شود:

الف) تنوع وندها[9] و واژه‌بست‌ها[10]

یکی از چالش‌های صرفی زبان فارسی در ستاک‌یابی، تنوع وندها و واژه‌بست‌هاست. زبان فارسی سرشار از وندها و واژه‌بست‌های متنوعی است که هریک می‌توانند معنی یا نقش دستوری کلمه را دگرگون کنند. گوناگونی وندها و واژه‌بست‌های زبان فارسی پیچیدگی‌های خاصی در ستاک‌یابی ایجاد می‌کنند. شیوهٔ نمایش زمان دستوری، نمود، وجه، حالت، شخص و شمار در نظام فعلی و شمار و درجه[11] برای اسم و صفت نمونه‌ای از این گوناگونی است. تسلط‌نداشتن کافی بر وندها و واژه‌بست‌های زبان ممکن است به کاهش بیش‌ازحد صورت کلمه[12] یا برعکس، ازقلم‌انداختن وند یا واژه‌بست در ستاک‌یابی[13] منجر شود.

واژه‌بست در زبان فارسی بار نقشی[14] مهمی دارد و شناخت دقیق نقش و رفتارش برای پردازش بهینهٔ فارسی ضروری است. واژه‌بست در بیان نقش‌های[15] دستوری گوناگون نقشی اساسی دارد و در بافت‌های[16] دستوری گوناگونی ظاهر می‌شود (شقاقی، ۱۳۹۴). سامان‌دادن به واژه‌بست در زبان فارسی می‌تواند به بهبود فرایند ستاک‌یابی و درنهایت، تحلیل صرفی در درک زبان طبیعی[17] و در تولید زبان طبیعی[18] بینجامد؛ به‌این‌دلیل، در پژوهش پیش‌رو چند مسئلهٔ اصلی پردازش واژه‌بست در زبان فارسی به‌اختصار معرفی می‌شوند.

1. lemma
2. stem
3. root
4. base
5. compound
6. reduplicated
7. homonymy
8. stem
9. affix
10. clitic
11. degree
12. overstemming
13. understemming
14. functional load
15. function
16. context
17. natural language understanding
18. natural language generation

گوناگونی[۱]، چندمعنایی[۲]، توالی[۳] و کثرت[۴] وندها و واژه‌بست‌ها ستاک‌یابی را پیچیده می‌کنند؛ برای نمونه، ـَم در «دفترم»، ضمیر ملکی و ـَم در «رفتم»، شناسهٔ فاعلی است.[۵] در صورت کلمه‌هایی مانند «ندیدمش»، «مورچه‌خوار»، «گل‌گلی»، «نارضایتی»، «نمی‌دیده است» و «جامدادی»، وندها و واژه‌بست‌ها هم‌زمان یا در چند مرحله به ریشه افزوده شده‌اند.

شماری از واژه‌بست‌های فارسی، ضمیری هستند که برای بیان شخص، شمار و حالت استفاده می‌شوند. این نوع واژه‌بست به بخش پایانی اسم، صفت، حرف اضافه و فعل متصل می‌شود تا مفهوم مالکیت[۶] یا مفعول صریح[۷] و مفعول غیرصریح[۸] را بیان کند. برای نمونه، در عبارت «کتابم»، «=م» پی‌بست ملکی و در «می‌بینمت»، «=ت» پی‌بست مفعول صریح است.

نقش دیگر واژه‌بست در زبان فارسی در ساخت اضافه مشاهده می‌شود. اضافهٔ اداتی[۹] با نمود آوایی /e/ پس از همخوان و /ye/ پس از واکه است که اسم را به توصیفگر[۱۰] بلافاصله بعد خود نظیر صفت بیانی، اسمی دیگر در رابطهٔ اضافه[۱۱] و بند موصولی[۱۲] متصل می‌کند. ساخت اضافه از مشخصه‌های[۱۳] برجستهٔ صرفی و نحوی زبان فارسی است (نساجیان و دیگران، ۱۳۹۸).

