Named Entities Recognition and Classification System for Persian Texts Based on Neural Network

Mojtaba Zali
MA Student; Department of Computer Engineering;
South Tehran Branch; Islamic Azad University; Tehran, Iran;
Email: Mojtaba.Zali@gmail.com

Mohsen Firoozbakht*
Assistant Professor; Department of Computer Engineering;
South Tehran branch; Islamic Azad University; Tehran, Iran;
Email: Firoozm@gmail.com

Iranian Journal of Information Processing and Management

Iranian Research Institute
for Information Science and Technology
(IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 34 | No. 1 | pp. 473-486
Autumn 2018

Received: 08, Nov. 2017 **Accepted: 28, Oct. 2017**

Abstract: Named Entity Recognition (NER) is a fundamental task in natural language processing and also known as a subset of information extraction. We seek to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, etc. Named Entity Recognition for English texts has been researched widely for the past years, however only a few limited researches have emphasized on Persian NER due to absence of resources for Persian named entities and the limited amount of progress made in Persian natural language processing in general. In this paper, a Persian named entity recognition system has been developed based on neural network with the study of researches conducted in other languages and benefiting from the latest methods in this area such as using the vector representation of words. The results from the proposed model show that word embedding features in Persian not only resolve the problem of feature selection, but also it could lead to the development of an efficient system with the least dependence to the domain.

Keywords: Natural Language Processing, Named Entity Recognition, Neural Network, Vector Representation of Words

* Corresponding Author

سیستم شناسایی و طبقه‌بندی موجودیت‌های اسمی در متون زبان فارسی بر پایهٔ شبکهٔ عصبی

مجتبی زالی
مهندسی نرم‌افزار کامپیوتر؛ دانشجوی کارشناسی ارشد؛ دانشگاه آزاد اسلامی؛ واحد تهران جنوب؛
پدیدآور رابط mojtaba.zali@gmail.com

محسن فیروزبخت
دکتری؛ مهندسی نرم‌افزار کامپیوتر؛ استادیار؛ دانشکدهٔ فنی و مهندسی؛ دانشگاه آزاد اسلامی؛ واحد تهران جنوب alidousti@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۵/۰۸/۰۶ | پذیرش: ۱۳۹۶/۰۵/۱۷ | مقاله برای اصلاح به مدت ۹ روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۴ | شمارهٔ ۱ | ص‌ص ۴۷۳-۴۸۶
پاییز ۱۳۹۷

چکیده: شناسایی موجودیت‌های اسمی به‌عنوان یک وظیفهٔ پایه‌ای در حوزهٔ پردازش زبان طبیعی و به‌طور کلی، زیرمجموعه‌ای از استخراج اطلاعات است. در فرایند شناسایی موجودیت‌های اسمی به‌دنبال مکان‌یابی عناصر اسمی در متن و دسته‌بندی آن‌ها به رده‌هایی از پیش تعیین‌شده از قبیل اسامی اشخاص، سازمان‌ها، مکان‌ها، عبارت‌های زمانی، و غیره هستیم. هرچند پژوهش‌هایی گسترده در توسعهٔ سیستم‌های شناسایی موجودیت‌های اسمی در حوزهٔ زبان انگلیسی درطی سال‌های پیشین انجام گرفته، متأسفانه با توجه به مشکلات موجود، مانند نبودِ پیکره‌های متنی نشانه‌گذاری‌شدهٔ استاندارد در زبان فارسی، پژوهش‌های بسیار محدودی در زبان فارسی وجود دارد. در این مقاله با بررسی پژوهش‌های انجام‌گرفته در دیگر زبان‌ها و با بهره‌گیری از روش‌های تازه در این حوزه، مانند استفاده از نمایش بردارهای عددی برای کلمات، به توسعهٔ سیستمی برای شناسایی موجودیت‌های اسمی بر پایهٔ شبکهٔ عصبی پرداخته شده است. نتایج به‌دست‌آمده از مدل پیشنهادی نشان‌دهندهٔ این واقعیت است که استفاده از مدل‌های نمایش بردارهای عددی برای کلمات در زبان فارسی، افزون بر مرتفع‌کردن مشکل انتخاب ویژگی‌ها، می‌تواند به توسعهٔ سیستمی کارآمد منجر شود که کمترین وابستگی را نیز به دامنه دارد.

کلیدواژه‌ها: پردازش زبان طبیعی، شناسایی موجودیت‌های اسمی، شبکهٔ عصبی، نمایش برداری کلمات

۱. مقدمه

شناسایی موجودیت‌های اسمی، یک زیروظیفه از استخراج اطلاعات است (Mihalcea and Moldovan 2001) و به معنی پردازش مستنداتی است که در آن به‌دنبال مکان‌یابی عناصر اسمی در متن و دسته‌بندی آن‌ها به رده‌هایی از پیش تعیین‌شده، مانند اسامی اشخاص، سازمان‌ها (شرکت‌ها، سازمان‌های دولتی و غیره)، مکان‌ها (شهرها، کشورها، رودخانه‌ها و غیره)، عبارت‌های زمانی، کمیت‌ها، مقدارهای پولی، درصدها و غیره هستیم. شناسایی موجودیت‌های اسمی یک وظیفهٔ پایه‌ای است و یکی از هسته‌های پردازشی زبان طبیعی محسوب می‌شود (Nadeau and Sekine 2007). در مورد کاربردهای شناسایی موجودیت‌های اسمی می‌توان به موارد مختلفی از جمله ترجمهٔ متون، بازیابی و درک متون، خلاصه‌سازی، جست‌وجوی معنایی و غیره اشاره کرد. به‌طور کلی، کاربردهای گوناگون و بسیار سودمندی را می‌توان برای سیستم‌های شناسایی موجودیت‌های اسمی و به‌طور عمومی‌تر، در پردازش زبان طبیعی متصور شد.

