

سیستم شناسایی و طبقه‌بندی موجودیت‌های اسمی در متون زبان فارسی بر پایه شبکه عصبی

مجتبی زالی *

مهندسی نرم‌افزار کامپیوتر

دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی واحد تهران جنوب

محسن فیروزبخت

دکتری مهندسی نرم‌افزار کامپیوتر

استادیار، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد تهران جنوب

پذیرش: ۹۶/۰۵/۱۷

دریافت: ۹۵/۰۸/۰۶

فصلنامه علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا(چاپی) ۲۲۵۱-۸۲۲۳
شاپا(الکترونیکی) ۲۲۵۱-۸۲۳۱
نمایه در SCOPUS, LISA و ISC
http://ijpm.irandoc.ac.ir
دوره XX | شماره X | صص XX-XX
۱۳XX X

نوع مقاله: پژوهشی

چکیده: شناسایی موجودیت‌های اسمی به عنوان یک وظیفه پایه‌ای در حوزه پردازش زبان طبیعی و به طور کلی زیرمجموعه‌ای از استخراج اطلاعات است. در فرآیند شناسایی موجودیت‌های اسمی به دنبال مکان‌یابی عناصر اسمی در متن و دسته‌بندی آن‌ها به رده‌هایی از پیش تعیین شده از قبیل اسامی اشخاص، سازمان‌ها، مکان‌ها، عبارت‌های زمانی، و غیره هستیم. هرچند پژوهش‌هایی گسترده در توسعه سیستم‌های شناسایی موجودیت‌های اسمی در حوزه زبان انگلیسی در طی سال‌های پیشین انجام گرفته است، متأسفانه با توجه به مشکلات موجود، مانند نبود پیکره‌های متنی نشانه‌گذاری شده استاندارد در زبان فارسی، پژوهش‌های بسیار محدودی در زبان فارسی وجود دارد. در این مقاله با بررسی پژوهش‌های انجام گرفته در دیگر زبان‌ها و با بهره‌گیری از روش‌های تازه در این حوزه همانند استفاده از نمایش بردارهای عددی برای کلمات، به توسعه سیستمی برای شناسایی موجودیت‌های اسمی بر پایه شبکه عصبی پرداخته شده است. نتایج بدست آمده از مدل پیشنهادی نشان دهنده این واقعیت است که استفاده از مدل‌های نمایش بردارهای عددی برای کلمات در زبان فارسی، افزون بر مرتفع کردن مشکل انتخاب ویژگی‌ها، می‌تواند به توسعه سیستمی کارآمد منجر شود که کم‌ترین وابستگی را نیز به دامنه دارد.

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

به این مقاله به شکل زیر استناد کنید:

دورن متن:

(زالی، زودآیند)

در فهرست منابع:

زالی، مجتبی. زودآیند. سیستم شناسایی و طبقه‌بندی موجودیت‌های اسمی در متون زبان فارسی بر پایه شبکه عصبی پژوهشنامه پردازش و مدیریت اطلاعات.

http://ijpm.irandoc.ac.ir (دسترسی در

روز/ماه/سال)

*پدیدآور mojtaba.zali@gmail.com

۱. مقدمه

شناسایی موجودیت‌های اسمی، یک زیروظیفه از استخراج اطلاعات است [۲] و به معنی پردازش مستندات است که در آن به دنبال مکان‌یابی عناصر اسمی در متن و دسته‌بندی آن‌ها به رده‌هایی از پیش تعیین شده 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 وظیفه پایه‌ای است و یکی از هسته‌های پردازشی زبان طبیعی محسوب می‌شود [۳]. در مورد کاربردهای شناسایی موجودیت‌های اسمی می‌توان به موارد مختلفی از جمله ترجمه متون، بازیابی و درک متون، خلاصه‌سازی، جست‌وجوی معنایی و غیره اشاره کرد. به طور کلی کاربردهای گوناگون و بسیار سودمندی را می‌توان برای سیستم‌های شناسایی موجودیت‌های اسمی و به طور عمومی‌تر در پردازش زبان طبیعی، متصور شد.

