مرجع‌گزینی در زبان فارسی با استفاده از شبکهٔ عصبی عمیق

حسین سهلانی[1*]، مریم حورعلی[2] و بهروز مینایی بیدگلی[3]

[1و2] دانشگاه صنعتی مالک اشتر، تهران، ایران

[3] دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

چکیده

در حال حاضر با توجه به کثرت شبکه‌های اجتماعی و شبکه‌های خبری تلویزیونی، رادیویی، اینترنتی و غیره، خواندن تمام متون مختلف و به‌تبع آن تحلیل آن‌ها و دست‌یابی به ارتباطات این متون نیازمند صرف هزینه زمانی و انسانی بسیار بالا است که در عصر کنونی با استفاده از فن‌های مختلف پردازش زبان طبیعی صورت می‌گیرد، یکی از چالش‌های موجود در این زمینه پایین‌بودن دقت سامانه‌های مرجع‌گزینی است که سبب کشف روابط ناصحیح و یا عدم کشف روابط صحیح می‌شود. مراحل کلی حل مسأله مرجع‌گزینی از سه‌گامِ شناسایی موجودیت‌های نامدار، استخراج ویژگی‌های موجودیت‌های نامدار و مرجع‌گزینی آن‌ها تشکیل شده است. موجودیت‌های نامدار ویژگی‌های فراوانی دارند، وجود ویژگی‌های مختلف (متناسب و متناقض با مرجع) در گراف‌ها این امکان را می‌دهند که بتوان حد آستانه‌ای را از ترکیب ویژگی‌های مختلف استخراج کرد. در مقاله ارائه‌شده ابتدا پیش‌پردازش‌های مختلف روی پیکره پژوهشگاه خواجه‌نصیر [1] انجام گرفت؛ سپس با استفاده از الگوریتم‌های مبتنی بر شبکه عصبی عمیق داده‌های موجود به بردارهای عددی تبدیل شدند و پس از آن با استفاده از گراف و با ویژگی‌هایی که در متن مقاله عنوان‌شده هرس اولیه انجام گرفت؛ درواقع رویکردهای مبتنی بر گراف، موجودیت‌ها را همچون مجموعه‌ای از عناصر مرتبط با یکدیگر می‌شناسد که تحلیل روابط میان موجودیت‌های اولیه در گراف و وزن‌دهی به این ارتباط‌ها، منجر به استخراج ویژگی‌های سطح بالاتر و مرتبط‌تری می‌شود و نیز تناقضات ایجادشده بر اساس کمبود اطلاعات را تا حدودی کاهش می‌دهد. سپس با استفاده از شبکه‌های عصبی، روی پیکره مورداشاره در [30] (پیکره آزمون اپسلا) مرجع‌گزینی انجام گرفت که نتایج حاصل بیان‌گر بهبود روش پیشنهادی (رسیدن به دقت ۶۲/۰۹) است که در متن مقاله به‌طور مشروح بیان‌شده است.

واژگان کلیدی: مرجع‌گزینی، گراف، شناسایی موجودیت نامدار، استخراج اطلاعات از متن، شبکه‌های عصبی عمیق.

Corefrence resolution with deep learning in the Persian Labnguage

Hossein Sahlani[1*], Maryam Hourali[2] & Behrouz Minaei-Bidgoli[3]

[1,2]Malek Ashtar University of Technology, Tehran, Iran

[3]School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

Abstract

Coreference resolution is an advanced issue in natural language processing. Nowadays, due to the extension of social networks, TV channels, news agencies, the Internet, etc. in human life, reading all the contents, analyzing them, and finding a relation between them require time and cost.

In the present era, text analysis is performed using various natural language processing techniques, one of the challenges in this field is the low accuracy in detecting name entities' reference, which detection process has been named as coreference resolution. Coreference resolution is finding all expressions that refer to a name entity, and two expressions are coreference together when these expressions located in the same coreference cluster.

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۷/۰۵/۳۰ • تاریخ پذیرش: ۱۳۹۸/۰۶/۱۱ • تاریخ انتشار: ۱۳۹۹/۰۶/۲۴ • نوع مطالعه: پژوهشی

Coreference resolution could be used in many natural language processing tasks such as question answering, text summarization, machine translation, information extraction, etc.

Coreference resolution methods are into two main categories; machine learning and rule-based approaches. In the rule-based approaches for detecting coreferences, a set of rich rule ordinary which written by a specialist is execued. These methods are quick, but these are language-dependent and necessary written to each language firstly again by a specialist. The machine learning method divides into supervised and unsupervised methods, in a supervised approach, it is require to have data labeled by a specialist.

Coreference resolution included three main phases: named entities recognition, features extraction of name entities, and analyzes the coreferences, in which the primary phase is feature extraction.

After corpus creation, name entities should be recognized in the corpus. This step depends on a corpus, in some corpora entities named as golden data, in this paper, we used RCDAT corpus, which determined name entities itself.

After the name entities recognition phase, the mention pairs are determined, and the features are extracted. The proposed method uses two categories of the features: the first is word embedding vector, the second is handcrafted features, which are the distance between the mentions, head matching, gender matching, etc.

This paper used a deep neural network to train the features extracted, in the analyze coreferences phase a Feed Forward Neural Network (FFNN) is trained by the candidate mention pairs (extracted features from them) and their labels (coreference / non-coreference or 1/0) so that the trained FFNN assigns a probability (between 0 and 1) to any given mention pair. Then used the graph technique with a threshold level to determine different or compatible name entities in the coreference resolution cluster. This step creates the graph by using the extracted mention pairs from the previous step. In this graph, nodes are the mention pairs that are clustered by using the agglomerative hierarchical clustering algorithm inorder to locate similar mention pairs in a group. The resulting clusters are considered as coreference resolution chains.

In this paper, RCDAT Persian language corpus is used for training the proposed coreference resolution approach and for testing the Uppsala Persian language dataset which is used and in the calculation of the accurate of system, different tools have been taken for features extraction which each of them effects on the accuracy of the whole system. The corpora, tools, and methods used in the system are standard. They are quite comparable to the ACE and Ontonotes corpora and tools used at the same time in the coreference resolution algorithm. The results of the improvements proposed method (F1 = 62.09) is expressed in the text of the paper.

Keywords: Coreference resolution, Deep neural networks, Graph, Named entities 2ecognition, Information extraction.

۱- مقدمه

با توجه به گسترش روزافزون اطلاعات و ماشینی‌شدن کارها، استخراج اطلاعات از متون، کاری پراهمیت خواهد شد بر این اساس استفاده و شناسایی روش‌هایی که درک قابل قبولی از متون داشته باشند از اهمیت برخوردار خواهد بود که دراین‌بین ابهاماتی درون متون وجود دارد که باید ماشین آن‌ها را رفع کند، به‌طورمعمول حین نوشتن یک متن، برای اشاره به یک موجودیت (یک شخص، سازمان، محل و غیره) تنها از نام آن استفاده نمی‌کنیم، بلکه بسته به شرایط و به‌دلیل جلوگیری از تکرار و بیان اطلاعات بیشتری در مورد آن موجودیت یا تأکید بر یک ویژگی خاص، از عبارات توصیفی مختلف و گاهی ضمایر برای اشاره به آن موجودیت استفاده می‌کنیم. برای مثال ممکن است، برای اشاره یک شخص نام کاملش بیان شود (حسن روحانی)، یا تنها نام خانوادگی (روحانی) او و در موارد غیررسمی‌تر تنها از نام کوچک او استفاده شود. حتی ممکن است شخص یا اشیا با ویژگی‌ها و یا کاربردشان توصیف شود (رئیس‌جمهور ایران و غیره) و یا با استفاده از یک ضمیر (او، ش و غیره). به چنین عباراتی که برای اشاره به یک موجودیت استفاده می‌شوند، موجودیت نامدار گویند؛ بنابراین می‌توان گفت همه موجودیت‌های نامداری که به یک موجودیت یکسان اشاره می‌کنند با یکدیگر هم‌مرجع هستند و با موجودیت‌های نامداری که به موجودیت دیگری اشاره می‌کنند ناهم‌مرجع هستند. مرجع‌گزینی یکی از چالش‌برانگیزترین مسائل حوزه پردازش زبان طبیعی است که عبارت است از تشخیص اینکه چه عبارات اسمی (NPs) یا موجودیت‌های نامداری در متن به یک موجودیت مشترک اشاره دارند [14].

مسأله مرجع‌گزینی یک نکته کلیدی در درک متن است و در مسائلی مثل استخراج اطلاعات، خلاصه‌سازی، پاسخ به سؤالات و ترجمه ماشینی که در آن‌ها درک متن از اهمیت بالایی برخوردار است، کاربرد فراوانی دارد.

در حالت کلی می‌توان روش‌های مرجع‌گزینی را به دو دستهٔ مبتنی بر قاعده و مبتنی بر یادگیری ماشین تقسیم کرد [26]. در روش‌های مبتنی بر قاعده، مجموعه‌ای از قواعد

که به‌صورت دست‌نویس توسط افراد خبره نوشته شده‌اند، به‌ترتیب اجرا می‌شوند تا موارد هم‌مرجعی درون‌متن مشخص شوند. از مزایای این روش می‌توان به‌دقت بالا و سادگی طراحی اشاره کرد، اما قابلیت انعطاف این روش پایین و لازم است برای هر زبان طبیعی مجزا، سامانه دوباره از ابتدا توسط افراد خبره طراحی شود [21].

روش‌های مبتنی بر یادگیری ماشین نیز به دو دستهٔ باناظر و بی‌ناظر تقسیم می‌شوند. در روش‌های باناظر، لازم است داده‌های آموزشی از قبل توسط افراد حاشیه‌نویسی شده باشد [28].

این مقاله با استفاده از ویژگی‌های استخراجی از پیکره مورد اشاره در [1] و همچنین ترکیب این ویژگی‌ها با ویژگی‌های استخراجی از قسمت یادگیری عمیق، سعی کرده که به ویژگی‌های استخراجی غنای بیشتری ببخشد؛ سپس با استفاده از رویکردهای مبتنی بر گراف به تحلیل روابط بین موجودیت‌های نامدار (گره‌های گراف) پرداخته و با اعمال وزن مناسب بین آن‌ها منجر به استخراج ویژگی‌های سطح بالاتر و مناسب‌تر و همچنین با هرس برخی از روابط هم‌مرجعی، تناقضات بین مجموعه موجودیت‌های نامدار هم‌مرجع را از بین برده است، درنهایت نتایج حاصل از اعمال روش پیشنهادی با استفاده از شبکه‌های عصبی، روی پیکره آزمون اپسلا مورد اشاره در [30] نشان داده که این روش نسبت به روش‌های مورد مقایسه در متن مقاله بهبود داشته است.