به‌طور کلی، عملیات شناسایی موجودیت‌های اسمی شامل دو وظیفه است: شناسایی اسامی مربوطه در متن و سپس، دسته‌بندی این اسامی به مجموعه‌های از پیش تعیین‌شده. به‌طور کلی، سیستم‌های شناسایی موجودیت‌های اسمی را می‌توان به سه کلاس اصلی دسته‌بندی کرد: روش‌های مبتنی بر قاعدهٔ دستی مشخص‌شده، سیستم‌های آماری مبتنی بر یادگیری ماشین و سیستم‌های ترکیبی. این سه کلاس، سه دستهٔ اصلی روش‌های شناسایی موجودیت‌های اسمی به شمار می‌روند. رهیافت‌های مبتنی بر قوانین، بر روی استخراج نام‌ها با استفاده از تعداد زیادی از مجموعه قوانین ساخته‌شده به‌صورت دستی تمرکز کرده است. به‌طور کلی، این سیستم‌ها مجموعه‌ای از الگوها هستند که از مشخصه‌های دستوری، نحوی و املایی در کنار ترکیبی از لغت‌نامه‌ها استفاده می‌شود. با توجه به این ویژگی‌ها، روش مبتنی بر قواعد بر خلاف کارایی بالا در برخی دامنه‌ها، دارای مشکل بالای غیرقابل انتقال بودن و نیز کمبود قابلیت اطمینان را دارد (Nadeau and Sekine 2007). در سیستم‌های مبتنی بر روش‌های آماری و یادگیری ماشین، هدف از رهیافت شناسایی موجودیت‌های اسمی، تبدیل مسئلهٔ شناسایی موجودیت‌ها به مسئلهٔ دسته‌بندی و بهره‌گیری از یک مدل آماری برای حل این مسائل است. در این روش‌ها، در صورت استفاده از ویژگی‌های مناسب می‌توان افزون بر ارائهٔ نتایج مطلوب، مدلی را توسعه داد که

کمترین وابستگی را به دامنه داشته باشد. از این رو، در این پژوهش تمرکز در به‌کارگیری روش‌های مبتنی بر یادگیری ماشین است.

ساختار این مقاله در ادامه به این قرار است: در بخش ۲، مروری بر روی کارهای پیشین در حوزهٔ شناسایی موجودیت‌های اسمی وجود دارد. در بخش ۳، به بررسی روش‌های استخراج ویژگی که در زمان بهره‌گیری از روش‌های آماری از مهم‌ترین فرایندها محسوب می‌شود، خواهیم پرداخت. در بخش ۴، پیکرهٔ متنی مورد استفاده در این پژوهش معرفی می‌گردد. بخش ۵، به ارائهٔ معماری شبکهٔ عصبی مدل پیشنهادی اختصاص داده شده است. در بخش ۶، معیار ارزیابی مورد استفاده در این پژوهش معرفی می‌شود و در نهایت، در بخش ۷، نتایج حاصل از مدل پیشنهادی با دیگر مدل‌های هم‌دامنه مورد ارزیابی قرار خواهند گرفت.

۲. مروری بر کارهای پیشین

در سال‌های اخیر تعداد زیادی از روش‌های آماری بر پایهٔ روش آموزش با ناظر ارائه شده است. «بایکل» و همکارانش یک اسم‌یاب آموزش‌پذیر به نام «نایمبل»[1] را مبتنی بر مدل مخفی «مارکوف» معرفی کرده‌اند (Bikel et al. 1997). «بورثویک» و همکارانش منابع دانش عظیمی را با استفاده از روش بیشترین آنتروبی در شناسایی موجودیت‌های اسمی مورد استفاده قرار داده‌اند (Borthwick et al. 1998). علامت‌گذاری نام‌های سادهٔ ناشناخته با استفاده از درخت تصمیم توسط «بچِت» و همکارانش پیشنهاد شد (Béchet, Nasr, and Genet 2000).

همان‌طور که پیش‌تر اشاره شد، در حوزهٔ زبان فارسی به‌دلیل محدودیت‌های موجود، پژوهش‌های بسیار کمتری بر روی شناسایی موجودیت‌های اسمی انجام شده است. «اصفهانی» و همکارانش با استفاده از ویژگی‌های گسترده و با به‌کارگیری شبکهٔ عصبی مبتنی بر چند لایه پرسپترون به شناسایی موجودیت‌های اسمی پرداخته‌اند (۱۳۸۸). «ناجی و نازلیا» نیز از یک شبکهٔ عصبی چندلایه به‌منظور ایجاد مدلی برای شناسایی موجودیت‌های اسمی در زبان عربی استفاده کرده است (Naji and Nazlia 2012). «ما» و همکاران در حوزهٔ زبان انگلیسی با استفاده از ویژگی‌های متعدد متنی همانند پیشوند و

1. Nymbel

پسوند هر کلمه، ادات سخن برای کلمات، ویژگی‌های تک-وزنی و دو-وزنی، ویژگی‌های سندی به همراه ویژگی‌های نمایش برداری متعدد از کلمات به توسعه مدل‌های خود پرداخته‌اند (Ma et al. 2016).