به طور کلی عملیات شناسایی موجودیت‌های اسمی شامل دو وظیفه است: شناسایی اسامی مربوطه در متن و سپس دسته‌بندی این اسامی به مجموعه‌های از پیش تعیین شده. به طور کلی سیستم‌های شناسایی موجودیت‌های اسمی را می‌توان به سه کلاس اصلی دسته‌بندی کرد. روش‌های مبتنی بر قاعده دستی مشخص شده، سیستم‌های آماری مبتنی بر یادگیری ماشین و سیستم‌های ترکیبی، سه دسته اصلی روش‌های شناسایی موجودیت‌های اسمی بشمار می‌روند. رهیافت‌های مبتنی بر قوانین، بر روی استخراج نام‌ها با استفاده از تعداد زیادی از مجموعه قوانین ساخته شده به صورت دستی تمرکز کرده است. به طور کلی این سیستم‌ها مجموعه‌ای از الگوها هستند که از مشخصه‌های دستوری، نحوی و املائی در کنار ترکیبی از لغت‌نامه‌ها استفاده می‌شود. با توجه به این ویژگی‌ها، روش مبتنی بر قواعد برخلاف کارایی بالا در برخی دامنه‌ها، دارای مشکل بالای غیرقابل انتقال بودن و نیز کمبود قابلیت اطمینان را داراست [۳]. در سیستم‌های مبتنی بر روش‌های آماری و یادگیری ماشین، هدف از رهیافت شناسایی موجودیت‌های اسمی، تبدیل مساله شناسایی موجودیت‌ها به مساله دسته‌بندی و بهره‌گیری از یک مدل آماری برای حل این مسائل است. در این روش‌ها، در صورت استفاده از ویژگی‌های مناسب می‌توان افزون بر ارائه نتایج مطلوب، مدلی را توسعه داد که کمترین وابستگی را به دامنه داشته باشد. از این رو در این پژوهش تمرکز در بکارگیری روش‌های مبتنی بر یادگیری ماشین است. ساختار این مقاله در ادامه به این قرار است: در بخش ۲ مروری بر روی کارهای پیشین در حوزه شناسایی موجودیت‌های اسمی وجود دارد. در بخش ۳ به بررسی روش‌های استخراج ویژگی که در زمان بهره‌گیری از روش‌های آماری از مهم‌ترین فرآیندها محسوب می‌شود، خواهیم پرداخت. در بخش ۴ پیکره

متنی مورد استفاده در این پژوهش معرفی می‌گردد. بخش ۵ به ارائه معماری شبکه 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 پایه روش آموزش با ناظر ارائه شده است. بایکل و همکارانش یک اسم‌یاب آموزش پذیر به نام نایمبل (Nymbel) را مبتنی بر مدل مخفی مارکوف معرفی کرده‌اند [۴]. بورثویک و همکارانش منابع دانش عظیمی را با استفاده از روش بیشترین انتروپی در شناسایی موجودیت‌های اسمی مورد استفاده قرار داده‌اند [۵]. علامت‌گذاری نام‌های ساده ناشناخته با استفاده از درخت تصمیم توسط بچت و همکارانش پیشنهاد شد [۶].

همان‌طور که پیش‌تر اشاره شد در حوزه زبان فارسی به دلیل محدودیت‌های موجود، پژوهش‌های بسیار کم‌تری بر روی شناسایی موجودیت‌های اسمی انجام شده است. سیدعبدالحمید اصفهانی و همکارانش با استفاده از ویژگی‌های گسترده و با بکارگیری شبکه عصبی مبتنی بر چند لایه پرسپترون به شناسایی موجودیت‌های اسمی پرداخته‌اند [۱]. ناجی نیز از یک شبکه عصبی چند لایه به منظور ایجاد مدلی برای شناسایی موجودیت‌های اسمی در زبان عربی استفاده کرده است [۷]. یوکان‌ما و همکاران [۸] در حوزه زبان انگلیسی با استفاده از ویژگی‌های متعدد متنی همانند پیشوند و پسوند هر کلمه، ادات سخن برای کلمات، ویژگی‌های تک-وزنی و دو-وزنی، ویژگی‌ها سندی به همراه ویژگی‌های نمایش برداری متعدد از کلمات به توسعه مدل‌های خود پرداخته‌اند.

