شناسایی موجودیت‌های اسمی با استفاده از یادگیری عمیق و رویکرد تقویتی

مهدی نقوی [۱]*، محمدرضا حسنی‌آهنگر [۲]، علی امیری‌جزه [۳]

استادیار دانشکده و پژوهشکدهٔ رایانه، شبکه و ارتباطات، دانشگاه جامع امام حسین (ع)، تهران، ایران [۱]*

استادتمام دانشکده و پژوهشکدهٔ هوش مصنوعی و علوم شناختی، دانشگاه جامع امام حسین (ع)، تهران، ایران [۲]

کارشناس ارشد دانشکده و پژوهشکدهٔ رایانه، شبکه و ارتباطات، دانشگاه جامع امام حسین (ع)، تهران، ایران [۳]

چکیده

مطالعات اخیر نشان می‌دهد که شناسایی موجودیت‌های اسمی به یکی از موضوعات پرطرفدار در حوزهٔ پردازش زبان طبیعی تبدیل شده‌است؛ این افزایش توجه بیشتر ناشی از رشد سریع روش‌های نوین در هوش مصنوعی، به‌ویژه شبکه‌های عصبی است. ظهور شبکه‌های عصبی عمیق به‌دلیل توانایی‌های برجستهٔ آن‌ها در حل مسائل پیچیدهٔ پردازش زبان طبیعی، فرصت‌هایی نوین برای بهبود عملکرد سامانه‌های شناسایی موجودیت‌های اسمی فراهم کرده‌است؛ یکی از عوامل کلیدی در افزایش دقت این سامانه‌ها، بهره‌گیری از ویژگی‌های پیشرفته‌ای است که تا پیش از ظهور یادگیری عمیق قابل دسترسی نبودند، از جمله استفاده از مدل‌های زبانی که توانایی درک معنای متن‌های بزرگ را دارند. در این مقاله، یک روش پیشنهادی مبتنی بر مدل زبانی معنایی در زبان فارسی و شبکهٔ عصبی عمیق طراحی شده‌است که با استفاده از منطق تکرار، مشکل وابستگی به داده‌های زیاد را کاهش می‌دهد. این روش روی مجموعه‌دادهٔ CoNLL 2003 انگلیسی و دو مجموعه‌دادهٔ فارسی آرمان و پیما ارزیابی شده‌است؛ نتایج ارزیابی به ترتیب مقادیر ۹۵.۳، ۹۶.۳۲ و ۹۴.۷۲ را برای معیار F نشان داده که حاکی از بهبود عملکرد نسبت به روش‌های پیشین است.

واژگان کلیدی: موجودیت اسمی، شناسایی موجودیت اسمی، مدل زبانی، یادگیری عمیق، رویکرد تقویتی.

Identify the Named Entities Using Deep Learning and Reinforcement Approach

Mehdi Naghavi*[1] ,Mohmmad Reza Hassani Ahangar[2] and Ali Amiri jezeh[3]

Assistant Professor, Faculty of Computer, Network, and Communications,
Imam Hossein University, Tehran, Iran[*1]

Professor, Faculty of Artificial Intelligence and Cognitive Sciences, Imam Hossein University, Tehran, Iran[2]

Master Student, Faculty of Computer, Network, and Communications, Imam Hossein University, Tehran, Iran[3]

Abstract

Named Entity Recognition (NER) has emerged as a critical and highly applicable task in the field of Natural Language Processing (NLP). Its significance stems from its essential role in numerous NLP applications, such as machine translation, question answering, text summarization, and information extraction. Recent studies highlight the substantial impact of advancements in Artificial Intelligence (AI), particularly Deep Neural Networks (DNNs), on improving the performance of NER systems.

Deep Neural Networks, with their ability to learn complex patterns and extract rich features, have opened new horizons in addressing NLP challenges. These methods leverage advanced language models like BERT and GPT to enable deeper comprehension of linguistic structures and semantic relationships. One of their prominent capabilities is to capture long-term dependencies in complex sentences while reducing the reliance on manually engineered features.

* **Corresponding author** *نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۴۰۰/۲/۲۵ • تاریخ پذیرش: ۱۴۰۲/۱۲/۶ • تاریخ انتشار: ۱۴۰۳/۱۰/۲۸ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.3.97]

This research introduces a novel hybrid approach for Named Entity Recognition in both Persian and English languages, based on deep neural networks and semantic language models. To address the dependency on large datasets, the proposed method employs an iterative logic mechanism that facilitates effective learning with limited data. The proposed system was evaluated on three datasets: The CoNLL 2003 dataset for English, Two Persian datasets, Arman and Peyma.

Experimental results demonstrate that the proposed method achieves F1-scores of 95.3, 96.32, and 94.72 on the CoNLL, Arman, and Peyma datasets, respectively. These scores reflect significant improvements over previous methods.

The findings of this study suggest that combining advanced language models with deep neural networks can significantly enhance the accuracy and efficiency of NER systems. These achievements pave the way for developing effective NLP tools for low-resource languages, particularly Persian, and enable the application of this technology in both industrial and research contexts.

Keywords: **Named Entity, Named Entity Recognition, Language Model, Deep Learning.**

۱-مقدمه

اهمیت شناسایی موجودیت‌های اسمی[1] از متون، بر متخصصان حوزهٔ پردازش زبان طبیعی[2] پوشیده نیست. اهمیت موجودیت‌های اسمی در مسائل حوزهٔ پردازش متن مانند بازیابی اطلاعات[3][۱, ۲]، خلاصه‌سازی خودکار متن[4][۳]، سامانه‌های پرسش‌وپاسخ[5][۴]، ترجمهٔ ماشینی[6][۵] و ساخت پایگاه دانش[7][۶] و همچنین اهمیت استفاده از شبکه‌های عصبی[8] و یادگیری عمیق[9] در شناسایی موجودیت‌های اسمی در سال‌های اخیر، پژوهش‌گران را بر آن داشته‌است تا با استفاده از روش‌های نوین مبتنی بر شبکه‌های عصبی و یادگیری عمیق، دقت[10]، صحت[11] و پوشش[12] نتایج سامانه‌های NER را افزایش دهند، تا جایی که بر اساس ادعای تارنمای Semantic Scholar [۷] تنها در سال ۲۰۲۴ بیش از ۶۰۰ مقاله با عنوان شناسایی موجودیت‌های اسمی چاپ شده است؛ مقالاتی اعم از ادبیات مروری در این حوزه [۱۰,۹,۸]، شناسایی موجودیت‌های اسمی در یک حوزهٔ خاص [۱۱] و یا یک زبان [۱۳,۱۲] که همگی اهمیت این حوزه را می‌رساند.

موجودیت‌های اسمی را می‌توان از روش‌های مختلفی تشخیص داد؛ یکی از این روش‌ها، روش‌های مبتنی بر فهرست است که در آن با داشتن فهرستی از موجودیت‌های اسمی و نوع آن‌ها، از داده‌های متنی، عین موجودیت‌ها را می‌توان استخراج کرد. از ضعف‌های این روش می‌توان به عدم پوشش تمام موجودیت‌های یک متن اشاره کرد؛ همچنین ساخت فهرستی از موجودیت‌ها کاری پرهزینه و زمان‌بر است. روش دیگر، روش مبتنی بر قاعده است که در آن قواعد توسط خبره استخراج می‌شوند، قواعد استخراج‌شده تمامی موجودیت‌های متن را پوشش نمی‌دهند، همچنین در این روش وجود خطا نیز محتمل است. روش بعدی، روش آماری است که به‌صورت خودکار از متون برچسب خوردهٔ مدل یادگیری استخراج می‌شود. به‌دلیل پایین‌بودن نرخ حضور موجودیت‌های اسمی نسبت به کل واژه‌های موجود در متن، به داده‌های آموزشی بسیاری برای یادگیری نیاز داریم؛ همچنین نمی‌توان تمام موجودیت‌های یک متن را استخراج کرد. روش بعدی روش ترکیبی است که در آن از ترکیب نتایج دیگر روش‌ها استفاده می‌شود که این خود باعث بروز یک خطای تصمیم‌گیری در انتخاب و یا عدم انتخاب موجودیت خواهد شد؛ برای مثال یک موجودیت مانند ایران در روش مبتنی بر فهرست، اسم مکان تشخیص داده شده‌است، اما در روش مبتنی بر قاعده به‌عنوان اسم شخص تشخیص داده می‌شود، حال تشخیص اینکه نوع موجودیتِ ایران چیست، خود یک مسئله است. روش پیشنهادی ما، یک روش ترکیبی و یادگیرندهٔ تقویتی مبتنی بر یادگیری عمیق است که می‌تواند بر چالش‌هایی مانند اندک‌بودن داده‌های آموزشی برای یادگیری، وابسته‌بودن به مهارت زبان‌شناختی در پردازش متون و وابستگی به زبان و مواردی از این قبیل فائق آید.

در این مقاله ابتدا مروری بر کارهای پیشین و مشابه صورت گرفته‌است، سپس روش رایج استخراج موجودیت اسمی بیان می‌شود و در ادامه راه‌حل پیشنهادی با بیان جزئیات به همراه نتایج کار ارائه خواهد شد.