در ادامه در بخش دوم تاریخچه‌ای از اقدامات صورت گرفته بیان خواهد شد؛ سپس در بخش سوم برای درک بهتر مطالب بعد از بیان مفاهیم و تعاریف ضروری، روش پیشنهادی بیان خواهد شد. در بخش چهارم نتایج حاصل از پیاده‌سازی روش پیشنهادی در مقایسه با سایر روش‌های مرتبط بازگو می‌شود و درنهایت در بخش پنجم جمع‌بندی مقاله صورت می‌پذیرد.

۲- مطالعات مرتبط

کارهای صورت‌گرفته در مرجع‌گزینی را می‌توان در دو دسته استخراج ویژگی برای یافتن ارتباط بین موجودیت‌های نامدار و چگونگی تحلیل و یادگیری ویژگی‌های استخراج‌شده برای هم‌مرجع یا غیر هم‌مرجع شناختن موجودیت‌های نامدار موجود دانست. امروزه در کاربردهای عملی و بر روی پایگاه داده‌های بزرگ سعی می‌شود برای رسیدن به بیشترین دقت از مجموع ویژگی‌ها استفاده شود و به‌طوراساسی مراحل کلی حل مسأله مرجع‌گزینی، از سه گام تشکیل شده است:

- ✓ شناسایی و انتخاب مجموعه‌ای از مرجع‌های برگزیده.
- ✓ اعمال محدودیت‌ها به مجموعه مرجع‌های برگزیده و پالایه‌کردن گزینه‌های غیرمحتمل و یا مرتب‌کردن مراجع بر مبنای مجموعه‌ای از قواعد تقدم.
- ✓ انتخاب محتمل‌ترین مرجع، برای مثال نزدیک‌ترین گزینه با توجه به یک معیار نزدیکی یا موجودیت نامداری با بالاترین رتبه بر مبنای امتیاز ترکیب تقدم‌ها.

مرحله ایجاد پیکره یا حاشیه‌نویسی متون، کاری است که نیاز به زمان و افراد متخصص در این حوزه دارد. همچنین پیشنهادهای زیادی برای نحوه کدکردن اطلاعات تفسیری ارائه شده است. درحقیقت هر پیکره از کدگذاری خاص خود استفاده می‌کند و استاندارد کلی برای آن وجود ندارد که در صورت وجود پیکره‌ای مناسب می‌توان از این قسمت از کار عبور کرد. در حال حاضر پیکره اشاره‌شده در [1] مورد مناسبی برای مرجع‌گزینی است که در ادامه این پیکره تشریح می‌شود.

اعمال محدودیت در مراحل آخر مرجع‌گزینی کاربرد دارد. به‌منظور مرجع‌گزینی یک ضمیر، قواعد محدودیت‌دار، مراجع ناسازگار را حذف کرده و قواعد تقدم مابقی گزینه‌ها را به‌ترتیب میزان مناسب بودنشان مرتب می‌کنند. این قواعد بر مبنای اطلاعات سطوح مختلف زبانی هستند و سعی در اعمال مهم‌ترین قواعد حاکم بر روابط مرجع-ضمیر داشتند [15, 21]. به‌هرحال بالابودن پیچیدگی مسأله مرجع‌گزینی، یکی از دلایل تبدیل‌شدن سامانه‌های مبتنی بر قواعد به سامانه‌های متکی بر یادگیری ماشین در دهه گذشته شد. اعمال قواعد یادگیری ماشین به مجموعه داده‌های بزرگ، باعث مرتب‌سازی و وزن‌گذاری مجموعه ویژگی‌های بزرگ به‌طور بسیار کارآمدتری در مقایسه با سامانه‌های متکی بر قواعد شد.

در شکل (۱) دسته‌بندی روش‌های مختلف مرجع‌گزینی نمایش داده شده است.

(شکل-۱): روش‌های مختلف مرجع‌گزینی
(Figure-1): corefrence resolution methods

برخلاف اختلاف‌های مهم، بیشتر سامانه‌های مرجع‌گزینی موجود (مبتنی بر قواعد دست‌نویس یا مبتنی بر داده) را می‌توان به‌عنوان نمونه‌ای از الگوریتم عمومی در نظر گرفت که در [34] بیان شده است. این الگوریتم در ابتدا یک سند خام D را دریافت کرده و مجموعه‌ای از اتصالات هم‌مرجعی LD را برای آن به‌طور گام‌به‌گام محاسبه می‌کند.

گام نخست الگوریتم سعی می‌کند موجودیت‌های نامدار وابسته موجود در سند D، یعنی مجموعه M را بیاید، بدان معنا که ضمایر و عباراتی که ارجاعی نیستند، حذف می‌شوند.

در گام دوم با استفاده از منابع دانشی مختلف (سامانه‌ها و کتابخانه‌های آماده) استخراج ویژگی موجودیت‌های نامدار انجام می‌شود. روش‌های مختلف مرجع‌گزینی در این زمینه با یکدیگر متفاوت هستند؛ ازاین‌جهت که برخی فرایند استخراج ویژگی موجودیت‌های نامدار را (با تکیه‌بر ماژول‌های پیش‌پردازشی مثل برچسب‌زن اجزای کلام، تشخیص موجودیت‌های نامدار، تجزیه‌گرها) به‌طورکامل خودکار انجام می‌دهند و برخی دیگر از اطلاعات استاندارد طلایی استفاده می‌کنند.

بعد از استخراج ویژگی باید وزن‌دهی ویژگی‌ها را به‌طور مطلوبی انجام داد که درنهایت آخرین گام الگوریتم خوشه‌بندی است. خروجی این گام انتخاب موجودیت‌های هم‌مرجع است [36].

در جهت بهبود عملکرد الگوریتم‌های یادگیری ماشین گام نخست ایجاد نمونه‌های آموزشی و ایجاد زنجیره‌های هم‌مرجعی است که بارزترین آن‌ها مدل‌های زوج اشاره[1] (دو موجودیت نامدار را یا به‌عنوان هم‌مرجع یا به‌عنوان غیر هم‌مرجع دسته‌بندی می‌کند)، موجودیت-اشاره[2] (به‌جای دسته‌بندی هر موجودیت نامدار با موجودیت نامدار دیگر، آن را با یک مرجع قبلی مقایسه می‌کند که درعمل از روش خوشه‌بندی استفاده می‌کند) و مدل رتبه‌ای[3] (به‌جای مقایسه تنها دو موجودیت نامدار با یکدیگر، با مجموعه‌ای از موجودیت‌های نامدار مقایسه کرده و نتیجه مقایسه، تعیین رتبه هر موجودیت نامدار است) است [28]. در گام دوم بهبود ویژگی‌هایی است که در تعیین موارد هم‌مرجعی به کار می‌روند که در بین آن‌ها ویژگی‌های معنایی توجه بیشتری را به خود جلب کرده‌اند. جدول (۱) سامانه‌های مرجع‌گزینی در دسته‌های مختلفی که در این قسمت بیان شده تقسیم‌بندی کرده است.

[1] Pairwise/mention-pair model
[2] entity-based/entity-mention model
[3] Ranking model

(جدول-۱): روش‌های مختلف مرجع‌گزینی
(Table-1): corefrence resolution method

رویکرد	مدل دسته‌بندی	فرایند اتصال	نحوه یادگیری
[31]	زوج-اشاره	بهینه‌سازی محلی	ناظر
[19]			
[28]			
[16]			
[23]	رتبه‌ای		
[12]			
[5]	موجودیت اشاره		
[41]		بهینه‌سازی سراسری	
[33]	زوج اشاره	خوشه‌بندی	
[9]			
[10]			
[13]		گراف	
[18]			
[11]			
[22]		خوشه‌بندی	نیمه‌ناظر
[24]	موجودیت اشاره		ناظر
[40]			
[38]			
[8]		ابر گراف	
[25]		گراف	
[39]			
[27]			
[14]		بهینه‌سازی سراسری	بی‌ناظر
[20]			
[17]			
[21]			

علاوه‌بر مواردی که در جدول (۱) مشاهده می‌شود که روش‌های مورداستفاده در مقالات را بررسی می‌کرد می‌توان از ویژگی‌های استفاده‌شده در مقالات نیز بهره برد مانند ویژگی تطبیق واژه سر در موجودیت‌های نامدار و غیره که در روش پیشنهادی نیز برخی از این روش‌ها مورداستفاده واقع‌شده ولی باید توجه داشت بسیاری از ویژگی‌ها وابسته به زبان هستند که در تمام زبان‌ها قابل‌استفاده نیستند که در قسمت مربوط به روش پیشنهادی به‌طور مشروح بیان شده است.

همان‌طور که مشاهده می‌شود در جدول (۱) یک دسته‌بندی کلی از روش‌های پیشین آورده شده که در این بین می‌توان برای برخی از روش‌ها چندین دسته را در نظر گرفت به‌عنوان‌مثال روش‌های [24, 40, 38] هم از روش موجودیت اشاره استفاده کرده‌اند و هم از خوشه‌بندی که

به‌نوعی مدل رتبه‌ای را شامل می‌شود؛ همچنین برای روش‌های [9, 10] علاوه‌براین روش‌های [9, 10] که مورد تأکید بیشتر این مقاله است هر دو از روش‌های یادگیری عمیق جهت بهبود کار خود استفاده کرده‌اند؛ بدین‌صورت که ابتدا با استفاده از بردار تعبیه واژگان استخراج‌شده از موجودیت‌های نامدار، زوج اشاره‌های استخراج‌شده را مشخص کرده‌اند؛ سپس با استفاده از خوشه‌بندی آن‌ها در دسته‌های هم‌مرجع سامانه ترکیبی و دقیقی را جهت مرجع‌گزینی ایجاد کرده‌اند. گفتنی است برخی از تنظیمات این مقاله با توجه به تنظیمات موجود در آن‌ها در نظر گرفته شده است.

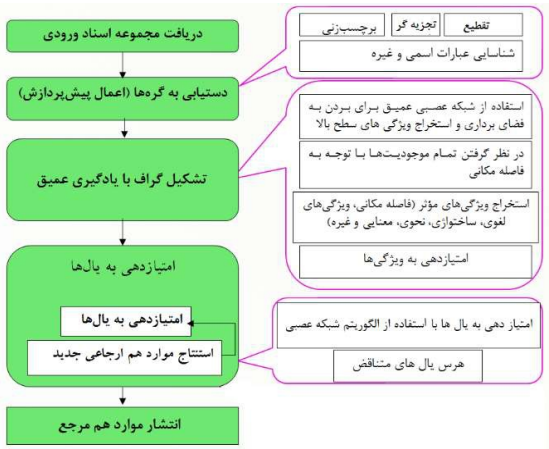
۳- روش پیشنهادی

در این بخش روشی مبتنی بر گراف و شبکه‌های عصبی عمیق برای مرجع‌گزینی و اقدامات انجام‌شده در این خصوص، بیان و معرفی می‌شوند. هر سند، جمله یا متن ورودی پس از اتمام گام پیش‌پردازش گره‌های گراف را شکل می‌دهند (با توجه به اینکه در برخی از پیکره‌ها موجودیت‌های نامدار به‌طور دستی مشخص شده‌اند، می‌توان در همان ابتدا گراف را تشکیل داد) و پس از اتمام این مرحله گرافی از موجودیت‌های نامدار تشکیل می‌شود که وزن یال‌های آن در ابتدا صفر در نظر گرفته شده است.