«احمدی و مرادی» نیز در استفاده از ویژگی‌های متنی و مدل مخفی «مارکوف» در زبان فارسی به توسعهٔ یک سیستم شناسایی موجودیت اسمی پرداختند. در این سیستم کارایی برابر با ۵۳/۱۴ بوده است (Ahmadi and Moradi 2015). در زبان فارسی بر خلاف زبان انگلیسی، با توجه به ویژگی‌های خاص این زبان، استفادهٔ تنها از ویژگی‌های متنی نمی‌تواند منجر به توسعهٔ یک سیستم کارا شود. از این رو، «احمدی و مرادی» با ترکیب این ویژگی‌ها با روش مبتنی بر قواعد، کارایی سیستم را به ۸۵/۹۳ درصد بهبود داده‌اند (همان).

«سیوک» و همکاران با استفاده از مجموعه‌ای از ویژگی‌های متنی مشابه با ویژگی‌های به‌کار گرفته شده توسط «ما» (Ma et al. 2016) و مدل‌های نمایش بردارهای عددی همانند بردار سراسری و Word2Vec، به توسعهٔ سیستمی پرداخته‌اند. با استفاده از ویژگی‌های متنی به‌همراه بردار سراسری کارایی برابر با ۷۹/۴۸ و در زمان استفاده از مدل Word2Vec برابر با ۸۰/۶۸ بوده است (Seok et al. 2016).

۳. استخراج ویژگی

در زمان استفاده از روش‌های آماری و مبتنی بر یادگیری ماشین، ارائهٔ یک نمایش مناسب برای کلمات و همچنین، استخراج ویژگی‌های مطلوب از مهم‌ترین بخش‌های توسعهٔ مدل است. به‌طور کلی، ویژگی‌های سیستم‌های شناسایی موجودیت‌های اسمی را می‌توان به ۳ دستهٔ کلی ویژگی‌های کلمه‌ای، ویژگی‌های سندی، و ویژگی‌های لیستی تقسیم‌بندی کرد. ویژگی‌های کلمه‌ای، توصیفگر شکل و ظاهر کلمه، ساختار کلمه، نقش کلمه در جمله و غیره خواهد بود. برای مثال، بزرگ بودن حرف اول کلمه در زبان انگلیسی، وجود نقطه و یا عدد مانند کلماتی نظیر W3C و .I.B.M می‌توانند نشانه‌های خوبی برای کاندید بودن یک کلمه در مجموعهٔ اسامی خاص باشند.

ویژگی‌های سندی مربوط به اطلاعاتی از کلمه است که در کل سند وجود دارد. این اطلاعات از طریق پردازش کل سند و یا بخشی از آن مطابق با مدل طراحی‌شده به‌دست می‌آید. در صورت پردازش کل اسناد و در مواردی که پیکرهٔ متنی و اسناد مورد استفاده

دارای حجم زیادی باشد، این ویژگی‌ها قوی‌تر استخراج شده و دارای اطلاعات بسیار خوبی برای سیستم شناسایی هستند. تعداد رخداد یک کلمه در جمله، پاراگراف و یا کل سند می‌تواند مثال‌هایی برای این نوع ویژگی‌ها باشد.

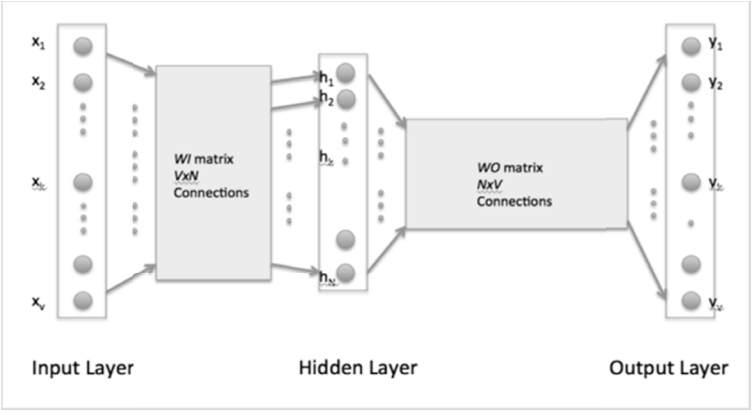
از دیگر ویژگی‌های موجود در این حوزه ویژگی‌های لیستی است. برای استخراج این ویژگی‌ها، از لیست‌هایی که شامل اسامی افراد، مکان‌ها، شهرها، کشورها، سازمان‌ها و غیره است، استفاده می‌شود. وجود کلمه‌ای در هر یک از این لیست‌ها، نشان‌دهندهٔ یکی از ویژگی‌های کلمه است. در برخی موارد وجود یک کلمه در یک لیست می‌تواند، با احتمال زیاد، طبقهٔ آن کلمه را مشخص کند، اما به‌دلیل ابهام در نقش و جایگاه اسامی، همواره امکان آن وجود دارد که اسمی در یک لیست قرار داشته باشد، اما به طبقه‌ای دیگر متعلق باشد. بزرگ بودن این لیست‌ها باعث می‌شود که ویژگی‌هایی با دقت بالاتر استخراج گردد. اما همواره این روش با مشکلاتی از قبیل، در دسترس نبودن، نگهداری، و به‌روزرسانی روبه‌رو بوده است. لازم به ذکر است که استفاده از ویژگی‌های لیستی قابل انتقال بودن مدل را بسیار کاهش می‌دهد.

