تشخیص موجودیت‌های نامدار در متون فارسی با استفاده از یادگیری عمیق

سعیده ممتازی* **و فرزانه ترابی**

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

شناسایی موجودیت‌های نامدار[1] یکی از فعالیت‌های زیربنایی در حوزهٔ پردازش زبان طبیعی[2] و به‌طورکلی زیرمجموعه‌ای از استخراج اطلاعات[3] است. در فرآیند شناسایی موجودیت‌های نامدار به‌دنبال یافتن عناصر اسمی در متن و دسته‌بندی آنها به رده‌هایی ازپیش‌تعیین‌شده از قبیل اسامی اشخاص، سازمان‌ها، مکان‌ها، مذاهب، عنوان کتاب‌ها، عنوان فیلم‌ها و غیره هستیم. در این مقاله با بهره‌گیری از روش‌های نوین در این حوزه مانند استفاده از دو بُردار مختلف بازنمایی معنایی واژگان برمبنای واژه و حروف تشکیل‌دهندهٔ آن برمبنای شبکه‌های عصبی و همچنین استفاده از روش‌های یادگیری عمیق[4] یک سامانهٔ تشخیص موجودیت‌های نامدار معرفی می‌شود؛ همچنین در راستای پژوهش حاضر، یک پیکره برچسب‌گذاری‌شده شامل سه‌هزار چکیده از ویکی‌پدیای فارسی که شامل نودهزار واژه است با استفاده از پانزده برچسب مختلف ارائه می‌شود که گام مهمی در ارتقای پژوهش‌های آینده این حوزه برداشته خواهد شد. نتایج حاصل از ارزیابی سامانهٔ پیشنهادی نشان می‌دهد که می‌توان با استفاده از دادهٔ معرفی‌شده به دقت ۷۲/۰۹ در معیار F رسید.

واژگان کلیدی: تشخیص موجودیت‌های نامدار، پردازش زبان طبیعی، بازنمایی معنایی واژه‌گان، یادگیری عمیق

Named Entity Recognition in Persian Text using Deep Learning

Saeede Momtazi* & Farzaneh Torabi

Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran

Abstract

Named entities recognition is a fundamental task in the field of natural language processing. It is also known as a subset of information extraction. The process of recognizing named entities aims at finding proper nouns in the text and classifying them into predetermined classes such as names of people, organizations, and places. In this paper, we propose a named entity recognizer which benefits from neural network-based approaches for both word representation and entity tagging.

In the word representation part of the proposed model, two different vector representations are used and compared: (1) the semantic representation of words based on their context using word2vec continues skip-gram model, and (2) the semantic representation of words based on their context as well as characters forming them using fasttext. While the former model captures the semantic concepts of words, the latter one considers the morphological similarity of words as well. For the entity identification, a deep Bidirectional Long Short Term Memory (BiLSTM) network is used. Using LSTM model helps to consider the history of

[1] Named Entity Recognition
[2] Natural Language Processing
[3] Information Extraction
[4] Deep Learning

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال: ۹۶/۱۱/۱۱ ● تاریخ پذیرش: ۹۸/۰۳/۲۹ ● تاریخ انتشار: ۹۸/۱۲/۲۸

text when predicting entities, while the BiLSTM model expands this idea by benefiting from the history from both sides of the context.

Moreover, inline of the present research, an annotated corpus containing 3000 abstracts (90000 tokens) from the Persian Wikipedia is provided. In contrast to the available datasets in the field, which includes up to 7 label types, the new dataset contains 15 different labels, namely person individual, person group, organizations, locations, religions, books, magazines, movies, languages, nationalities, events, jobs, dates, fields, and other. Developing this dataset will be an important step in promoting future research in this field, especially for the tasks such as question answering that need wider range of entity types. The results of the proposed system show that by using the introduced model and the provided data, the system can achieve 72.92 F-measure.

Keywords: Name entity recognition, natural language processing, word embedding, deep learning

۱- مقدمه

شناسایی موجودیت‌های نامدار در پردازش زبان طبیعی به عملیاتی گفته می‌شود که در آن کلیه اسامی خاص موجود در متن شناسایی و استخراج می‌شوند و به مقولات ازپیش تعریف‌شده‌ای مانند اسم افراد، سازمان‌ها، مکان‌ها و... دسته‌بندی می‌شوند. به این صورت که متن را بر اساس واژگان قطعه‌بندی و عبارات حاوی موجودیت نامدار را با برچسب‌زنی مشخص می‌کنیم.

این مفهوم برای نخستین‌بار در ششمین کنفرانس Message Understanding در سال ۱۹۹۵ مطرح شد. درواقع مسأله تشخیص موجودیت‌های نامدار در متن، به‌طورعمومی به دو زیر مسأله تشخیص و دسته‌بندی موجودیت‌ها تقسیم می‌شود. اسامی خاصی که تشخیص داده می‌شوند و همچنین قالبی که برای دسته‌بندی آنها به‌کار می‌رود، وابسته به نوع کاربرد آن خواهد بود. در سامانه‌های تشخیص موجودیت‌های نامدار، بیشتر بر روی تشخیص اسامی اشخاص، مکان‌ها و سازمان‌هایی که در یک متن آورده شده است تمرکز می‌شود [1].

نیاز به شناسایی موجودیت‌های نامدار در دنیای امروز که عصر ارتباطات و اطلاعات است، به‌شدت احساس می‌شود. شناسایی موجودیت‌های نامدار برای جستجوهای معنادار[1]، ترجمه خودکار[2]، استخراج اطلاعات از متن، کشف ارجاعات در متن[3]، سامانه‌های پرسش و پاسخ[4]، سامانه‌های خبره[5]، کشف دانش[6]، مدیریت دانش[7]، نظرکاوی[8]، بازیابی اطلاعات[9]، تحلیل خبر[10] و بسیاری دیگر از شاخه‌های مرتبط با پردازش زبان طبیعی کاربرد دارد. اینکه سامانه چه نوع موجودیتی را تشخیص دهد و یا به بیان دیگر دسته‌های معنایی مورد نظرش چه باشند، وابسته به زمینه کاربردی سامانه است [2]. برای مثال شناسایی موجودیت‌های نامدار در علم زیست‌شناسی می‌تواند تشخیص اسامی وابسته به انواع پروتئین ،دی‌ان‌ای، نوع سلول و غیره، در حوزه پزشکی تشخیص انواع بیماری، دارو، و مراکز درمانی، در حوزهٔ تجارت نام شرکت‌ها و مؤسسات، تراکنش‌های مالی، بورس و غیره باشد؛ یا به‌صورت خیلی خاص‌منظوره به‌عنوان مثال تنها برای تشخیص اسامی شرکت‌های تولیدکننده فولاد به‌کار رود.

یکی از کاربردهای تشخیص موجودیت‌های نامدار در ترجمه ماشینی رفع ابهام از ترجمه و افزایش دقت آن است. به‌عنوان مثال، اگر در متنی واژه Apple به‌عنوان موجودیت نامدار شناخته شده باشد و دارای برچسب باشد، در این صورت در هنگام ترجمه به‌عنوان شرکت اپل شناخته می‌شود و معنای سیب نخواهد داشت. در مثالی دیگر، در ترجمه فارسی به انگلیسی می‌توان به واژهٔ زیبا اشاره کرد. اگر اسم فرد و موجودیت نامدار باشد، نیاز به ترجمه ندارد، و در غیراین‌صورت باید به واژه beautiful ترجمه شود[3].

به‌طورکلی از دو روش قاعده‌مند و آماری برای تشخیص موجودیت‌های نامدار استفاده می‌شود [4]. در روش‌های قاعده‌مند، قوانینی تعریف می‌شود که براساس آنها موجودیت‌های نامدار تشخیص داده می‌شود. در روش آماری، از روش‌های یادگیری ماشین برای دسته‌بندی موجودیت‌های نامدار به هر مقوله استفاده می‌شود. استفاده از روش‌های دسته‌بندی بانظارت[11] در این قسمت سبب می‌شود که با استفاده از پیکره‌ای که برچسب‌های موجودیت‌های نامدار دارد، بتوان مدلی را آموزش داد، و با آن مدل متن بدون برچسب را برچسب‌گذاری کرد. در مقاله حاضر، برای ساخت

[1] Semantic Search
[2] Automatic Translation
[3] Co-reference Resolution
[4] Question Answering Systems
[5] Expert Systems
[6] Knowledge Discovery
[7] Knowledge Management
[8] Sentiment Analysis
[9] Information Retrieval
[10] News Analysis
[11] Supervised

سامانه تشخیص موجودیت‌های نامدار از روش‌های یادگیری ماشین بانظارت و مبتنی بر یادگیری عمیق استفاده می‌شود.

یادگیری عمیق یک زیرشاخه بر مبنای مجموعه‌ای از الگوریتم‌ها است که در تلاش است، مفاهیم انتزاعی سطح بالا در دادگان را مدل کنند. این فرایند را با استفاده از یک شبکه عمیق که دارای چندین لایه پردازشی متشکل از چندین لایه تبدیلات خطی و غیرخطی است، مدل می‌کنند؛ به بیان دیگر، پایه آن بر یادگیری نمایش دانش و ویژگی‌ها در لایه‌های مدل است که در آن داده‌های بدون برچسب به سامانه داده می‌شود و سامانه با استخراج واژه‌های آن برای هر واژه بردار ویژگی آن را بر اساس ویژگی‌های مفهومی از متن استخراج می‌کند. این عملیات را می‌توان در دو مرحله انجام داد. ابتدا برای هر واژه بردار مربوطه استخراج می‌شود و سپس با استفاده از الگوریتم شبکه عصبی[1] به‌کمک بردارهای به‌دست‌آمده و داده‌های آموزش (پیکره برچسب‌خورده) داده‌های آزمون برچسب‌دهی می‌شوند.

برای این منظور، در مقاله حاضر، دو مدل مختلف بازنمایی برای یادگیری عمیق شبکه استفاده و هر واژه با دو نوع بردار بازنمایی جایگزین می‌شود. بدین ترتیب، به‌ازای هر واژه یک بردار بازنمایی واژه و یک بردار بازنمایی ویژگی حروف آن واژه را خواهیم داشت. با کمک این بردارها و داده‌های آموزش می‌توان به تشخیص موجودیت‌های نامدار پرداخت.

ساختار مقاله حاضر بدین شرح است: در بخش ۲، مروری بر روش‌های تشخیص موجودیت‌های نامدار استفاده‌شده خواهیم داشت؛ در بخش ۳، درخصوص شبکه عصبی و نحوهٔ به‌کارگیری آن در این مقاله بحث خواهد شد. در بخش ۴، مدل پیشنهادی توصیف و در بخش ۵، توضیحاتی از دادگان تهیه‌شده ارائه می‌شود. در بخش ۶، به بررسی آزمایش‌های مربوطه می‌پردازیم، و درنهایت نتیجه‌گیری کلی در بخش ۷ ارائه می‌شود.

۲- کارهای انجام‌شده

در زبان‌های مختلف بر روی مبحث "تشخیص و دسته‌بندی موجودیت‌های نامدار" به‌عنوان زیرشاخه‌ای از پردازش زبان طبیعی پژوهش‌های متعددی انجام شده است. اگرچه نسبت بالایی از پژوهش‌ها متعلق به زبان انگلیسی است، فعالیت‌هایی برای زبان‌هایی چون آلمانی، یونانی، ژاپنی، فرانسوی، ایتالیایی، هلندی، چینی، روسی، کره‌ای، رومانیایی و ترکی نیز انجام شده است. بیش‌تر کارهای انجام‌شده متعلق به زبان، دامنه و یا گونه نوشتاری خاص است و بزرگ‌ترین مشکل این‌گونه سامانه‌ها نیز مربوط به انتقالشان به دامنه جدید است [5].