فرید احمدی و حامد مرادی نیز در [۹] با استفاده از ویژگی‌های متنی و مدل مخفی مارکوف در زبان فارسی به توسعه یک سیستم شناسایی موجودیت اسمی پرداختند. در این سیستم کارایی برابر با ۵۳,۱۴ بوده است. در زبان فارسی برخلاف زبان انگلیسی، با توجه به ویژگی‌های خاص این زبان، استفاده تنها از ویژگی‌های متنی نمی‌تواند منجر به توسعه یک سیستم کارا شود. از این رو احمدی با ترکیب این ویژگی‌ها با روش مبتنی بر قواعد کارایی سیستم را به ۸۵,۹۳ درصد بهبود داده است.

میران سیوک و همکاران در [۱۰] با استفاده از مجموعه‌ای از ویژگی‌های متنی مشابه با [۸] و مدل‌های نمایش بردارهای عددی همانند بردار سراسری و Word2Vec به توسعه سیستمی پرداخته‌اند. با استفاده از ویژگی‌های متنی به همراه بردار سراسری کارایی برابر با ۷۹,۴۸ و در

زمان استفاده از مدل Word2Vec برابر با ۸۰,۶۸ بوده است.

۳. استخراج ویژگی

در زمان استفاده از روش های آماری و مبتنی بر یادگیری ماشین، ارائه یک نمایش مناسب برای کلمات و همچنین استخراج ویژگی های مطلوب، از مهم ترین بخش های توسعه مدل است. به طور کلی ویژگی های سیستم های شناسایی موجودیت های اسمی را می توان به ۳ دسته کلی ویژگی های کلمه ای، ویژگی های سندی و ویژگی های لیستی تقسیم بندی کرد. ویژگی های کلمه ای، توصیف گر 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 I.B.M و W3C می توانند نشانه های خوبی برای کاندید بودن یک کلمه در مجموعه اسامی خاص باشد.

ویژگی های 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 طراحی شده، بدست می آید. در صورت پردازش کل اسناد و در مواردی که پیکره متنی و اسناد مورد استفاده دارای حجم زیادی باشد، این ویژگی ها قوی تر استخراج شده و دارای اطلاعات بسیار خوبی برای سیستم شناسایی هستند. تعداد رخداد یک کلمه در جمله، پاراگراف و یا کل سند می تواند مثال هایی برای این نوع ویژگی ها باشد.

از دیگر ویژگی های موجود در این حوزه ویژگی های لیستی است. برای استخراج این ویژگی ها، از لیست هایی که شامل اسامی افراد، مکان ها، شهرها، کشورها، سازمان ها و غیره است، استفاده می شود. وجود کلمه ای در هر یک از این لیست ها، نشان دهنده یکی از ویژگی های کلمه است. در برخی موارد وجود یک کلمه در یک لیست می تواند با احتمال زیاد طبقه آن کلمه را مشخص کند اما به دلیل ابهام در نقش و جایگاه اسامی، همواره امکان آن وجود دارد که اسمی در یک لیست قرار داشته باشد اما به طبقه ای دیگر متعلق باشد. بزرگ بودن این لیست ها باعث می شود تا ویژگی هایی با دقت بالاتر استخراج گردد اما همواره این روش با مشکلاتی از قبیل، در دسترس نبودن، نگهداری، و به روز رسانی روبه رو بوده است. لازم به ذکر است که استفاده از ویژگی های لیستی قابل انتقال بودن مدل را بسیار کاهش می هد.

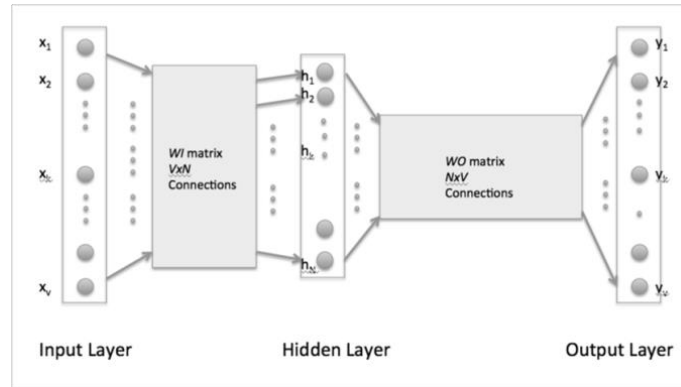
۳-۱. نمایش کلمات

یکی از روش های بسیار کارا برای بهره گیری در یادگیری ماشین در حوزه پردازش زبان طبیعی استفاده از نمایش کلمات به صورت بردارهای عددی است. در فرآیند استخراج ویژگی،