۲- مروری بر کارهای پیشین

برای نخستین‌بار اصطلاح موجودیت اسمی توسط گریشمن و ساندهیم[۱۴] در ششمین کنفرانس درک پیام

[1] Named Entity Recognition (NER)
2 Natural Language Processing (NLP)
3 Information Retrieval (IR)
4 Automatic Summarization
5 Question Answering
6 Machine Translation
7 Knowledge Base (KB)
8 Neural Network
9 Deep Learning
10 Accuracy
11 Precision
12 Recall

[DOI: 10.61186/jsdp.21.3.97]

در سال ۱۹۹۶ به‌عنوان یک امر مهم در شناسایی اسامی سازمان‌ها ، اشخاص و مکان در متن و همچنین ارز و مقادیر زمان و درصد استفاده شد. بعد از این کنفرانس بود که علاقه‌مندی پژوهش‌گران در حوزهٔ پردازش زبان طبیعی به موضوع شناسایی موجودیت‌های اسمی زیاد شد و در رویدادهای مختلف علمی اعم از CoNLL03[۱۵]، ACE[۱۶]، IREX[۱۷] و TREC Entity Track[12] تلاش‌های زیادی برای این موضوع انجام پذیرفت. در همین سال‌ها بود که کارهای بسیاری در این زمینه انجام گرفت و به‌منظور بهبود دقت، این سامانه‌ها با یکدیگر مقایسه و بررسی می‌شدند.

یکی از این بررسی‌ها در سال ۲۰۰۷ توسط نادائو و سکین[۱] منتشر شد. در این بررسی انواع سامانه‌های NER تحت نظارت، نیمه‌نظارتی و بدون نظارت به همراه ویژگی‌های مشترک و معیارهای ارزیابی مورد مطالعه و بررسی قرار گرفتند. این پژوهش مروری کلی از روند تکنیک از قوانین ماشینی به سمت یادگیری ماشین ارائه می‌دهد.

مارزو و همکاران [۱۹] در سال ۲۰۱۳ خلاصه کارهای NER از منظر اشتباهات، چالش‌ها و فرصت‌ها را ارائه کردند، سپس پاتاور و پوتی [۲۰] یک بررسی کوتاه در سال ۲۰۱۵ ارائه دادند. در چند بررسی کوتاه اخیر به ترتیب سیر تکامل موجودیت‌های اسمی از سال ۲۰۰۰ تا ۲۰۲۴ [۹, ۲۱]، شناسایی موجودیت‌های اسمی با یادگیری عمیق [۲۲, ۲۳] و شناسایی موجودیت‌های اسمی چندوجهی [۲۵, ۲۴] مورد بررسی قرار گرفته‌اند.

زالی و فیروزبخت [۲۶] به دنبال مکان‌یابی موجودیت‌های اسمی در متن و دسته‌بندی آن‌ها به رده‌هایی از پیش تعیین‌شده، یک سامانهٔ شناسایی و طبقه‌بندی موجودیت اسمی را بر پایهٔ شبکهٔ عصبی بر روی اسناد فارسی با کمترین وابستگی به دامنه ارائه کردند. آن‌ها در روش پیشنهادی با استفاده از ویژگی‌های بازنمایی برداری به روش word2vec، برچسب اجزای کلام و طول واژه و استفاده از پیکرهٔ ویکی‌پدیای فارسی با بهره‌گیری از شبکهٔ عصبی عمیق مدلی برای شناسایی موجودیت‌های اسمی ارائه کرده‌اند. در این پژوهش معیار F به‌دست‌آمده ۷۹.۲۴ درصد بیان شده‌است.

شهشهانی و همکاران[۲۷] با ساخت پیکره‌ای برچسب‌خورده از موجودیت‌های اسمی با نام پیما به طراحی سامانه‌ای ترکیبی (آماری و قاعده محور) با استفاده از مدل CRF و سامانهٔ مبتنی بر یادگیری عمیق از نوع LSTM[1] پرداخته‌اند. در روش پیشنهادی برای آموزش سامانهٔ CRF از ویژگی‌های برچسب عبارت اسمی[2]، برچسب اجزای کلام، بن‌واژه[3]، ویژگی‌های مبتنی بر دیکشنری به‌صورت تطابق جزئی (حضور و عدم حضور واژه در فهرست موجودیت‌های اسمی بدون ابهام) و n-gramهای سطح نویسه واژه تا سقف شش نویسه در ساخت مدل استفاده شده‌است. در طراحی سامانهٔ مبتنی بر یادگیری عمیق از ویژگی‌های بن‌واژهٔ کلمه، برچسب اجزای کلام و برچسب عبارت اسمی در ساخت مدل بهره گرفته شده‌است. در این پژوهش برای روش CRF از ابزار تشخیص موجودیت‌های اسمی[4]، گروه پردازش زبان طبیعی دانشگاه استنفورد[5]، برای استخراج ویژگی‌های بن‌واژه، برچسب عبارت اسمی و برچسب ادات سخن از ابزار Persianp[6] و برای اجرای شبکهٔ عصبی از ابزار openNMT[7] استفاده شده‌است. در این پژوهش نتایج آزمایش‌ها نشان می‌دهد که سامانهٔ ترکیبی مبتنی بر مدل CRF معیار F ۸۴ درصد را به‌دست می‌دهد.

محسنی و طبی‌فخر [۲۸] در روشی با استفاده از BERT و تحلیل ریخت‌شناسی کلمات، روشی جدید در شناسایی موجودیت‌های اسمی ارائه کرده‌اند. آن‌ها مدل BERT را روی حجم زیادی از متون فارسی آموزش داده‌اند تا بازنمایی دقیقی از توکن‌ها به‌دست آورند، سپس از LSTM دوطرفه بر روی بازنمایی برداری به‌دست‌آمده برای برچسب‌گذاری توکن‌ها استفاده کرده‌اند. در این روش با تجزیه‌وتحلیل متن از منظر ریخت‌شناسی، ویژگی‌های بن‌واژهٔ واژه و ضمایر واژگان را استخراج کرده‌اند و سپس مدل را بر روی این ویژگی‌ها آموزش داده‌اند. نتایج حاصل از این پژوهش نشان می‌دهد که معیار F به‌دست‌آمده ۸۵.۴ درصد است.

طاهر و همکاران [۲۹] با استفاده از یک شبکهٔ عصبی دو طرفهٔ عمیق آموزش‌دیده‌شده با روش BERT، اقدام به ارائهٔ روشی برای شناسایی موجودیت‌های اسمی از متون فارسی کرده‌اند؛ آن‌ها در این پژوهش بعد از بازنمایی برداری توکن‌ها با روش BERT با استفاده از یک لایه شبکهٔ به‌طور کامل متصل[8] با بهره‌گیری از روش CRF[9]

[1] Long Short-Term Memory (LSTM)
2 NP-Chunk
3 Lemma
4https://nlp.stanford.edu/software/CRF-NER.html (Accessed: 2020)
5 https://nlp.stanford.edu (Accessed: 2020)
6 https://persianp.ir/toolbox.html (Accessed: 2020)
7 http://opennmt.net (Accessed: 2020)
[8] Fully Connected Layer

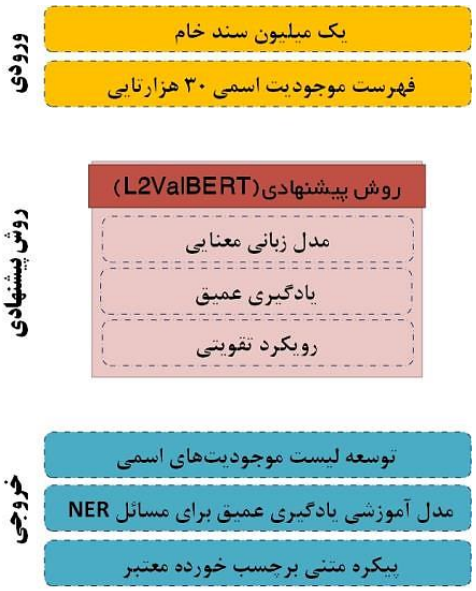
[DOI: 10.61186/jsdp.21.3.97]

مدل نهایی را آموزش داده‌اند. نتایج ارزیابی بر اساس معیار F در سطح واژه و عبارت به ترتیب ۸۴ درصد و ۸۷.۹ درصد به‌دست آمده‌است.

بکائی و محمودی [۳۰] از پیکرهٔ برچسب‌خوردهٔ آرمان برای آموزش مدلی استفاده کرده‌اند که در آن ویژگی‌ها با استفاده از شبکه‌های عصبی بازگشتی و کانولوشن استخراج و سپس بردارهای ویژگی استخراج‌شده به یک شبکهٔ به‌طور کامل متصل ساده منتقل می‌شوند؛ درنهایت یک لایه CRF برای یافتن بهترین برچسب دنباله برای توالی واژهٔ ورودی استفاده می‌شود. در این پژوهش ادعا شده‌است که نتایج بهبود قابل توجهی را نسبت به نتایج موجود بر روی پیکره نشان می‌دهند و معیار F به‌دست‌آمده در سطح واژه و عبارت به‌ترتیب ۸۱.۵۰ درصد و ۷۶.۷۹ درصد بیان‌شده‌است. آن‌ها در این پژوهش یک تحلیل خطا بر روی خروجی بهترین مدل انجام داده‌اند و نشان داده‌اند که پیکرهٔ زبانی آرمان خطاهای زیادی دارد و باید برای بهبود عملکرد، مورد بازبینی قرارگیرد.