۳-۱. نمایش کلمات

یکی از روش‌های بسیار کارا برای بهره‌گیری در یادگیری ماشین در حوزهٔ پردازش زبان طبیعی استفاده از نمایش کلمات به‌صورت بردارهای عددی است. در فرایند استخراج ویژگی، افزون بر استفاده از ویژگی‌هایی که در بخش پیشین مورد بررسی قرار گرفت، می‌توان از نمایش عددی کلمات به‌عنوان ویژگی که حاوی اطلاعاتی ارزشمند از خود کلمه است، استفاده کرد. پژوهش‌ها نشان‌دهندهٔ این موضوع است که استفاده از نمایش کلمات می‌تواند در بهبود عملکرد سیستم‌های شناسایی موجودیت‌های اسمی بسیار مفید باشد (Turian, Ratinov, and Bengio 2010). مدل‌های متعددی برای نمایش کلمات به‌صورت بردارهای عددی در حوزهٔ پردازش زبان طبیعی در طی سال‌های پیشین معرفی شده‌اند. از جلمهٔ مدل‌های بسیار کارا در برداری‌سازی کلمات، word2vec است که توسط «میکولوف» در شرکت گوگل طراحی و معرفی شده است (Mikolov et al. 2013). الگوریتم‌های word2vec برای نمایش برداری کلمات با ابعاد بالا مورد استفاده قرار می‌گیرد. اطلاعات استخراج‌شده توسط این مدل به حدی است که می‌تواند ارتباط معنایی میان کلمات را نیز حفظ کند.

این مدل از یک شبکهٔ عصبی ۳-لایه پرسپترون به‌منظور تولید بردارها استفاده می‌کند که در شکل ۱، معماری کلی این شبکه نمایش داده شده است. تمامی نورون‌های شبکهٔ عصبی از توابع خطی تشکیل شده‌اند. برای نمایش برداری کلمات در ابتدای کار، واژه‌نامه‌ای از کلمات تشکیل‌دهندهٔ پیکرهٔ‌متنی ایجاد می‌شود. تعداد نورون‌ها در لایهٔ ورودی و لایهٔ خروجی برابر با تعداد کلمات واژه‌نامه و همچنین، تعداد نورون‌ها در لایهٔ پنهان برابر با طول بردار مطلوب مورد نیاز است. اتصالات میان لایهٔ ورودی و لایهٔ پنهان و همچنین، لایهٔ پنهان و لایهٔ خروجی به‌صورت اتصالات کامل است. بنابراین، اگر طول واژه‌نامه را V در نظر بگیریم و طول مطلوب برای بردار کلمات N باشد، وزن‌های اتصالات میان لایهٔ ورودی به لایهٔ پنهان می‌توانند با ماتریسی به اندازهٔ V*N و به همین ترتیب، وزن‌های لایهٔ پنهان به لایهٔ خروجی با اندازهٔ N*V نگهداری و نمایش داده شوند. به‌طور کلی، ماتریس لایهٔ ورودی را با عنوان WI و همچنین، ماتریس وزن‌های لایهٔ خروجی WO نامیده می‌شود (شکل ۱).

برای آموزش مدل word2vec در این پژوهش از پیکرهٔ متنی «ویکی‌پدیا» استفاده شد. از این پیکره در حدود ۶/۵ میلیون کلمه استخراج و برای فرایند آموزش مورد استفاده قرار گرفت. برای بهبود فرایند آموزش شبکه‌های عصبی word2vec و شبکهٔ عصبی عمیق پشنهادشده در این پژوهش، در مرحلهٔ پیش‌پردازش اطلاعات، داده‌های موجود در پیکرهٔ متنی که مربوط به کلمات بازدارنده[1] بوده‌اند، حذف شد.



شکل ۱. معماری شبکهٔ عصبی مدل word2vec (Mikolov et al. 2013)

1. stop words

۳-۲. ویژگی‌های مدل پیشنهادی

تمرکز برای به‌کارگیری ویژگی‌های مناسب به‌منظور بالا بردن نرخ بازشناسی و کارایی در سیستم شناسایی موجودیت‌های اسمی در این پژوهش از مهم‌ترین بررسی‌های انجام گرفته است. همان‌طور که پیش‌تر اشاره شد، هدف از این پژوهش ارزیابی و پیشنهاد مدلی است که بتواند در حالی که دارای نرخ بازشناسی مناسبی است، از ویژگی‌ها و فرایندهایی استفاده نماید که کمترین وابستگی را به دامنه و یا حتی زبان داشته باشد و محدویت‌های موجود در این حوزه را مرتفع کند. از این رو، در ادامه، ویژگی‌های به‌کار گرفته‌شده برای آموزش مدل با در نظر داشتن اهداف اشاره‌شده، معرفی شده است.