افزون بر استفاده از ویژگی‌هایی که در بخش پیشین مورد بررسی قرار گرفت، می‌توان از نمایش عددی کلمات به عنوان ویژگی که حاوی اطلاعاتی ارزشمند از خود کلمه است، استفاده کرد. پژوهش‌ها نشان دهنده این موضوع است که استفاده از نمایش کلمات می‌تواند در بهبود عملکرد سیستم‌های شناسایی موجودیت‌های اسمی بسیار مفید باشد [۱۱]. مدل‌های متعددی برای نمایش کلمات به صورت بردارهای عددی در حوزه پردازش زبان طبیعی در طی سال‌های پیشین معرفی شده‌اند. از جمله مدل‌های بسیار کارا در برداری‌سازی کلمات، word2vec است که توسط میکولوف در شرکت گوگل طراحی و معرفی شده است [۱۲]. الگوریتم‌های word2vec برای نمایش برداری کلمات با ابعاد بالا مورد استفاده قرار می‌گیرد. اطلاعات استخراج شده توسط این مدل به حدی است که می‌تواند ارتباط معنایی میان کلمات را نیز حفظ کند.

این مدل از یک شبکه عصبی ۳-لایه پرسپترون به منظور تولید بردارها استفاده می‌کند که در شکل ۱ معماری کلی این شبکه نمایش داده شده است. تمامی نورون‌های شبکه عصبی از توابع خطی تشکیل شده‌اند. برای نمایش برداری کلمات در ابتدای کار، واژه‌نامه‌ای از کلمات تشکیل دهنده پیکره‌متنی ایجاد می‌شود. تعداد نورون‌ها در لایه ورودی و لایه خروجی برابر با تعداد کلمات واژه‌نامه و همچنین تعداد نورون‌ها در لایه پنهان برابر با طول بردار مطلوب مورد نیاز است. اتصالات میان لایه ورودی و لایه پنهان و همچنین لایه پنهان و لایه خروجی به صورت اتصالات کامل است. بنابراین اگر طول واژه‌نامه را V در نظر بگیریم و طول مطلوب برای بردار کلمات N باشد، وزن‌های اتصالات میان لایه ورودی به لایه پنهان می‌تواند با ماتریسی به اندازه $V \times N$ و به همین ترتیب وزن‌های لایه پنهان به لایه خروجی با اندازه $N \times V$ نگهداری و نمایش داده شوند. به طور کلی ماتریس لایه ورودی را با عنوان W_I و هم‌چنین ماتریس وزن‌های لایه خروجی را W_O نامیده می‌شود (شکل ۱).

برای آموزش مدل word2vec در این پژوهش از پیکره‌متنی ویکی‌پدیا استفاده شد. از این پیکره در حدود ۶٫۵ میلیون کلمه استخراج و برای فرآیند آموزش مورد استفاده قرار گرفت. برای بهبود فرآیند آموزش شبکه‌های عصبی word2vec و شبکه عصبی عمیق پیشنهاد شده در این پژوهش، در مرحله پیش‌پردازش اطلاعات، داده‌های موجود در پیکره‌متنی که مربوط به کلمات بازدارنده (stop words) بوده‌اند، حذف شد.



شکل ۱. معماری شبکه عصبی مدل word2vec [۱۲]

۲-۳. ویژگی‌های مدل پیشنهادی

تمرکز برای بکارگیری از ویژگی‌های مناسب به منظور 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 درحالی 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 محدودیت‌های موجود در این حوزه را مرتفع کند. از این رو در ادامه، ویژگی‌های بکارگرفته شده برای آموزش مدل با در نظر داشتن اهداف اشاره شده، معرفی شده است.