۳- راه‌حل پیشنهادی

در روش پیشنهادی که آن را با نام L2ValBERT می‌شناسیم، ما برای استخراج موجودیت‌های اسمی، یک روش ترکیبی با رویکرد تقویتی، مبتنی بر یادگیری عمیق را ارائه کرده‌ایم که می‌تواند بر چالش‌هایی مانند اندک‌بودن داده‌های آموزشی برای یادگیری، وابسته‌بودن به مهارت زبان‌شناختی در پردازش متون و وابستگی به زبان و مواردی از این قبیل چیره شود.



(شکل-۱): مدل مفهومی روش پیشنهادی
(Figure-1): Conceptual model of the proposed method

۳-۱-مدل مفهومی روش پیشنهادی

در روش پیشنهادی ورودی که شامل یک میلیون سند خام است به همراه فهرست سی‌هزارتایی موجودیت‌های اسمی به روش پیشنهادی داده می‌شود و درنهایت خروجی‌های مدنظر حاصل می‌شود. در (شکل-۱) به‌خوبی این مراحل نشان داده شده‌است.

۳-۲-معماری روش پیشنهادی

روش L2ValBERT دارای دو مرحلهٔ اصلی است. هرکدام از این مرحله‌ها عبارت‌اند از:

۱. مرحلهٔ جمع‌آوری و آماده‌سازی مجموعه‌دادهٔ مناسب
- جمع‌آوری داده از تارنماهای خبری
- برچسب‌گذاری مبتنی بر فهرست موجودیت اسمی در اسناد
- انتخاب و گزینش اسناد برچسب‌گذاری‌شده

۲. مرحلهٔ یادگیری عمیق و رویکرد تقویتی
- صحت‌سنجی اسناد برچسب‌گذاری‌شده
- تبدیل پیکرهٔ متنی به پیکرهٔ برداری با مدل زبانی
- یادگیری و ساخت مدل با شبکهٔ عصبی عمیق
- برچسب‌گذاری مبتنی بر مدل

در روش L2ValBERT با استفاده از یادگیری عمیق و یک رویکرد تقویتی، مدلی را ارائه خواهیم کرد که علاوه بر شناسایی انواع موجودیت‌های اسمی، خروجی‌های مطلوب‌تری ارائه می‌کند.

همان‌طور که در (شکل-۲) و در مرحلهٔ نخست مشاهده می‌شود، ابتدا مجموعه‌دادهٔ مناسب با روش حاشیه‌نویسی مبتنی بر فهرست اولیهٔ موجودیت‌ها تولید می‌شود و با برخی معیارهای ارزیابی تعدادی از اسناد برای فرایند یادگیری در روش پیشنهادی انتخاب می‌شوند.

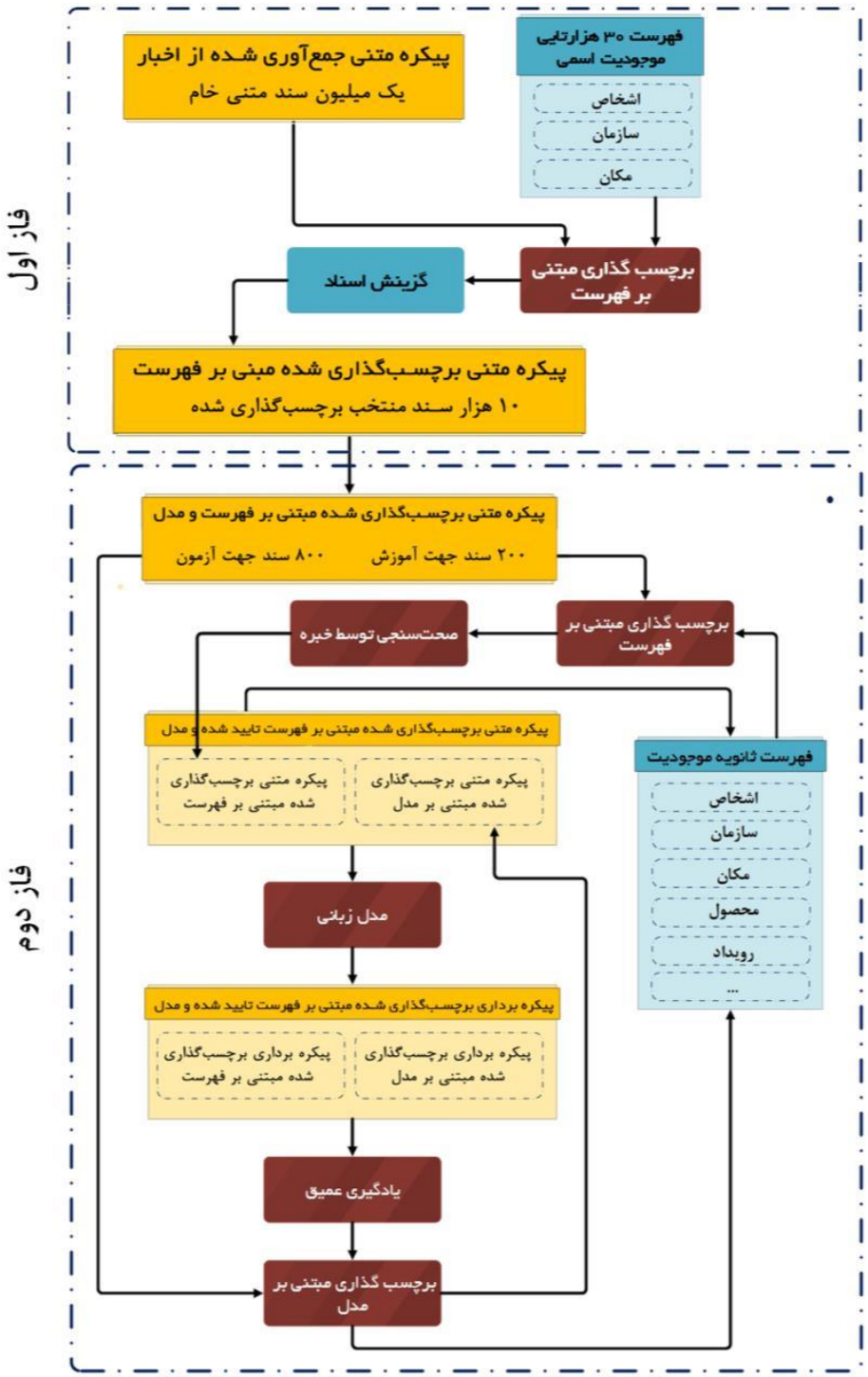
در (شکل-۲) و در مرحلهٔ دوم از روش پیشنهادی، بخشی از مجموعه‌دادهٔ تولیدشده انتخاب می‌شود و با نسبت‌های بیست و هشتاد درصد، داده‌ها به ترتیب به داده‌های آموزش و آزمون تقسیم می‌شوند. داده‌های آموزش با استفاده از فهرست ثانویهٔ موجودیت‌ها حاشیه‌نویسی می‌شود و موجودیت‌های جدیدی که در این داده‌ها برچسب نخورده‌اند برچسب می‌خورند؛ در این مرحله اولویت در برچسب‌گذاری یک موجودیت با فهرست ثانویه است. در مرحلهٔ نخست از اجرای روش پیشنهادی فهرست ثانویه فاقد موجودیت است؛ بعد از اینکه اسناد حاشیه‌نویسی شدند، توسط خبره صحت‌سنجی صورت می‌گیرد و برچسب‌ها اصلاح می‌شوند، بعد از تأیید خبره، موجودیت‌های اسمی جدید شناسایی و به فهرست ثانویهٔ

9 Conditional Random Field

[DOI: 10.61186/jsdp.21.3.97]

موجودیت‌ها افزوده می‌شوند؛ سپس اسنادی که مورد تأیید قرار گرفته‌اند با استفاده از مدل زبانی معنایی تولیدشده، به مجموعه‌دادهٔ برداری برچسب‌خورده تبدیل می‌شوند. این مجموعه‌دادهٔ برداری به یک شبکهٔ عصبی عمیق داده می‌شود، بعد از اتمام فرایند آموزش شبکه، با استفاده از مدل آموزشی، داده‌هایی را که در ابتدای روش پیشنهادی با نسبت هشتاد درصد به‌عنوان دادهٔ آزمون انتخاب شده بودند، حاشیه‌نویسی می‌کنیم.

موجودیت‌های اسمی شناسایی‌شده در این اسناد به فهرست ثانویهٔ موجودیت‌ها اضافه می‌شوند و مجموعه‌دادهٔ برچسب‌خورده برای مراحل بعدی در فرایند آموزش مورداستفاده قرار می‌گیرد. این فرایند آن‌قدر تکرار می‌شود تا مدل ضعیفی که در ابتدا توسط الگوریتم یادگیری عمیق بر روی مجموعه‌داده‌های اندک تولید شده‌است را به یک مدل قوی و مورد اعتماد تبدیل کند.