مرحله بعدی، استفاده از ویژگی‌های استخراجی از موجودیت‌های نامدار است که دراین‌بین هم از ویژگی‌های مهم و مورداستفاده در مقالات مختلف و هم از شبکه‌های عصبی عمیق استفاده شده است. خروجی حاصل از شبکه عصبی عمیق یک بردار عددی است که اعداد آن بین صفر تا یک است که در کنار سایر ویژگی‌هایی که از موجودیت‌های نامدار استخراج می‌شود، مشخص می‌کند که جفت موجودیت نامدار معرفی‌شده با چه احتمال یا امتیازی می‌تواند در یک خوشه هم‌مرجع قرار گیرند. همان‌طورکه عنوان شد مرحله استخراج ویژگی تنها منوط به شبکه‌های عصبی عمیق نیست و سایر ویژگی‌های مهم را نیز شامل می‌شود که دراین‌بین از ویژگی‌های فاصله‌ای نقش مهمی را ایفا می‌کنند که در قسمت مربوط به ویژگی‌ها به‌طور کامل بیان شده است.

به‌بیان‌دیگر مرحله بعد از استخراج ویژگی برای هر یک از موجودیت‌های نامدار، مقایسه هر یک از آن‌ها با سایر موجودیت‌های نامدار دیگر است که موجودیت‌های نامداری که فاصله کمتری از هم دارند، در یک خوشه قرار می‌گیرند و به‌عنوان هم‌مرجع شناخته می‌شوند.

با استفاده از ویژگی‌های استخراج‌شده از موجودیت‌های نامدار و تعیین میزان شباهت هر یک از آن‌ها به هم می‌توان گرافی را تشکیل داد که گره‌های آن، موجودیت‌های نامدار و وزن یال‌های آن میزان شباهت و یا میزان اختلاف هر یک از آن‌ها باشد، پس از تخصیص وزن‌ها و تشکیل گراف، حال نوبت به هرس گره‌های تکی یا غیر وابسته می‌رسد. در نمودار جعبه‌ای شکل (۲) می‌توان مراحل الگوریتم پیشنهادی را بهتر درک کرد.



(شکل-۲): نمودار جعبه‌ای روش پیشنهادی
(Figure-2): block diagram of proposed method

روش پیشنهادی در چهار بخش کلی تقسیم‌بندی می‌شود که عبارتند از:

مرحلهٔ نخست، در این مرحله پیکره موردنظر انتخاب و بارگذاری می‌شود، دست‌یابی به پیکره در برخی زبان‌ها با مشکلات عدیده‌ای روبه‌رو است و تاحدودی پژوهش‌گر باید در ابتدا پیکره موردنظر را ایجاد کند ولی با توجه به وجود پیکره اشاره‌شده در [1] برای زبان فارسی، در این مقاله از این پیکره استفاده شده است.

مرحله دوم مربوط به اعمال پیش‌پردازش‌های اولیه است که خروجی این مرحله دست‌یابی به عبارات اسمی در متون ورودی است که اقداماتی که در این مرحله صورت می‌گیرد، بسته به پایگاه داده ورودی متفاوت و تاحدودی می‌توان در برخی از پایگاه داده‌ها از این مرحله صرفه نظر کرد.

مرحله سوم مربوط به تشکیل گراف است، گراف اولیه با توجه به عبارات اسمی شناسایی‌شده از مرحله قبل و

با توجه به فاصله مکانی عبارات اسمی از هم تشکیل می‌شود و ویژگی‌های مربوط به هر یک را استخراج می‌کند.

تعیین وزن بین گره‌ها یا وزن بین موجودیت‌های نامدار یا گره‌ها، بر اساس ارتباطشان با سایر گره‌ها، مشخص می‌شود که روابط معنایی و مشابهت‌های بین هر دو گره بر اساس شبکه‌های عصبی عمیق و سایر ویژگی‌ها محاسبه شده و بر اساس آن به هر یال معتبر (میان دو گره) یک وزن نسبت می‌دهد.

مرحله چهارم، مرحله اصلی الگوریتم است و در آن با توجه به قواعد موجود در زبان گراف تشکیل‌شده هرس می‌شود؛ سپس سایر ویژگی‌های استخراج‌شده در مرحله قبل را با درنظرگرفتن امتیازشان مرحله‌به‌مرحله لحاظ می‌کنند و امتیاز (وزن) یال‌های ایجادشده را بروز می‌کنند، ممکن است در به‌روزرسانی وزن یال‌ها برخی ویژگی‌ها باعث هرس برخی از یال‌ها شوند. نمودار جعبه‌ای روش کلی در شکل (۲) آورده شده است. در ادامه قسمت‌های مختلف روش پیشنهادی را تشریح خواهیم کرد.

۱-۳- دریافت مجموعه اسناد ورودی[1]

در امر مرجع‌گزینی مشاهده شده است که استفاده از روش‌های یادگیری ماشینی نتایج بهتری داشته اما بزرگ‌ترین مشکل این روش‌ها برای زبان‌هایی مانند زبان فارسی کمبود دادهٔ برچسب‌گذاری شده است. برای رفع این مشکل لازم است که پیکره‌ای با داده‌های مناسب و حجم مناسب برچسب‌گذاری شود. پیکره‌های مرجع‌گزینی مختلفی برای مرجع‌گزینی ایجاد شده است که می‌توان پیکره‌های زبان انگلیسی را جزء کامل‌ترین پیکره‌های موجود در نظر گرفت؛ اما در زبان فارسی پیکره‌های ایجادشده جامعیت نداشته و در همین اواخر پیکره اشاره‌شده در [1] ایجاد شده که قابل‌مقایسه با پیکره‌های مطرحی چون پیکره CoNLL است.

این پیکره شامل بیش از یک‌میلیون واژه فارسی است که برچسب دستی و خودکار هم‌مرجعی و موجودیت نامدار را دارا است (آمار نهایی پیکره، در جدول ۲ مشاهده می‌شود)، این برچسب‌ها شامل، برچسب اجزای کلام (۱۰۰ برچسبی و برچسبی ۱۶)، برچسب موجودیت نامدار (۱۳ برچسبی) و برچسب قطعه است. در شکل (۳) مثالی از اسناد پیکره مشاهده می‌شود.

(جدول-۲): اطلاعات آماری پیکره مرجع‌گزینی زبان فارسی موجود در [1]

(Table-2): KNT corefrence resolution corpus [1]

موضوع	شمارش داده‌ها
تعداد تکواژ	۱۰۵۲۶۳۷
تعداد اسناد	۱۵۹۹
کل کلمات برچسب خورده	۴۰۸۱۳۸
تعداد موجودیت‌های نامدار برچسب خورده	۸۶۹۶۰
تعداد موجودیت‌های نامدار جاندار	۴۲۶۶
تعداد موجودیت‌های نامدار ضمیر	۲۶۸۷
تعداد موجودیت‌های نامدار اسامی خاص	۳۴۱۲

همان‌طورکه در جدول (۲) مشاهده می‌شود و در شکل (۳) نشان داده‌شده است، منظور از تکواژ هر واژه یا بخش مجزایی است برچسب اجزای کلام داشته باشد، به ازای هر تکواژ یک سطر مجزایی در پیکره وجود دارد (نمونه آن در شکل (۳) نشان داده‌شده است). سند متنی است که مورد تحلیل واقع شده (اولین ستون در شکل (۳))، واژه هر بخش مجزایی است که در پیکره با یک فاصله از هم جدا شده‌اند، موجودیت‌های نامدار موجود در پیکره در سه دسته ضمایر، اسامی خاص و موجودیت‌های نامدار جاندار تقسیم‌شده‌اند که تعداد هر یک در جدول (۲) مشخص شده است. برچسب‌های مشاهده‌شده در شکل (۳) به‌ترتیب شامل موارد زیر هستند. نام سند، شماره جمله، واژه، برچسب اجزای کلام (شانزده برچسبی)، ریشه واژه یا خود واژه، اصل واژه (بدون در نظر گرفته پیشوند و ...) یا خود واژه، موجودیت نامدار (برچسب سه تایی طلایی)، اندیس واژه در زنجیره هم‌مرجعی خود (برچسب طلایی)، شماره زنجیره هم‌مرجعی که واژه در آن واقع است (برچسب طلایی)، نوع موجودیت نامدار (برچسب طلایی)، جانداری (برچسب طلایی)، نوع عبارت (برچسب طلایی)، برچسب اجزای کلام صدتایی.

گفتنی است، تمامی اسناد پیکره قبل از برچسب‌گذاری با ابزارهای پیش‌پردازشی نرمال‌سازی، غلط‌یابی و تقطیع شده‌اند و مرز جمله‌ها نیز در آن‌ها مشخص‌شده است.