۱-۲-۳. بکارگیری نمایش بردار word2vec

همان‌طور که در بخش پیش به صورت 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 شبکه عصبی می‌تواند در فهم کلمات، جایگاه و ارتباطات آن‌ها کمک‌کننده باشد. از آنجا که محدوده انتخاب ویژگی‌ها در زبان فارسی بسیار محدودتر از زبان‌های دیگر مانند انگلیسی است، نمایش بردار عددی به عنوان مهم‌ترین ویژگی در نظر گرفته می‌شود. طول بردار برای مدل پیشنهادی با توجه به بررسی‌های انجام شده بر روی پژوهش‌های موجود در این حوزه، برای هر

کلمه ۱۰۰ در نظر گرفته شده است [۸]. 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 تاثیر مرزی، برداری با مقادیر تشکیل دهنده صفر در نظر گرفته شده است [۱۳]. در زمان بکارگیری از مدل‌های پیوسته و به خصوص در حوزه شبکه‌های عصبی، استفاده از این تکنیک یکی از متداول‌ترین روش‌ها در حل مسئله مرزی است.

۲-۲-۳. 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 ارائه شده، شاخص‌ترین ویژگی آن‌ها استفاده از برچسب‌های ادات سخن بوده است [۱]. با توجه به اینکه در مدل پیشنهادی این پژوهش، از ویژگی‌های پیوسته بودن 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 انجام شده بر روی پیکره‌های متنی، مشخص شده است که به صورت معمول میانگین طول کلماتی که اسامی خاص هستند از میانگین طول کلمات دیگر بیشتر است. با پردازش‌های انجام شده در پیکره متنی که در این پژوهش مورد استفاده قرار گرفته است نیز این نتیجه حاصل می‌شود که طول کلمات خاص به طور میانگین در حدود ۱٫۵ نویسه بیشتر از طول کلمات دیگر است (جدول ۱).

جدول ۱. میانگین طول کلمه در پیکره 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

۴. پیکره متنی

ویکی‌پدیای فارسی نام یکی از دانشنامه‌های فارسی زبان در اینترنت است. ویکی‌پدیای فارسی یکی از نسخه‌های ویکی‌پدیا، از پروژه‌های بنیاد ویکی‌مدیا بشمار می‌آید. این دانشنامه فارسی هم‌اکنون بیش از ۵۰۰ هزار مقاله دارد. با بهره‌گیری از مقالات موجود در دانشنامه

ویکی‌پدیا، در حال حاضر پیکره متنی در برگیرنده ۱۶۶۱۷۶ واژه، توسط بهرامی در دانشگاه شریف فراهم شده است. این پیکره از متون ویکی‌پدیا فارسی استخراج و بر اساس استانداردهای MUC-6 برجسب‌گذاری شده است. برجسب‌گذاری 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 جز موارد بالا نیست.

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

جدول ۲. شمار واژه‌های موجود در پیکره متنی

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

۵. معماری شبکه عصبی

با توجه به بررسی‌های انجام شده بر روی مدل‌های مطرح شده در حوزه شناسایی موجودیت‌های اسمی بر پایه شبکه‌های عصبی، استفاده از دو مدل از شبکه عصبی عمیق و شبکه عصبی بازگشتی، از جمله متداول‌ترین مدل‌ها در این حوزه هستند. این دو مدل از شبکه عصبی و به طور کلی شبکه‌های عمیق، قابلیت‌های ارزشمندی را در شناسایی الگوهای بسیار پیچیده دارند [۱۴]. در حوزه پردازش زبان طبیعی و به خصوص حوزه 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 ارائه خروجی مطلوب کمک کننده باشد. با توجه به پیچیدگی‌های بسیار بالا در آموزش شبکه‌های عصبی بازگشتی و به منظور مقایسه روش‌های بکارگرفته شده در این پژوهش با پژوهش‌های موجود در حوزه زبان فارسی، تمرکز مدل پیشنهادی در بهره‌گیری از شبکه عصبی عمیق است.

برای حل مشکل انطباق بیش از حد در شبکه عصبی عمیق مورد استفاده در این پژوهش، از روش حذف تصادفی استفاده شده است. بررسی‌ها نشان دهنده این موضوع است که روش حذف تصادفی می‌تواند با کم‌ترین هزینه و بدون سربار زمانی اضافی برای مرتفع کردن مشکل انطباق بیش از حد در شبکه‌های عصبی، به خصوص در شبکه‌های عمیق و بزرگ، مورد استفاده قرار گیرد [۱۶].