(شکل-۲): معماری روش پیشنهادی

(Figure-2): Architecture of the proposed method

[DOI: 10.61186/jsdp.21.3.97]

۳-۳-جمع‌آوری مجموعه‌داده

در روش L2ValBERT با استفاده از خزش تارنماهای خبری، حدود یک میلیون سند فارسی در دسته‌های مهم خبری اعم از سیاسی، اقتصادی، فرهنگی، ورزشی و... جمع‌آوری شد. علت انتخاب متون خبری برای روش پیشنهادی را می‌توان موارد زیر بیان کرد:

- محاوره‌ای‌نبودن متن اسناد
- کمتربودن خطاهای املایی در متون
- رعایت ساختاربندی متن اعم از جمله و پاراگراف
- پوشش دامنه و حوزه‌های مختلف

۳-۴-ساخت فهرست موجودیت‌ها

در روش L2ValBERT از فهرست سی‌هزارتایی موجودیت‌های اسمی که شامل سه نوع مکان، اشخاص و سازمان است، استفاده شده‌است. این فهرست توسط کارشناس از پیکرهٔ متنی ویکی‌پدیا استخراج و توسط خبره بهبودیافته و تکمیل شده‌است. همان‌طور که در (جدول-۱) مشاهده می‌شود، جزئیات این فهرست بیان‌شده‌است.

(جدول-۱): مشخصات فهرست موجودیت‌های اسمی
(Table-1): Specifies the list of Named Entities

نوع موجودیت	تعداد
سازمان	۱۹۵۸
مکان	۳۰۴۴
اشخاص	۲۴۸۸۲

۳-۵-پیش‌پردازش اسناد

در روش پیشنهادی هیچ‌گونه پیش‌پردازشی روی متون انجام نمی‌گیرد، تنها به‌منظور صحت داده‌های متنی خزش‌شده، برخی روش‌های پیش‌پردازشی به‌منظور صحت وجود متن فارسی در خزش صفحات اخبار اعمال می‌شود. گفتنی است به‌منظور افزایش دقت در امر برچسب‌گذاری، قبل از تطبیق یک موجودیت از فهرست با کلمات و عبارات یک متن، آن دسته از کاراکترهایی که از لحاظ یونیکد متفاوت‌اند، اما نوشتار یکسانی دارند نرمال‌سازی می‌شوند، برای مثال «آ» به «آ» تبدیل می‌شود؛ همچنین کاراکتر «ی» با توجه به تنوعی که در یونیکد دارد، همگی به یک یونیکد واحد تبدیل می‌شوند.

۳-۶-آماده‌سازی اسناد

با توجه به این‌که برای ساخت مدل زبانی باید از بخش‌های معنادار متن استفاده و متن را باید بخش‌بندی کنیم؛ کوچک‌ترین واحد را جملات در نظر می‌گیریم. در زبان فارسی به‌منظور تشخیص پایان جملات از نشانه‌های (.!؟:؛) استفاده می‌شود. در روش پیشنهادی پایان جملات نقطه (.) درنظر گرفته شده‌است.

اسنادی که جمع‌آوری شده‌اند باید ماشینی برچسب‌گذاری شوند؛ برای برچسب‌گذاری اسناد از فهرست سی‌هزارتایی موجودیت‌های اسمی استفاده می‌کنیم. در روش پیشنهادی ابتدا در هر سند موجودیت‌های اسمی فهرست با واژه یا عبارات سند تطبیق داده می‌شود و یک مرحله برچسب‌گذاری بر روی سند صورت می‌پذیرد. در مرحلهٔ بعد با استفاده از روش برچسب‌گذاری Inside/Outside بار دیگر سند برچسب‌گذاری می‌شود؛ درواقع در مرحلهٔ نخست، واژگان و عباراتی که موجودیت اسمی‌اند، شناسایی و برچسب نوع موجودیت آن‌ها مشخص می‌شود. در مرحلهٔ دوم با یک مرحله شکست، ابتدا و انتهای موجودیت با برچسب IOB مشخص می‌شود. در (شکل-۳) به‌خوبی مراحل کار نشان داده شده‌است. درعمل مرحلهٔ نخست و دوم هم‌زمان صورت می‌پذیرد.

در ادامه اسناد بر اساس تعداد موجودیت برچسب‌گذاری‌شده در هر سند، مرتب‌سازی می‌شوند. برای روش پیشنهادی ده‌هزار سند درنظر گرفته شده‌است. برای تهیهٔ این اسناد این‌گونه عمل می‌شود که هشتاد درصد اسناد از بیست درصد ابتدایی لیست (یعنی از بین دویست هزار سند) انتخاب و بیست درصد مابقی به‌صورت تصادفی از هشتاد درصد باقی‌مانده (یعنی از بین هشتصدهزار سند) انتخاب می‌شوند.

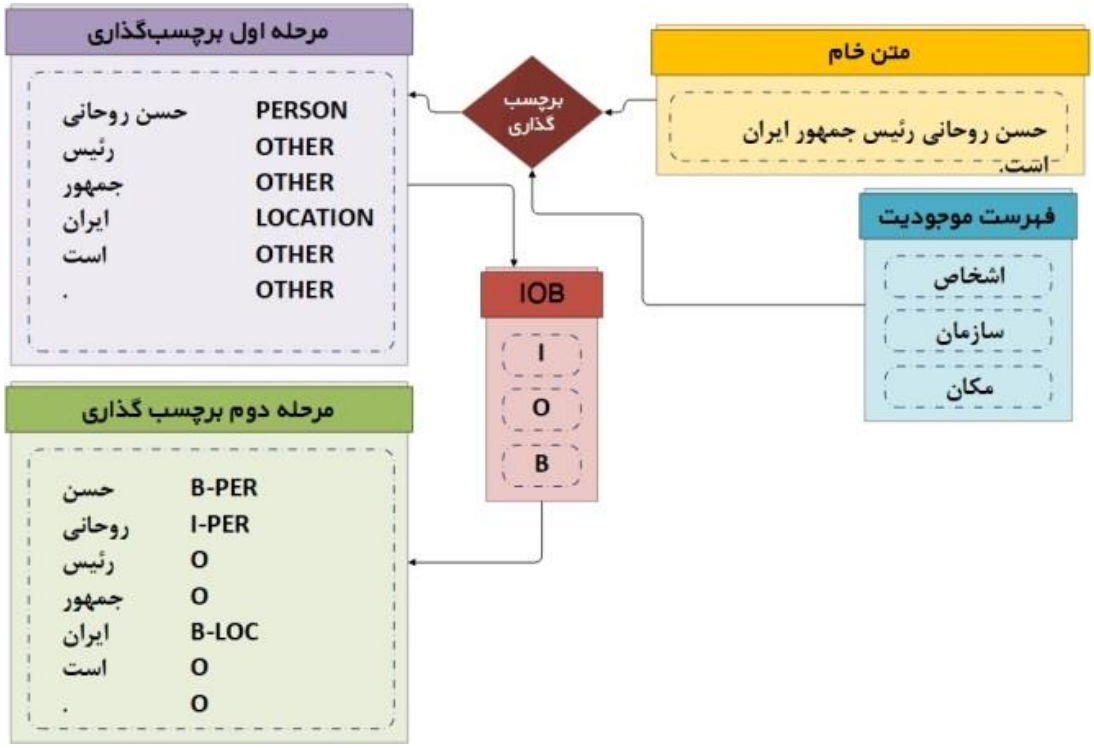
۳-۷-مدل زبانی

معماری مورداستفاده برای ساخت مدل زبانی فارسی بر اساس معماری مدل پایه BERT [۳۱] پیاده‌سازی شده‌است. مؤلفه‌های اصلی این معماری دوازده لایهٔ اصلی، ۷۶۸ لایهٔ پنهان و دوازده تابع Attention است. در ساخت مدل زبانی از یک میلیون سند جمع‌آوری‌شده از تارنماهای خبری استفاده شده‌است.

۳-۸-مراحل اجرا

همان‌گونه که در بخش ۲-۳-معماری روش پیشنهادی ذکر شد، مراحل اجرای کار شامل دو مرحلهٔ اصلی است. گفتنی است که مرحلهٔ نخست یک مرتبه و مرحلهٔ دوم در هر مرحله از اجرای روش پیشنهادی تکرار می‌شود، با احتساب پیکرهٔ متنی در دسترس ده بار الگوریتم پیشنهادی اجرا خواهد شد.

[DOI: 10.61186/jsdp.21.3.97]



(شکل-۳): نحوهٔ برچسب‌گذاری اسناد با استفاده از فهرست موجودیت‌ها

(Figure-3): How to label documents using the list of entities

۳-۸-۱-صحت‌سنجی دادهٔ آموزش

در بخش ۶-۳-آماده‌سازی اسناد نحوهٔ انتخاب ده هزار سند از پیکرهٔ متنی جمع‌آوری‌شده بیان شده‌است. این مجموعه‌دادهٔ متنی گزینش‌شده به دو مجموعه‌داده دو هزار و هشت هزارتایی تقسیم می‌شود؛ درواقع هشتاد درصد اسناد که شامل هشت هزار سند می‌شود، برای فرایند آزمون و بیست درصد باقی‌مانده یعنی دو هزار سند، برای فرایند آموزش در روش پیشنهادی مورد استفاده قرار می‌گیرند. با احتساب این‌که روش پیشنهادی ده مرتبه تکرار می‌شود، داده‌های آموزش و آزمون به ده قسمت تقسیم می‌شوند و در هر مرحله بخشی از این دو مجموعه‌داده در روش پیشنهادی مورداستفاده قرار می‌گیرند. در هر مرحله از اجرای روش پیشنهادی به‌صورت تصادفی، دویست سند برای فرایند آموزش مدل از بین دو هزار سند و هشتصد سند برای آزمون روش پیشنهادی از بین هشت هزار سند انتخاب می‌شوند.