[1] پیکره مرجع‌گزینی

```
2.coref.txt 5   با  P   O   با   با   -   -   -   -   B-PP    P
2.coref.txt 5   توجه   N   O   توجه   توجه   -   -   -   -   B-NP    N-SING-COM
2.coref.txt 5   به  P   O   به   به   -   -   -   -   B-PP    P
2.coref.txt 5   تاثير   N   O   تاثير   تاثير   -   -   -   -   B-NP    AJ-COMP
2.coref.txt 5   آمار   N   O   آمار   آمار   -   -   -   -   I-NP    N-SING-COM
2.coref.txt 5   تجاري   N   O   تجاري   تجاري   -   -   -   -   I-NP    AJ-COMP
2.coref.txt 5   كشور   N   O   كشور   كشور   Location(*  10(*    1(* Entity(*    NO(*    I-NP    N-SING-COM
2.coref.txt 5   چين N   O   چين چين Location*)  10*)    1*) Entity*)    NO*)    I-NP    AJ-COMP
2.coref.txt 5   بر  P   O   بر   بر   -   -   -   -   B-PP    P
2.coref.txt 5   بازار   N   O   بازار   بازار   -(* 11(*    11(*    Other(* NO(*    B-NP    N-SING-COM
2.coref.txt 5   فلزات   N   O   فلزات   فلزات   *   *   *   *   I-NP    N-PL-COM
2.coref.txt 5   اساسي   PUNC   O   اساسي   اساسي   -*) 11*)    11*)    Other*) NO*)    O   AJ-COMP
2.coref.txt 5   همچون   ADV O   همچون   همچون   -   -   -   -   -   B-ADVP  ADV-EXM
2.coref.txt 5   مس  N   O   مس   مس   -(* 13(*    13(*    Other(* NO(*    B-NP    N-SING-COM
2.coref.txt 5   ،   PUNC   O   ،   ،   *   *   *   *   O   DELM
2.coref.txt 5   فولاد   N   O   فولاد   فولاد   *   *   *   *   B-NP    N-SING-COM
2.coref.txt 5   و   CONJ   O   و   و   *   *   *   *   B-CONJP CON
2.coref.txt 5   ديگر   N   O   ديگر   ديگر   *   *   *   *   B-NP    AJ-COMP
2.coref.txt 5   فلزات   N   O   فلزات   فلزات   *   *   *   *   I-NP    N-PL-COM
2.coref.txt 5   پيوسته   N   O   پيوسته   پيوسته   -*) 13*)    13*)    Other*) NO*)    I-NP    AJ-COMP
2.coref.txt 5   به  P   O   به   به   -   -   -   -   B-PP    P
2.coref.txt 5   اين  N   O   اين   اين   -(* 14(*    13(*    Other(* NO(*    B-NP    AJ-COMP
2.coref.txt 5   كالاها   N   O   كالاها   كالاها   -*) 14*)    13*)    Other*) NO*)    I-NP    N-PL-COM
2.coref.txt 5   ،   PUNC   O   ،   ،   -   -   -   -   O   DELM
2.coref.txt 5   رشد N   O   رشد رشد   -   -   -   -   B-NP    N-SING-COM
2.coref.txt 5   اقتصادي   V   O   اقتصادي   اقتصادي   -   -   -   -   I-NP    AJ-COMP
2.coref.txt 5   چين CONDET   O   چين چين   -   -   -   -   I-NP    V-SUB-NEG
2.coref.txt 5   عامل   N   O   عامل   عامل   -   -   -   -   I-NP    N-SING-COM
2.coref.txt 5   اصلي   PUNC   O   اصلي   اصلي   -   -   -   -   -   I-NP    AJ-COMP
2.coref.txt 5   حركت   N   O   حركت   حركت   -   -   -   -   I-NP    N-SING-COM
2.coref.txt 5   اين PUNC   O   اين   اين   -(* 12(*    11(*    Other(* NO(*    I-NP    AJ-COMP
2.coref.txt 5   بازارها N   O   بازارها   بازارها   -*) 12*)    11*)    Other*) NO*)    I-NP    N-PL-COM
2.coref.txt 5   خواهد   V   O   خواهد   خواهد   -   -   -   -   -   B-VP    V-AUX-FUT-POS
2.coref.txt 5   بود V   O   بود بود   -   -   -   -   I-VP    V-COP-PA-POS
2.coref.txt 5   .   PUNC   O   .   .   -   -   -   -   O   DELM
```

(شکل-۳): مثالی از اسناد پیکره با برچسب‌های دستی و خودکار موجود در [1]

(Figure-3): an example of corpus files [1]

داده‌های مورداستفاده در پیکره از وبگاه‌های پربازدید خبری فارسی در بازه ماه ۱۲ میلادی سال ۲۰۱۶ تهیه‌شده‌اند، گفتنی است که داده‌های انتخاب‌شده برای پیکره دارای تنوع موضوعی بوده تا نماینده مناسبی برای زبان فارسی باشد. ازاین‌رو وبگاه‌های خبری و موضوعات مورداستفاده در پیکره عبارت‌اند از:

«اقتصادی، فنّاوری، سیاسی، ورزشی، اجتماعی، فرهنگی و هنری»

وبگاه‌های خبری موردبحث عبارت‌اند از:

«خبرگزاری فارس، خبرگزاری مهر، خبرگزاری جمهوری اسلامی (ایرنا)، خبرگزاری دانشجویان ایران (ایسنا)، همشهری آنلاین، تابناک، فرارو، ورزش ۳، انتخاب، باشگاه خبرنگاران جوان»

نسبت انتخاب متون از خبرگزاری‌ها یا وبگاه‌های خبری با توجه به تعداد اخبار و گستردگی پوشش خبری متفاوت بوده و از برخی از آن‌ها تعداد کمی سند انتخاب‌شده است.

۲-۳- دست‌یابی به گره‌ها (یافتن موجودیت‌های نامدار)

نخستین گام در حل مسأله مرجع‌گزینی یافتن موجودیت‌های نامداری است که بایستی مرجع آن‌ها مشخص شود. نوع عباراتی که به‌عنوان موجودیت نامدار شناسایی می‌شوند به چندین عامل بستگی دارد ازجمله: کاربرد موردنظر، زبان متن، نوع و دامنه متن؛ برای مثال در یک سامانه که هدف آن مرجع‌گزینی ضمیرها است، موجودیت-های نامدار عبارت‌اند از: ضمیر سوم شخص، ضمیر ملکی و برای سامانه‌ای که هدف آن مرجع‌گزینی است (در حالت کلی) موجودیت‌های نامدار عبارت‌اند از عبارات اسمی و ضمایر (ضمیر شخصی، ضمیر اشاره، ضمیر انعکاسی و ضمیر نسبی و ضمیر مالکیت و نسبیت).

برای بازشناسی موجودیت‌های نامدار در متون حاشیه‌نویسی‌نشده در مرحله آزمون مشکلاتی وجود دارد که علت اصلی آن‌ها عدم هم‌خوانی موجودیت‌های نامدار به‌دست‌آمده و موجودیت‌های نامدار استاندارد طلایی است. برای مثال، موجودیت‌های نامدار، اغلب تودرتو هستند (برای رفع این مشکل باید استاندارد MUC7 در نظر گرفته شود)، مرز آن‌ها با موارد موجود در استاندارد طلایی متمایز است و ممکن است مواردی از آن‌ها یافته نشده یا مواردی یافته شود که در استاندارد طلایی وجود ندارد [24].

به‌منظور تولید ورودی‌های یک الگوریتم دسته‌بندی‌کننده موجودیت‌های نامدار هم‌مرجع، بایستی موجودیت‌های نامدار موجود در متن شناسایی و استخراج

شوند. برای این منظور بایستی عملیات پیش‌پردازش روی متن انجام شود. خروجی این مراحل مرزهای خوش‌تعریف برای موجودیت‌های نامدار و اطلاعاتی در مورد موجودیت‌های نامداری است که قرار است در مرحله استخراج ویژگی‌ها استفاده شوند. مراحل دست‌یابی به گره‌ها در ادامه تشریح می‌شود.

۱-۲-۳- پیش‌پردازش

در مرحله پیش‌پردازش، با استفاده از یک ماژول پیش‌پردازش، متون انتخاب‌شده یکنواخت، تقطیع و غلط یابی می‌شوند. این ماژول برای یکسان‌سازی نویسه‌ها در متن، تغییر رمزگذاری[1]، حذف علائم، اشکال و اضافات متن و به‌طورکلی ایجاد یک متن تمیز و آماده برای اعمال پردازش‌های متنی است. این ماژول‌ها با روش‌های مبتنی بر قانون تهیه شده‌اند و اقداماتی که در این پیکره بر روی متن ورودی انجام شده عبارت است از:

۱- نرمال‌سازی: تصحیح رمز (کدگذاری)، یکسان‌سازی نویسه‌ها و حذف نویسه‌های ناشناخته؛

۲- تصحیح نقطه‌گذاری: تصحیح نشانه‌گذاری‌ها، جداسازی علائم از حروف و تصحیح یکی بودن علائم جفت، مثل پرانتز؛

۳- تقطیع؛

۴- شکستن متن به جملات؛

۵- شکستن جملات به واژگان.

دقت هر یک از مراحل پیش‌پردازشی مورداستفاده در پیکره [1] در جدول (۳) آورده شده است.

(جدول-۳): دقت ابزارهای پیش‌پردازشی مورداستفاده در پیکره [1]

(Table-3): Preprocessing tools accuracy in corpus [1]

ردیف	نوع ابزار	دقت تقریبی
1	POS	98
2	NP-chunker	70
3	NER	85
4	Paragraph Splitter	98
5	Sentence Splitter	98
6	Tokenizer and Spellchecker	93

همان‌طورکه در جدول (۳) نشان داده شده، دقت ابزارهای پیش‌پردازشی مورداستفاده در حد کامل نیست که این خود سبب کاهش دقت نهایی سامانه مرجع‌گزینی خواهد شد.

[1] encoding

۲-۲-۳- تشخیص موجودیت نامدار

در این پیکره برای تشخیص موجودیت‌های نامدار از دو ویژگی اصلی برچسب اجزای کلام و فهرست واژگان (شامل پیکره اعلام [2] و پیکره ایجادشده در [29] است، این فهرست شامل اسامی اشخاص و مکان‌ها و ... است) برای تشخیص موجودیت‌های نامدار استفاده شده که در سه بند زیر تشریح شده است.

مهم‌ترین منابع مورداستفاده برای تشخیص موجودیت‌های نامدار عبارت است از:

۱- استفاده از پیکره متنی زبان فارسی اشاره شده در [1] برای تولید برچسب‌گذار اجزای کلام که برچسب‌گذارهای تولیدشده به شرح زیر است [1, 3]:

- ✓ برچسب‌گذار پایه: این برچسب‌گذار از مجموعه ۱۵ برچسبی استفاده می‌کند. این برچسب‌ها شامل برچسب‌های درشت‌دانه موجود در پیکره متنی مانند اسم، فعل، صفت و ...
- ✓ برچسب‌گذار کسره اضافه: این برچسب‌گذار از مجموعه ۲۶ برچسبی استفاده می‌کند. در این برچسب‌گذار علاوه بر پانزده برچسب مورداستفاده در برچسب‌گذار پایه، یک برچسب کسرهٔ اضافه نیز به مجموعه برچسب‌ها اضافه‌شده که تعداد برچسب‌ها را از ۱۵ به ۲۶ برچسب افزایش داده است.
- ✓ برچسب‌گذار با مجموعه ۳۳ برچسب: این برچسب‌ها شامل برچسب‌های درشت‌دانه موجود در پیکره هستند که در هر مقوله جزئیات بیشتری به آن‌ها اضافه‌شده است. به‌عنوان مثال برای اسامی علاوه‌بر مشخص‌کردن اسم، جمع و مفردبودن اسم نیز مشخص شده است. در افعال، زمان فعل و در قید و صفت نوع قید و صفت مشخص شده است.
- ✓ برچسب‌گذار با مجموعه یکصد برچسب: این برچسب‌گذار علاوه‌بر مقوله کلی (که در سایر مدل‌های برچسب‌گذاری آمده است) برای هر واژه جزئیات بیشتری را در برمی‌گیرد. در این مجموعه برچسب تأکید بر تشخیص جزئیات فعل است. این مدل برچسب‌گذاری قابلیت تشخیص زمان، نوع و شخص فعل، پیشوند و پسوند فعل و ... را دارد.