ب) فعل مرکب

از دیگر چالش‌های صرفی زبان فارسی در ستاک‌یابی، فعل مرکب است. در زبان فارسی فعل مرکب نیز مسائل خاصی را در فرایند ستاک‌یابی ایجاد می‌کند. فعل مرکب از ویژگی‌های بارز ساخت‌واژهٔ فارسی است. «فعل مرکب به فعلی اطلاق می‌شود که ساختمان واژهٔ آن بسیط نیست، بلکه از پیوند یک سازهٔ غیرفعلی همچون اسم، صفت، اسم مفعول، گروه حرف اضافه‌ای یا قید با یک سازهٔ فعلی تشکیل شده است» (دبیرمقدم، ۱۳۸۴: ۱۵۰). به‌باور دبیرمقدم (۱۳۸۴)، ازآنجاکه فعل مرکب حاصل پیوند دو سازهٔ (عمدتاً) مستقل است که ماحصل آن واژه‌ای مرکب است، درنتیجه، ما در اینجا با فرایندی ساخت‌واژی، مربوط به حوزهٔ واژگان[۱۴]، روبه‌رو هستیم. ازنظر او، شناسایی و جداسازی سازهٔ غیرفعلی از سازهٔ فعلی در ستاک‌یابی همواره آسان نیست. دبیرمقدم معتقد است در زبان فارسی گاهی افعال مرکب حاصل بسط استعاری هستند. ازاین‌رو، معنی این افعال با استخراج معنی هر سازه به دست نمی‌آید. در ستاک‌یابی مطلوب معنای کلی درک و حفظ می‌شود. به‌عبارت دیگر، ستاک‌یاب مطلوب و کارآمد فعل مرکب را فقط به سازه‌هایش نمی‌شکند، بلکه ملاحظات معنایی را رعایت و تمامیت معنایی را حفظ می‌کند. برای نمونه، می‌توان به «دوست‌داشتن»، «دست‌داشتن»، «هواخوردن»، «پیچ‌خوردن»، «برباددادن» و «ورآمدن» اشاره کرد (دبیرمقدم، ۱۳۸۴).

مسئلهٔ دیگری که در ستاک‌یابی افعال مرکب زبان فارسی مطرح است، ناهماهنگی‌ها در رسم‌الخط است. بسته به ترجیح نویسنده، افعال مرکب ممکن است بافاصله، بی‌فاصله یا سرهم نوشته شوند.

ج) چالش‌های اسم

علاوه‌بر «تنوع وندها و واژه‌بست‌ها» و «فعل مرکب»، اسم از چالش‌های صرفی زبان فارسی در ستاک‌یابی به شمار می‌رود. اسم در زبان فارسی مسائل ستاک‌یابی خاص خود را دارد. صورت جمع اسم فارسی باید به صورت مفرد کاهش یابد. تنوع پسوندهای جمع فارسی و استفاده از پسوندهای جمع فارسی برای وام‌واژه‌های عربی، بازتحلیل[۱۵] وام‌واژه‌های عربی در صورت جمع و استفاده از پسوندهای جمع عربی برای کلمه‌های فارسی باید در فرایند ستاک‌یابی مدیریت شوند.

---

1. variety
2. polysemy
3. order
4. multiplicity

۵. برای نمایش واژه‌بست از علامت = و برای نمایش وند از علامت – استفاده شده است.

6. possession
7. direct object
8. indirect object
9. particle
10. modifier
11. genitive relationship
12. relative clause
13. feature

۱۴. اشاره به اختلافِ‌نظر میان دو دیدگاه واژگان‌محور (lexicon-centric) و صرف‌محور (morphology-centric) ضرورت دارد. از چشم‌انداز زبان‌شناسی نظری، در دیدگاه واژگان‌محور واژه‌ها در واژگان ذخیره می‌شوند و در دیدگاه صرف‌محور حوزهٔ صرف محل ذخیرهٔ واژه است.

15. reanalysis

با درنظرگرفتنِ اِعمال سلیقه در خط فارسی، موانعی در ستاک‌یابی اسم مرکب ایجاد می‌شود. حفظ تمامیت معنایی اسم مرکب درعینِ حذف وندهای تصریفی به دقت و ظرافت نیاز دارد. ستاک‌یاب نباید کلمه‌هایی مانند «دانشجو»، «دنیادیده»، «برفی»، «گندیده»، «دست‌زنان» و «پیشتاز» را تجزیه کند. ازسوی دیگر، ستاک‌یابی اسم‌های مرکب در صورت جمع نظیر «بخت‌برگشته‌ها» و «تازه‌به‌دوران‌رسیده‌ها» نیازمند تمهیدات زبان‌ویژه است.

به واموواژه‌های موجود در زبان فارسی، به‌ویژه واموواژه‌های عربی، باید به‌طور خاص توجه شود. شناسایی صورت‌های جمع مکسر عربی، فرایندهای تصریفی وام‌گرفته‌شدهٔ زایا و چگونگی شرکت واموواژه‌ها در فرایندهای تصریفی ضروری است. برای ستاک‌یابی صورت جمع‌های مکسر می‌توان از فنونی نظیر جدول جست‌وجو استفاده کرد؛ در این صورت، ستاک‌یاب هنگام مواجهه با کلمه‌هایی مانند «کتب»، «دفاتر» و «مزارع» که نشانهٔ جمع فارسی ندارند، آن‌ها را مفرد نمی‌انگارد و با فهرستی مطلوب صورت‌های مفرد را به دست می‌دهد. این امکان برای فرایندهای وام‌گیری‌شده هم مفید است؛ مثلاً کلمهٔ فارسی «میدان» که به قیاس دارای صورت جمع «میادین» است یا پسوند جمع «-ات» که در «پیشنهادات» به کار رفته است.