در مدل شبکه عصبی پیشنهادی و در فرآیند به روزرسانی وزن نورون‌ها نیز از روش‌های mini-batch استفاده می‌شود. استفاده از این روش در شبکه‌های عصبی عمیق و بزرگ با تعداد داده‌های بالا که حوزه کاری این پژوهش است، توصیه می‌شود. به منظور حل مشکل کمینه محلی نیز از تکنانه استفاده شده است. بهره‌گیری از تکنانه به همراه نرخ آموزش پایین در روش‌های بروزرسانی mini-batch در شبکه‌های عمیق می‌تواند منجر به پرهیز از کمینه محلی و به روزرسانی دقیق وزن‌ها شود [۱۵].

۶. معیار ارزیابی

با توجه به متوازن نبودن کلاس‌ها در پیکره‌های متنی که از جمله ویژگی‌های زبان طبیعی به شمار می‌آید، استفاده از معیار دقت (accuracy) در سیستم‌های شناسایی موجودیت‌های اسمی نمی‌تواند معیار دقیقی برای ارزیابی سیستم باشد. یک مدل امتیازدهی که به طور استاندارد توسط MUC برای شناسایی موجودیت‌های اسمی مورد استفاده قرار می‌گیرد، دو معیار دقت یا به اختصار P و بازخوانی یا R است. این معیارها در حقیقت از تعاریف موجود در حوزه‌ی بازیابی اطلاعات گرفته شده‌اند و تعریف آن‌ها به صورت 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، T_p پاسخ‌های مثبت است که به درستی مثبت تشخیص داده شده، F_p پاسخ‌های اشتباهی است که به اشتباه درست تشخیص داده شده و در نهایت F_n پاسخ‌های درستی بوده است که سیستم آن‌ها را به اشتباه، غلط تشخیص داده شده 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 می گویند:

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

۷. ارزیابی نتایج

پس از ارزیابی بردار word2vec به عنوان ویژگی شاخص، شبکه عصبی پیشنهادی را با در نظر گرفتن تمامی ویژگی‌ها، مورد آموزش و ارزیابی قرار داده‌ایم. با استناد به جدول ۲ و مقایسه شمار واژه‌های عمومی (کلاس O) نسبت به واژه‌هایی که اسامی خاص هستند (کلاس‌های LOC، ORG، و PERS)، می‌توان به مشکل نامتوازن بودن پیکره‌متنی پی برد. این مشکل تنها محدود به پیکره‌متنی مورد استفاده ما نبوده و از جمله ویژگی‌های زبان طبیعی محسوب می‌شود. به طوری که همواره اسامی خاص، درصد بسیار کمی از متون را به خود اختصاص می‌دهند. از این رو برای آموزش شبکه عصبی، سعی بر آن شد، برای ایجاد یک توازن نسبی، جملاتی که حداقل شامل یک اسم خاص هستند در مجموعه داده‌های تست بکار گرفته شود. به طور کلی ۶۰ درصد از حجم داده برای آموزش و ۴۰ درصد باقی مانده برای تست استفاده شده است. متوازن کردن واژه‌ها و همچنین استفاده از تکنیک‌های آموزش شبکه عصبی با ترتیب ورودی تصادفی، از جمله روش‌های بکار گرفته شده برای بهبود فرایند آموزش بوده است. برای صحت درستی نتایج بدست آمده از مدل پیشنهادی از کلاس‌های موجود در پیکره‌متنی 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 94.6 درصد بوده است.

جدول ۳. نتایج حاصل از اجرای مدل

Name-Entity	Precision	Recall	F-measure
Person	82.89	90.53	86.54
Location	79.04	82.91	80.92
Organization	73.16	67.04	69.96
Overall	78.36	80.16	79.24

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 هم دامنه بودن و استفاده از روش‌های مبتنی بر یادگیری ماشین توجه شده است. مدل ارائه شده توسط سیوک [۱۰] 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. علاوه بر استفاده از ویژگی‌های ادات سخن و طول کلمه در بخش ویژگی‌های مدل پیشنهادی، بهره‌گیری از بردارهای پیشین و پسین word2vec به همراه بکارگیری از شبکه عصبی عمیق، مدل توسعه داده شده در این پژوهش توانسته است، کارایی مشابه با مدل سیوک را ارائه کند. لازم به ذکر است که پیکره متنی مورد استفاده در این پژوهش بر پایه ویکی‌پدیا بوده و در مقایسه با پیکره‌های متنی استاندارد زبان انگلیسی از کیفیت مطلوبی برخوردار نیست.