۳-۲-۱. به‌کارگیری نمایش بردار word۲vec

همان‌طور که در بخش پیش به‌صورت کامل مورد بررسی قرار گرفت، نمایش برداری کلمات به کمک مدل word2vec می‌تواند اطلاعات بسیار ارزشمندی را از کلمات استخراج نماید. بردارهای عددی حاصل از این مدل شامل تمامی اطلاعات زبان‌شناسی و ارتباطات معنایی است. از این رو، برای انتخاب ویژگی برای یک کلمه، با استفاده از بردارهای کلمات پیشین و پسین می‌توان اطلاعات بیشتری مانند جایگاه یک کلمه استخراج کرد. این اطلاعات در زمان آموزش شبکهٔ عصبی می‌تواند در فهم کلمات، جایگاه و ارتباطات آن‌ها کمک‌کننده باشد. از آنجا که محدودهٔ انتخاب ویژگی‌ها در زبان فارسی بسیار محدودتر از زبان‌های دیگر مانند انگلیسی است، نمایش بردار عددی به‌عنوان مهم‌ترین ویژگی در نظر گرفته می‌شود. طول بردار برای مدل پیشنهادی با توجه به بررسی‌های انجام‌شده بر روی پژوهش‌های موجود در این حوزه، برای هر کلمه ۱۰۰ در نظر گرفته شده است (Ma 2016). به این ترتیب و با در نظر گرفتن بردار کلمه‌های پیشین و پسین برای هر کلمه، طول نهایی بردار ۳۰۰ خواهد بود. لازم به ذکر است که برای کلماتی که در ابتدا و یا انتهای جمله قرار گرفته‌اند، برای حل مشکل تأثیر مرزی، برداری با مقادیر تشکیل‌دهندهٔ صفر در نظر گرفته شده است (Mesnil et al. 2015). در زمان به‌کارگیری مدل‌های پیوسته و به‌خصوص در حوزهٔ شبکه‌های عصبی، استفاده از این تکنیک یکی از متداول‌ترین روش‌ها در حل مسئلهٔ مرزی است.

۳-۲-۲. برچسب ادات سخن

استفاده از برچسب ادات سخن می‌تواند در یادگیری ماشین برای کشف ارتباط معنایی

و ساختار کلمه بسیار کمک‌کننده باشد. ادات سخن اطلاعات بسیار ارزشمندی را در ارتباط با یک واژه فراهم می‌کند. در زبان فارسی با توجه به محدود بودن ویژگی‌ها، در برخی از مدل‌های ارائه‌شده، شاخص‌ترین ویژگی آن‌ها استفاده از برچسب‌های ادات سخن بوده است (اصفهانی، قوچانی، و جهانگیری ۱۳۸۸). با توجه به این‌که در مدل پیشنهادی این پژوهش از ویژگی‌های پیوسته‌بودن مدل بهره می‌بریم، در استفاده از ویژگی برچسب ادات سخن نیز برای هر کلمه برچسب‌های پیشین و پسین آن نیز لحاظ خواهد شد.

۳-۲-۳. طول کلمه

با بررسی‌های انجام‌شده بر روی پیکره‌های متنی مشخص شده است که به‌صورت معمول، میانگین طول کلماتی که اسامی خاص هستند از میانگین طول کلمات دیگر بیشتر است. با پردازش‌های انجام‌شده در پیکرهٔ متنی که در این پژوهش مورد استفاده قرار گرفته است نیز این نتیجه حاصل می‌شود که طول کلمات خاص به‌طور میانگین در حدود ۱/۵ نویسه بیشتر از طول کلمات دیگر است (جدول ۱).

جدول ۱. میانگین طول کلمه درپیکرهٔ متنی مورد استفاده در پژوهش

	O	LOC	ORG	PERS
حجم داده	۱۴۶۸۲۸	۵۲۷۶	۸۶۲۶	۴۷۴۶
میانگین طول کلمات	۴/۰۹	۵/۴۴	۵/۳۴	۵/۲

٤. پیکرهٔ متنی

«ویکی‌پدیا»ی فارسی نام یکی از دانشنامه‌های فارسی زبان در اینترنت است. «ویکی‌پدیا»ی فارسی یکی از نسخه‌های «ویکی‌پدیا»، از پروژه‌های بنیاد «ویکی‌مدیا» به‌شمار می‌آید. این دانشنامهٔ فارسی هم‌اکنون بیش از ۵۰۰ هزار مقاله دارد. با بهره‌گیری از مقالات موجود در دانشنامهٔ «ویکی‌پدیا»، در حال حاضر پیکرهٔ متنی دربرگیرندهٔ ۱۶۶۱۷۶ واژه، توسط «بهرامی» در دانشگاه شریف فراهم شده است. این پیکره از متون «ویکی‌پدیا»ی فارسی استخراج و بر اساس استانداردهای MUC-6 برچسب‌گذاری شده است. برچسب‌گذاری کلمات متن در این پیکره به قرار زیر است:

- B-PERS: شروع نام یک شخص؛
- I-PERS: ادامه یا انتهای نام یک شخص؛

- B-LOC: شروع نام یک مکان؛
- I-LOC: ادامه یا انتهای نام یک مکان؛
- B-ORG: شروع نام یک سازمان؛
- I-ORG: ادامه یا انتهای نام یک سازمان؛
- O: یک واحد اسمی که جزء موارد بالا نیست.