همچنین برای مدیریت صورت‌های بی‌قاعده و واموواژه‌های موجود در زبان فارسی می‌توان از دستاوردهای پژوهش‌های پردازش زبان‌های طبیعی مانند زبان انگلیسی در مدیریت صورت‌های جمع بی‌قاعده استفاده کرد. نظام‌های قاعده‌بنیاد، رویکردهای واژگانی و الگوهای یادگیری ماشینی که در زبان‌شناسی رایانشی انگلیسی کارساز بوده‌اند، می‌توانند راه‌حل‌هایی مؤثر و سازگار با ویژگی‌های زبانی خاص فارسی به دست دهند. این رویکرد تطبیقی ممکن است مسیر توسعهٔ الگوریتم‌های ستاک‌یابی قدرتمند و مطمئن برای زبان فارسی را هموار کند.

۴- تحلیل داده‌ها

رویکرد یادگیری عمیق، قابلیت بالایی برای ارتقای شیوه‌های ستاک‌یابی دارد. نخستین تغییر مثبتی که این رویکرد می‌تواند ایجاد کند افزایش دقت است. اگر داده‌های آموزشی کافی موجود باشد، الگوی مبتنی‌بر یادگیری عمیق می‌تواند الگوهای پیچیدهٔ موجود در هر متن را شناسایی کند؛ درنتیجه، فرایند ستاک‌یابی را با دقت بیشتری به‌ویژه در مقایسه با روش‌های قدیمی‌تر قاعده‌بنیاد انجام دهد (چارنیاک[1]، ۲۰۱۹).

شیوه‌های مبتنی‌بر یادگیری عمیق در مواجهه با صورت‌های بی‌قاعده ابزارهای کارآمدتری هستند. در بسیاری از زبان‌ها صورت‌های بی‌قاعده‌ای وجود دارند که نمی‌توان از هیچ قاعدهٔ صرفی زبان‌ویژه‌ای برای ستاک‌یابی استفاده کرد. الگوهای یادگیری عمیق پس از استفاده از داده‌های آموزشی مطلوب در مواجهه با چنین بی‌قاعدگی‌هایی کارآمدترند. جالب این است که برخلاف الگوریتم‌های ستاک‌یابی سنتی که عموماً زبان‌ویژه هستند، الگوی ستاک‌یابی مبتنی‌بر یادگیری عمیق با آموزش خوب ممکن است بتواند با تغییرات حداقلی برای ستاک‌یابی در زبانی دیگر هم به کار آید (چارنیاک، ۲۰۱۹).

افزون‌براین، بعضی از الگوهای یادگیری عمیق مانند شبکه‌های عصبی بازگشتی[2] و مبدل‌ها[3] قادرند هنگام ستاک‌یابی به بافت[4] کلمه هم توجه کنند. به‌این‌ترتیب، ستاک‌یابی مقید به بافت[5] هم سامان می‌یابد. در گذر زمان، کلمه‌هایی به زبان افزوده می‌شوند یا کلمه‌هایی گسترش معنایی پیدا می‌کنند. الگوهای ستاک‌یابی مبتنی‌بر یادگیری عمیق بهتر از الگوریتم‌های قاعده‌بنیاد با این تغییرات سازگار می‌شوند؛ زیرا با یادگیری داده‌های به‌روز می‌توانند بر تغییرات جدید نیز نظارت کنند.

الگوهای شبکه‌های عصبی و الگوهای یادگیری عمیق درزمینهٔ یادگیری ماشینی پیشرفتی چشمگیر داشته‌اند. این الگوها قادرند الگوهای پیچیدهٔ داده‌های آموزشی را فرابگیرند. ازاین‌رو، در بسیاری از امور پردازش زبان طبیعی نظیر ستاک‌یابی می‌توانند گزینه‌ای مطلوب باشند.

۴-۱ یادگیری عمیق، شبکه‌های عصبی، ستاک‌یابی

کوشش‌های اخیر پژوهشگران در به‌کارگیری شبکه‌های عصبی و شیوه‌های یادگیری عمیق برای ستاک‌یابی نتایج امیدوارکننده‌ای

1. E. Charniak
2. Recurrent Neural Networks (RNN)
3. transformer
4. context
5. context sensitive stemming

دربرداشته‌اند. این دستاوردها می‌توانند برخی از محدودیت‌های روش‌های ستاک‌یابی آماری قاعده‌بنیاد را جبران کنند.