جدول ۴. مقایسه نتایج حاصل از مدل پیشنهادی با مدل‌های موجود در حوزه مشابه

	Naji [7]	Yukun [8]	Ahmadi [9]	Seok [10]	Proposed Method
Person	69.90	81.7	59.99	NA ¹	86.54
Location	43.30	80.9	57.54	NA	80.92
Organization	59.20	55.7	41.91	NA	69.96
Overall	57.46	72.76	53.14	80.72	79.24

۸. نتیجه‌گیری

نتایج بدست آمده از مدل پیشنهادی نشان دهنده این واقعیت است که استفاده از مدل‌های نمایش بردارهای عددی برای کلمات که به تازگی 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 برخلاف زبان‌های لاتین با محدودیت‌های زیادی همراه است. از این رو مدل‌های نمایش بردار عددی برای کلمات می‌تواند برای مرتفع کردن این مشکلات بسیار موثر باشد. استفاده از این ویژگی‌ها افزون بر افزایش کارایی در سیستم‌های شناسایی موجودیت‌های اسمی، باعث توسعه سیستم‌هایی می‌شود که کم‌ترین وابستگی را به دامنه دارند، حال آنکه بزرگ‌ترین چالش حاضر در توسعه این سیستم‌ها، وابستگی بسیار بالای آن‌ها به دامنه و پیکره‌متنی است که در آن توسعه داده شده‌اند. با استناد به نتایج بدست آمده و مقایسه 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

¹ Not available

اسمی موثر و کارآمد باشد.

فهرست منابع

۱. اصفهانی، سیدعبدالحمید، سعید راحتی قوچانی، و نادر جهانگیری. ۱۳۸۸. «استخراج ویژگی برای یک سیستم شناسایی و طبقه بندی اسامی فارسی» مقاله ارائه شده در پانزدهمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران، تهران.
2. Mihalcea, R. and Moldovan, D.I., 2001. Document indexing using named entities. *Studies in Informatics and Control*, 10(1).
3. Nadeau, D. and Sekine, S., 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), pp.3-26.
4. Appelt, D. and Martin, D., 1999, June. Named entity extraction from speech: Approach and results using the TextPro system. In *Proceedings of the DARPA Broadcast News Workshop* (pp. 51-54).
5. Borthwick, A., Sterling, J., Agichtein, E. and Grishman, R., 1998, August. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proc. of the Sixth Workshop on Very Large Corpora* (Vol. 182).
6. Béchet, F., Nasr, A. and Genet, F., 2000, October. 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.
7. Mohammed, N.F. and Omar, N., 2012. Arabic named entity recognition using artificial neural network. *Journal of Computer Science*, 8(8), p.1285.
8. Ma, Y., Kim, J.J., Bigot, B. and Khan, T.M., 2016, March. Feature-enriched word embeddings for named entity recognition in open-domain conversations. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on* (pp. 6055-6059). IEEE.
9. Ahmadi, F. and Moradi, H., 2015, May. A hybrid method for Persian Named Entity Recognition. In *Information and Knowledge Technology (IKT), 2015 7th Conference on* (pp. 1-7). IEEE.
10. Seok, M., Song, H.J., Park, C.Y., Kim, J.D. and Kim, Y.S., 2016. Named Entity Recognition using Word Embedding as a Feature. *International Journal of Software Engineering and Its Applications*, 10(2), pp.93-104.
11. 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.
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
13. Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D. and Zweig, G., 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3), pp.530-539.
14. Sutskever, I., Martens, J., Dahl, G.E. and Hinton, G.E., 2013. On the importance of initialization and momentum in deep learning. *ICML* (3), 28, pp.1139-1147.
15. Collobert, R. and Weston, J., 2008, July. 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.
16. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), pp.1929-1958.

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,
Mojtaba.Zali@gmail.com

Mohsen Firoozbakht, Assistant Professor, Department of computer engineering, south Tehran branch, Islamic Azad University, Tehran, Iran, Firoozm@gmail.com

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 the 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