دو رویکرد اصلی که برای تشخیص موجودیت‌های نامدار استفاده می‌شوند، عبارتند از روش‌های قاعده‌مند و روش‌های آماری. در روش‌های قاعده‌مند، قوانینی تعریف می‌شود که از روی این قوانین می‌توانیم موجودیت‌های نامدار را تشخیص دهیم [6]. به‌عنوان مثال هرگاه عبارت "در سال...." در متن باشد، این عبارت می‌تواند آغاز یک موجودیت نامدار از نوع تاریخ باشد. به بیان دیگر با ظاهرشدن عبارت "در سال ۹۵۰ هجری قمری" در متن دنباله واژگان "سال ۹۵۰ هجری قمری" در این متن به‌عنوان موجودیت تاریخ تشخیص داده می‌شود. همچنین می‌توان در این رویکرد فهرستی تهیه کرد که شامل نوع خاصی از موجودیت‌های نامدار باشد، برای مثال فهرستی شامل تمام شهرهای جهان. اگر در یک متن، یکی‌از شهرهای موجود در این فهرست قرار داشت، آن شهر به‌عنوان موجودیت نامدار مکان در نظر گرفته شود.

ایده استفاده از این روش از وجود گنج‌واژه[2] در علم جغرافیا الهام گرفته شده است. گنج‌واژه یک واژه‌نامهٔ جغرافیایی است که در رابطه با یک نقشه یا یک اطلس استفاده می‌شود. آنها به‌طورمعمول شامل اطلاعات مربوط به آرایش جغرافیایی، آمار اجتماعی و ویژگی‌های فیزیکی کشور، منطقه یا قاره هستند. محتوای گنج‌واژه می‌تواند شامل یک موقعیت هدف، ابعاد یک قله یا راه آبی، میزان جمعیت و میزان سواد باشد. این اطلاعات به‌طورکلی به موضوعات با نوشته‌های یادشده و به‌ترتیب حروف الفبا مرتب می‌شوند.

ایده اصلی استفاده از گنج‌واژه در بحث جغرافیا و به‌منظور جمع‌آوری اطلاعات جغرافیایی مکان‌های مختلف بوده است که جغرافی‌دان‌ها از آن استفاده می‌کرده‌اند و بعدها از این فرهنگ‌های جغرافیایی در زمینه‌های مربوط به پردازش زبان طبیعی و متن کاوی استفاده شد. هم‌اکنون اصطلاحات Lexicon، Gazatteer و Dictionary اغلب با اصطلاح فهرست جایگزین می‌شوند. گنج‌واژه به‌طورعمومی به فهرست بزرگی از نام مکان‌ها گفته می‌شود؛ اما این اصطلاح در زمینه تشخیص موجودیت‌های نامدار عمومی شده است و

[1] Neural Network

[2] Gazetteer

حتی از نام گنج‌واژه برای فهرست واژگان دیگر هم استفاده می‌شود.

این رویکرد در بسیاری از موارد به‌درستی عمل می‌کند؛ ولی در دو صورت با ابهام مواجه می‌شود: ۱) زمانی که واژه موجود در واژه‌نامه به‌صورت کامل استفاده نشده باشد؛ به‌عنوان مثال، اگر در واژه‌نامه واژهٔ "وزارت علوم، تحقیقات و فناوری" موجود باشد، این روش در تشخیص یک موجودیت که در آن تنها واژه "وزارت علوم" آورده شده است، ضعیف عمل می‌کند. مورد دیگر ابهام زمانی خواهد بود که یک واژه، معانی متفاوتی داشته باشد و در واژه‌نامه به‌اشتباه به معنی دیگری مرتبط شود. برای مثال، ممکن است واژه "هما" در یک متن به‌عنوان اسم شخص استفاده شده باشد، اما به‌اشتباه به‌عنوان مخفف "هواپیمایی ملی ایران" تشخیص داده شود.

در رویکرد دوم که در سامانه‌های مبتنی بر یادگیری ماشینی مورد استفاده قرار می‌گیرد، هدف از ره‌یافت تشخیص واحدهای اسمی موجودیت‌های نامدار تبدیل مسأله تشخیص به مسأله دسته‌بندی است و از یک مدل آماری دسته‌بندی برای حل آن استفاده می‌شود. در این روش، مدل به‌دنبال تشخیص الگوها و یافتن رابطه آنها با متن و ساختن یک مدل آماری و الگوریتم یادگیری ماشین است. این سامانه‌ها اسامی خاص را یافته و آنها را براساس مدل به‌دست‌آمده با استفاده از روش‌های یادگیری ماشین به مقوله‌های ازپیش‌تعیین‌شده مانند اشخاص، مکان‌ها، زمان‌ها، مذاهب، سازمان‌ها و ... تقسیم می‌کند.

روش‌های مبتنی بر یادگیری که به‌صورت بانظارت عمل می‌کنند، نیاز به حجم زیادی از متون برچسب‌گذاری‌شده دارند؛ و با دراختیارداشتن این داده‌ها، به استفاده از الگوریتم‌هایی نظیر درخت تصمیم‌گیری[1] و مدل مخفی مارکوف۲، آنتروپی بیشینه[3]، میدان تصادفی شرطی[4] و غیره می‌پردازند. در روش‌های یادگیری بانظارت ابتدا سامانه توسط پیکره‌ای به‌عنوان دادهٔ آموزش که به‌صورت دستی و به‌وسیله انسان برچسب‌گذاری شده است، آموزش می‌بیند و با یادگیری از طریق این داده‌ها، به تشخیص خودکار اسامی خاص در متون دیده‌نشده می‌پردازد. گفتنی است، که روش‌های بانظارت به داده‌های برچسب‌گذاری‌شده برای ساخت یک مدل آماری نیاز دارد. برخلاف ره‌یافت‌های مبتنی بر قاعده، روش‌های یادگیری ماشین مستقل از حوزه و زبان عمل می‌کنند [7].

در پژوهش انجام‌شده توسط موروال[5] و همکاران مدل مخفی مارکوف برای تشخیص موجودیت‌های نامدار مورد استفاده قرار گرفته است [8]. در پژوهشی دیگر که توسط لی[6] و همکاران انجام گرفته روش میدان تصادفی شرطی برای تشخیص موجودیت‌های نامدار دانه‌ریز به‌کار گرفته شده است [9]. این روش همچنین برای اندازهٔ زبان‌ها نیز مورد استفاده قرار گرفته است؛ که از جمله آن می‌توان به زبان‌های ترکی، عربی، بنگالی و هندی اشاره کرد [10,11,12]. الگوریتم آنتروپی بیشینه نیز در این حوزه توسط کوران و کلارک[7] مورد استفاده قرار گرفته و توانسته است در یک چارچوب مستقل از زبان برای تشخیص موجودیت‌های نامدار در زبان‌های مختلف مورد استفاده قرار گیرد [13].

علاوه‌بر الگوریتم‌های یادشده، برای تشخیص موجودیت‌های نامدار، از شبکه‌های عصبی مختلفی نیز استفاده شده است که روند تشخیص را بهبود می‌دهد و دقت کار را بالا می‌برد. شبکه‌های عصبی معروفی که در این زمینه به‌کار می‌رود، شبکه عصبی بازگشتی[8] و شبکه عصبی حافظه طولانی کوتاه مدت[9] است. از جمله پژوهش‌های این حوزه می‌توان به پژوهش لی[10] و همکاران اشاره کرد که از شبکه عصبی بازگشتی برای تشخیص موجودیت‌های نامدار در حوزه پزشکی استفاده کرده‌اند [14]. گفتنی است، استخراج موجودیت‌های نامدار از متون پزشکی تنها به روش‌های مبتنی بر شبکه عصبی خلاصه نمی‌شود و کارهای انجام‌شده در این حوزه با استفاده از روش‌های مدل مخفی مارکوف، آنتروپی بیشینه و میدان تصادفی شرطی نیز صورت گرفته است [15,16].

شبکه‌های عصبی یادشده به‌تفضیل در بخش ۳ توضیح داده خواهد شد.

۱-۲- کارهای انجام‌شده برای زبان فارسی

نخستین کارهای انجام‌شده برای تشخیص موجودیت‌های نامدار فارسی برمبنای روش‌های مبتنی بر قاعده صورت گرفته است.

[1] Decision Tree
[2] Hidden Markov Model
[3] Maximum Entropy
[4] Conditional Random Field

[5] Morwal
[6] Lee
[7] Curran و Clark
[8] Recurrent Neural Network (RNN)
[9] Long Short Term Memory (LSTM)
[10] Li

در روش ارائه‌شده توسط مرتضوی و شمس‌فرد بعد از انجام پیش‌پردازش‌های لازم از جمله نرمال‌سازی و همچنین برچسب‌زنی مقوله‌های نحوی، با تعریف قوانین دستی به تشخیص موجودیت‌های نامدار پرداختند [17].

کلالی و بذرافکن از یک رویکرد مبتنی بر فرهنگ لغت برای تشخیص موجودیت‌های نامدار استفاده کرده‌اند [18]. در این پژوهش از برچسب‌های مقولهٔ نحوی پیکرهٔ بی‌جن‌خان بهره برده شده است [19]. همچنین مقالات ویکی‌پدیا برای ساخت فرهنگ لغت مورد استفاده قرار گرفته است.

مرادیان‌نسب و ممتازی از ترکیبی از قواعد تعریف‌شده و همچنین فرهنگ لغت برای تشخیص موجودیت‌های نامدار استفاده کرده‌اند [6].

در پژوهشی دیگر عبدوس و مینایی بیدگلی نشان داده‌اند که با درنظرگرفتن ویژگی کسره اضافه در واژگان در چهارچوب یک ساختار قاعده‌محور می‌توان باعث بهبود تشخیص موجودیت‌های نامدار شد [20].

معیار F ۷۴٪، ۸۳٪ ، ۸۹٪ و ۸۲٪ به‌ترتیب توسط مقالات بالا گزارش شده است، اما به‌دلیل یکسان‌نبودن دادگان آزمون، این نتایج قابل مقایسه نیستند. گفتنی است در تمامی موارد ذکرشده تعداد برچسب‌های مورد استفاده بسیار محدود بوده و در بیش‌تر موارد این برچسب‌ها به سه موجودیت شخص، سازمان و مکان خلاصه شده است.

پس از روش‌های مبتنی بر قاعده، روش‌های ترکیبی که از هر دو مدل مبتنی بر قاعده و مبتنی بر یادگیری ماشین استفاده می‌کنند، روی کار آمد.

احمدی و مرادی برای تشخیص موجودیت‌های نامدار از یک رویکرد ترکیبی بهره برده‌اند که در آن علاوه‌بر روش مبتنی بر قاعده، روش یادگیری ماشین با استفاده از الگوریتم مدل مخفی مارکوف مورد استفاده قرار گرفته است [21].

پوستچی و همکاران از روش‌های مختلف یادگیری ماشین از جمله بردار تصادفی شرطی، ماشین بردار پشتیبان و مدل مخفی مارکوف برای این هدف استفاده کرده‌اند و بهترین دقت را با استفاده از مدل ترکیبی ماشین بردار پشتیبان و مدل مخفی مارکوف به‌دست آورده‌اند. در این پژوهش همچنین یک پیکره تحت عنوان «پیکره واحدهای اسمی آرمان» ارائه شده که شامل برچسب‌های شخص، سازمان، مکان، امکانات، محصولات و رویدادها است. با توجه به این‌که برچسب «محصولات» بسیار کلی است، هیچ‌گونه تفکیکی بین موارد مانند اسم کتاب، اسم مجله و یا اسم فیلم وجود ندارد؛ ضمن این‌که اطلاعاتی مانند شغل، زبان، ملیت، مذهب در این پیکره پوشش داده نشده است [22].