الگوریتم‌های ستاک‌یابی مبتنی‌بر یادگیری عمیق، یادگیری صورت‌های صرفی گوناگون یک کلمه و ریشهٔ آن را دربرمی‌گیرند. این فرایند یک کاربرد یادگیری تحت‌نظارت[1] است که به داده‌های برچسب‌گذاری‌شدهٔ فراوانی نیاز دارد. در داده‌های آموزشی، هر صورت کلمه به شکل توالی یا دنباله‌ای[2] از نویسه‌ها نشان داده می‌شود و الگوی ستاک‌یابی یاد می‌گیرد باتوجه‌به هر دادهٔ ورودی، ریشهٔ صحیح را پیش‌بینی کند. بحث دنباله در همین فرصت بررسی خواهد شد؛ اما پیش‌ازآن به چند مرحله از فرایند ستاک‌یابی با رویکرد مبتنی‌بر یادگیری عمیق با استفاده از شبکه‌های عصبی اشاره می‌شود (کِلِهِر[3]، ۲۰۱۹):

≠ **آماده‌سازی داده‌ها[4]:** نخست پیکرهٔ بزرگی از کلمه‌ها و ریشه‌های آن‌ها فراهم می‌شود. این پیکره به دو مجموعه تقسیم می‌شود: مجموعهٔ یادگیری[5] برای آموزش مدل و مجموعهٔ اعتبارسنجی[6] برای آزمودن دقت مدل.

≠ **ساخت مدل[7]:** برای کاربردهایی که با هر نوع توالی مثلاً متن سروکار دارند از شبکه‌های عصبی موسوم به شبکه‌های عصبی بازگشتی استفاده می‌شود. این شبکه‌ها قادرند الگوهای توالی نویسه‌ها[8] در کلمه را یاد بگیرند که برای درک واژه‌سازی و درپی آن، کاهش کلمه‌ها به ریشه بسیار مفید است.

≠ **آموزش[9]:** آموزش مدل ستاک‌یابی از این قرار است که صورت کلمه‌ای از مجموعهٔ آموزشی به‌طور آزمایشی به مدل نشان داده می‌شود، مدل ریشه را پیش‌بینی می‌کند؛ سپس پیش‌بینی مدل با ریشهٔ صحیح موجود در مجموعهٔ آموزشی مقایسه می‌شود. چنانچه پیش‌بینی اشتباه بوده باشد، مدل پارامترهایی را تنظیم می‌کند تا نوبت بعد بهتر پیش‌بینی کند. تا هنگامی که پیش‌بینی‌های مدل به دقیق‌ترین حدِّممکن برسد، این فرایند بارهاوبارها تکرار می‌شوند.

≠ **کاربرد[10]:** مدل پس از آموزش می‌تواند با پردازش توالی نویسه‌های هر صورت کلمهٔ جدید، ریشهٔ کلمه را پیش‌بینی کند.

۴-۲ برتری‌ها و محدودیت‌های ستاک‌یابی مبتنی‌بر یادگیری عمیق

شبکه‌های عصبی و الگوهای یادگیری عمیق در بسیاری از حوزه‌های پژوهشی همچون پردازش زبان طبیعی انقلابی به پا کرده‌اند. این مدل‌های قدرتمند که می‌توانند از پیکره‌های بزرگ داده‌ها الگوهای پیچیده‌ای بیاموزند، نویدبخش بهبود الگوریتم‌های ستاک‌یابی هستند.

از امتیازهای ستاک‌یابی مبتنی‌بر یادگیری عمیق، توانایی یادگیری تمام‌خودکار قواعد پیچیده و استثناهای صرفی از داده‌های آموزشی است. باوجودِ این امتیاز ضرورتی ندارد قواعد به‌صورت دست‌نویس یا ابتکاری[11] تهیه شوند. به‌علاوه، در صورت دسترسی به داده‌های آموزشی کافی، این مدل‌ها قادرند قواعدی را که آموخته‌اند به کلمه‌های جدید تعمیم دهند؛ درنتیجه، با صورت‌های استثنا سازگار شوند.