داده‌های آموزش با استفاده از یک فهرست ثانویه از موجودیت‌های اسمی، مجدد حاشیه‌نویسی می‌شود و موجودیت‌هایی که در این مجموعه‌داده برچسب نخورده‌اند، برچسب‌گذاری خواهند شد. در این روش درصورتی‌که یک موجودیت در متن از قبل برچسب مشخصی خورده باشد، اولویت برچسب‌گذاری با فهرست ثانویه خواهد بود. فهرست ثانویه در طول مراحل اجرای روش پیشنهادی بعد از صحت‌سنجی داده‌ها توسط خبره و بعد از استفاده از مدل در برچسب‌گذاری دادهٔ آزمون تکمیل می‌شود.

تمام داده‌های آزمون حاشیه‌نویسی‌شده که در این مرحله تولید شده‌اند، توسط خبره صحت‌سنجی و اصلاح می‌شوند.

بیان این نکته حائز اهمیت است که در ابتدای کار اسناد با سه نوع موجودیت مکان، اشخاص و سازمان حاشیه‌نویسی می‌شوند، اما در ادامهٔ کار موجودیت‌هایی از نوع رویداد، تاریخ، زمان، تارنما، رایانامه و مبالغ مالی نیز توسط خبره برچسب‌گذاری می‌شوند که در فرایند آموزش شبکهٔ عصبی مورد استفاده قرار می‌گیرند.

آنچه بعد از تأیید خبره به‌دست می‌آید یک پیکرهٔ برچسب‌گذاری‌شدهٔ مورد تأیید و فهرست جدیدی از موجودیت‌های اسمی است. موجودیت‌های جدید شناسایی‌شده و موجودیت‌های تأییدشده بعد از تطبیق با فهرست ثانویهٔ موجودیت‌ها و حذف موجودیت‌های تکراری، به فهرست ثانویهٔ موجودیت‌ها اضافه می‌شوند. تمام اسنادی که در این مرحله توسط خبره مورد تأیید قرار می‌گیرند در یک پیکرهٔ متنی حاشیه‌نویسی‌شده مبتنی بر فهرست ذخیره می‌شوند. تعداد اسناد این پیکره

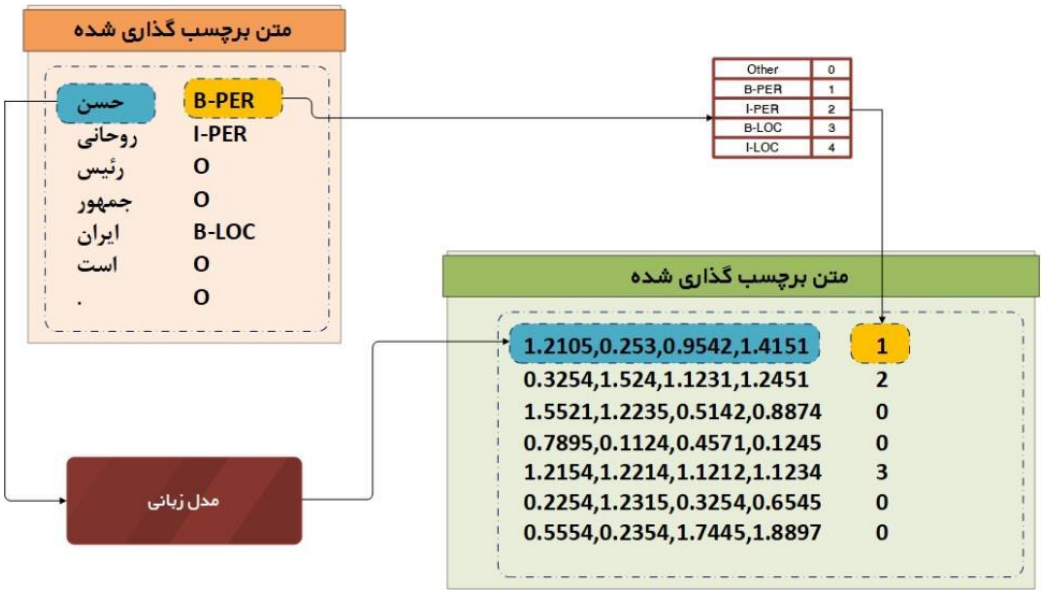
[DOI: 10.61186/jsdp.21.3.97]

به‌مرور زمان در هر مرحله از اجرای روش پیشنهادی افزایش می‌یابد.

۳-۸-۲ ساخت پیکرهٔ برداری با مدل زبانی

استفاده از BERT برای یک کار خاص، کم و بیش ساده است، مدل BERT می‌تواند برای طیف گسترده‌ای از کارهای زبانی مورد استفاده قرار گیرد، درحالی‌که فقط یک لایهٔ کوچک به مدل اصلی اضافه می‌کند. برای استفاده از مدل آموزش‌دیدهٔ BERT در سامانهٔ شناسایی موجودیت اسمی، کافی است متن برچسب‌خورده به بردار برچسب‌خورده تبدیل شود.

در این روش هر توکن از پیکرهٔ متنی دارای یک برچسب IOB است، تنها کافی است که به جای هر توکن، بردار مربوط به آن را از مدل زبانی BERT احصا و جایگزین کرد؛ همچنین برچسب‌ها به‌صورت یک شماره که هر شماره به یک برچسب اختصاص دارد، نشان داده می‌شوند. در (شکل-۴) به‌خوبی نحوهٔ کار نشان داده شده‌است.



(شکل-۴): ساخت پیکرهٔ برداری برچسب‌خوردهٔ موجودیت اسمی با مدل زبان

(Figure-4): Manufacturing vector data set of a tagged named entity with language model

۳-۸-۳-یادگیری و ساخت مدل

در حل مسائل پردازش متن اعم از پرسش‌وپاسخ و ترجمهٔ ماشینی با استفاده از شبکه‌های عصبی، ترتیب ورودی‌ها و خروجی‌های مسئله بسیار مهم است. برای حل چنین مسائلی به روش‌های مبتنی بر توالی نیاز است. ورودی به‌صورت توالی از کلمات در نظر گرفته می‌شوند؛ ازآنجاکه شبکه‌های عصبی پیش‌خور برای حل چنین مسائلی مناسب نیستند و توالی و ترتیب را درنظر نمی‌گیرند، شبکه‌های عصبی بازگشتی ارائه شدند. شبکه‌های عصبی بازگشتی توالی از واژگان را می‌گیرند و قابلیت ارائهٔ توالی از خروجی‌ها را دارند، اما نکتهٔ قابل‌توجه در این است که این شبکه‌ها یک توالی با طول ثابت را می‌گیرند و یک توالی خروجی با همان طول را تولید می‌کنند. درصورتی‌که برای مسائل حوزهٔ پردازش زبان طبیعی، طول ورودی‌ها متغیر است. شبکهٔ عصبی LSTM برای حل چنین مسائلی به وجود آمدند. با استفاده از این شبکه می‌توان مسائلی را که در آن‌ها توالی مهم و طول ورودی متغیر است، حل کرد[۳۲, ۳۳].

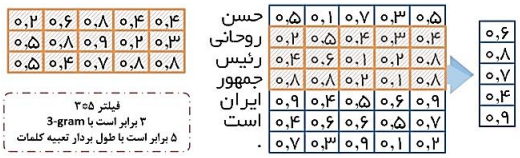
تعیین ویژگی‌های مجموعه‌دادهٔ آموزشی با استفاده از الگوریتم CNN صورت پذیرفته‌است. در روش L2ValBERT ویژگی‌ها همان n-gramها هستند که با اعمال پالایه بر روی بردار ورودی به‌وسیلهٔ الگوریتم CNN استخراج می‌شوند. با بررسی‌هایی که در فهرست موجودیت‌های اسمیِ تولیدشده صورت پذیرفت، طول موجودیت‌ها حداقل یک و به‌ندرت پنج بودند. با توجه به این نکته اندازهٔ پالایه‌ای که بر روی بردار ورودی اعمال می‌شود در اندازهٔ چهار دنظر گرفته می‌شود.

در (شکل-۵) تعیین ویژگی‌ها با پالایه‌ای به ابعاد ۵*۳ بر روی بردار ورودی نشان داده شده‌است.

در روش L2ValBERT برای بهینه‌سازی پارامترهای تولید مدل از روش توقف زودهنگام[1] استفاده شده‌است. در (شکل-۶) مقادیر بهینه برای این پارامترها نشان داده شده‌است.