دشتی‌پور و همکاران نیز از ترکیبی از روش‌های مبتنی بر قاعده که از فرهنگ لغت و قواعد زبان‌شناسی بهره برده است و همچنین الگوریتم ماشین بردار پشتیبان استفاده کرده‌اند. در این پژوهش برچسب‌های شخص، مکان و تاریخ مد نظر قرار گرفته است [23].

شاکری و همکاران نیز ترکیبی از روش‌های مبتنی بر فرهنگ لغت، بردار تصادفی شرطی و شبکه عصبی را به‌کار برده‌اند. همچنین اطلاعات زبان‌شناختی از جمله برچسب مقوله نحوی، ریشه واژگان و برچسب عباراتِ حاصل از تجزیه سطحی زبان، مورد استفاده قرار گرفته است [24]. اگرچه استفاده از اطلاعات زبان‌شناختی می‌تواند نقش مثبتی در تشخیص موجودیت‌های نامدار داشته باشد؛ اما این اطلاعات باعث وابستگی کامل سامانه به زبان می‌شود. «پیکره موجودیت‌های نامدار پیما» محصول این پژوهش است. برچسب‌های درنظرگرفته‌شده در این پیکره نیز همانند پیکره آرمان محدود به شش مورد است و قابلیت استخراج بسیاری از اطلاعات مرتبط با موجودیت‌های نامدار را ندارد.

۳- شبکه‌های عصبی

به‌طورکلی از شبکه‌های عصبی در تشخیص موجودیت‌های نامدار در دو بخش اصلی استفاده می‌شود: ۱) در بازنمایی کلیه واژگان متن، ۲) در تعیین برچسب واژگان. در این مقاله، برای هر دو مرحله یادشده از شبکه‌های عصبی استفاده می‌شود.

۳-۱- شبکه‌های عصبی برای بازنمایی واژگان

برای بسیاری از روش‌های پردازش متن، نیاز به نمایش عددی واژگان و متون است تا بتوان از انواع روش‌های عددی حوزه یادگیری ماشین مانند بیش‌تر الگوریتم‌های دسته‌بندی روی واژه‌ها و اسناد استفاده کرد. یکی از ره‌یافت‌هایی که در این حوزه بسیار رایج شده، نمایش برداری واژگان و جملات است.

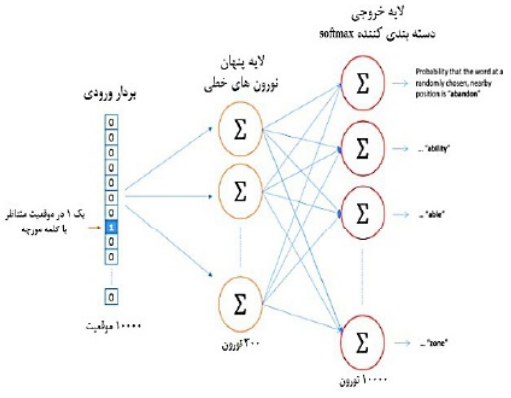
فرض کنید، فرهنگ لغتی با N واژه داریم که به‌ترتیب الفبا مرتب شده‌اند و هر واژه یک مکان مشخص در این فرهنگ لغت دارد؛ حال برای نمایش هر واژه، برداری را در نظر می‌گیریم با طول N که هر خانه آن، متناظر با یک لغت در فرهنگ لغت است که برای راحتی کار فرض می‌کنیم، شماره آن خانه بردار، همان اندیس لغت مربوطه در این فرهنگ لغت خواهد بود؛ با این پیش‌فرض، برای هر لغت یک بردار به طول N داریم که همه خانه‌های آن به‌جز خانه متناظر با آن واژه صفر خواهد بود؛ در خود ستون متناظر با واژه بسامد آن ذخیره خواهد شد [25].

اگرچه روش بالا کاربرد زیادی در پردازش زبان طبیعی داشته امّا دارای نواقص زیادی است. یکی از نواقص اصلی این روش بازنمایی عدم درنظرگرفتن ارتباط میان واژگان است. برای رفع این مشکل استفاده از روش‌های مبتنی بر شبکه عصبی پیشنهاد شده است.

برای ایجاد بردارهای بازنمایی واژگان بر مبنای شبکه عصبی از الگوریتم word2vec استفاده شده است که این الگوریتم توسط تیمی از پژوهش‌گران به رهبری توماس میکولوف[1] در گوگل ایجاد شد [26]. این الگوریتم سپس توسط سایر پژوهش‌گران مورد تجزیه و تحلیل قرار گرفت. استفاده از الگوریتم word2vec دارای مزایای بسیاری در مقایسه با الگوریتم‌های قبلی مانند تجزیه و تحلیل معنایی نهان است.

word2vec گروهی از مدل‌های مرتبط است که برای تولید بازنمایی واژگان استفاده می‌شود[2]. این مدل‌ها مبتنی بر شبکه‌های عصبی دارای دو لایهٔ کم‌عمق هستند که برای بازنمایی محتوای واژگان زبان آموزش داده می‌شوند. در تبدیل واژه به بردار به‌عنوان ورودی، یک متن با حجم بالا در نظر گرفته می‌شود و یک فضای برداری را که به‌طور معمول از چندصد قطعه تشکیل شده تولید می‌کند. با هر واژه منحصربه‌فرد در قسمت پیکره، یک بردار در فضا ایجاد می‌شود. بردارهای واژه در فضای بردار قرار می‌گیرند؛ به‌طوری‌که واژه‌هایی که دارای محتوای مشترکی در پیکره هستند، بردار آنها در فضا در نزدیکی یکدیگر قرار می‌گیرند [26].

در شکل (۱) معماری شبکه عصبی مورد استفاده در این روش نمایش داده شده است [27].



(شکل-۱): نمونه‌ای از معماری شبکه عصبی بازنمایی واژگان

(Figure-1): Neural network architecture for word representation

[1] Tomas Mikolov

[2] مدل‌هایی مانند پرش نکاشت و کیسه واژگان پیوسته ، همچنین نسخه گسترش داده شده از آنها از جمله این مدل‌ها هستند.

در ساختار این شبکه هیچ تابع فعال‌سازی غیرخطی در نورون‌های پنهان وجود ندارد؛ اما نورون‌های خروجی از تابع سافت مکس[3] استفاده می‌کنند. این شبکه بر روی دو واژه آموزش داده می‌شود، به‌نحوی که ورودی یک بردار یک-روشن[4] نشان‌گر واژه ورودی و خروجی آموزش نیز یک بردار یک-روشن نشان‌گر واژه خروجی است؛ اما هنگامی که شبکه آموزش‌دیده روی یک واژه ورودی ارزیابی می‌شود، بردار خروجی درواقع یک توزیع احتمالاتی است.

همچنین با ایجاد تغییراتی در الگوریتم بردارهای بازنمایی واژگان می‌توان از ویژگی‌های حروف واژگان نیز بهره گرفت و برای هر واژه برداری از ویژگی‌های حروف آن را استخراج کرد که به آن تعبیه واژگان در سطح نویسه[5] گفته می‌شود [28].

در این راستا، از الگوریتم CharWNN استفاده می‌شود که معماری شبکه عصبی را برای طبقه‌بندی تکراری گسترش می‌دهد و با اضافه‌کردن یک لایه پیچش[6] برای استخراج نمایه‌های سطوح حروف مورد استفاده قرار می‌گیرد. با توجه به جملهٔ داده‌شده، شبکه به‌ازای هر واژه امتیازی برای هر واژه در بافت آن واژه در نظر می‌گیرد و بنا بر امتیازات یک واژه، شبکه به‌عنوان ورودی یک پنجره با اندازه ثابتی از واژگان متمرکز در واژه هدف را می‌گیرد. ورودی از طریق دنباله‌ای از لایه‌ها که ویژگی‌هایی با افزایش سطوح پیچیده دارند، انتقال می‌یابد؛ سپس خروجی برای همه جملات با استفاده از الگوریتم ویتربی پردازش می‌شود [28].

۲-۳- شبکه‌های عصبی برای برچسب‌زنی متن

همان‌طورکه گفته شد، برای برچسب‌زنی موجودیت نامدار در بین شبکه‌های عصبی، شبکه عصبی بازگشتی و به‌طور خاص شبکه عصبی حافظه طولانی کوتاه‌مدت دو طرفه[7] بیشتر مورد استفاده قرار می‌گیرد که در این قسمت به توضیح آنها پرداخته می‌شود.

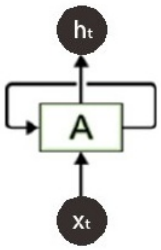
نحوه تفکر انسان‌ها در هرلحظه به‌گونه‌ای است که از اطلاعات گذشتهٔ خود بهره می‌برند. به‌عنوان مثال، هنگام مطالعه یک مطلب، معنی هر واژه را می‌توان با توجه به دانشی که از خواندن واژه‌های قبلی کسب‌شده درک کرد.

[3] Soft-max
[4] One-hot
[5] character level word embedding
[6] Convolution
[7] Bidirectional LSTM

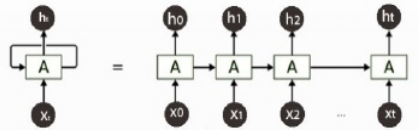
به‌عبارتی دیگر، هنگام مطالعه یک متن، درک و فهمی که در مورد آن متن با توجه به خواندن واژگان قبل کسب شده است از بین نمی‌رود؛ بلکه به‌صورت پیوسته با خواندن هر واژهٔ جدید، نسبت به آن متنی که خوانده می‌شود، درک بیشتری کسب می‌شود.

شبکه‌های عصبی متداولی که متخصصان یادگیری ماشین از آنها استفاده می‌کردند، نمی‌توانستند به این شیوه شبیه انسان عمل کنند و این یک نقص بزرگ برای این شبکه‌ها محسوب می‌شد [29].

برای برطرف‌کردن این مشکل شبکه‌های عصبی بازگشتی طراحی شدند. این شبکه‌ها در داخل خود دارای یک حلقه بازگشتی هستند که منجر می‌شود اطلاعاتی که از لحظات قبلی به‌دست‌آمده از بین نرود و در شبکه باقی بماند. در شکل (۲)، بخش A، مقدار x_t به‌عنوان ورودی دریافت و مقدار h_t را به خروجی می‌برد. حلقه، باعث می‌شود که اطلاعات از یک مرحله به مرحله بعد ارسال شود[1]. درواقع شبکه‌های عصبی بازگشتی را می‌توان به صورت چندین رونوشتِ یکسان از یک شبکه عصبی در نظر گرفت که هر کدام اطلاعاتش را به شبکه بعدی منتقل می‌کند. در شکل (۳)، وضعیت شبکه عصبی بازگشتی در صورت بازکردن حلقه نمایش داده شده است.



(شکل-۲): شبکه‌های عصبی بازگشتی دارای حلقه
(Figure-2): Recursive neural networks with a ring



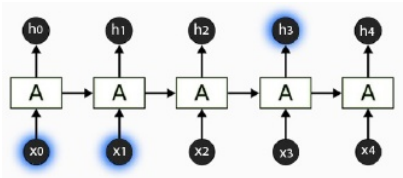
(شکل-۳): شبکه عصبی بازگشتی باز شده
(Figure-3): Expanded recursive neural network

با توجه به زنجیره‌واربودن شبکه‌های عصبی بازگشتی، می‌توان تشخیص داد که این شبکه‌ها در حد زیادی به دنباله‌ها و فهرست‌ها مرتبط هستند. در حقیقت شبکه‌های عصبی بازگشتی نخستین انتخاب برای کار با چنین داده‌هایی است.