بااین‌حال، به‌رغم امتیازهایی که ذکر شد، رویکردهای ستاک‌یابی مبتنی‌بر یادگیری عمیق، محدودیت‌هایی هم دارند. آموزش این مدل‌ها به حجم زیادی از داده‌های آموزشی برچسب‌گذاری‌شده نیاز دارد که ممکن است همیشه، به‌ویژه برای زبان‌های با منابع کم[12]، در دسترس نباشد. به‌علاوه، این مدل‌ها ازنظر محاسباتی فشرده‌اند[13] و در مقایسه با دیگر رویکردهای ستاک‌یابی به قدرت پردازش و حافظهٔ بیشتری نیاز دارند (کلارک[14] و دیگران، ۲۰۲۲). ازنظر امکانات کاربردی، برای طراحی این ستاک‌یاب، الگوی دنباله‌به‌دنباله[15]

1. supervised learning task
2. sequence
3. J. D. Kelleher
4. data preparation
5. training set
6. validation set
7. model building
8. character
9. training
10. application
11. heuristic
12. low resource language
13. computationally intensive
14. J. H. Clark
15. sequence to sequence, seq2seq

«دنباله‌به‌دنباله» برابرنهادی برای sequence to sequence در بسیاری از متن‌های علوم رایانه است؛ به‌این‌دلیل، در این مقاله نیز از همین برابرنهاد استفاده شده است؛ اما لازم است ذکر شود که ازنظر معنایی، کلمه‌های «توالی» و «زنجیره» برابرنهادهای مناسب‌تری هستند.

پیشنهاد می‌شود.

۴-۳ ستاک‌یابی با الگوی دنباله‌به‌دنباله

دنباله‌به‌دنباله الگویی در یادگیری ماشینی به‌ویژه در پردازش زبان طبیعی است. این الگو عموماً در جاهایی کاربرد دارد که دنبالهٔ خروجی[1] به‌طور مستقیم به دنبالهٔ ورودی[2] مرتبط نیست؛ نظیر ترجمهٔ ماشینی، تشخیص گفتار و خلاصه‌سازی متن (فؤاد[3] و دیگران، ۲۰۲۰). برای نمونه، یک مترجم ماشینی فرضی میان زبان انگلیسی و زبان فرانسه در نظر گرفته می‌شود، هدف ترجمهٔ جمله‌ای از زبان انگلیسی به زبان فرانسه است. جملهٔ انگلیسی دنباله/توالی ورودی و جملهٔ فرانسه دنبالهٔ/توالی خروجی است. این دو جمله، هر دو، دنباله‌هایی از کلمات هستند؛ اما کلمه‌به‌کلمه نظیر یکدیگر نیستند. ازاین‌رو، الگویی مطلوب است که بتواند کل دنباله/توالی جملهٔ انگلیسی را بفهمد؛ سپس یک دنباله/توالی کامل برای جملهٔ فرانسه تولید کند. الگوی دنباله‌به‌دنباله مناسب است؛ زیرا از دو بخش رمزگذار[4] و رمزگشا[5] تشکیل شده است.

≠ **رمزگذار:** رمزگذار دنبالهٔ ورودی ـجملهٔ انگلیسیـ را دریافت می‌کند و آن را به بازنمایی از اطلاعات جمله یا به‌اصطلاح بُردار بافت[6] تبدیل می‌کند.

≠ **رمزگشا:** رمزگشا بُردار بافت را دریافت می‌کند و دنبالهٔ خروجی ـجمله فرانسهـ را تولید می‌کند.

الگوی دنباله‌به‌دنباله قدرتمند و منعطف است و در مطالعات پردازش زبان طبیعی نقشی مهم دارد. یادگیری این الگو ازطریق جفت دنباله‌هایی است که با یکدیگر جفت[7] شده‌اند. این الگو یاد می‌گیرد از دنبالهٔ ورودی به دنبالهٔ خروجی نگاشت[8] کند (فؤاد و دیگران، ۲۰۲۰).

۴-۴ نگاشت در الگوی دنباله‌به‌دنباله

نگاشت به یادگیری رابطهٔ میان یک ورودی[9] و یک خروجی[10] دلالت دارد؛ به‌بیان ساده‌تر، نگاشت همان ایجاد یا یافتن رابطهٔ دو چیز است. برای نمونه، در فهرست زیر افراد و میوه‌های موردعلاقه‌شان مشاهده می‌شود:

مریم> سیب

سپهر> پرتقال

مهدی> زردآلو

در فهرست یادشده، نگاشتی میان نام‌ها و میوه‌های موردعلاقه ایجاد شده است. در این الگو، اگر نام «مریم» داده شود، با استفاده از این نگاشت می‌توان گفت میوهٔ موردعلاقهٔ مریم «سیب» است. به‌این‌ترتیب، الگوی یادگیری ماشینی برای ستاک‌یابی یاد می‌گیرد، دنباله‌های نویسه‌ها را به ریشهٔ درست نگاشت کند.