در این پیکره برای تشخیص موجودیت‌های نامدار از برچسب اجزای کلام یکصدبرچسبی برای تشخیص موجودیت‌های نامدار استفاده‌شده است.

۲-علاوه بر موارد یادشده از سایر منابع لغوی قوی نیز استفاده شده است؛ مانند پیکره ایجادشده در [29] (برای تعیین طبقه موجودیت‌های نامدار از جهت موجود زنده‌بودن و ...) که در تعیین موجودیت‌های نامدار، هرس موجودیت‌های موجود در گراف و به‌تبع آن تشخیص درست هم‌مرجعی می‌تواند کارساز باشد.

۳-استفاده از پیکره اعلام [2]، برای تولید برچسب موجودیت‌های نامدار: داده‌های مورداستفاده در این پیکره بیش از ۵۵۰ هزار تکواژ است. تعداد برچسب‌های موجودیت نامدار به‌کاررفته در پیکره، سیزده موجودیت، شامل: شخص، مکان، سازمان، رخداد، تاریخ، بازه، زمان، عدد اصلی، عدد ترتیبی، درصد، پول و اندازه است.

۳-۳- تشکیل گراف

در مرحله تشکیل گراف فرض بر این است که تمام موجودیت‌های نامدار بازشناسی شده به‌عنوان گره‌های گراف هستند و تمام موجودیت‌ها به هم وصل هستند با فرض این‌که امتیاز هر یال در ابتدا صفر است. در جهت امتیازدهی و به‌روزرسانی وزن یال‌ها لازم است از قواعد دستور زبان و سایر ویژگی‌های استخراجی دیگر نیز استفاده شود.

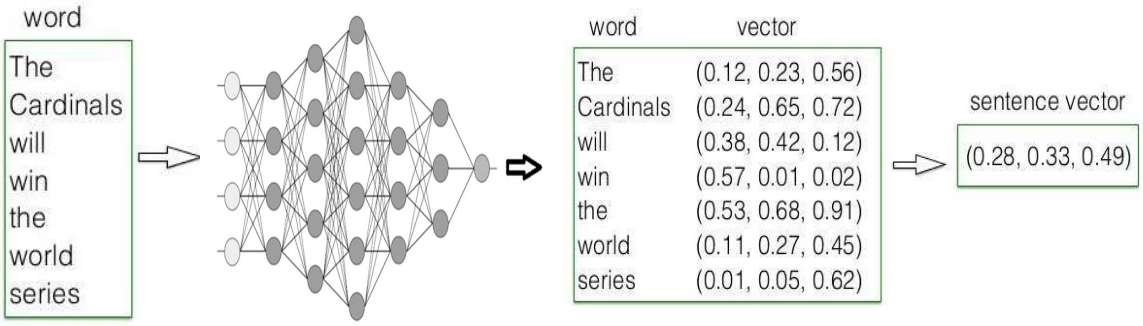
مرجع‌گزینی به‌کمک گراف این امکان را فراهم می‌کند که از بروز تناقضات و خطاهایی که در اثر ناکافی‌بودن اطلاعات در مسائل مرجع‌گزینی به‌وجود می‌آیند، جلوگیری کند و امکان به‌کارگرفتن اطلاعات جدید در این مدل فراهم می‌شود.

موجودیت‌های نامدار گره‌های تشکیل‌دهندهٔ گراف هستند و گره‌ها یا موجودیت‌های وابسته در خوشه‌های یکسان قرار می‌گیرند و در هر گراف گره‌ها به نسبت به یک گره موجودیت نامدار (انتخاب شده) دو صورت هستند: دستهٔ نخست گره‌هایی هستند که ارتباط چندانی با این موجودیت نامدار ندارند و درواقع با اطلاعات و ویژگی‌های این موجودیت نامدار تناقض دارند و دستهٔ دوم گره‌هایی هستند که اطلاعات آن‌ها با اطلاعات این موجودیت نامدار (گرهٔ انتخاب‌شده) همخوانی داشته و باهم به‌اصطلاح هم‌مرجع هستند. زمانی که یک موجودیت نامدار با یک موجودیت نامدار یا مجموعه‌ای از موجودیت‌های نامدار ارتباط دارد، بین این دو گروه یال وزن‌داری قرار می‌گیرد که وزن این یال با استفاده از الگوریتم‌ها و مدل‌های مختلف محاسبه می‌شود (ممکن است این ارتباط اشتباه تشخیص داده شود و در مراحل بعدی هرس شود). نحوه وزن‌دهی به یال‌ها با استفاده از ویژگی‌هایی است که در ادامه تشریح می‌شود.

۱-۳-۳- استفاده از شبکه‌های عصبی برای ایجاد بردار تعبیه کلمات

شبکه‌های عصبی به‌دلیل قابلیت ترکیب ویژگی‌ها و ایجاد ویژگی‌های جدید در لایه‌های بالاتر به‌طور گسترده در زمینه‌های مختلف یادگیری ماشین مورداستفاده قرار گرفته‌اند. مدل‌های شبکه عمیق توسعه‌یافته مدل‌های شبکه عصبی برای یادگیری تبدیل‌های غیرخطی روی داده‌ها هستند و به‌نوعی تلاش می‌کند مفاهیم انتزاعی سطح بالا را با استفاده یادگیری در سطوح و لایه‌های مختلف مدل کنند.

هدف استفاده از الگوریتم‌های یادگیری عمیق در شاخه پردازش متن ارائه تحلیل کاملی از موارد مشخص‌شده در متن است و در این بین مرجع‌گزینی نیاز به تحلیل پیشرفته موجودیت‌های نامداری است که در مرحله قبل از مرجع‌گزینی یافت شده‌اند. به همین سبب از شبکه‌های عصبی عمیق برای دست‌یابی به‌دقت بالاتر در نتایج استفاده شده است. یکی از ره‌یافت‌ها برای اینکه بتوان از انواع روش‌های عددی حوزه یادگیری ماشین مانند بیش‌تر الگوریتم‌های دسته‌بندی روی لغات و اسناد استفاده کرد، نمایش برداری واژگان و جملات است. به‌عنوان‌مثال فرض می‌شود فرهنگ لغتی با N واژهٔ مرتب‌شده به‌ترتیب الفبایی، وجود دارد و برای نمایش هر واژه، برداری با طول N که شماره آن خانه بردار، همان اندیس لغت مربوطه در این فرهنگ لغت است. با این پیش‌فرض، برای هر لغت یک بردار به طول N که همه خانه‌های آن به‌جز خانه متناظر با آن لغت صفر است، وجود دارد (خود ستون متناظر با لغت عدد یک است). با این ره‌یافت، هر متن یا سند را هم می‌توان با یک بردار نشان داد که به‌ازای هر واژه و لغتی که در آن به‌کار رفته است، ستون مربوطه از این بردار برابر تعداد تکرار آن لغت خواهد بود و تمام ستون‌های دیگر که نمایان‌گر لغاتی از فرهنگ لغت هستند که در این متن به کار نرفته‌اند، برابر صفر خواهد بود. با وجود سادگی، این روش هم نیاز به فضای ذخیره‌سازی زیادی دارد و هم پیچیدگی الگوریتم و زمان اجرای آن‌ها را بسیار بالا است. از طرف دیگر در این روش فقط واژگان و تکرار آن‌ها مهم است؛ ولی ترتیب واژگان یا موضوع متن (اقتصادی، علمی، سیاسی و ...) تأثیری در مدل ندارد.



(شکل-۴): مثالی از تبدیل متن به بردار [32]

(Figure-4): example of text to vector convertor

روش دیگر برای این منظور استفاده از الگوریتم Word2Vec گوگل [32] است که روشی کارآمد و مناسب برای نمایش لغات و متون و پردازش آن‌ها است. در این روش به‌کمک شبکه عصبی عمیق یک بردار با اندازه ثابت برای نمایش تمام لغات و متون در نظر گرفته شده، اعداد مناسب در مرحله آموزش برای هر لغت محاسبه می‌شود و هر لغت در این فضا یک نمایش منحصربه‌فرد می‌گیرد. بعد از ایجاد بردارهای مرتبط با هر لغت، برای نمایش برداری هر متن یا خبر، می‌توان بردار تک‌تک واژگان به‌کاررفته در آن را یافته و میانگین اعداد هر ستون را به دست آورد که نتیجه آن، یک بردار برای هر متن یا سند خواهد بود (شکل ۴). بدین منظور در این مقاله از پیکره همشهری و بی‌جن‌خان [4, 7] برای آموزش و بردار سازی لغات استفاده شده است. پیکره همشهری و بی‌جن‌خان هم از حیث اعتبار، پیکره مناسبی هستند و هم اینکه لغات به‌کاررفته در آن‌ها از جامعیت خوبی برای زبان فارسی برخوردار هستند؛ ضمن این‌که در صورت استفاده از لغتی که در پیکره وجود نداشته باشد می‌توان در حین آموزش برای آن بردار متناسب با آن لغت برایش ساخت. نتایج حاصل از اجرای الگوریتم Word2Vec روی پیکره همشهری ایجاد بردارهای منحصربه‌فردی با طول پنجاه است (با تنظیمات مختلف می‌توان این عدد را تغییر داد، با توجه کارهای انجام‌شده در زبان انگلیسی [9, 10]. در جهت استفاده از بردار تعبیه واژگان و کاهش پیچیدگی‌های الگوریتم این تعداد پنجاه در نظر گرفته شده است).

۲-۳-۳- توابع ویژگی

در این قسمت توابع و یا ویژگی‌های رایجی که مورداستفاده قرارگرفته توصیف می‌شود که برای استخراج هر ویژگی یک یا دو موجودیت نامدار به‌منظور ارزیابی یک ویژگی بررسی می‌شود. در سامانه‌های یادگیری باناظر، نمونه‌های یادگیری با جفت‌کردن موجودیت‌های نامدار mi و mj و برچسب‌گذاری آن‌ها با برچسب‌های درست (هم‌مرجع) یا نادرست (ناهم‌مرجع) ایجاد می‌شدند؛ برای مثال نمونه (mi, mj, Boolean) درست است اگر و تنها اگر mi و mj هم‌مرجع باشند (البته نه‌تنها دو موجودیت نامدار بلکه چندین موجودیت نامدار می‌توانند باهم هم خوشه و هم‌مرجع باشند و بیان دوبه‌دو برای فهم بهتر مطلب است). زوج‌های (mi, mj) با بردارهای ویژگی که شامل ویژگی‌های تکی یا توصیفی باشد، برای مثال اطلاعاتی در مورد یکی از موجودیت‌های نامدار، (توصیف یک موجودیت نامدار با برشمردن خواصی مثل دسته واژگان، شیء است و یا موجود زنده و تعداد آن) و یا مقایسه‌ای با زوج ویژگی (اطلاعاتی در مورد ارتباط بین دو موجودیت نامدار، برای مثال هم‌خوانی در تعداد و جنسیت) است نمایش داده می‌شوند.