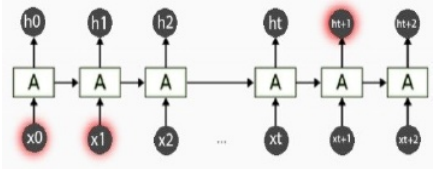
این شبکه‌ها در سال‌های اخیر مکرراً مورد استفاده قرار گرفته‌اند که منجر به موفقیت‌های بسیار چشم‌گیری در حوزه‌های مختلف از جمله تشخیص صدا، مدل‌کردن زبان، ترجمه، درج خودکار توضیح برای تصویر و... شده است [29].

اگرچه به‌نظر می‌رسد در بسیاری مواقع برای به‌دست‌آوردن اطلاعات مورد نیاز، تنها مراجعه به اطلاعات گذشتهٔ نزدیک کافی است، در برخی موارد اطلاعات بیشتری مورد نیاز است. به‌عنوان مثال، با درنظرگفتن جمله "من زبان فرانسه را خیلی راحت صحبت می‌کنم... من متولد کشور فرانسه هستم." با توجه به اطلاعات اخیر (یعنی چهار پنج واژه قبل از واژه فرانسه)، می‌توان گفت که این واژه به‌احتمال اسم یک کشور است، ولی اگر مدل بخواهد به‌طوردقیق نام کشور را تشخیص دهد، نیازمند اطلاعات دورتر (یعنی تا ده یا بیست واژه قبل از آخرین واژه) است. به بیان دیگر، ممکن است، فاصله بین اطلاعات مرتبط و قسمتی که به این اطلاعات نیاز داریم، زیاد باشد.

متأسفانه، هر چه این فاصله افزایش پیدا می‌کند، شبکه‌های عصبی بازگشتی قدرت‌شان را در به‌یادآوردن و استفاده از اطلاعاتی که در گذشته دورتر یاد گرفته‌اند، از دست می‌دهند و به‌عبارتی توانایی استفاده از اطلاعات گذشتهٔ دورتر را ندارند. شکل‌های (۴ و ۵) به نمایش این مشکل پرداخته‌اند. برای حل این مشکل شبکه‌های عصبی حافظه طولانی کوتاه‌مدت پیشنهاد شده‌اند [30].



(شکل-۴): امکان استفاده شبکه عصبی بازگشتی از اطلاعات گذشتهٔ نزدیک
(Figure-4): The possibility of using short past information in a recursive neural network

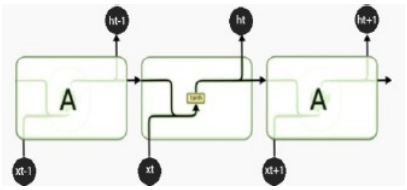


(شکل-۵): مشکل شبکه عصبی بازگشتی در به یادآوردن اطلاعات گذشتهٔ دور
(Figure-5): The problem of recurring neural network in recalling long past information

[1] شکل‌های (۲-۷) از آدرس زیر مورد استفاده قرارگرفته‌اند:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

شبکه‌های حافظه طولانی کوتاه‌مدت نوع خاصی از شبکه‌های عصبی بازگشتی هستند که توانایی یادگیری وابستگی‌های بلندمدت را دارند. این شبکه‌ها برای نخستین‌بار توسط هاچریتر[1] و اسمیدهابر[2] در سال ۱۹۹۷ معرفی شدند [31]. البته تعداد زیادی از پژوهش‌گران در بهبود این شبکه‌ها نقش داشتند.

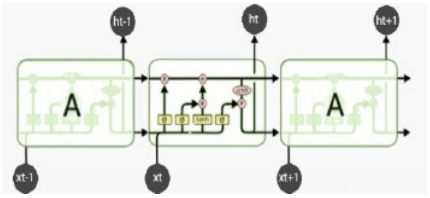
هدف از طراحی شبکه‌های حافظهٔ طولانی کوتاه‌مدت، حل‌کردن مشکل وابستگی بلندمدت بود. به‌عبارت دیگر، به‌یادسپاری اطلاعات برای بازه‌های زمانی بلندمدت، رفتار پیش‌فرض و عادی شبکه‌های حافظه طولانی کوتاه‌مدت است؛ و ساختار آنها به‌صورتی است که اطلاعات خیلی دور را به‌خوبی یاد می‌گیرند که این ویژگی در ساختار آنها نهفته است [32].



(شکل-۶): ماژول‌های تکرارشونده در شبکه‌های عصبی بازگشتی استاندارد فقط دارای یک لایه هستند.

(Figure-6): Routine modules in a standard recursive neural network only have one layer

شبکه‌های حافظه طولانی کوتاه‌مدت ساختار دنباله‌دار یا زنجیره‌مانندی دارند؛ ولی بخش تکرارشونده، ساختار متفاوتی دارد که به‌جای داشتن تنها یک لایه شبکه عصبی، چهار لایه دارند و طبق ساختار ویژه‌ای با یکدیگر در تعامل و ارتباط هستند. تفاوت ساختار داخلی دو شبکه در شکل‌های (۶ و ۷) نمایش داده شده است.



(شکل-۷): ماژول‌های تکرار شونده در LSTMها دارای ۴ لایه‌اند که با همدیگر در تعاملند

(Figure-7): Repetitive modules in LSTMs have 4 layers that interact with each other

همچنین نوع توسعه‌یافته شبکه‌های حافظه طولانی کوتاه‌مدت شبکه‌های حافظه طولانی کوتاه‌مدت دوطرفه هستند که علاوه‌بر آن که وابستگی دورتری را در دنباله واژگان در نظر می‌گیرند، به اطلاعات هر دو جهت دسترسی دارند که باعث بهبود عملکردشان شده است. در این مقاله نیز با توجه به ویژگی یادشده از این نوع شبکه‌های عصبی بازگشتی استفاده می‌شود.

۴- مدل پیشنهادی

به‌طورکلی "سامانه تشخیص موجودیت‌های نامدار زبان فارسی" در بخش‌های زیر قابل بررسی و پیاده‌سازی است:

در این سامانه یک متن به‌عنوان ورودی ارائه می‌شود و اسامی به تفکیک طبقهٔ مربوط به آن به‌عنوان خروجی سامانه به‌دست می‌آید. مراحل لازم تشخیص موجودیت‌های نامدار یک متن عبارتند از:

۱- متن ورودی به سامانه ارائه می‌شود، مانند جملهٔ: " محمد جمعه به دانشگاه نرفت ."

۲- با استفاده از یک نشان‌گذار، واژگان در متن ورودی از یکدیگر جدا می‌شوند. (محمد/جمعه /به/ دانشگاه /نرفت)

۳- با استفاده از روش یادگیری بدون‌نظارت[3] مبتنی بر شبکه‌های عصبی، برای هر یک از واژگان بردار ویژگی مورد نیاز بر مبنای واژه و همچنین حروف تشکیل‌دهندهٔ واژه استخراج می‌شود.

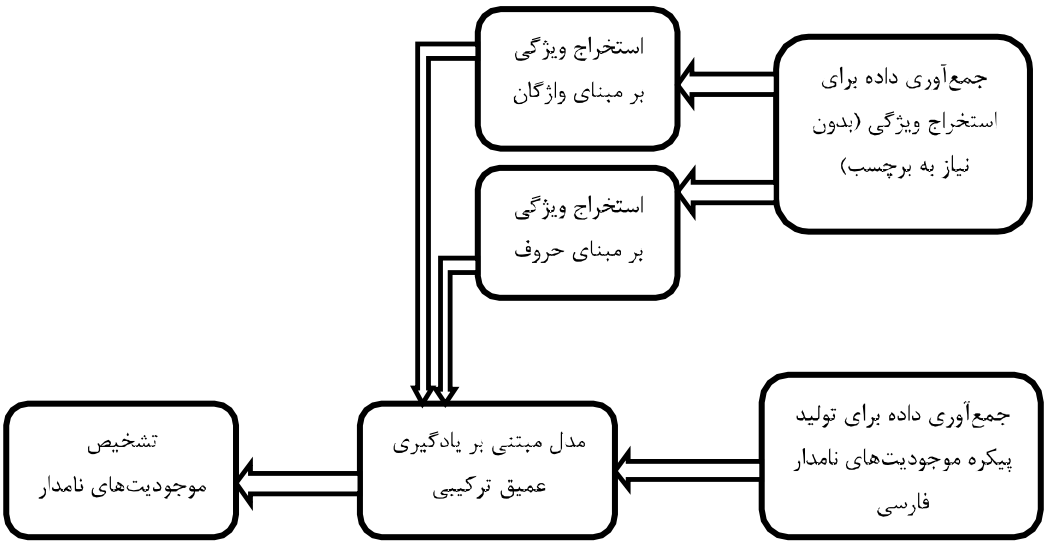
۴- با استفاده از روش یادگیری بانظارت مبتنی‌بر شبکه‌های عمیق و دادگان آموزش، برچسب‌دهی دادگان آزمون انجام می‌شود.

۵- معیار ارزیابی دقت[4]، صحت[5]، فراخوانی[6] و معیار F[7] برروی خروجی سیستم محاسبه می‌شود.

همان‌طور که گفته شد، برای انجام این پروژه از الگوریتم شبکه عصبی حافظه طولانی کوتاه‌مدت دوطرفه استفاده شده که برگرفته از شبکه‌های عصبی بازگشتی است و می‌تواند اطلاعات گذشته دورتری از بردارها را در حافظه خود نگه دارد. همچنین با بهره‌گیری از الگوریتم word2vec دو بردار تعبیه واژگان در سطح واژه و در سطح نویسه ایجاد شده‌اند که بردار بازنمایی واژگان و بردارهای بازنمایی بر مبنای حروف متشکل هر واژه هستند. در این مقاله، میزان دقت و کارایی روش با استفاده از این دو بردار نیز مورد بررسی و مقایسه قرار می‌گیرد. در شکل (۸) ساختار کلی تشخیص موجودیت‌های نامدار در پژوهش حاضر نمایش داده شده است.

[1] Hochreiter
[2] Schmidhuber

[3] Unsupervised
[4] Accuracy
[5] Precision
[6] Recall
[7] F-Measure



(شکل-۸): ساختار روال کلی انجام پروژه

(Figure-8): Overview of the overall project execution

۵- دادگان

برای یادگیری مدل، نیازمند پیکره‌ای برای آموزش شبکه عصبی هستیم. از جمله دادگان موجود در زمینه تشخیص موجودیت‌های نامدار در زبان فارسی دو پیکرهٔ موجویت‌های نامدار آرمان رایان شریف و بانک درختی هسته‌بنیان فارسی هستند [33,34]. در این دو منبع تنها موجودیت‌های نامدار مربوط به نام اشخاص، موقعیت‌های مکانی و اسامی مربوط به سازمان‌ها برچسب خورده‌اند. در پیکره نخست اسامی رویدادها، امکانات و محصولات نیز مشخص شده است؛ این در حالی است که در بسیاری از کاربردهای تشخیص موجودیت‌های نامدار، مانند سامانه‌های پرسش و پاسخ و یا تحلیل اخبار و ...، نیازمند پوشش موجودیت‌های نامدار بیشتری مانند زبان‌ها، ملیت‌ها، رخدادها، مشاغل، کتاب‌ها، اسامی فیلم‌ها، تاریخ‌ها، مذاهب، زمینه‌های علمی و دانش‌ها، روزنامه‌ها و سایر اسامی خاص در زبان هستیم.