[1] Early Stopping

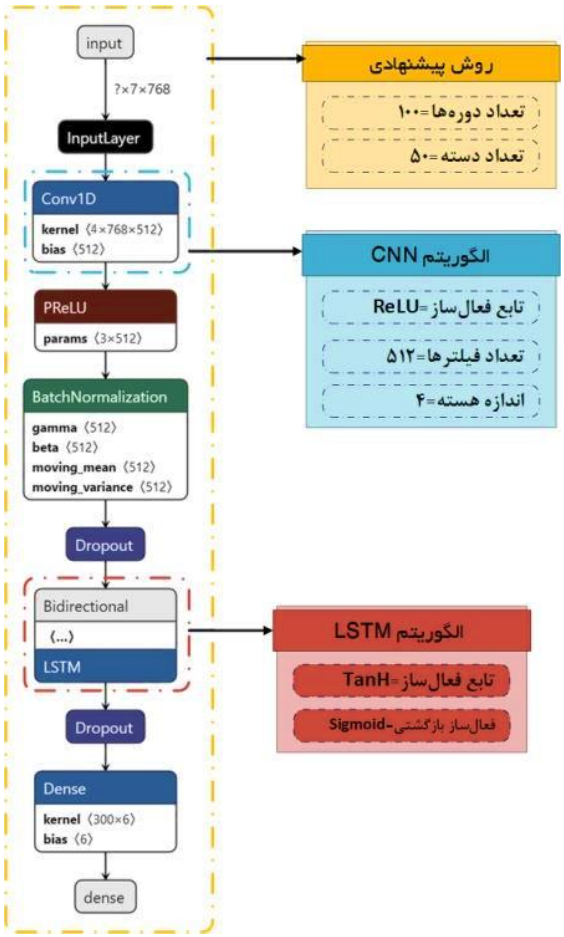
[DOI: 10.61186/jsdp.21.3.97]



(شکل-۵): تعیین ویژگی‌ها در روش پیشنهادی با استفاده از الگوریتم CNN

(Figure-5): Determining the feature in the proposed method using CNN algorithm

همان‌طور که در (شکل-۶) مشاهده می‌شود، در روش L2ValBERT از یک لایه کانولوشن به منظور کاهش ابعاد استفاده شده‌است. این لایه یک هستهٔ کانولوشن ایجاد می‌کند که به لایهٔ ورودی با یک بُعد متصل می‌شود و یک تنسور[1] خروجی تولید می‌کند. خروجی به یک تابع فعال‌سازی ReLU داده می‌شود، در مرحلهٔ بعد با استفاده از نرمال‌سازی دسته‌ای[2] [۳۴] با رویکرد تسریع در آموزش شبکهٔ عصبی از یک لایه BatchNormalization استفاده شده‌است.



(شکل-۶): معماری شبکهٔ عصبی عمیق پیشنهادی برای تشخیص موجودیت اسمی

(Figure-6): Proposed deep neural network architecture for named entity recognition

[1] Tensor
2 BatchNormalization

آموزش شبکه‌های عصبی عمیق با این واقعیت که توزیع ورودی‌های هر لایه در طول آموزش تغییر می‌کند، با تغییر پارامترهای لایه‌های قبلی، پیچیده است. این امر به‌دلیل نیاز به میزان یادگیری پایین‌تر و تنظیم دقیق پارامتر، آموزش را کند و آموزش مدل‌های غیرخطی را بسیار دشوار می‌کند. نرمال‌سازی دسته‌ای این امکان را می‌دهد تا از میزان یادگیری بسیار بالاتری استفاده کنیم و نسبت به مقداردهی اولیه دقت کمتری داشته باشیم و در بعضی موارد نیاز به حذف تصادفی[3] را [۳۵] که برای جلوگیری از Overfitting از آن استفاده می‌شود، برطرف می‌کند. در Dropout، واحدها تصادفی به همراه اتصالات آن‌ها در شبکه حذف می‌شوند و در آموزش شبکه لحاظ نمی‌شوند. خروجی نرمال‌سازی دسته‌ای برای چیره‌شدن به مشکل حفظ داده‌ها به‌وسیلهٔ شبکهٔ عصبی و بروز خطای Overfiting به یک Dropout با نرخ ۰.۶ داده می‌شود. خروجی به یک لایهٔ Bidirectional داده می‌شود، Bidirectional یک wrapper برای شبکهٔ عصبی بازگشتی RNN است، در این لایه یک شبکهٔ عصبی بازگشتی دو طرفه[4] [۳۶] ایجاد شده‌است. در این لایه می‌توان بدون محدودیت در استفاده از اطلاعات ورودی، تا فریم بعدی از پیش تعیین‌شده را آموزش داد.

ازآنجاکه در مسئلهٔ شناسایی موجودیت اسمی، توالی واژگان مهم و طول ورودی‌ها متغیر است، پیش از آنکه خروجی به شبکهٔ عصبی بازگشتی دو طرفه داده شود، اطلاعات از یک لایه LSTM عبور داده می‌شوند؛ درنهایت نتایج شبکه از یک Dropout با نرخ ۰.۵ عبور داده و سپس به یک لایهٔ متراکم[5] [۳۷] داده می‌شود؛ این لایه، یک لایهٔ عمیقاً متصل است، به این معنی که هر نورون در لایهٔ متراکم، ورودی را از تمام سلول‌های عصبی لایهٔ قبلی خود دریافت می‌کند. از لایهٔ متراکم برای تغییر ابعاد بردار استفاده می‌شود.

برای اطمینان از این‌که مدل بتواند در داده‌های دیده نشده عملکرد خوبی داشته باشد از روش اعتبارسنجی متقابل[6] K-fold [۳۸] استفاده کرده‌ایم، در روش L2ValBERT مقدار K را ده در نظر گرفته‌ایم. در این روش که ده مدل آموزش داده می‌شود، به ازای آموزش هر مدل، ترتیب داده‌ها به هم می‌خورد و در اصطلاح عمل Shuffling روی‌داده‌ها صورت می‌پذیرد. هنگام

[3] Dropout
[4] Bidirectional Recurrent Neural Network (BRNN)
5 Dense Layer
6 Cross-Validation

[DOI: 10.61186/jsdp.21.3.97]

استفاده از مدل، هر ده مدل در حافظه بارگذاری می‌شود و معیار F نهایی با میانگین‌گیری نتایج تمام مدل‌ها به‌دست می‌آید.

۳-۸-۴-برچسب‌گذاری مبتنی بر مدل

بعد از ساخت مدل و اتمام فرایند یادگیری، با استفاده از مجموعه‌دادهٔ آزمون و مدل آموزشی یک پیکرهٔ برچسب‌خوردهٔ جدید را تولید می‌کنیم. این مجموعه‌دادهٔ حاشیه‌نویسی شده را در یک پیکرهٔ حاشیه‌نویسی شدهٔ مبتنی بر مدل ذخیره می‌کنیم. همانند پیکرهٔ متنی حاشیه‌نویسی شدهٔ مبتنی بر فهرست برای داده‌های آموزش، تعداد اسناد برچسب‌گذاری شدهٔ این پیکره با اجرای هر مرحله از روش پیشنهادی افزایش خواهد یافت. این پیکره در مراحل آموزش مدل مورد استفاده قرار خواهد گرفت؛ درواقع بعد از اجرای یک مرحله از ده مرحله در روش پیشنهادی، تعداد اسناد برچسب‌گذاری شدهٔ ورودی به شبکهٔ عصبی عمیق، مجموع اسناد مورد تأیید خبره به همراه اسناد برچسب‌خورده به روش مبتنی بر مدل است. موجودیت‌های اسمی جدیدی که به‌وسیلهٔ مدل آموزشی در اسناد آزمون شناسایی شده‌اند به فهرست ثانویهٔ موجودیت‌ها اضافه خواهند شد. گفتنی است در هنگام تطبیق یک موجودیت با فهرست ثانویه، در صورت تکراری‌بودن یک موجودیت، اولویت با فهرست ثانویه است.

۴- نتایج و بحث

ما در این مقاله با ارائهٔ یک روش پیشنهادی که برگرفته از الگوی یادگیری در شبکه‌های عصبی است، به‌خوبی نشان دادیم که با استفاده از مقادیر داده‌ای کم برچسب‌گذاری‌شده، چگونه می‌توان بر مشکل داده در شبکه‌های عصبی عمیق چیره شد.. در روش L2ValBERT تداوم یادگیری و استفاده از روش اعتبارسنجی متقابل در ساخت مدل، در کنار رویکردی تقویتی در افزایش دقت شناسایی موجودیت اسمی کمک فراوانی کرد.

نتایج حاصل از اجرای روش پیشنهادی بر روی مجموعه‌داده‌های ConNLL 2003، آرمان و پیما با سایر روش‌ها که نتایج خود را بر روی این مجموعه‌داده‌ها ارائه کرده‌اند مورد بررسی قرار گرفته‌است.