این توابع را می‌توان به‌منظور ارزیابی گروهی از موجودیت‌های نامدار نیز به کار برد. به‌عنوان‌مثال تابع GENDER سازگاری جنسیت دو موجودیت نامدار را بررسی کرده و نتایج y در صورت همخوان‌بودن، n در صورت ناسازگاربودن و u در صورت نامشخص‌بودن جنسیت (دست‌کم یکی از) موجودیت‌های نامدار را برمی‌گرداند. می‌توان به‌راحتی این تابع را به‌منظور ارزیابی یک موجودیت جزئی (گروهی از موجودیت‌های نامدار) تعمیم داد.

گفتنی است که برخی از ویژگی‌ها که در زبان‌های مختلف به‌خصوص زبان انگلیسی مورداستفاده قرار می‌گیرد، ویژگی تطبیق جنسیت و تطبیق عددی است که در مرجع‌گزینی بسیار حائز اهمیت است؛ برای مثال تطبیق جنس ضمیر و مرجع آن (به‌خصوص ضمایر زبان انگلیسی)، در حالی در زبان فارسی چنین ویژگی‌ای در رابطه با جنس

وجود ندارد و یا دارای اهمیت کمتری است. ویژگی مهم دیگری که در زبان انگلیسی به‌کار می‌رود، ویژگی تطبیق عدد است (ضمایر مفرد به یک عبارت اسمی مفرد و ضمایر جمع به یک عبارت اسمی جمع اشاره دارند) درصورتی‌که ضمایر مفرد زبان فارسی می‌توانند به یک عبارت اسمی جمع اشاره داشته باشند و گاهی جهت احترام به اشخاص، به‌جای ضمیر مفرد از ضمیر جمع استفاده می‌شود.

با توجه به موارد عنوان‌شده و چالش‌هایی که در زبان فارسی وجود دارد، ویژگی‌های استخراجی برای مرجع‌گزینی درروش پیشنهادی با توجه به فرضیات زیر در نظر می‌گیریم:
الف) فرض می‌شود که m=(m1,...,mn) مجموعه‌ای از موجودیت‌های نامدار درون‌سندی با n موجودیت نامدار است و (mi,mj) زوج موجودیت نامداری است که در آن داریم j>i.
ب) ویژگی‌ها استخراج‌شده در یک بردار ویژگی قرار می‌گیرند که هر بردار ویژگی بیان‌گر ویژگی‌های یک جفت موجودیت نامدار است.

بردار ویژگی‌ای که در مقاله برای هر جفت موجودیت نامدار به‌کار می‌رود عبارت است از:
«موجودیت نامدار نخست، موجودیت نامدار دوم، فاصله دو موجودیت نامدار (برحسب جمله)، در یک‌جمله‌بودن دو موجودیت نامدار، ضمیربودن موجودیت نامدار نخست، ضمیربودن موجودیت نامدار دوم، مطابقت هسته یا ریشه دو موجودیت نامدار، تطبیق نخستین واژه، تطبیق دقیق موجودیت‌های نامدار، مطابقت کلی موجودیت‌های نامدار (بعد از حذف اضافات، پسوندها و ...)، زیررشته‌بودن موجودیت نامدار دوم برای موجودیت نامدار نخست، تعریف‌کننده+ گروه اسمی، تطابق شمار (مفرد و جمع)، تطابق جانداری، مطابقت وابسته موجودیت نامدار دوم با هسته موجودیت نامدار نخست، داشتن ضمیر اشاره در موجودیت نامدار دوم، تطابق نوع موجودیت نامدار دو عبارت اسمی»

علاوه‌بر شانزده ویژگی بالا پنجاه ویژگی نیز از مرحله یادگیری عمیق استخراج می‌شود که درمجموع تعداد ویژگی‌ها برای هر جفت موجودیت نامدار به ۶۶ عدد می‌رسد. در جدول (۴) ویژگی‌های مورداستفاده به‌صورت دسته‌بندی‌شده آورده شده است.

برای مثال در عبارت موجود در شکل (۳): «با توجه به تأثیر آمار تجاری کشور چین بر بازار فلزات اساسی همچون مس، فولاد و دیگر فلزات پیوسته به این کالاها، رشد اقتصادی چین عامل اصلی حرکت این بازارها خواهد بود.» که دو موجودیت نامدار «بازار فلزات اساسی» و «این بازارها» که باهم، هم‌مرجع نیز هستند، بردار ویژگی برای این دو موجودیت نامدار به‌طور زیر است.

بازار فلزات اساسی، این بازارها،

feat۵۰ + ۱،۰ ،۰ ،۰،۰ ،۰ ،۰ ،۰ ،۰ ،۱ ،۱ ،۱ ،۱ ،۱،۱،۱

+ ۵۰ عدد از مرحله شبکه عصبی عمیق (بردار تعبیه واژگان) که درمجموع تعداد ویژگی‌ها به ۶۶ عدد می‌رسد.

(جدول-۴): توابع ویژگی
(Table-4): features

نوع ویژگی	انواع مقادیر برای هر ویژگی					
تشابه رشته‌ای	تطبیق هسته	تطبیق نخستین واژه	تطبیق کلی	تطبیق دقیق	زیررشته بودن موجودیت نامدار دوم برای موجودیت نامدار نخست	تطابق وابسته عبارت موجودیت نامدار دوم با هسته موجودیت نامدار نخست
نحوی	تعریف‌کننده+ گروه اسمی	تطابق شمار (مفرد و جمع)	وجود ضمیر اشاره در موجودیت نامدار دوم	تطابق نوع موجودیت نامدار	ضمیر بودن موجودیت نامدار نخست	ضمیربودن موجودیت نامدار دوم
معنایی	تطابق جنسیت	تطابق جانداری				
گفتمان	فاصله موجودیت‌های نامدار	در یک جمله بودن دو موجودیت نامدار				

در رابطه با بردار تعبیه واژگان و چگونگی تشکیل این بردار، باید عنوان کرد که به‌ازای هر لغت یا موجودیت نامدار یک بردار پنجاه‌تایی ایجاد می‌شود و زمانی که یک موجودیت نامدار بیش از یک واژه باشد، برای تشکیل بردار تعبیه لغات

آن همان‌طور که در شکل (۴) نشان داده شده است، میانگین بردار واژگان، بردار تعبیه عبارت جدید را ایجاد می‌کند و زمانی که نیاز است بردار تعبیه واژگان برای هر جفت موجودیت نامدار محاسبه شود، اتفاقی که رخ می‌دهد اختلاف بردار تعبیه واژگان هر دو موجودیت نامدار با استفاده از روش‌های مختلف (در این مقاله اقلیدسی) محاسبه و نمایش داده می‌شود.

جدول (۳) توابع مورداستفاده در ارزیابی مرجع‌گزینی موجودیت‌های نامدار را درون سند نمایش می‌دهد. ویژگی فاصله، از مهم‌ترین ویژگی‌های مرجع‌گزینی بوده و زمانی که با ویژگی‌های دیگر ترکیب شود، اثرگذاری بیشتری خواهد داشت: به‌عنوان‌مثال، دو موجودیت نامدار با نامی به‌طوردقیق مشابه بدون توجه به فاصله بین ایشان هم‌مرجع هستند (البته زمانی که آن دو هم‌نگاره یا هم نویسه نباشند، برای مثال ایران هم می‌تواند نام کشور باشد و هم نام شخص)، اما سازگاری جنسیت بین یک ضمیر و یک اسم خاص، تنها درون‌جمله‌ای یکسان یا بافاصله یک جمله از هم قابل‌قبول است و نه برای فاصله‌های بیشتر [14]. پس ویژگی فاصله، یک ویژگی مکمل برای بسیاری از دیگر ویژگی‌ها به‌حساب می‌آید و به همین خاطر در جدول (۳) با عناوین مختلفی بیان‌شده است. به عبارتی انتخاب مرجع برگزیده به‌طورمعمول با توجه به نظریه نزدیکی مکانی انجام می‌شود؛ زیرا فرض بر این است که محدوده مرجع یک موجودیت نامدار نزدیک به خود موجودیت نامدار است؛ سپس برای بهبود عملکرد و بالابردن دقت، سایر ویژگی‌ها نیز مطرح می‌شوند.

۴-۳- امتیازدهی به یال‌ها (اتصال یا انفصال گره‌ها)

در فرایند اتصال گره‌ها به هم به‌طورمعمول یک گراف بدون جهت در نظر گرفته می‌شود و هر گره به یک گروه وابسته (یک موجودیت یا گروهی از موجودیت‌ها) با استفاده از یک یال وصل می‌شود. وزن یال‌ها با استفاده از الگوریتم‌ها و ویژگی‌های مختلفی بیان می‌شود. ویژگی‌های متعددی در دو دسته موافق و مخالف تقسیم‌بندی می‌شوند که هرکدام باعث افزایش یا کاهش وزن یال‌ها می‌شوند که عبارت‌اند از [37]:

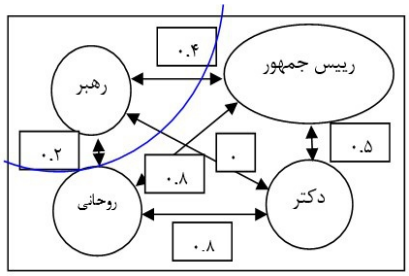
الف) ویژگی‌های مربوط به فاصله، تطابق موجودیت‌های نامدار، ویژگی‌های مستخرج از شبکه عصبی عمیق و غیره برای بالابردن وزن یال‌ها مورداستفاده قرار گیرند.

ب) ویژگی‌هایی مانند عدم رعایت جنسیت، عدم رعایت تعداد در ضمایر و غیره برای پالایه‌کردن یال موردنظر مورداستفاده قرار می‌گیرد.

بعد از وزن‌دهی یال‌ها گراف‌ها با توجه به وزن یال‌ها، برش داده می‌شود تا محتمل‌ترین بخش‌بندی پیدا شود. به‌بیان‌دیگر الگوریتم یال‌هایی را به‌منظور جداکردن گروه‌هایی که موجودیت‌های مجزا را نشان می‌دهند، قطع می‌کند.

نکته حائز اهمیت دیگر چگونگی هرس‌کردن یال‌ها است که درروش پیشنهادی بر اساس یک‌میزان آستانه[1] مشخص می‌شود. این نرخ را می‌توان با آموزش اولیه سامانه و نتایج اولیه آن‌ها تشخیص داد. درنهایت بخش‌بندی نهایی مشخص می‌کند که کدام موجودیت‌های نامدار هم‌مرجع و کدام غیر هم‌مرجع هستند که این اطلاعات برای بهبود کارایی سامانه مرجع‌گزینی مفید خواهند بود.