ازآنجاکه در پیکره‌های یادشده این موارد پوشش داده نشده است، برای طراحی یک سامانه تشخیص موجودیت‌های نامدار جامع، در راستای پژوهش حاضر، پیکره‌ای از دادگان مربوط به ویکی‌پدیا جمع‌آوری شده و پانزده برچسب موجودیت‌های نامدار یادشده برای حدود سه‌هزار چکیده استخراج و برای این کار به صفحات متنوعی از ویکی‌پدیا مراجعه شده است.

جمع‌آوری داده یکی از اصلی‌ترین بخش‌ها در یک کار پژوهشی است. در صورت تهیهٔ دادهٔ با کیفیت بالا، می‌توان تجزیه و تحلیل و نتیجه‌گیری از داده‌ها را با سرعت و دقت بالاتری انجام داد؛ درنتیجه، در این مقاله توجه خاصی به کیفیت تحلیل داده شده و تا جای ممکن دادگان از بهترین و متنوع‌ترین داده‌های ممکن جمع‌آوری شده است. در ادامه این بخش، به‌صورت اجمالی در مورد نحوه برچسب‌زنی داده‌ها می‌پردازیم.

۵-۱- نمایش مدل‌های IO و IOB

برای اختصاص برچسب به هر واژه، ابتدا متن را به واژگان سازندهٔ آن تقطیع کرده و سپس به‌کمک دو روش می‌توان نوع موجودیت نامدار یافته‌شده را معین کرد. در روش نخست به‌ازای هر اسم خاص موجود میان واژگان متن، برچسب متناسب با نوع اسم خاص مربوطه را به آن اختصاص می‌دهیم و به‌ازای سایر واژگانی که اسم خاص تشخیص داده نشده‌اند برچسب O را که مخفف واژه Out است، تخصیص می‌دهیم. به این روش برچسب‌گذاری مدل IO گفته می‌شود. که در جدول (۱) نمونه‌ای از آن مشخص شده است.

در روش دوم، کار برچسب‌گذاری را با دقت بالاتری انجام می‌دهیم؛ بدین صورت که با علائمی ابتدا و انتهای هر واژه خاص را که ممکن است، دنباله‌دار باشد، معین می‌کنیم. برای این منظور، شروع هر واژه خاص را با اضافه‌کردن حرف b به ابتدای برچسب آن مشخص و برای نمایش دنباله واژه مورد نظر، برچسب ادامه آن را با افزودن حرف i به آن تعیین می‌کنیم که به این روش برچسب‌گذاری موجودیت‌های نامدار مدل IOB می‌گوییم [35]. مثالی از این مدل در جدول (۲) آورده شده است.

(جدول ۱): نمونه برچسب‌گذاری با مدل IO
(Table-1): Labeling with IO model

واژگان	برچسب موجودیت نامدار
ملا	PER
محمدمحسن	PER
فیض	PER
کاشانی	PER
۹۷۷–۱۰۵۸	DTE
هجری	DTE
شمسی	DTE
حکیم	JOB
،	O
فیلسوف	JOB
و	O
عارف	JOB
دوره	DTE
صفوی	DTE
و	O
از	O
دانشمندان	PEG
شیعه	REL
است	O
.	O

(جدول–۲): نمونه برچسب‌گذاری با مدل IOB
(Table-2): Labeling with IOB model

واژگان	برچسب موجودیت نامدار
ملا	b-PER
محمدمحسن	i-PER
فیض	i-PER
کاشانی	i-PER
۹۷۷–۱۰۵۸	b-DTE
هجری	i-DTE
شمسی	i-DTE
حکیم	b-JOB
،	O
فیلسوف	b-JOB
و	O
عارف	b-JOB
دوره	b-DTE
صفوی	i-DTE
و	O
از	O
دانشمندان	b-PEG
شیعه	b-REL
است	O
.	O

اگرچه روش IOB در برچسب‌گذاری موجودیت‌های نامدار دقیق‌تر است، پژوهش‌های انجام‌شده بر روی زبان‌هایی مانند انگلیسی، چک، اسپانیایی و هلندی [36] نشان داده است که تفاوت معناداری بین نتایج به‌دست‌آمده از این دو روش وجود ندارد؛ اما با توجه به ایراد مهمی که به مدل برچسب‌گذاری IO به‌دلیل عدم تشخیص مرز بین موجودیت‌های نامداری که پشت سر هم در متن ظاهر می‌شود، وارد است، روش IO برای زبان فارسی توصیه نمی‌شود. به‌عنوان نمونه در مثال جدول (۳) دو موجودیت نامدار محمد علی صادقی و فرهاد در مدل IO قابل تمایز نیستند. یک دلیل عمده برای این اشکال این است که ساختار دستور زبان فارسی به‌صورت فاعل–مفعول–فعل است. درنتیجه میزان رخداد این‌گونه ابهامات در متن بسیار بالاتر از سایر زبان‌ها است؛ زیرا موجودیت‌های نامدار زیادی در قالب فاعل و مفعول پشت سر هم در جمله ظاهر می‌شوند؛ در حالی‌که در سایر زبان‌ها مانند انگلیسی ساختار دستوری زبان به‌صورت فاعل–فعل–مفعول است که درصد رخداد باهم‌آیی موجودیت‌ها را کاهش می‌دهد. این امر به لزوم استفاده از مدل IOB در زبان فارسی تأکید دارد. بنابراین برای رسیدن به دقت بالا در کار تشخیص موجودیت‌های نامدار، در مقاله حاضر، از روش برچسب‌گذاری IOB استفاده شده است.

(جدول–۳): مثالی از رفع ابهام برچسب‌گذاری به روش IOB
(Table-3): An example of removing ambiguity with IOB labeling

واژگان	برچسب موجودیت نامدار IOB		واژگان	برچسب موجودیت نامدار IO
محمد	**b-PER**		محمد	**PER**
علی	**i-PER**		علی	**PER**
صادقی	**i-PER**		صادقی	**PER**
فرهاد	**b-PER**		فرهاد	**PER**
را	**O**		را	**O**
شاعر	**b-JOB**		شاعر	**JOB**
دانسته	**O**		دانسته	**O**
است	**O**		است	**O**
.	**O**		.	**O**

۲-۵- مجموعه برچسب‌های پیشنهادی

واژگان گردآوری‌شده از ویکی‌پدیای فارسی برای آموزش سامانه نیازمند برچسب‌گذاری با نمادهایی است که هر کدام به یک نوع از موجودیت‌های نامدار اشاره دارد. تعداد کل برچسب‌های استفاده‌شده در دادگان ۳۱ برچسب است که برای مشخص‌کردن پانزده نوع موجودیت متفاوت که شامل اسم شخص مفرد، اسم شخص جمع، موقعیت مکانی، اسم سازمان، زبان، ملیت، رخداد، شغل، کتاب، اسم فیلم، تاریخ، مذهب، عنوان علمی و دانش، روزنامه و سایر اسامی خاص یادنشده در زبان فارسی است، استفاده شده است. همچنین برای واژگانی که جزو موجودیت‌های نامدار نیستند، نیز علامتی در نظر گرفته شده است. گفتنی است این دادگان نخستین دادگان موجودیت‌های نامدار زبان فارسی با این تعداد تنوع در انواع موجودیت است. با توجه به این‌که استخراج اطلاعات از متن [37] و همچنین سامانه‌های پرسش و پاسخ [38] از جمله مهم‌ترین کاربردهای شناسایی موجودیت‌های نامدار است، درنظرگرفتن تنوع بیشتر در نوع موجودیت‌های نامدار کمک به‌سزایی در بهبود کیفیت و کارایی این سامانه‌ها خواهد داشت که این مهم با به‌کارگیری دادگان بالا محقق می‌شود.

(جدول-۴): راهنمای برچسب واژگان
(Table-4): Vocabulary label Guide

برچسب	تعریف انگلیسی برچسب	تعریف فارسی برچسب	توضیحات	مثال
O	Out	هیچ	واژه مورد نظر از موجودیت‌های نامدار نیست.	ممکن کوتاه اوقات ثبت‌نام
PE1	Person Individual	شخص مفرد	این علامت به نام شخص اشاره دارد.	سلطان محمود غزنوی حسن روحانی
PEG	person Group	اسم خاص گروه	این علامت به موجودیت نامداری اشاره دارد که به صورت جمع آمده است.	اعراب شعبان صفاریان
LOC	Location	موقعیت مکانی	واژه مورد نظر این برچسب به موقعیت مکانی خاصی اشاره دارد.	ایران اراک استان مرکزی
ORG	Organization	سازمان	اسامی سازمان‌ها و مؤسسات مختلف با این برچسب نشان داده می‌شوند.	فرهنگستان زبان و ادب فارسی وزارت آموزش و پرورش نیروی انتظامی بانک مرکزی ایران
LAN	Language	زبان	این برچسب نشان‌دهنده زبان‌های مختلف است.	فارسی یونانی عربی
NAT	Nationality	ملیت	ملیت‌های مختلف با این علامت تعیین می‌شوند.	ایرانی یونانی
EVN	Events	رخدادها	رخدادها و وقایع خاص را با این برچسب نشان دادیم.	جام جهانی ۲۰۱۸ جنگ جهانی دوم جشنواره فیلم فجر
JOB	Job	شغل	این برچسب نمایان‌گر مشاغل است.	اخترشناس نویسنده فیلسوف معمار
BOK	Book	کتاب	اسامی کتاب‌ها با این برچسب مشخص می‌شوند.	کتاب تاریخ بلعمی لغت‌نامه دهخدا منظومه ویس و رامین
FLM	Film	فیلم	نام فیلم‌های مختلف با این برچسب تعیین می‌شوند.	فیلم کارتونی شیرشاه بادیگارد

DTE	Date	تاریخ	برای نشان دادن تاریخ و دوره‌های مختلف ازین برچسب استفاده شده است.	۱۰ محرم ۶۱ قمری سال ۱۳۷۰ قرن پنجم دوره هخامنشیان
REL	Religion	مذهب	این برچسب تعیین‌کننده مذاهب مختلف است.	اسلام کاتولیک زرتشت آیین بودا
FLD	Field	زمینه	زمینه‌ها و دانش‌های مختلف با این برچسب تعیین گردیده است.	اقتصادی زمین‌شناسی ریاضی
MAG	Magazine	روزنامه و مجله	اسامی روزنامه‌ها و مجلات با این برچسب آمده است.	روزنامه صوراسرافیل روزنامه گاردین تایمز
OTH	Other	سایرین	اگر واژه‌ای به‌عنوان موجودیت نامدار باشد، ولی در بین موجودیت‌های معرفی‌شده در بالا نباشد با این برچسب مشخص می‌شود.	هیدروژن عطارد ساکسوفون سیستم عامل لینوکس

در این مقاله، برای تعیین اسم خاص اشخاص از دو برچسب مختلف استفاده شده است که یکی برای تعیین اسامی خاص مفرد و معمول به‌کار می‌رود و با علامت PEI[1]مشخص شده است و دیگری برای اسامی خاص که به‌صورت جمع استفاده می‌شوند، کاربرد دارد. این موارد با برچسبPEG[2] مشخص شده است.

اسم شخص جمع به تمامی اسامی‌ای گفته می‌شود که اسم مفرد آن نوعی موجودیت باشد. به‌عنوان مثال واژگان مسلمان، معلم و ایرانی به‌ترتیب برچسب‌های *مذهب*، *شغل* و *ملیت* می‌گیرند و اسم جمع این واژگان یعنی مسلمانان، معلمان و ایرانیان PEG محسوب می‌شوند. همین‌طور واژه ابوالفضل بلعمی دارای برچسب PEI است و به‌طبع آن واژگان خاندان بلعمی و یا بلعمیان PEG هستند.