برای ارزیابی روش پیشنهادی به‌دلیل هزینه‌بربودن فرایند آموزش بر روی داده‌های انگلیسی، از اجرای روش پیشنهادی بر روی یک مجموعه‌دادهٔ انگلیسی صرف‌نظر شد. برای ارزیابی روش پیشنهادی از مدل زبانی آمادهٔ BERT [۳۱] استفاده شده‌است. با این تفاوت که در فرایند آموزش هشتاد درصد پیکره CoNLL 2003 انتخاب و با استفاده از مدل زبانی BERT به پیکرهٔ برداری برچسب‌گذاری‌شده تبدیل شده‌است. این مجموعه‌دادهٔ جدید به شبکهٔ عصبی عمیق پیشنهادی داده می‌شود و با استفاده از مدل‌های نهایی به‌دست‌آمده از روش اعتبارسنجی متقابل، دقت مدل‌ها بر روی بیست درصد باقی‌ماندهٔ پیکره CoNLL 2003 محاسبه می‌شود. نتیجهٔ ارائه‌شده در (جدول-۲) برای روش پیشنهادی از روشی که بیان شد به‌دست آمده‌است.

(جدول-۲): مقایسهٔ روش پیشنهادی با سایر روش‌ها بر روی مجموعه‌دادهٔ CoNLL 2003

(Table-2): Comparison of the proposed method with other methods on the CoNLL 2003 dataset

عنوان روش	معیار F1
روش پیشنهادی (L2ValBERT)	۹۵.۳
LUKE[۳۹]	۹۴.۳
CNN Large + fine-tune[۴۰]	۹۳.۵
Biaffine-NER[۴۱]	۹۳.۵
RNN-CRF+Flair[۴۲]	۹۳.۴۷
CrossWeigh + Pooled Flair[۴۳]	۹۳.۴۳
+Flair[۴۴]LSTM-CRF+ELMo+BERT	۹۳.۳۸
Hierarchical + BERT[۴۵]	۹۳.۳۷
BERT-MRC[۴۶]	۹۳.۰۴
BERT Large[۳۱]	۹۲.۸

در (جدول-۳) معیار F1 به‌دست‌آمده برای روش پیشنهادی بر روی مجموعه‌دادهٔ آرمان با سایر روش‌ها که نتایج خود را در مقالات مربوطه بر روی این مجموعه‌داده ارائه کرده‌اند، مقایسه شده‌است. عملکرد روش پیشنهادی نسبت به سایر روش‌ها بهبود قابل توجهی داشته‌است، اما در مقایسه با روش ParsBERT [۴۱] ضعیف‌تر عمل کرده‌است.

(جدول-۳): مقایسهٔ روش پیشنهادی با سایر روش‌ها بر روی مجموعه‌دادهٔ آرمان

(Table-3): Comparison of the proposed method with other methods on the Arman dataset

عنوان روش	معیار F1
روش پیشنهادی (L2ValBERT)	۹۶.۳۲
ParsBERT[۴۷]	۹۹.۸۴
LSTM-CRF[۴۸]	۸۶.۵۵
mBERT[۲۹]	۸۴.۰۳
Deep-CRF[۳۰]	۸۱.۵۰
Deep-Local[۳۰]	۷۹.۱۹
BiLSTM-CRF[۴۹]	۷۷.۴۵
SVM-HMM[۵۰]	۷۲.۵۹

[DOI: 10.61186/jsdp.21.3.97]

(جدول-۴): مقایسهٔ روش پیشنهادی با سایر روش‌ها بر روی مجموعه‌دادهٔ پیما

(Table-4): Comparison of the proposed method with other methods on the Peyma dataset

عنوان روش	معیار F1
روش پیشنهادی (L2ValBERT)	۹۴.۷۲
ParsBERT[۴۷]	۹۳.۴۰
mBERT[۲۹]	۹۰.۵۹
Rule-Based-CRF[۲۷]	۸۴.۰

در (جدول-۴) معیار F1 به‌دست‌آمده برای روش پیشنهادی بر روی مجموعه‌دادهٔ پیما با سایر روش‌ها که نتایج خود را در مقالات مربوطه بر روی این مجموعه‌داده ارائه کرده‌اند، مقایسه شده‌است. عملکرد روش پیشنهادی نسبت به سایر روش‌ها بهبود داشته‌است.

نتایج به‌خوبی نشان می‌دهد که استفاده از روش پیشنهادی، نسبت به سایر روش‌ها عملکرد مطلوب‌تری دارد. استفاده از روش اعتبارسنجی متقابل و همچنین استفاده از رویکرد تقویتی نتایج نهایی را نسبت به سایر روش‌ها بهبود می‌بخشد. استفاده از یک مدل زبانی معنایی این امکان را می‌دهد تا ویژگی‌هایی به‌وسیلهٔ روش پیشنهادی استخراج شود که تا حد مطلوبی دقت سامانه شناسایی موجودیت اسمی را افزایش دهد. استفاده از مدل زبانی معنایی باعث می‌شود بین کلماتی که نوشتار یکسانی دارند، اما معنای متفاوتی دارند تمایز قائل شد، یکی از دلایل افزایش دقت در روش پیشنهادی استفاده از این ویژگی است.

نتایج حاصل از این پژوهش، طراحی و پیاده‌سازی یک سامانهٔ شناسایی موجودیت اسمی است که با استفاده از روش‌های یادگیری عمیق نتایج مطلوب‌تری را نسبت به برخی ابزارهای موجود ارائه می‌دهد. علاوه بر این می‌توان به دستاوردهای زیر نیز اشاره کرد:

- تولید پیکرهٔ متنی برچسب‌خوردهٔ موجودیت‌های فارسی
- تهیهٔ فهرست بزرگی از موجودیت‌های اسمی
- ارائهٔ روشی برای ساخت مدل یادگیری عمیق به منظور شناسایی موجودیت‌های اسمی با استفاده از شبکهٔ عصبی
- ارائهٔ مدل یادگیری عمیق برای شناسایی موجودیت اسمی در زبان فارسی
- تولید مدل زبانی معنایی در زبان فارسی

۵- نتیجه‌گیری و کارهای آینده

در این پژوهش، با توجه به اهمیت و جایگاه شناسایی موجودیت‌های اسمی در حوزهٔ پردازش زبان طبیعی، ابتدا در قالب بیان مسئله این اهمیت بیان شده‌است، ضرورت و اهمیت این پژوهش با توجه به انتشار مقالات جدید در ماه‌های گذشته مشهود است. در ادامه، مفاهیم و تعاریف اولیه با بیان برخی جزئیات به‌صورت مفصل بیان شده‌است. در سال‌های اخیر کارهای بسیاری در حوزه NER صورت پذیرفته‌است که در بخش کارهای مرتبط با پژوهش، در قالب چهار روش این کارها بیان‌شده‌اند. با درنظرگرفتن کارهای انجام‌شده و مزایا و معایب روش‌های مطرح، روش پیشنهادی ارائه شده‌است. در روش پیشنهادی مجموعه‌دادهٔ مورد نیاز از خزش تارنماهای خبری به‌دست آمده‌است که پیکره‌ای با یک میلیون سند را تشکیل می‌دهد. به‌منظور تهیهٔ فهرستی از موجودیت‌های اسمی، ابزاری برای این کار تولید و در اختیار کارشناسان قرار داده شد، در این ابزار با استفاده از مقالات ویکی‌پدیا، موجودیت‌ها با برچسب مکان، سازمان، اشخاص و غیره مشخص می‌شوند. در ادامه برای برچسب‌گذاری ماشینی پیکرهٔ متنی خزش شده از این فهرست استفاده شده‌است. اسناد برچسب‌گذاری‌شده در این مرحله توسط خبره اعتبارسنجی و اصلاح می‌شوند. برای سهولت کار و افزایش سرعت در بررسی این پیکره، ابزاری تولید شد که خبره با استفاده از آن به‌راحتی بتواند برچسب‌های اسناد را بررسی و اصلاح کند. با استفاده از مدل زبانی آموزش‌دیده شده با روش BERT، پیکرهٔ متنی برچسب‌گذاری‌شده به یک پیکرهٔ برداری برچسب‌گذاری‌شده تبدیل می‌شود. از این پیکره در فرایند آموزش شبکهٔ عصبی با استفاده از روش اعتبارسنجی متقابل در ساخت مدل استفاده شده‌است. مدل نهایی تولیدشده بر روی متون خام اجرا و خروجی نهایی فهرست جدیدی از موجودیت‌های اسمی و پیکرهٔ برچسب‌گذاری‌شدهٔ جدید است که در مراحل بعدی آموزش مورداستفاده قرار می‌گیرد. با توجه به نتایج ارزیابی، روش پیشنهادی نسبت به سایر روش‌ها از عملکرد بهتری برخوردار است.

با توجه به اهمیت شناسایی موجودیت‌های اسمی در اسناد، ارائهٔ روش‌های جدید در راستای بهبود نتایج الزامی است، در روش پیشنهادی این انتظار می‌رود که می‌توان با برخی تغییرات و کارهای جدید، به نتایج مطلوب‌تری دست‌یافت. از جمله پیشنهادها برای کارهای آینده عبارت‌اند از:

- استفاده از ویژگی‌های برچسب عبارت اسمی، برچسب اجزای کلام، بن‌واژه، ویژگی‌های حضور و عدم حضور واژه در فهرست موجودیت، n-

[DOI: 10.61186/jsdp.21.3.97]

gramهای سطح کاراکتر واژه و هم رخدادی موجودیت‌ها در ساخت مدل

- حذف خبره و تبدیل روش نیمه‌نظارتی ارائه‌شده به یک روش بدون نظارت

6-Reference ۶-مراجع

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3-26, 2007.