به‌عنوان‌مثال شکل (۵) یک مثال از نحوه نمایش موجودیت‌های نامدار و ارتباطاتشان توسط گراف را نمایش می‌دهد. در این مثال الگوریتم مرجع‌گزینی، تصمیم به قطع یال‌های موجودیت نامدار رهبر و درنتیجه زنجیره‌ای حاوی موجودیت‌های نامدار دکتر، رئیس‌جمهور و روحانی ایجاد می‌کند (اقتباس‌شده از [11]).



(شکل-۵): مثالی از هرس گراف بر مبنای وزن یال‌ها.
(Figure-5): example of edges prune

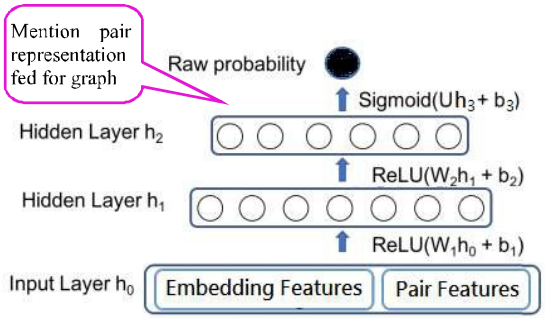
برای امتیازدهی به یال‌ها در روش پیشنهادی از شبکه عصبی عمیق استفاده شد؛ بدین منظور در ابتدا با استفاده از داده‌های آموزش (پیکره اشاره‌شده در [1]) موجودیت‌های نامدار مثبت و منفی (هم‌مرجع و ناهم‌مرجع) با ۶۶ ویژگی استخراج‌شده به‌عنوان ورودی به شبکه عصبی داده شد و سپس با استفاده از مدل ایجادشده داده‌های آزمون مورد ارزیابی قرار گرفتند. در شکل (۶) معماری شبکه عصبی عمیق مورداستفاده در روش پیشنهادی نشان داده شده است.

[1] threshold

لایه ورودی: برای هر یک از موجودیت‌های نامدار یک بردار ورودی (W) در نظر گرفته‌شده است. این بردار ورودی شامل دو دسته ویژگی، بردار تعبیه واژگان (در قسمت ۳-۳-۱ توضیح داده شده) و بردار ویژگی استخراج‌شده (در قسمت ۳-۳-۲ توضیح داده شده)، است که طول این بردار نیز ۶۶ است.

لایه‌های پنهان: بردار ویژگی‌ها به‌عنوان ورودی به لایه پنهان نخست داده می‌شود. تعداد لایه‌های پنهان دو عدد در نظر گرفته‌شده و هر لایه پنهان به‌صورت اتصال کامل[1] به لایه قبلی متصل شده و تابع مورداستفاده در لایه‌های پنهان (ReLU[2]) است. تعداد نرون‌های لایه‌های پنهان به‌ترتیب ۵۰ و ۲۵ در نظر گرفته شده است.

لایه خروجی: خروجی آخرین لایه پنهان به لایه خروجی داده می‌شود که خروجی نهایی این لایه عددی بین صفر تا یک است (با توجه به اینکه در لایه خروجی از سیگموید[3] استفاده شده است، خروجی شبکه عصبی به‌صورت احتمالی، نشان می‌دهد که آیا زوج اشاره‌ها می‌توانند باهم، هم‌مرجع باشند و یا خیر) که با توجه به آستانه ۰/۸ (شکل ۷) موارد هم‌مرجع و ناهم‌مرجع را مشخص کرده است.

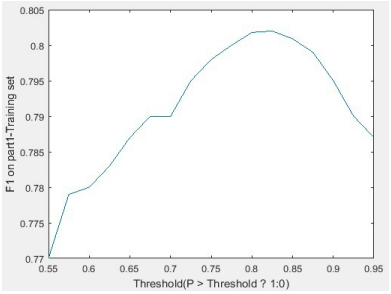


(شکل-۶): معماری شبکه عصبی

(Figure-6): structure of neural network

برای به‌دست‌آوردن حد آستانه بهینه با درنظرگرفتن مقدار F1 بر اساس حد آستانه نمودار شکل (۷) نشان داده شده است. برای این منظور تنها از معیار MUC استفاده شده است (با توجه به پیچیدگی زمانی پایین این معیار ارزیابی) و تنها روی بخش نخست (برای کاهش پیچیدگی زمانی) پیکره آزمون محاسبات صورت گرفته است.

[1] fully connected
[2] rectified linear units
[3] sigmoid



(شکل-۷): مقدار F1_MUC برحسب آستانه‌های مختلف

(Figure-7): F1_muc value vs different Threshold

همان‌طورکه در شکل (۷) نشان داده شده، مقدار ۰/۸ با توجه به سایر مقادیر در حد مطلوب است.

۴- ارزیابی و نتایج

برای ارزیابی روش پیشنهادی از پیکره آزمون اپسلا [30] استفاده شده است، این پیکره درمجموع شامل ۶۱۴ جمله و ۱۶۲۷۴ واژه (متوسط ۲۶/۵ واژه در هر جمله) است که به‌طورکامل قابل‌مقایسه با بخش آزمون و توسعه پیکره‌های MUC-6 (پیکره آزمون رقابت MUC-6 دارای ۱۳ هزار تکواژ است) و MUC-7 (پیکره توسعه رقابت MUC-7 دارای ۱۷ هزار تکواژ است) است. این پیکره طبق دستورالعمل مرجع‌گزینی CoNLL2012، برچسب‌گذاری شده و دارای برچسب‌های دقیق است؛ این مجموعه به چهار قسمت تقسیم‌بندی شده که اساس این تقسیم‌بندی بر این است که در هر سند یک روایت وجود داشته باشد تا مرجع‌گزینی معنادار باشد. برای آموزش مدل نیز همان‌طور که قسمت روش پیشنهادی اشاره شد، از پیکره موجود در [1] استفاده می‌شود.

۴-۱- معیارهای ارزیابی

در سامانه‌های مرجع‌گزینی، مشکل اصلی در تعریف یک معیار کارایی مناسب، مشخص‌نبودن تعداد کامل موجودیت‌های نامدار موجود درون پیکره است. این مشکل زمانی شدیدتر می‌شود که موجودیت‌های نامدار به‌دست‌آمده توسط سامانه[4] با موجودیت‌های نامدار مشخص‌شده توسط استاندارد طلایی[5] منطبق نباشند؛ به‌علاوه موجودیت‌های نامدار متفاوتی که توسط تفاسیر متفاوت در نظر گرفته‌شده‌اند، تأثیر مستقیمی روی پیچیده‌شدن ارزیابی یک متن خاص خواهد داشت؛ بنابراین، نتایج به‌دست‌آمده در

[4] system mentions
[5] true mentions

پیکره‌های مختلف تفاوت زیادی با یکدیگر خواهد داشت. ضمن این‌که بیشتر الگوریتم‌ها از ویرایش پس از اجرای نتایج بهره می‌برند که تأثیر به‌سزایی روی نتایج دقت و بازخوانی به‌دست‌آمده خواهد داشت. در این مقاله سعی شده از معیارهای استانداردی که در مقالات مختلف دیده‌شده استفاده شود و فرضیات در نظر گرفته‌شده همگی استاندارد و موردقبول عام باشند. در ادامه این معیارها تعریف‌شده‌اند.

اجلاس MUC نخستین‌بار معیار ارزیابی مرجع‌گزینی را با عنوان معیارهای ارزیابی MUC تعریف کرد [35] برای حل نقاط ضعف معیار MUC معیارهای B^3 [6] و CEAF [26] به‌عنوان پراستفاده‌ترین جایگزین‌ها معرفی شدند. تعاریف این معیارها به‌صورت زیر است:

در معیار MUC، خوشه‌های هم‌مرجع استخراجی از سامانه و موجود در استاندارد را با هم مقایسه می‌کند. مقدار بازیابی، نسبت تعداد پیوندهای مشترک در سامانه و استاندارد بر تعداد پیوندهای استاندارد محاسبه و دقت، نسبت تعداد پیوندهای مشترک استخراج‌شده در سامانه و موجود در استاندارد بر تعداد پیوندهای سامانه محاسبه می‌شود.

$$R = \frac{\#\text{ common links in true and system partition}}{\#\text{ minimum links for true partition}}$$

$$P = \frac{\#\text{ common links in true and system partition}}{\#\text{ minimum links for system partition}}$$

در معیار B^3 به‌جای اینکه به پیوندهای بین عبارات نگاه شود، به خود عبارات و حضور و یا عدم حضور آن‌ها در یک کلاس هم‌ارزی توجه می‌شود. درنتیجه مقدار بازیابی و دقت برای هر عبارت محاسبه و سپس باهم ادغام می‌شوند که بازیابی و دقت نهایی را تولید کنند.

$$R = \frac{\sum_{i=1}^{n} \frac{\#\text{ common mentions in true and system entity of mention}_i}{\#\text{ mentions in true entity of mention}_i}}{n}$$

$$P = \frac{\sum_{i=1}^{n} \frac{\#\text{ common mentions in true and system entity of mention}_i}{\#\text{ mentions in system entity of mention}_i}}{n}$$

در معیار B^3 ممکن است عبارات حاصل از سامانه یا عبارات موجود در استاندارد بیش از یک‌بار در محاسبه دقت نقش داشته باشند. برای رفع این مشکل در معیار CEAF تلاش شده است که رابطه بهینه یک‌به‌یکی میان عبارات سامانه و استاندارد پیدا شود. به‌دلیل این‌که این رابطه یک‌به‌یک است، در این معیار از تمام عبارات موجود در مجموعه سامانه و استاندارد استفاده نمی‌شود. معیار CEAF در دو دسته مبتنی بر موجودیت[1] و مبتنی بر موجودیت نامدار[2] تقسیم‌بندی می‌شود.

$$R/P = \frac{\#\text{ common mentions in best one-to-one aligned true and system entities}}{\#\text{ mentions in true/system partition}}$$

معیار CoNLL نیز میانگین حسابی (بدون وزن) سه معیار پرکاربرد MUC، B^3 و CEAF محاسبه می‌شود.

$$Conll = (muc+b^3+ceafm+ceafe)/4$$

برخلاف تعریف معیارهای ارزیابی کامل‌تر نسبت به معیار MUC، این معیار همچنان مورداستفاده قرار می‌گیرد و دلایل این کار عبارت است از: (۱) مقایسه با سامانه‌های قدیمی که تنها از معیارهای MUC برای ارزیابی استفاده کرده بودند و (۲) عدم وجود یک معیار استاندارد، زیرا هیچ‌یک از معیارهای ارزیابی دارای برتری محسوسی نیستند.