گفتنی است درنظرگرفتن اسامی مشاغل، مذاهب، اسم شخص جمع و ... از چالش‌های اصلی این پژوهش بوده است. اگرچه در برخی از دادگان این موارد لحاظ نمی‌شود، اما برای وسعت‌بخشیدن به کاربرد تشخیص موجودیت‌های نامدار در مواردی مانند سامانه‌های استخراج اطلاعات و یا سامانه‌های پرسش و پاسخ درنظرگرفتن این موارد مورد نیاز است تا حجم بیشتری از اطلاعات تحت پوشش این سامانه‌ها قرار گیرند.

جدول (۴) به تفضیل به‌شرح و بررسی برچسب‌های استفاده‌شده برای دادگان می‌پردازد.

برای نمونه، مثال‌هایی از جملات برچسب‌خورده در دادگان براساس برچسب‌های معرفی‌شده در جداول (۵ و ۶) نمایش داده شده‌اند.

(جدول-۵): مثال اول از جمله برچسب خورده دادگان

(Table-5): First example from tagged data

برچسب‌ها	واژگان
کتاب	b-BOK
ویس	i-BOK
و	i-BOK
رامین	i-BOK
از	O
شاهکارهای	O
ادب	O
فارسی	b-LAN
و	O
اثر	O
فخرالدین	b-PEI
اسعد	i-PEI
گرگانی	i-PEI
شاعر	b-JOB
قرن	b-DTE
پنجم	i-DTE
هجری	i-DTE
است	O

[1] Person Individual
[2] person Group

(جدول-۶): مثال دوم از جمله برچسب خورده دادگان
(Table-6): second example from tagged data

واژگان	برچسب‌ها
محمد	b-PEI
قاضی	i-PEI
اسداللهی	i-PEI
)زاده	O
مهر	b-DTE
۱۳۰۳	i-DTE
در	O
تهران	b-LOC
-	O
درگذشته	O
۲۰	b-DTE
آذر	i-DTE
1382	i-DTE
در	O
تهران)	b-LOC
هنرمند	O
ایرانی،	b-NAT
نقاش	b-JOB
)پرتره(	O
مجسمه‌ساز	b-JOB
و	O
خوشنویس	b-JOB
معاصر	O
است	O

آمار تعداد عبارتی که برچسب موجودیت‌های نامدار را به خود تخصیص داده‌اند در جدول (۷) گزارش شده است. گفتنی است این آمار در سطح عبارت است و در سطح واژه تعداد بالاتر است. به‌عنوان مثال تعداد عبارتی که برچسب سازمان دارند در جدول ۳۱۷ گزارش شده است، اما با توجه به اینکه بسیاری از عبارات از بیش از یک واژه تشکیل شده‌اند این تعداد عبارت معادل ۹۰۸ واژه است. درمجموع تعداد عبارات دارای برچسب موجودیت نامدار ۱۰۲۷۷ و تعداد کل واژه‌های دارای برچسب ۱۹۴۲۶ است. تعداد کل واژه‌های بدون برچسب موجودیت نیز ۷۳۸۷۶ است.

۳-۵- داده‌های ویکی‌پدیا

ویکی‌پدیا دانش‌نامه‌ای همگانی و آزاد است؛ بدین معنی که همه می‌توانند به نوشتن و ویرایش نوشتارهای موجود در آن بپردازند. البته این نوشتارها و ویرایش‌ها باید مطابق با اساس‌نامهٔ ویکی‌پدیا باشد؛ یعنی مطالب بی‌طرفانه و بدون پایمال‌کردن حق نشر دیگران نوشته شده باشند. مدیریت بررسی نوشتارها توسط خودِ کاربران انجام می‌شود. کسانی که در امر تکمیل این پروژه بی‌پایان مشارکت می‌کنند، به هم‌زبانان خود یاری می‌رسانده تا در گردآوری بی‌همتاترین دانشنامه جهانی سهمی داشته باشند.

(جدول-۷): آمار برچسب‌های پیکره تهیه شده
(Table-7): Statistics of the annotated dataset

برچسب	تعداد
PEI	۱۴۶۰
PEG	۵۸۸
LOC	۴۰۵۲
ORG	۳۱۷
LAN	۵۵۵
NAT	۳۳۶
EVN	۹۱
JOB	۴۴۹
BOK	۲۶۵
FLM	۱
DTE	۱۰۱۰
REL	۱۱۱
FLD	۷۰۳
MAG	۲۱
OTH	۳۱۷

ویکی‌پدیای فارسی دو سال پس از شروع پروژه ویکی‌پدیای انگلیسی، در ۲۸ آذر ۱۳۸۲ (۱۹ دسامبر ۲۰۰۳) فعالیت خود را آغاز کرد و اکنون در رده هفدهم ویکی‌پدیاها قرار دارد و بزرگترین ویکی‌پدیا در میان زبان‌های خاورمیانه و زبان‌های راست به چپ و بزرگترین دانشنامه فارسی محسوب می‌شود.

ویکی‌پدیا فارسی تا اول سال ۲۰۱۷ میلادی شامل مقالاتی است که توسط بسیاری از افراد در موضوعات مختلف منتشر شده است. پس از حذف مقالات با طول کمتر از پنجاه واژه، حدود ۳۶۱۰۰۰ سند در این دانشنامه موجود است و تعداد کل واژگان متمایز در این مجموعه حدود ۲۱۲۰۰۰ است [39].

کل دادگان یادشده برای آموزش الگوریتم word2vec مورد استفاده قرار گرفته است و از روی این دادگان بردار بازنمایی واژگان برمبنای واژه و حروف واژه ساخته شده است. همچنین برای فراهم‌کردن مجموعه دادگان آموزش مورد نیاز این پژوهش از داده‌های ویکی‌پدیای فارسی

استفاده شده است. بدین صورت که قسمت چکیده تعداد ۲۹۱۳ سند مختلف ویکی‌پدیا، به‌صورت تصادفی انتخاب شده و متن آنها به واژگان تشکیل‌دهنده آن شکسته و برای تک-تک واژگان برچسب‌های موجودیت نامدارشان معین می‌شود. تعداد دادهٔ برچسب‌گذاری‌شده در این پژوهش ۹۰۷۹۳ واژه است.

همان‌طور که پیش‌تر گفته شد، برای گردآوری دادگان برچسب‌گذاری‌شده از ویکی‌پدیای فارسی استفاده شده است، چون در این نوع داده ساختار جملات و متن به‌گونه‌ای است که اطلاعات مورد نیاز موجودیت‌های نامدار و اسامی خاص را بیشتر در بردارد، و یا به‌نحوی بیش‌تر به توضیح اطلاعات مرتبط با اسامی خاص می‌پردازد و به‌کمک آن دادگان غنی از موارد مختلف موجودیت‌های نامدار شناسایی می‌شود. همچنین علت انتخاب ویکی‌پدیا به‌عنوان دادگان مورد استفاده برای تعبیه واژگان نیز هم‌راستابودن نثر آن با داده برچسب‌گذاری‌شده است که براساس نتایج ارائه‌شده توسط هادیفر و ممتازی این هم‌خوانی نقش به‌سزایی در نتایج خواهد داشت [40].

۶- آزمایش‌های صورت‌گرفته

برای بررسی و ارزیابی مدل استفاده‌شده در این مقاله، از الگوریتم‌ها و روش‌هایی استفاده شده است که به توضیح آن می‌پردازیم. همان‌طورکه پیش‌تر به آن اشاره شد، ابتدا از الگوریتم word2vec برای به‌دست‌آوردن دو مدل بردار واژه به‌طول سیصد و بردار ویژگی حروف هر واژه به‌طول دویست استفاده شده است. طول این بردارها بر مبنای پیشنهادهای پروژه‌های مختلف پیشین که از این بازنمایی‌ها استفاده کرده بودند، انتخاب شده است [41]. در این قسمت واژگان مشابه و هم‌ریشه، بردارهای شبیه به هم خواهند داشت. این الگوریتم ازجمله روش‌های بدون‌نظارت است که برای آموزش از دادگان بدون برچسب ویکی‌پدیا استفاده کرده است؛ سپس از الگوریتم شبکه عصبی حافظه طولانی کوتاه‌مدت دوطرفه استفاده شده است. برای آموزش این شبکه از دادگان برچسب‌خورده و بردارهای مربوط به هر واژه و حروف که از قسمت قبلی به‌دست آمده است، استفاده می‌شود. بدین ترتیب، شبکه آموزش داده می‌شود و برای آزمایش شبکه عصبی پیاده‌سازی‌شده از داده‌های بدون برچسب و بردارهای به‌دست‌آمده مرحله نخست استفاده می‌کنیم.

با توجه به دراختیارداشتن حدود نودهزار واژه برچسب‌گذاری‌شده برای ارزیابی سامانه حاضر، از هفتادهزار واژه این پیکره برای آموزش مدل استفاده شده است. ده‌هزار واژه به‌عنوان دادهٔ اعتبارسنجی[1] (جهت یافتن مقادیر مناسب برای پارامترهای شبکه عصبی) و ده‌هزار واژه دادگان نیز جهت آزمون استفاده شده است.

۶-۱- ابزارهای استفاده‌شده

در راستای انجام این پژوهش، از زبان برنامه‌نویسی پایتون[2] و کتابخانه‌های مربوط به آن استفاده شده است. برای تبدیل واژگان به بردار (word2vec) از کتابخانه Gensim استفاده شده که از ابزارهای مدل‌سازی فضای بردار مبتنی بر پایتون هستند و توسط ریهرک و پتر[3] در سال ۲۰۱۰ به‌وجود آمدند[42]. همچنین Fasttext، یک کتابخانه قدرتمند C++ است که برای آموزش نمایندگی واژه استفاده شده است که بوجانسکی[4] و همکارانش آن را در سال ۲۰۱۶ ابداع کردند.

Gensim انتخاب خوبی است؛ زیرا اجازه می‌دهد مدل‌های مختلف در یک چارچوب مستقر شوند و آن‌ها را در فرمت سازگار با پیاده‌سازی word2vec اصلی ذخیره کنند. همچنین از FastText استفاده شده است که یک کتابخانه کارآمد برای یادگیری نمایش واژگان با توجه به ویژگی‌های صرفی آنها است [28].

برای اجرای الگوریتم حافظه طولانی کوتاه‌مدت دوطرفه نیز یک ساختار سه‌لایه با یکصد سلول ایجاد و همچنین از ابزارهای مختلفی در راستای پیاده‌سازی مدل استفاده شده است که برخی از آن‌ها عبارتند از tensorflow, zlib, python و scipy.

۶-۲- معیارهای ارزیابی

برای ارزیابی الگوریتم ارائه‌شده، از چهار معیار دقت[5]، صحت[6]، فراخوانی[7] و معیار F[8] استفاده شده است. این چهار معیار ازجمله معیارهای معتبر در ارزیابی روش‌های مختلف یادگیری ماشین هستند.

یکی‌از نکات مهم در ارزیابی الگوریتم‌های مختلف یادگیری درنظرگرفتن تقابل میان معیارهای صحت و فراخوانی است: با افزایش یکی از این دو معیار دیگری کاهش

[1] Validation Set
[2] python
[3] Rehurek and Petr
[4] Bojanowski
[5] Accuracy
[6] Precision
[7] Recall
[8] F-Measure

خواهد یافت. از طرفی برای محاسبه میانگین این دو معیار میانگین ریاضی مناسب نیست و اطلاعات کافی از آن استخراج نخواهد شد. به همین جهت از میانگین هارمونی یا معیار F برای ارزیابی بهتر استفاده می‌شود.