پیکرهٔ متنی اشاره‌شده، شامل ۹۸۷۶ جمله است که ۵۱۰۶ جمله از آن حداقل دارای یک واژه از گروه غیر از O است. آمار کامل واژه‌ها و دسته‌بندی آن‌ها در جدول ۲، نمایش داده شده است.

جدول ۲. شمار واژه‌های موجود در پیکرهٔ متنی

کلاس	شمار واژه‌ها
O	۱۴۶۸۲۸
B_ORG	۳۸۴۱
I_ORG	۵۴۸۵
B_LOC	۴۱۳۴
I_LOC	۱۱۴۲
B_PERS	۳۰۱۶
I_PERS	۱۷۳۰

۵. معماری شبکهٔ عصبی

با توجه به بررسی‌های انجام‌شده بر روی مدل‌های مطرح‌شده در حوزهٔ شناسایی موجودیت‌های اسمی بر پایهٔ شبکه‌های عصبی، استفاده از دو مدل از شبکهٔ عصبی عمیق و شبکهٔ عصبی بازگشتی، از جلمه متداول‌ترین مدل‌ها در این حوزه هستند. این دو مدل از شبکهٔ عصبی و به‌طور کلی، شبکه‌های عمیق، قابلیت‌های ارزشمندی را در شناسایی الگوهای بسیار پیچیده دارند (Sutskever et al. 2013). در حوزهٔ پردازش زبان طبیعی و به‌خصوص حوزهٔ شناسایی موجودیت‌های اسمی، با توجه به پیچیدگی بالای حاکم بر این حوزه، بهره‌گیری از شبکه‌های عصبی عمیق می‌تواند به کشف این الگوهای پیچیده و در نهایت، ارائهٔ خروجی مطلوب کمک‌کننده باشد. با توجه به پیچیدگی‌های بسیار

بالا در آموزش شبکه‌های عصبی بازگشتی و به‌منظور مقایسهٔ روش‌های به‌کاررفته در این پژوهش با پژوهش‌های موجود در حوزهٔ زبان فارسی، تمرکز مدل پیشنهادی در بهره‌گیری از شبکهٔ عصبی عمیق است.

برای حل مشکل انطباق بیش از حد در شبکهٔ عصبی عمیق مورد استفاده در این پژوهش از روش حذف تصادفی استفاده شده است. بررسی‌ها نشان‌دهندهٔ این موضوع است که روش حذف تصادفی می‌تواند با کمترین هزینه و بدون سربار زمانی اضافی برای مرتفع‌کردن مشکل انطباق بیش از حد در شبکه‌های عصبی، به‌خصوص در شبکه‌های عمیق و بزرگ مورد استفاده قرار گیرد (Srivastava et al. 2014).

در مدل شبکهٔ عصبی پشنهادی و در فرایند به‌روزرسانی وزن نورون‌ها نیز از روش‌های mini-batch استفاده می‌شود. استفاده از این روش در شبکه‌های عصبی عمیق و بزرگ با تعداد داده‌های بالا، که حوزهٔ کاری این پژوهش است، توصیه می‌شود. به‌منظور حل مشکل کمینهٔ محلی نیز از تکانه استفاده شده است. بهره‌گیری از تکانه به‌همراه نرخ آموزش پایین در روش‌های به‌روزرسانی mini-batch در شبکه‌های عمیق می‌تواند منجر به پرهیز از کمینهٔ محلی و به‌روز رسانی دقیق وزن‌ها شود (Collobert and Weston 2008).

٦. معیار ارزیابی

با توجه به متوازن نبودن کلاس‌ها در پیکره‌های متنی که از جمله ویژگی‌های زبان طبیعی به‌شمار می‌آید، استفاده از معیار دقت[1] در سیستم‌های شناسایی موجودیت‌های اسمی نمی‌تواند معیار دقیقی برای ارزیابی سیستم باشد. یک مدل امتیازدهی که به‌طور استاندارد توسط MUC برای شناسایی موجودیت‌های اسمی مورد استفاده قرار می‌گیرد، دو معیار دقت یا به‌اختصار P و بازخوانی یا R است. این معیارها در حقیقت از تعاریف موجود در حوزهٔ بازیابی اطلاعات گرفته شده‌اند و تعریف آن‌ها به‌صورت زیر است:

$$\text{Precision} = \frac{T_p}{T_p + F_p} \quad (1)$$

$$\text{Recall} = \frac{T_p}{T_p + F_n} \quad (2)$$

که در این فرمول‌ها، Tp پاسخ‌های مثبت است که به‌درستی مثبت تشخیص داده

1. accuracy

شده، Fp پاسخ‌های اشتباهی است که به اشتباه درست تشخیص داده شده و در نهایت، Fn پاسخ‌های درستی بوده است که سیستم آن‌ها را به اشتباه، غلط تشخیص داده است. از ترکیب این دو معیار از کارایی، معیاری حاصل می‌شود که در حقیقت متوسط میانگین همساز این دو معیار است و به آن معیار F یا F-measure می‌گویند:

$$F = \frac{2PR}{P + R} \quad (3)$$

۷. ارزیابی نتایج

پس از ارزیابی بردار word2vec به‌عنوان ویژگی شاخص، شبکهٔ عصبی پیشنهادی را با در نظر گرفتن تمامی ویژگی‌ها، مورد آموزش و ارزیابی قرار داده‌ایم. با استناد به جدول ۲، و مقایسهٔ شمار واژه‌های عمومی (کلاس O) نسبت به واژه‌هایی که اسامی خاص هستند (کلاس‌های ORG، LOC، و PERS)، می‌توان به مشکل نامتوازن بودن پیکرهٔ متنی پی‌برد. این مشکل تنها محدود به پیکرهٔ متنی مورد استفادهٔ ما نبوده و از جمله ویژگی‌های زبان طبیعی محسوب می‌شود؛ به‌طوری که همواره اسامی خاص، درصد بسیار کمی از متون را به خود اختصاص می‌دهند. از این رو، برای آموزش شبکهٔ عصبی سعی بر آن شد، برای ایجاد یک توازن نسبی، جملاتی که حداقل شامل یک اسم خاص هستند، در مجموعه داده‌های تست به‌کار گرفته شود. به‌طور کلی، ۶۰ درصد از حجم داده برای آموزش و ۴۰ درصد باقی‌مانده برای تست استفاده شده است. متوازن کردن واژه‌ها و همچنین، استفاده از تکنیک‌های آموزش شبکهٔ عصبی با ترتیب ورودی تصادفی، از جمله روش‌های به‌کاررفته برای بهبود فرایند آموزش بوده است. برای درستی نتایج به‌دست‌آمده از مدل پیشنهادی از کلاس‌های موجود در پیکرهٔ متنی «بهرامی» استفاده شده است. نتایج حاصل از این اجرا، در جدول ۳، نمایش داده شده است. لازم به ذکر است که میزان دقت در این روش ۹۴/۶ درصد بوده است.

جدول ۳. نتایج حاصل ازاجرای مدل

موجودیت اسمی	دقت	بازخوانی	معیار-اف
شخص	۸۲/۸۹	۹۰/۵۳	۸۶/۵۴
مکان	۷۹/۰۴	۸۲/۹۱	۸۰/۹۲
سازمان	۷۳/۱۶	۶۷/۰۴	۶۹/۹۶

نتایج حاصل از مدل پیشنهادی با مدل‌های موجود در این حوزه در جدول ۴، مورد مقایسه قرار گرفته است. در انتخاب مدل‌های مورد مقایسه به هم‌دامنه بودن و استفاده از روش‌های مبتنی بر یادگیری ماشین توجه شده است. مدل ارائه‌شده توسط Seok et al. (2016) همان‌طور که در بخش ۲، اشاره شد، با توجه به ویژگی‌های زبان انگلیسی، مجموعه‌ای گسترده از ویژگی‌ها را در کنار بردار نمایش word2vec استفاده کرده است. به همین منظور، مشاهده می‌شود که مقدار نهایی معیار f در حدود ۲ درصد از مدل پیشنهادی بیشتر است. علی‌رغم استفاده از ویژگی‌های ادات سخن و طول کلمه در بخش ویژگی‌های مدل پیشنهادی، بهره‌گیری از بردارهای پیشین و پسین word2vec به همراه به‌کارگیری از شبکهٔ عصبی عمیق، مدل توسعه داده‌شده در این پژوهش توانسته است کارایی مشابه با مدل «سیوک و همکاران» را ارائه کند. لازم به ذکر است که پیکرهٔ متنی مورد استفاده در این پژوهش بر پایهٔ «ویکی‌پدیا» بوده و در مقایسه با پیکره‌های متنی استاندارد زبان انگلیسی از کیفیت مطلوبی برخوردار نیست.

جدول ٤. مقایسهٔ نتایج حاصل از مدل پیشنهادی با مدل‌های موجود در حوزهٔ مشابه

	Ahmadi	Naji	Yukun	Seok	روش پیشنهادی
شخص	۵۹/۹۹	۶۹/۹۰	۸۱/۷	ناموجود	۸۶/۵۴
مکان	۵۷/۵۴	۴۳/۳۰	۸۰/۹	ناموجود	۸۰/۹۲
سازمان	۴۱/۹۱	۵۹/۲۰	۵۵/۷	ناموجود	۶۹/۹۶
جمع کلی	۵۳/۱۴	۵۷/۴۶	۷۲/۷۶	۸۰/۷۲	۷۹/۲۴