ساختن ستاک‌یاب با الگوی دنباله‌به‌دنباله برای زبان فارسی مستلزم گردآوری مجموعهٔ بزرگی از کلمه‌های فارسی همراه با ریشه‌های صحیح آن‌ها است. پس از آموزش، این الگوی ستاک‌یاب به‌احتمال می‌تواند ریشهٔ یک کلمهٔ جدید را با پردازش نویسه‌ها یا زیرکلمه‌های آن پیش‌بینی کند. شبکه‌های عصبی و یادگیری عمیق فرصت‌های نوظهوری برای پیشرفت ستاک‌یابی زبان‌های طبیعی ارائه می‌دهند؛ اما ازنظر فنّی، تحقق ظرفیت الگوی دنباله‌به‌دنباله در گروی به‌کارگیری رمزگذاری جفت بایت[11] است. رمزگذاری جفت بایت و الگوی دنباله‌به‌دنباله از فنون پردازش زبان طبیعی هستند که می‌توانند یکدیگر را تکمیل کنند. رمزگذاری جفت بایت، متن را به واحدهای زیرکلمه واحدسازی می‌کند و الگوی دنباله‌به‌دنباله، دنباله‌ای را به دنباله‌ای دیگر نگاشت می‌کند.

1. output sequence
2. input sequence
3. M. M. Fouad
4. encoder
5. decoder
6. context vector
7. match
8. mapping
9. input
10. output
11. Byte Pair Encoding (BPE)

رمزگذاری جفت بایت، داده‌های ورودی را آماده‌سازی می‌کند؛ این کارکرد در کاربردها با گستردگی و تنوع واژگان نظیر ترجمهٔ ماشینی مفید است. ازسوی دیگر، الگوی دنباله‌به‌دنباله مانند مدل‌هایی که در ترجمهٔ ماشینی، خلاصه‌سازی متن و بازشناسی گفتار به کار می‌روند، این ورودی‌های واحدسازی‌شده را دریافت و دنباله‌های خروجی تولید می‌کند.

۵- بحث و نتیجه‌گیری

بهره‌گیری از هوش مصنوعی و یادگیری عمیق برای پژوهش‌های آینده ستاک‌یابی پیشنهاد می‌شود. افق امکانات پژوهش‌های هوش مصنوعی گسترده و امیدوارکننده است و یکی از مسیرهای پژوهشی ستاک‌یابی برای آینده، پژوهش در فنون هوش مصنوعی توصیف‌پذیر[1] است. ازآنجاکه الگوهای هوش مصنوعی درحال پیچیده‌ترشدن هستند، شفاف‌سازی فرایندهای تصمیم‌گیری این الگوها ضرورت دارد.

همچنین امکانات یادگیری عمیق می‌توانند توانایی تشخیص الگوهای زبانی و ظرایف بافتی را برای مدل‌های زبانی فراهم آورند؛ این توانایی به تولید ابزارهای ستاک‌یابی حساس‌تر کمک می‌کند. توسعهٔ الگوهای ترکیبی، مشخصاً ترکیب روش‌های قاعده‌بنیاد با قابلیت‌های یادگیری عمیق پیشنهاد می‌شود.

نقش زبان‌شناسان نظری در این مسیر پژوهشی دیجیتال، فنّاورانه و بسیار نوین، مهم و تعیین‌کننده است. در این فرصت به مشارکت‌های زبان‌شناسان نظری طی پژوهش‌های فنّاورانهٔ پردازش زبان طبیعی اشاره شد. دانش نظری زبان مسیر را برای نوآوری‌هایی هموار می‌کند که با طبیعت پویای زبان در دوران دیجیتال هماهنگ است. ماهیت بین‌رشته‌ای پژوهش‌های هوش مصنوعی همکاری میان زبان‌شناسان، دانشمندان رایانه و اخلاق‌شناسان را می‌طلبد تا پژوهش‌ها و پیشرفت‌ها در فنّاوری با ملاحظات اخلاقی و نیازهای اجتماعی همسو باشند.