معیارهای بالا نیاز به محاسبه مقادیر چهارگانه جدول آشفتگی دارند که این مقادیر برای هر موجودیت بدین صورت محاسبه می‌شود: مقدار مثبت درست[1] عبارت است از تعداد دفعاتی که یک نمونه به‌صورت صحیح برچسب موجودیت مد نظر را دریافت کرده است. مقدار مثبت نادرست[2] عبارت است از تعداد دفعاتی که یک نمونه برچسب موجودیت را به‌اشتباه دریافت کرده است (خواه آن نمونه به هیچ وجه موجودیت نبوده، خواه موجودیت بوده اما از نوع دیگری). مقدار منفی درست[3] به تعداد دفعاتی اشاره دارد که یک نمونه که موجودیت نامدار نیست برچسب موجودیت نامدار را دریافت نکرده است و مقدار منفی نادرست[4] به تعداد دفعاتی اشاره دارد که یک نمونه می‌بایستی برچسب موجودیت هدف را دریافت می‌کرده اما به‌اشتباه یا به هیچ وجه به‌عنوان موجودیت تشخیص داده نشده یا نوع دیگری از موجودیت به آن تخصیص یافته است.

فرمول‌های محاسبه هر کدام از این چهار معیار یادشده در زیر آمده است:

$$Precision = \frac{TP}{TP + FP} \quad (۱)$$

$$Recall = \frac{TP}{TP + FN} \quad (۲)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (۳)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۴)$$

گفتنی است در محاسبات انجام‌شده، برچسب یک نمونه تنها درصورتی صحیح قلمداد می‌شود که علاوه‌بر تشخیص موجودیت نامدار، نوع آن نیز به‌صورت صحیح تشخیص داده شود.

[1] True Positive(TP)
[2] False Positive (FP)
[3] True Negative (TN)
[4] False Negative (FN)

۳-۶- نتایج نهایی آزمایش‌ها

نتایج به‌دست‌آمده از آزمایش‌های صورت‌گرفته برروی داده‌ها به‌کمک روش شبکه عصبی حافظه طولانی کوتاه‌مدت دوطرفه و با استفاده از روش ارزیابی‌ای که از هفتادهزار داده برای آموزش مدل، ده‌هزار داده برای اعتبارسنجی و ده‌هزار داده نیز جهت آزمون استفاده شده است در جدول (۸) نمایش داده شده است. ضمن اینکه آزمایش‌ها بر مبنای مدل اعتبارسنجی متقاطع[5] تکرار شده‌اند.

(جدول-۸): نتایج بر اساس استفاده از هفتادهزار داده برای آموزش

(Table-8)Results based on the use of 70000 data for training

	Word Embedding	Character Embedding
F-measure	70.53	72.09
Precision	73.28	74.74
Recall	67.98	69.63
Accuracy	91.58	92.23

همان‌طورکه در نتایج مندرج در این جدول مشاهده می‌شود، نتایج حاصل از استفاده از بردارword2vec با ویژگی حروف بهتر از بردار برمبنای واژه بوده است. این امر نشان می‌دهد، شباهت معنایی واژگان به‌تنهایی کافی نیست و درنظرگرفتن شباهت واژگان بر مبنای ویژگی‌های صرفی نیز از اهمیت ویژه‌ای برخوردار است.

برای تحلیل خطاهای مطرح در داده جدول آشفتگی نتایج نیز برای تمام برچسب‌های موجود استخراج شده است که در جدول (۹) نمایش داده شده است.

همان‌طورکه در نتایج این جدول مشخص است، در بیشتر موارد روش پیشنهادی برچسب صحیح را انتخاب کرده است. بیشترین خطاها در تشخیص شروع و میانه مکان‌ها (b-LOC, i-LOC) و شروع و میانه اسامی اشخاص (b-PEI, i-PEI) مشاهده می‌شود. یکی از دلایل این خطا پیچیدگی ساختار زبانی در تشخیص محل صحیح شروع عبارات اسمی است. به‌عنوان مثال عبارت "جهان پهلوان رستم" در داده‌های این پژوهش برچسب اسم شخص گرفته است. به این‌صورت که واژه "جهان" شروع عبارت و واژه‌های "پهلوان" و "رستم" میانه عبارت هستند؛ درحالی‌که مدل

[5] Cross-Validation

پیشنهادی تنها واژه "رستم" را به‌عنوان موجودیت تشخیص داده و به آن برچسب شروع عبارت داده است. و یا در جمله "آنتاریو از لحاظ جمعیت قویترین استان کانادا است"، برخلاف این‌که واژه "کانادا" به‌تنهایی یک عبارت اسمی مکان می‌باشد، اما روش پیشنهادی به‌اشتباه عبارت "استان کانادا" را به‌عنوان اسم مکان تشخیص داده است. در نتیجه برچسب صحیح شروع عبارت برای واژه "کانادا" به‌اشتباه به‌صورت میانه عبارت مشخص شده است. مشکل دیگری که در نتایج مشاهده می‌شود، عدم تشخیص برچسب موجودیت و پیش‌بینی برچسب O برای برخی موجودیت‌ها است. علت اصلی این مشکل عدم تعادل در داده‌های آموزشی میان داده‌های داری برچسب موجودیت و داده‌های غیر موجودیت است که از چالش‌های اصلی در حوزه تشخیص موجودیت‌های نامدار است.

(جدول-۹): جدول آشفتگی نتایج

(Table-9): Confusion Matrix of the Results

####	b-BOE	i-BOE	b-DOE	i-DOE	b-EVN	i-EVN	b-FLD	i-FLD	b-FLW	i-FLW	b-JOB	i-JOB	b-LAN	i-LAN	b-LOC	i-LOC	b-MAG	i-MAG	b-MAT	i-MAT	b-ORG	i-ORG	b-OTH	i-OTH	b-PEC	i-PEC	b-PEI	i-PEI	b-REL	i-REL	o	total
b-BOE	23	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	23	51
i-BOE	4	28	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	4	2	0	0	31	75
b-DOE	0	0	122	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	26	152
i-DOE	2	0	6	217	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	34	261
b-EVN	0	0	0	14	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	2	17
i-EVN	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	36
b-FLD	0	0	0	0	0	0	93	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	27	123
i-FLD	0	0	0	0	0	0	1	23	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	8	34
b-FLW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i-FLW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b-JOB	0	0	0	0	0	0	0	0	0	0	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	71
i-JOB	0	0	0	0	0	0	0	0	0	0	0	52	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	58
b-LAN	0	0	0	0	0	0	0	0	0	0	0	0	59	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	65
i-LAN	0	0	0	0	0	0	0	0	0	0	0	0	2	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	33
b-LOC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	526	25	1	0	0	0	0	1	0	0	0	0	1	0	0	0	48	617
i-LOC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	487	0	0	0	0	0	2	0	0	0	0	0	0	0	0	44	576
b-MAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
i-MAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b-MAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	1	0	0	0	0	0	0	0	0	0	0	1	51
i-MAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b-ORG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	1	0	0	0	0	0	0	0	0	9	37
i-ORG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	46	0	0	0	0	0	0	0	0	12	66
b-OTH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	7	13
i-OTH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	6	7
b-PEC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	3	0	0	0	0	9	71
i-PEC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	11	45
b-PEI	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	224	2	0	0	11	270
i-PEI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	245	0	0	1	310
b-REL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	10
i-REL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
o	1	1	7	5	2	0	23	21	0	0	3	5	1	1	48	44	0	0	1	0	9	12	7	6	9	11	17	28	3	0	9360	9823
total	30	32	140	226	30	28	121	54	0	0	64	65	33	617	576	3	0	51	37	66	13	1	71	45	270	310	10	1	9616	12707		

در بخش دیگری از آزمایش‌ها، به بررسی اهمیت حجم دادهٔ تهیه‌شده در این پژوهش پرداخته شده است. برای این بررسی، ارزیابی کارایی سامانه طراحی‌شده با تعداد کمتری داده آموزش تکرار شده است. به این معنا که به‌جای استفاده از هفتادهزار داده آموزش در یک مرحله، تنها چهل‌هزار واژه و در مرحله دیگر تنها بیست‌هزار واژه به‌عنوان داده آموزش در نظر گرفته شده است. جداول (۱۰ و ۱۱) حاصل این بررسی را نمایش می‌دهند.

(جدول-۱۰): نتایج بر اساس استفاده از چهل‌هزار داده برای آموزش

(Table-10): Results based on the use of 40,000 data for training

	Word Embedding	Character embedding
F-measure	65.33	68.37
Precision	69.11	72
Recall	61.93	65.09
Accuracy	90.32	90.72

(جدول-۱۱): نتایج بر اساس استفاده از بیست‌هزار داده برای آموزش

(Table-11): Results based on the use of 20,000 data for training

	Word embedding	Character embedding
F-measure	61.34	60.99
Precision	65.69	66.93
Recall	57.53	56.02
Accuracy	88.79	88.55

باتوجه‌به نتایج به‌دست‌آمده در دو جدول (۱۰ و ۱۱)، می‌توان به این نتیجه دست یافت که افزایش تعداد دادگان برچسب‌خورده و بالابودن حجم دادگان تولیدشده می‌تواند در تشخیص موجودیت‌های نامدار مؤثر باشد و دقت بالاتری را به‌دست آورد. این امر خود مبین این است که با قراردادن تلاش بیشتر بر روی مقوله برچسب‌دهی دادگان می‌توان دقت سامانه حاضر را باز هم افزایش داد.

در گام آخر ارزیابی سامانه، به بررسی تأثیر تنوع بالای برچسب‌ها در دقت سامانه تشخیص موجودیت‌های نامدار پرداخته شده است. برای بررسی این امر در دادگان یادشده، به جای پانزده برچسب موجودیت‌های نامدار، تنها از سه برچسب PER، LOC و ORG استفاده شده است و سایر برچسب‌ها حذف شده است. همچنین در ارزیابی از هفتادهزار داده به‌منظور آموزش شبکه، ده‌هزار داده برای اعتبارسنجی و ده‌هزار داده به‌عنوان آزمون استفاده شده و نتایج به‌دست‌آمده برای مدل مبتنی بر بردار حروف در جدول (۱۲) آماده است.

از مقایسه اعداد به‌دست‌آمده در جدول بالا با اعداد موجود در جدول (۸) می‌توان دریافت که افزایش تعداد برچسب‌های موجودیت نامدار سبب افزایش پیچیدگی مدل می‌شود و درنتیجه میزان دقت تشخیص موجودیت‌های نامدار در متون مورد پردازش کمتر می‌شود. به بیان دیگر با تعداد برچسب‌های کمتر می‌توان با دقت بالاتری موجودیت‌های متن را تشخیص داد. البته بدیهی است که سایر برچسب‌های متنوع استفاده‌شده در این دادگان از اهمیت به‌سزایی برخوردارند و در کاربردهای مختلف تشخیص موجودیت‌های نامدار این برچسب‌ها می‌توانند تأثیر مهمی در کارایی سامانه داشته باشند. درنهایت با استناد به این نتایج می‌توان نتیجه‌گیری کرد که بنا بر کاربرد مدل پیشنهادی حاضر در سامانه‌های مختلف پردازش زبان طبیعی می‌توان در مورد انتخاب ساختار برچسب‌گذاری دادگان تصمیم‌گیری مناسب داشت.