۸. نتیجه‌گیری

نتایج به‌دست‌آمده از مدل پیشنهادی نشان‌دهندهٔ این واقعیت است که استفاده از مدل‌های نمایش بردارهای عددی برای کلمات که به‌تازگی معرفی شده‌اند، در زبان فارسی نیز می‌تواند در بهبود سیستم‌های شناسایی موجودیت‌های اسمی بسیار کارآمد باشد. با توجه به مشکلات گسترده در شناسایی موجودیت‌های اسمی در زبان فارسی که حاصل رسم‌الخط و قواعد این زبان است، انتخاب ویژگی‌های شناسایی بر خلاف زبان‌های لاتین با محدودیت‌های زیادی همراه است. از این رو، مدل‌های نمایش بردار عددی برای کلمات می‌تواند برای مرتفع کردن این مشکلات بسیار مؤثر باشد. استفاده

از این ویژگی‌ها افزون بر افزایش کارایی در سیستم‌های شناسایی موجودیت‌های اسمی، باعث توسعهٔ سیستم‌هایی می‌شود که کمترین وابستگی را به دامنه دارند، حال آن‌که بزرگ‌ترین چالش حاضر در توسعهٔ این سیستم‌ها، وابستگی بسیار بالای آن‌ها به دامنه و پیکرهٔ متنی است که در آن توسعه داده شده‌اند. با استناد به نتایج به‌دست‌آمده و مقایسهٔ آن‌ها با پژوهش‌های مرتبط در این حوزه، این نتیجه حاصل می‌شود که بهره‌گیری از شبکه‌های عصبی عمیق با توجه به ویژگی‌های آن‌ها در کشف روابط پیچیده می‌تواند برای شناسایی و طبقه‌بندی موجودیت‌های اسمی مؤثر و کارآمد باشد.

فهرست منابع

اصفهانی، سیدعبدالحمید، سعید راحتی قوچانی، و نادر جهانگیری. ۱۳۸۸. استخراج ویژگی برای یک سیستم شناسایی و طبقه‌بندی اسامی فارسی. مقاله ارائه‌شده در پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران، تهران.

References

Ahmadi F, and H. Moradi. 2015. A hybrid method for Persian named entity recognition. In Information and Knowledge Technology (IKT), 7th Conference on 2015 May 26. IEEE, Amirkabir University, Tehran, Iran.

Béchet, F., A. Nasr, and F. Genet. 2000. Tagging unknown proper names using decision trees. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (pp. 77-84). Association for Computational Linguistics. Hong Kong.

Bikel, Daniel M., Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. a High-Performance Learning Name finder,. In Proceedings of the fifth international conference on Applied Natural Language Processing. Washington, DC, USA.

Borthwick, Andrew, John Sterling, Eugene Agichtein, and Ralph Grishman..1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In Sixth Workshop on Very Large Corpora. Montreal, Quebec, Canada.

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167). ACM. New York, NY, USA.

Ma, Y, J. J. Kim, B. Bigot, and T. M. Khan. 2016. Feature-enriched word embeddings for named entity recognition in open-domain conversations. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on 2016 Mar 20 (pp. 6055-6059). IEEE, Shanghai, China.

Mesnil, G., Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). NJ, USA: IEEE Advancing Technology for Humanity. 23 (3): 530-539.

Mihalcea, Rada, and D. L. Moldovan. 2001. Document indexing using named entities. *Studies in Informatics and Control* 10 (1): 21-28. Bucharest, Romanian: National Institute for R&D in Informatics

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. *Distributed representations of words*

and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Nevada, USA.

Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*. Netherlands: John Benjamins Publishing Company. 30 (1): 3-26.

Naji F, Mohammed, and Nazilia, Omar. 2012. Arabic named entity recognition using artificial neural network. *Journal of Computer Science*. USA: Science Publications 8 (8): 1285.

Seok, M., H. J. Song, C. Y. Park, J. D. Kim, and Y. S. Kim. 2016. Named Entity Recognition using Word Embedding as a Feature. *International Journal of Software Engineering and Its Applications*. South Korea: Science and Engineering Research Support Society. 10 (2): 93-104.

Srivastava, N., G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*. USA: MIT Press. 15 (1): 1929-1958.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G.2013. On the importance of initialization and momentum in deep learning. In the 30th International conference on machine learning: 1139-1147. Atlanta, GA, USA.

Turian, J., Ratinov, L. and Bengio, Y., 2010, July. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 384-394). Association for Computational Linguistics. Uppsala, Sweden.

مجتبی زالی

متولد سال ۱۳۶۷ و دانش‌آموختهٔ کارشناسی ارشد از دانشگاه آزاد اسلامی واحد تهران جنوب در رشتهٔ مهندسی کامپیوتر گرایش نرم‌افزاراست. هوش مصنوعی، یادگیری ماشین و بازشناسی الگو، پردازش زبان طبیعی، داده‌کاوی، و تحلیل کلان داده از جمله علایق پژوهشی وی است.

محسن فیروزبخت

متولد سال ۱۳۵۰ دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی کامپیوتر از دانشگاه کینگستون لندن است. ایشان هم‌اکنون استادیار گروه فنی و مهندسی دانشگاه آزاد اسلامی واحد تهران جنوب است. یادگیری ماشین، شبکهٔ حسگر بیسیم، امنیت شبکه، فشرده‌سازی تصاویر پزشکی، و داده‌کاوی از جمله علایق پژوهشی وی است.