زبان‌شناسان با تسلط بر جنبه‌های گوناگون زبان فارسی می‌توانند در پیشرفت و توسعهٔ الگوریتم‌های ستاک‌یابی پیچیده‌تر و دقیق‌تر برای این زبان نقش تأثیرگذاری داشته باشند. اکنون در فهرستی خلاصهٔ فرصت‌های مشارکت زبان‌شناسان در ستاک‌یابی مرور می‌شود:

≠ **تحلیل صرفی:** زبان‌شناسان می‌توانند برای قواعد صرفی گوناگون زبان فارسی نظیر «الگوهای وندافزایی»، «تغییرات واکه و همخوان» و «صورت‌های بی‌قاعده»، تحلیل و مستندات دقیق ارائه دهند. این اطلاعات می‌توانند مبنای طراحی الگوریتم‌های ستاک‌یابی دقیق باشند. به‌علاوه، صورت‌بندی جامع و صورت‌گرایانهٔ زبان‌شناسان از قاعده‌ها و الگوهای زبان می‌تواند در هر چهار رویکرد ستاک‌یابی که ذکر شد به کار آید؛ در ستاک‌یابی قاعده‌بنیاد برای تهیهٔ قاعده‌های دست‌نویس، در جدول جست‌وجو برای ارزیابی صحت صورت کلمه‌های فهرست‌شده، در ستاک‌یابی آماری برای ارزیابی برونداد و در ستاک‌یابی مبتنی‌بر یادگیری ماشینی برای برچسب‌گذاری داده‌های آموزشی.

≠ **توسعهٔ قواعد:** مطالعات زبان‌شناسان دربارهٔ ویژگی‌های زبان‌ویژهٔ رده‌شناختی و صرفی فارسی برای فراهم‌آوردن قاعده‌های ستاک‌یابی جامع‌ومانع ضروری است. «شناسایی»، «تعیین معنی‌ها» و «چگونگی تعامل وندها با ریشه» در حوزهٔ تخصصی زبان‌شناسان قرار دارد. ازاین‌رو، زبان‌شناسان می‌توانند برای حذف وندها قاعده‌هایی صورت‌بندی کنند یا قاعده‌های موجود را بهبود بخشند. زبان‌شناسان با تسلط بر قواعد واجی و صرفی‌واجی نظیر هماهنگی واکه‌ای، گونه‌های گویشی و دگرگونی‌های تاریخی کلمه‌ها می‌توانند در طراحی الگوریتم‌های هرچه دقیق‌تر و موجزتر مؤثر باشند.

≠ **رسیدگی به استثناها:** زبان فارسی مانند دیگر زبان‌ها استثناهایی هم دارد. فارغ از نوع رویکرد ستاک‌یابی، سامان‌دادن به استثناهای زبان فارسی از دشواری‌های برجستهٔ ستاک‌یابی است. در مرحلهٔ نخست، زبان‌شناسان می‌توانند با مطالعهٔ دقیق صرف فارسی فهرستی از استثناها تهیه کنند. سپس می‌توانند تعامل هر الگوریتم ستاک‌یابی با استثناها را ارزیابی و استثناها را در گروه‌هایی نظیر صورت‌های جمع، صورت‌های مکمل[2] و تصریف فعل طبقه‌بندی کنند و برای سامان‌دادن به این صورت‌ها تمهیداتی بیندیشند.

≠ **منابع زبانی:** زبان‌شناسان می‌توانند در تهیه و تدوین منابع زبانی لازم برای آموزش و ارزیابی الگوریتم‌های ستاک‌یابی نقشی مهم ایفا کنند. آن‌ها قادرند پیکره‌هایی گرد هم آورند که تا جایی که ممکن است همهٔ ویژگی‌های زبان فارسی را دربرداشته باشند؛ پیکره‌های

1. explainable AI
2. suppletive forms

برچسب‌گذاری‌شده‌ای که اطلاعات دربارهٔ صورت‌های کلمه‌ها و ریشه‌ها را شامل می‌شود. این منابع زبانی برای آموزش ستاک‌یاب‌های مبتنی‌بر یادگیری ماشینی ضروری هستند؛ زیرا آموزش بهتر پیش‌بینی‌های دقیق‌تری به‌همراه می‌آورد.

≠ **ارزیابی و تحلیل الگوریتم‌های ستاک‌یابی:** زبان‌شناسان در ارزیابی دقت و کارایی الگوریتم‌های ستاک‌یابی می‌توانند نقش مؤثری داشته باشند. تسلط زبان‌شناسان بر جنبه‌های مختلف زبان می‌تواند در شناسایی خطاها و کاستی‌های عملکرد الگوریتم‌های ستاک‌یابی کارآمد باشد. زبان‌شناسان پس از تجزیه‌وتحلیل نظام‌مند پیش‌بینی‌های ستاک‌یاب و مقایسهٔ ریشه‌های خروجی با ریشه‌های مطلوب زبان فارسی نقاط ضعف الگوریتم را شناسایی می‌کنند.