[2] J. Guo, G. Xu, X. Cheng, and H. Li, "Named entity recognition in query," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 267-274 .

[3] D. Petkova and W. B. Croft, "Proximity-based document representation for named entity retrieval," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 731-740 .

[4] B. Larsen, "A trainable summarizer with knowledge acquired from robust NLP techniques," *Advances in automatic text summarization*, vol. 71, 1999.

[5] D. Mollá, M. Van Zaanen ,and D. Smith, "Named entity recognition for question answering," 2006.

[6] B. Babych and A. Hartley, "Improving machine translation quality with automatic named entity recognition," in *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*, 2003 .

[7] S. Scholar."Semantic Scholar." https://www.semanticscholar.org (accessed 2024).

[8] I. Keraghel, S. Morbieu, and M. Nadif, "A survey on recent advances in named entity recognition," arXiv preprint, arXiv:2401.10825, 2024.

[9] J. Yang, T. Zhang, C.-Y. Tsai, Y. Lu, and L. Yao, "Evolution and emerging trends of named entity recognition: Bibliometric analysis from 2000 to 2023," *Heliyon*, 2024.

[10] Z. Hu, W. Hou, and X. Liu, "Deep learning for named entity recognition: A survey," *Neural Computing and Applications*, vol. 36, no. 16, pp. 8995–9022, 2024.

[11] Y. Park, G. Son, and M. Rho, "Biomedical flat and nested named entity recognition: Methods, challenges, and advances," Applied Sciences, vol. 14, no. 20, 2024.

[12] J. Liu, M. Sun, W. Zhang, G. Xie, Y. Jing, X. Li, and Z. Shi, "Dae-ner: Dual-channel attention enhancement for Chinese named entity recognition," Computer Speech & Language, vol. 85, p. 101581, 2024.

[13] P. Deshmukh, N. Kulkarni, S. Kulkarni, K. Manghani, P. A. Khadkikar, and R. Joshi, "Named entity recognition for Indic languages: A comprehensive survey," in 2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST), pp. 1–6, IEEE, 2024.

[14] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996 .

[15]]E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[16] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation," in *Lrec* ,۲۰۰۴ ,vol. 2, no. 1: Lisbon, pp. 837-840 .

[17] G. Demartini, T. Iofciu, and A. P. De Vries, "Overview of the INEX 2009 entity ranking track," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*, 2009: Springer, pp. 254-264 .

[18] K. Balog, P. Serdyukov, and A. P. de Vries, "Overview of the TREC 2011 Entity Track," in *TREC*, 2011, vol. 2011, p. 11 .

[19] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, "Named entity recognition: fallacies, challenges and opportunities," *Computer Standards & Interfaces*, vol. 35, no. 5, pp. 482-489, 2013.

[20] M. L. Patawar and M. Potey, "Approaches to named entity recognition: a survey," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 3, no. 12, pp. 12201-12208, 2015.

[21] H. Pu and W. Chen, "Review of multimodal named entity recognition studies," Data Analysis and Knowledge Discovery, vol. 8, no. 4, pp. 50–63, 2024.

[22] Momtazi S, Torabi F. Named Entity Recognition in Persian Text using Deep Learning. JSDP 2020; 16 (4) :93-112

[23] O. Khade, S. Jagdale, G. Takalikar, M. Inamdar, R. Joshi, and A. S. Ghotkar, "Enhancing code-mixing in named entity recognition: A comprehensive survey of deep learning models," in 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), pp. 1–6, IEEE, 2024.

[24] H. Pu and W. Chen, "Review of multimodal named entity recognition studies," Data Analysis and Knowledge Discovery, vol. 8, no. 4, pp. 50–63, 2024.

[25] H. Wang, X. Xu, T. Wang, and B. Jing, "Research progress of multimodal named entity recognition," Journal of Zhengzhou University: Engineering Science, vol. 45, no. 2, 2024.

[26] M. Zali and M. Firoozbakht, "Named Entities Recognition and Classification System for Persian Texts Based on Neural Network," *Iranian Research Institute for Information Science and Technology*, vol. 34, pp. 473-486, 2018.

[27] Shahshahani M S, Mohseni M, Shakery A, Faili H. PAYMA: A Tagged Corpus of Persian Named Entities. JSDP 2019; 16 (1) :91-110

[28] M. Mohseni and A. Tebbifakhr, "MorphoBERT: a Persian NER system with BERT and morphological

[DOI: 10.61186/jsdp.21.3.97]

analysis," in *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, 2019, pp. 23-30 .

[29] E. Taher, S. A. Hoseini, and M. Shamsfard, "Beheshti-NER: Persian named entity recognition Using BERT," *arXiv preprint arXiv:2003.08875*, 2020.

[30] M. H. Bokaei and M. Mahmoudi, "Improved deep persian named entity recognition," in *2018 9th International Symposium on Telecommunications (IST)*, 2018: IEEE, pp. 381-386 .

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[32] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv preprint arXiv:1409.3215*, 2014.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015: PMLR, pp. 44 ۴۵۶-۸.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[36] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708 .

[38] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108-132, 2000.

[39] I. Yamada, A. Asai, H. Shindo, H .Takeda, and Y. Matsumoto, "LUKE: deep contextualized entity representations with entity-aware self-attention," *arXiv preprint arXiv:2010.01057*, 2020.

[40] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, and M. Auli, "Cloze-driven pretraining of self-attention networks," *arXiv preprint arXiv:1903.07785*, 2019.

[41] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," *arXiv preprint arXiv:2005.07150*, 2020.

[42] Y. Jiang, C. Hu, T. Xiao, C. Zhang, and J. Zhu, "Improved differentiable architecture search for language modeling and named entity recognition," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP۲* ,pp. 3576-3581, 2019.

[43] Z. Wang, J. Shang, L. Liu, L. Lu, J. Liu, and J. Han, "Crossweigh: Training named entity tagger from imperfect annotations," *arXiv preprint arXiv:1909.01441*, 2019.

[44] J. Straková, M. Straka, and J. Hajič, "Neural architectures for nested NER through linearization," *arXiv preprint arXiv:1908.06926*, 2019.

[45] Y. Luo, F. Xiao, and H. Zhao, "Hierarchical contextualized representation for named entity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 05, pp. 8441-8448 .

[46] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A unified mrc framework for named entity recognition," *arXiv preprint arXiv:1910.11476*, 2019.

[47] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "ParsBERT: Transformer-based Model for Persian Language Understanding," *arXiv preprint arXiv:2005.12515*, 2020.

[48] L. Hafezi and M. Rezaeian, "Neural architecture for persian named entity recognition," in *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 2018: IEEE, pp. 61-64 .

[49] H. Poostchi, E. Z. Borzeshi, and M. Piccardi, "Bilstm-crf for persian named-entity recognition armanpersonercorpus: The first entity-annotated persian dataset," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018 .

H. Poostchi, E. Z. Borzeshi, M. Abdous, and M. Piccardi, "PersoNER: Persian named-entity recognition," in *COLING 2016-26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2016 .

مهدی نقوی استادیار مهندسی نرم‌افزار دانشکدهٔ مهندسی کامپیوتر دانشگاه جامع امام حسین (ع) است. ایشان دکتری خود را در رشتهٔ مهندسی کامپیوتر، گرایش نرم‌افزار از دانشگاه علم‌وصنعت ایران در سال ۱۳۹۴ دریافت کرد. رسالهٔ دکتری او در موضوع خزش برخط وب‌نوشت‌های فارسی جهت رصد مستمر فضای وب است. زمینه‌های پژوهشی مورد علاقهٔ ایشان نرم‌افزار سامانه، سامانه‌های توزیع‌شده، رایانش ابری، داده‌های حجیم، بازیابی اطلاعات و وب‌کاوی است.

نشانی رایانامهٔ ایشان عبارت‌است از:

mnaghavi@ihu.ac.ir

[DOI: 10.61186/jsdp.21.3.97]

محمدرضا حسنی‌آهنگر استاد تمام هوش مصنوعی و رباتیک دانشکده و پژوهشکدهٔ هوش مصنوعی و علوم‌شناختی دانشگاه جامع امام حسین (ع)، همچنین عضو حقیقی شورای ملی راهبری هوش مصنوعی و عضو هیئت‌امنای بنیاد ملی نخبگان کشور هستند. از سال ۱۳۹۱ ریاست دانشگاه جامع امام حسین (ع) به عهدهٔ ایشان است.
نشانی رایانامهٔ ایشان عبارت‌است از:
mrhassani@ihu.ac.ir

علی امیری‌جزه کارشناسی‌ارشد خود را در رشتهٔ هوش مصنوعی از دانشگاه امام حسین (ع) در سال ۱۳۹۹ دریافت کرد. حوزهٔ مورد علاقهٔ ایشان پردازش زبان طبیعی، تحلیل‌های زبانی و داده‌های حجیم است و پایان‌نامهٔ کارشناسی‌ارشد او در زمینهٔ شناسایی موجودیت‌های اسمی با استفاده از یادگیری عمیق و رویکرد تقویتی است.
نشانی رایانامهٔ ایشان عبارت‌است از:
aamirij@ihu.ac.ir

[DOI: 10.61186/jsdp.21.3.97]