(جدول-۱۲): نتایج بر اساس استفاده از هفتادهزار داده با در نظرگرفتن سه برچسب PER، LOC و ORG

(Table-12): Results based on the use of 70000 data with the consideration of the three PER, LOC and ORG labels

	Character Embedding
F-measure	74.28
Precision	77.75
Recall	71.1
Accuracy	94.88

۷- نتیجه‌گیری

در مقاله حاضر، روشی برای تشخیص موجودیت‌های نامدار در زبان فارسی ارائه شده است که مبتنی بر تبدیل واژه به بردار و ایجاد بردارهای بازنمایی واژگان بر مبنای واژه و همچنین حروف واژه بر مبنای شبکه عصبی است. در روش پیشنهادی که بر مبنای الگوریتم شبکه عصبی حافظه طولانی کوتاه‌مدت دوطرفه عمل می‌کند، شبکه با دادگان برچسب‌خورده و بردارهای به‌دست‌آمده از الگوریتم word2vec آموزش داده می‌شود. نتایج حاصل از ارزیابی به‌کمک معیارهای دقت، صحت، فراخوانی و معیار F نشان می‌دهد، استفاده از روش بردار بازنمایی ویژگی حروف واژه دقت بالاتری را در بردارد که به بهبود روش تشخیص

موجودیت نامدار کمک می‌کند. بر اساس آزمایش‌های انجام‌گرفته مشاهده می‌شود که افزایش تعداد دادگان نقش به‌سزایی در بالابردن دقت سامانه دارد. همچنین بالابودن تعداد برچسب‌های مورد استفاده باعث افزایش پیچیدگی مدل می‌شود.

گفتنی است که در این مقاله گامی بزرگ در راستای تشخیص موجودیت‌های نامدار در زبان فارسی برداشته شده است؛ چرا که با برچسب‌زنی مجموعه سه‌هزار چکیده از پیکره ویکی‌پدیای فارسی که شامل نودهزار واژه است با مشخص‌کردن پانزده موجودیت مختلف در دادگان، به غنی‌سازی دادگان فارسی می‌پردازد.

با توجه به نامتوازن‌بودن داده آموزشی در تشخیص موجودیت‌های نامدار، استفاده از مدل‌های پیشرفته‌تر از جمله مدل‌هایی که در آن تمایل به تابع هزینه شبکه عصبی اضافه می‌شود از جمله کارهای آینده این پژوهش خواهد

۸- مراجع 8- References

[1] Y. Chen, T.A. Lasko, Q. Mei, J.C. Denny and H. Xu, "A study of active learning methods for named entity recognition in clinical text", *Journal of Biomedical Informatics*, 2015.

[2] Z.S. Abdallaha, M.Carmana and Gh. H7affari, "Multi-domain evaluation framework for named entity recognition tools",Computer Speech & Language, 2017.

[3] S. Hussain, J.J. K. and G.C. Hazarika, "The First Step Towards Named Entity Recognition in Missing Language", International Conference on Electrical, Electronics, and Optimization Techniques, 2016.

[4] D.Jurafsky, and M.James H, "Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition", Prentice Hall, 2009.

[5] C.Santos, B.Zadrozny, "Learning Character-level Representations for Part-of-Speech Tagging", International Conference on Machine Learning, 2014.

[6] O.Moradiannasab, S.Momtazi and A.Palmer, "A Named Entity Recognition Tool for Persian", In Proceedings of the 3rd Iranian Conference on Computational Linguistics, 2014.

[7] F.Erik, K.S.Tjong, & D.M.Fien, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition", *Proceedings of CoNLL-2003*, pp. 142-147, 2003.

[8] S. Morwal, N. Jahan and D. Chopra, "Named Entity Recognition using Hidden Markov Model (HMM)" *International Journal on Natural Language Computing (IJNLC)*, Vol.1, No.4, pp. 15-23, 2012.

[9] C. Lee, Y. Hwang, H. Oh, S. Lim, J. Heo, C. Lee, H. Kim, J. Wang, M. Jang, "Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering", *Asia Information Retrieval Technology, Lecture Notes in Computer Science*, vol 4182, 2006.

[10] S. Özkaya and B. Diri, "Named Entity Recognition by Conditional Random Fields from Turkish informal texts", *IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, pp. 662-665, 2011.

[11] Y. Benajiba and P. Rosso, "Arabic Named Entity Recognition using Conditional Random Fields", Workshop on HLT&NLP within the Arabic World, LREC, 2008.

[12] A. Ekbal and S. Bandyopadhyay, "A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi". *Linguistic Issues in Language Technology (LiLT)*, Volume (2:1), pp. 1-44, CSLI Publication, 2009.

[13] J.R. Curran and S. Clark. 2003, "Language independent NER using a maximum entropy tagger", *Seventh conference on Natural language learning at HLT-NAACL, Association for Computational Linguistics*, pp. 164-167, 2003.

[14] L. Li, L. Jin, Z. Jiang, D. Song and D. Huang, "Biomedical named entity recognition based on extended Recurrent Neural Networks", *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 649-652, 2015.

[15] N. Suakkaphong, Z. Zhang, H. Chen, "Disease named entity recognition using semisupervised learning and conditional random fields", *Journal of the American Society for Information Science and Technology*, Vol. 62, Issue 4, 2011.

[16] Q. Wei, T. Chen, R. Xu, Y. He, L. Gui, "Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks", *journal of biological data-bases and curation*, 2016.

[۱۷] پ. مرتضوی و م. شمس‌فرد، «شناسایی موجودیت‌های نامدار در متون فارسی». پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران.

تهران. انجمن کامپیوتر. مرکز توسعه فناوری نیرو. ۱۳۸۸

[17] P. Mortazavi and M. Shamsfard. "Named Entity Recognition in Persian Texts", The 15th National CSI Computer Conference, 2009.

[18] M. Kolali Khormuji and M. Bazrafkan, "Persian named entity recognition based with local filters", International Journal of Computer Applications 100(4), 2014.

[19] M. Bijankhan, J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi, "Lessons from building a Persian written corpus: Peykare", *Language resources and evaluation*, 45(2), pp. 143–164, 2011.

[۲۰]م. عبدوس، ب. مینایی بیدگلی. "بهبود شناسایی موجودیت‌های نامدار فارسی با استفاده از کسره اضافه". پردازش علائم و داده‌ها. ۱۴ (۴) :۴۳-۵۴. ۱۳۹۶

[20] M. Abdoos, B. Manaei, "Improving Named Entity Recognition Using Izafe in Farsi",*JSDP*, vol.14 (4), pp.43-54, 2018.

[21] F. Ahmadi and H. Moradi, "A hybrid method for Persian named entity recognition", *The IEEE Conference on Information and Knowledge Technology (IKT)*, pp. 1–7, 2015.

[22] H. Poostchi, E.Z. Borzeshi, M. Abdous, and M. Piccardi, "PersoNER: Persian named-entity recognition", International Conference on Com-putational Linguistics (COLING), 2016.

[23] K. Dashtipour, M. Gogate, A. Adeel, A. Algarafi, N. Howard, and A. Hussain, "Persian Named Entity Recognition", *IEEE International Con-ference on Cognitive Informatics & Cognitive Computing*, pp. 79–83, 2017.

[۲۴]آ. شاکری، م. شهشهانی، ه. فیلی، م. محسنی، م. ملاعباسی، «تشخیص موجودیت‌های اسمی در زبان فارسی». گزارش فنی، SE-P18-MGT-PRS-01-v2، مرکز تحقیقات مخابرات ایران، ۱۳۹۷

[24] A. Shakeri, M. Shahshahani, H. Feili, M. Mohseni, and M. Molla Abbasi., "Persian Language Processing Tools (Research on Named Entity Recognition Tools in Natural Language and Presentation of a Laboratory Instance for Persian)", Technical Report SE-P18-MGT-PRS-01-v2.0. Iran Telecommunication Research Center, 2017.

[25] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning", *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pp. 384-394, 2010.

[26] T. Mikolov, K.Chen, G.Corrado, J.Dean, "Efficient Estimation of Word Representations in Vector Space",arXiv preprint arXiv:1301.-3781, 2016.

[27] C.N.dos Santos and V.Guimaraes, "Boosting Named Entity Recognition with Neural Character Embeddings",Proceedings of the Fifth Named Entity Workshop, 2015.

[28] P.Bojanowski, E.Grave, A.Joulin, and T.Mikolov, "Enriching Word Vectors with Subword Information", Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[29] R. Socher, Ch. D. Manning and A.Y. Ng, "Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks", Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop, 2010.

[30] Phong Le and Willem H. Zuidema, "Compositional Distributional Semantics with Long Short Term Memory", Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, 2015.

[31] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory. Neural Computing, vol.9 (8), 1997.

[32] G.Lample, M.Ballesteros, Sandeep Subramanian, K.Kawakami and Ch.Dyer, "Neural Architectures for Named Entity Recognition", Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[33] Poostchi, Hanieh, Ehsan Zare Borzeshi, Mohammad Abdous and Massimo Piccardi. "PersoNER: Persian Named-Entity Recognition." COLING, 2016.

[34] M. Ghayoomi, "Bootstrapping the development of an HPSG-based treebank for Persian," Linguistic Issues in Language Technology, vol.7, no.1, 2012.

[35] J.R. Finkel, T.Grenager and Ch.Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pp. 363-370, 2005.

[36] M. Konkol, M. Konopík, "Segment Representations in Named Entity Recognition". International Conference on Text, Speech, and Dialogue (TSD), Lecture Notes in Computer Science, vol 9302. Springer, 2015.

[37] S. Momtazi and O. Moradianmasab, "A Statistical Approach for Knowledge Discovery: Bootstrapped Analysis of Language Models for Knowledge base Population from Unstructured Text". Scientia Iranica 26 (Special Issue on: Socio-Cognitive Engineering), pp. 26-39, 2019.

[38] S. Momtazi and D. Kalkow, "Bridging the Vocabulary Gap between Questions and Answer Sentences". Information Processing & Management, 51 (5), 2015.

[39] https://fa.wikipedia.org,96.05.25.

[40] A. Hadifar, S. Momtazi, "The Impact of Corpus Domain on Word Representation: a Study on Persian Word Embeddings", Lang Resources & Evaluation, 52(4), pp. 997–1019, 2018.

[41] Z. Bairong, W. Wenbo, L. Zhiyu, Z. Chonghui, T. Shinozaki, "Comparative Analysis of Word Embedding Methodsfor DSTC6 End-to-End Conversation Modeling Track", Proceddings of International Conference of Dialog System Technology Challenges, 2017.

[42] R.Rehurek and S.Petr, "Software Framework for Topic Modelling with Large Corpora", *In Proceedings of LREC2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 2010.

سعیده ممتازی در حال حاضر استادیار دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر است. وی سرپرستی آزمایشگاه پردازش زبان طبیعی را برعهده دارد. ایشان دوره کارشناسی و کارشناسی ارشد خود را در دانشگاه صنعتی شریف گذارنده و سپس دوره دکترا را در دانشگاه زارلند آلمان سپری کرده است. وی به‌عنوان فرصت مطالعاتی بخشی از تحصیل دکترای خود را در دانشگاه جان‌هاپکینز آمریکا گذرانده و پس از اتمام دوره دکترا به‌عنوان پژوهش‌گر پسادکترا در دانشگاه پتسدام آلمان و همچنین مؤسسه تحقیقات آموزشی آلمان به پژوهش پرداخته است. زمینه اصلی فعالیت او پردازش زبان طبیعی است.

نشانی رایانامه ایشان عبارت است از:

momtazi@aut.ac.ir

فرزانه ترابی دارای مدرک مهندسی کامپیوتر از دانشگاه صنعتی شاهرود و فارغ‌التحصیل کارشناسی ارشد هوش مصنوعی دانشگاه صنعتی امیرکبیر است. وی پایان‌نامه کارشناسی ارشد خود را در آزمایشگاه پردازش زبان طبیعی دانشگاه امیرکبیر و در زمینه تشخیص موجودیت‌های نامدار فارسی با استفاده از روش‌های یادگیری عمیق به انجام رسانده است.

نشانی رایانامه ایشان عبارت است از:

f9torabi@gmail.com