≠ **پردازش زبان فارسی محاوره‌ای:** برای پردازش زبان فارسی محاوره‌ای، دانش واج‌شناسی و آواشناسی ضرورت بنیادی دارد. زبان‌شناسان با درکی عمیق از سطح آوایی زبان فارسی و تسلط به قاعده‌های واجی می‌توانند صورت‌بندی‌هایی برای این سطح زبان فارسی تهیه کنند و اطلاعات به‌دست‌آمده را در الگوهای زبانی بگنجانند. این اطلاعات از مرحلهٔ پیش‌پردازش تا مرحلهٔ ارزیابیِ خروجیِ الگوی زبانی قابل‌استفاده و بسیار مفید هستند؛ به‌طور مشخص، به‌کارگیری این اطلاعات به‌کمک رمزگذاری جفت بایت در ستاک‌یابی فارسی محاوره‌ای برای نظارت بر فرایندها و دگرگونی‌های آوایی محاوره‌ای توصیه می‌شود.

منابع

اسلامی، محرم (۱۳۸۱). *دشواری‌های پردازش رایانه‌ای خط فارسی*. نشر دانش، ۱۹ (۳)، ۲۸ـ۳۲.

دبیرمقدم، محمد (۱۳۸۴). *پژوهش‌های زبان‌شناختی فارسی (مجموعه‌مقالات)*. تهران: مرکز نشر دانشگاهی.

شقاقی، ویدا ([۱۳۸۶] ۱۳۹۲). *مبانی صرف*. تهران: سمت.

شقاقی، ویدا (۱۳۹۴). *فرهنگ توصیفی صرف*. تهران: علمی.

نساجیان، مینو؛ شجاعی، راضیه؛ بحرانی، محمد (۱۳۹۸). ساخت اضافه در زبان فارسی: بررسی پیکره‌بنیاد. *پژوهش‌های زبانی*، ۱۰ (۱)، ۱۶۱ـ۱۸۲.

References

Amtrup, J. W., Mansoori Rad, H., Magerdoomian, K., & Zajac, R. (2000). *Persian-English machine translation: An overview of the Shiraz project*. Las Cruces, NM: Computing Research Laboratory, New Mexico State University.

Chai, C. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509-553. http://dx.doi.org/10.1017/S1351324922000213

Charniak, E. (2019). *Introduction to deep learning*. Cambridge, MA: MIT Press.

Clark, J. H., Garrette, D., Turc, I., & Wieting, J. (2022). Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, *10*, 73–91.

Eslami, M. (2003). Difficulties of computational processing of Persian script. *Nashr-e Danesh*, *19*(3), 28-32. (In Persian)

Fouad, M. M., Mahany, A., & Katib, I. (2020, Sep.). Masdar: A novel sequence-to-sequence deep learning model for Arabic stemming. *Paper presented at the Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys)*. London. Retrieved from: https://link.springer.com/chapter/10.1007/978-3-030-29513-4_26

Kelleher, J. D. (2019). *Deep learning*. Cambridge, MA: MIT Press.

Kurdi, M. Z. (2016). *Natural language processing and computational linguistics: Speech, morphology and syntax*. Hoboken, NJ=John Wiley & Sons.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11*(1-2), 22–31.

Matthews, P. H. (2007). *Oxford concise dictionary of linguistics*. Oxford: Oxford University Press.

Mustafa, M., Eldeen, A. S., Bani-Ahmad, S., & Elfaki, A. O. (2017). A comparative survey on Arabic stemming: Approaches and challenges. *Intelligent Information Management*, *9*(2), 39-78. http://dx.doi.org/10.4236/iim.2017.92003

Nassajian, M., Shojaie, R., & Bahrani, M. (2020). The corpus-based study of ezafe construction in Persian. *Pajooheshhaye Zabani*, *10*(1), 161-182. https://doi.org/10.22059/jolr.2019.72007 (In Persian).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *14*(3), 130–137.

Salton, G., & Lesk, M. E. (1965). The SMART automatic document retrieval systems—an illustration. Commun. ACM 8, 6 (June 1965), 391–398. https://doi.org/10.1145/364955.364990

Shaghaghi, V. (2014). *Principles of morphology*. Tehran: Samt. (In Persian)

Shaghaghi, V. (2016). *Dictionary of morphology*. Tehran: Elmi. (In Persian)

Sharifloo, A. A., & Shamsfard, M. (2008). A bottom up approach to Persian stemming. Paper presented at the *Proceedings of the Third International Joint Conference on Natural Language Processing*, Hyderabad. Retrieved from: https://www.semanticscholar.org/paper/A-Bottom-Up-approach-to-Persian-Stemming-Sharifloo-Shamsfard/8a327139e9795eec55ddbb23e55f87cb1d2c8c36

Sharma, D. (2012). Stemming algorithms: A comparative study and their analysis. *International Journal of Applied Information Systems*, *4*(3), 7–12. http://dx.doi.org/10.5120/ijais12-450655