۲-۴- نتایج حاصل از پیاده‌سازی

در ادامه نتایج حاصل از پیاده‌سازی روش پیشنهادی در مقایسه با روش (پژوهشگاه خواجه‌نصیر) [1] در جدول (۵) آورده شده است که نتایج نشان می‌دهد هرس برخی از موجودیت‌های نامدار سبب شده که مقدار فراخوان پایین‌تر بیاید و درنتیجه مقدار نهایی F1 در هر دو معیار MUC و B3 مقدار محدودی بهبود یابد.

(جدول-۵): مقایسه روش پیشنهادی
(Table-5): result of proposed method

	معیار	muc	b3	ceafe	ceafm	conll
روش [1]	بخش ۱	77.27	54.9	64.56	56.78	61.62
	بخش ۲	63.3	56.81	59.17	58.78	59.51
	بخش ۳	77.01	51.16	52.64	46.07	56.72
	بخش ۴	66.42	55.93	59.88	59.36	60.39
	میانگین	69.25	54.7	59.06	55.24	**59.56**
روش پیشنهادی	بخش ۱	80.02	56.86	66.97	58.04	65.47
	بخش ۲	64.89	58.73	61.05	60.66	61.33
	بخش ۳	78.9	52.79	54.45	50.09	59.05
	بخش ۴	68.59	57.45	61.97	61.98	62.49
	میانگین	73.1	56.45	61.11	57.69	**62.09**

در جدول (۵) نتایج حاصل از مقایسه روش پیشنهادی و روش [1] با استفاده از پیکره آزمون اپسالا [30] آورده شده است.

[1] entity-based CEAF (CEAFe)
[2] mention-based CEAF (CEAFm)

(جدول-۶): مقایسه روش پیشنهادی

(Table-6): compare proposed method

الگوریتم	Deep Learning used (Y/N)	Neural Network architecture(s) used	Pre-trained Word Embeddings Used	Graph based (Y/N)
[1]	N	-	-	no
روش پیشنهادی	Y	FFNN	Persian:50d word2vec	yes

در جدول (۶) نشان می‌دهد که روش پیشنهادی نسبت به روش [1] چه نوآوری‌هایی داشته است، همان‌طور که مشاهده می‌شود، استفاده از شبکه عصبی عمیق و بردار تعبیه واژگان و همچنین ارائه روشی مبتنی بر گراف از نوآوری‌های روش پیشنهادی بوده که سبب بهبود دقت روش پیشنهادی شده است.

گفتنی است که روش [1] با استفاده از ویژگی‌های استخراجی به مرجع‌گزینی پرداخته است. برخی از این ویژگی‌ها در جدول (۴) آورده شده است و در روش پیشنهادی که از شبکه‌های عصبی عمیق به همراه ویژگی‌های منتخب به مرجع‌گزینی پرداخته دقت سامانه مرجع‌گزینی [1] را به‌طورتقریبی سه درصد بهبود داده است. روش‌های قابل‌مقایسه دیگری در حوزه زبان فارسی برای هم‌مرجعی وجود نداشت [1]، (روش‌های قبلی موجود از پیکره‌های ضعیف‌تری استفاده کرده‌اند و یا هیچ منبعی برای آن‌ها وجود ندارد که قابل‌مقایسه باشند و در [1] به‌طور کامل تشریح شده است) به همین خاطر روش پیشنهادی با تنها روش موجود قابل‌مقایسه در زبان فارسی مورد مقایسه قرارگرفته و نتایج بیان‌گر بهبود دقت در روش پیشنهادی است.

۵- نتیجه‌گیری و جمع‌بندی

در این مقاله، ابتدا در رابطه با مرجع‌گزینی، مشکلات موجود در حل مسائل مرجع‌گزینی مطالبی بیان و علاوه‌بر کلیات مربوط به مرجع‌گزینی پیکره فارسی مورداستفاده در [1] (که به‌منظور مرجع‌گزینی ایجادشده) تشریح و سپس مراحلی که برای رسیدن به موجودیت‌های نامدار باید طی کرد بیان شد. در مرحله بعدی چگونگی قرارگرفتن دو یا چند موجودیت نامدار در یک خوشه و یا دسته عنوان شد که برای این مهم از روش مبتنی بر گراف به همراه ویژگی‌های استخراج‌شده در قسمت مربوطه و شبکه‌های عصبی عمیق استفاده شد. درنهایت نتایج حاصل از روش پیشنهادی در مقایسه با روش [1] مورد ارزیابی قرار گرفت که نتایج نشان می‌دهند که روش پیشنهادی توانسته دقت تشخیص مرجع‌گزینی را تا حدود سه درصد بالاتر ببرد؛ ولی بااین‌وجود به نظر می‌رسد، می‌توان با به‌کارگیری ویژگی‌های پیچیده‌تری مانند ویژگی‌های نحوی، کارایی را بهبود داد.

سپاس‌گزاری

از مرکز تحقیقات مخابرات ایران به خاطر دراختیارگذاشتن مستندات و پیکره فارسی مربوط به پروژه مرجع‌گزینی تقدیر و سپاس‌گزاری می‌شود.

۶- مراجع 6- References

[۱] رحیمی زینب، حسین نژاد شادی. هم‌مرجع‌یابی مبتنی بر پیکره در متون فارسی. پردازش علائم و داده‌ها. ۱۳۹۹; ۱۷ (۱) :۹۸-۷۹

[1] Z. Rahimi, S. HosseinNejad "Corpus based coreference resolution for Farsi text", *JSDP*, vol. 17 (1), pp. 79-98, 2020.

[۲] حسین نژاد، شادی؛ شکفته، یاسر و امامی آزادی، طاهره. «پیکره اعلام، یک پیکره استاندارد موجودیت‌های نامدار فارسی»؛ پردازش علائم و داده‌ها، دوره ۱۴، شماره ۳؛ صص. ۱۲۷-۱۴۲، ۱۳۹۶.

[2] Y. Shekofteh, T. Emami Azadi, "A'laam Corpus: A Standard Corpus of Named Entity for Persian Language", *JSDP*, Vol. 14 (3), pp.127-142, 2017.

[۳] سادات‌مرتضوی، پونه؛ شمس‌فرد، مهرنوش. «شناسایی موجودیت‌های نامدار در متون فارسی.» پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر، تهران، ۱۳۸۸.

[3] P. S. Mortazavi and M.Shamsfard "Recognition of named entities in Persian texts," *in 15-th annual conference of computer society of Iran*, Tehran, 2009.

[4] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A Standard Persian text collection", *Knowledge-Based Systems*, Vol. 22(5), pp.382-387, 2009.

[5] A. Rahman, Ng. Vincent, "Coreference resolution with world knowledge," *49th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 2013.

[6] B. Amit, B. Breck, "Algorithms for scoring coreference chains", *In Proceedings of the LREC Workshop on Linguistic Coreference*, pp. 563–566, 1998.

Conference on Artificial Intelligence (IJCAI-18) putational Linguistics: Human Language Technologies, pp. 1148–1158, 2011.

[17] J. Shanshan, Y. Li, T. Qin, Q. Meng, and B. Dong, "SRCB entity discovery and linking (EDL) and event nugget systems for TAC 2017", In Proceedings of the Text Analysis Conference, 2017.

[18] P. Haoruo, Y. Song, and D. Roth, "Event detection and co-reference with minimal supervision", *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 392–402, 2016.

[19] P. S. Paolo, S. Michael, "Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution," main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2014.

[20] H. Poon, P. Domingos, "Joint unsupervised coreference resolution with markov logic", *In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics,* pp 650 659, 2008.

[21] L. Heeyoung, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules" *Computational Linguistics*, Vol. 39(4), pp.885–916, 2013.

[22] L. Zhengzhong, J. Araki, E. Hovy, and T. Mitamura, "Supervised within-document event coreference using information propagation", *In Proceedings of the Ninth Language Resources and Evaluation Conference*, pp. 4539– 4544, 2014.

[23] L. Jing and V. Ng, "Joint learning for event coreference resolution", *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol.1, pp. 90–101, 2017.

[24] L. Jing and V. Ng, "Learning antecedent structures for event coreference resolution", *In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications*, pp. 113–118, 2017.

[25] L. Jing and V. Ng, "UTD's event nugget detection and coreference system at KBP 2017", In Proceedings of the Text Analysis Conference, 2017.

[26] L. Xiaoqiang, "On coreference resolution performance metrics" *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 25–32, 2005.

[7] M. Bijankhan, J.Sheykhzadegan, M. Bahrani, and M.Ghayoomi, "Lessons from Building a Persian Written Corpus: Peykare", *Language Resources and Evaluation*, Vol. 45(2), pp.143–164, 2011.

[8] Ch.Prafulla, Ch. Kumar and R. Huang, "Event coreference resolution by iteratively unfolding inter-dependencies among events", *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2124–2133, 2017.

[9] C. Kevin and Ch. D. Manning, "Deep Reinforcement Learning for Mention-Ranking Coreference Models," *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2256-2262, 2016.

[10] C. Kevin and Ch. D. Manning, "Improving Coreference Resolution by Learning Entity-Level Distributed Representations," *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* Vol.1, pp. 643-653. 2016.

[11] C. Nicolae, G. Nicolae, "Bestcut: A graph algorithm for coreference resolution," conference on empirical methods in natural language processing, 2014.

[12] D. Pascal and B. Jason, "Specialized models and ranking for coreference resolution," *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 660–669, 2008.

[13] D. Chase, L. Chan, H. Peng, H. Wu, Sh. Upadhyay, N. Gupta, C. Tsai, M. Sammons, and D. Roth, "UI CCG TAC-KBP2017 submissions: Entity discovery and linking, and event nugget detection and co-reference," In Proceedings of the Text Analysis Conference, 2017.

[14] A. Haghighi, and D. Klein, "Simple coreference resolution with rich syntactic and semantic features," *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol.3, pp. 1152-1161, Association for Computational Linguistics, 2009.

[15] Lee. Heeyoung, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules". Computational Linguistics, 2013.

[16] J.Heng and R. Grishman, "Knowledge base population: Successful approaches and challenges", *In Proceedings of the 49th Annual Meeting of the Association for ComProceedings of the Twenty-Seventh International Joint*

[27] A. McCallum, B.Wellner, "Conditional models of identity uncertainty with application to noun coreference", *In: Advances in neural information processing systems*, pp. 905–912, 2005.

[28] V. Ng, "Supervised noun phrase coreference research", *The first fifteen years. In: Proceedings of the 48th annual meeting of the association for computational linguistics, Association for Computational Linguistics*, 2010, pp. 1396-1411.

[29] M. Rasooli, M. Kouhestani, and A. Moloodi, "Development of a Persian Syntactic Dependency Treebank", *In The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, Atlanta, USA, 2013, pp. 306-314.

[30] M. Seraji, B.Megyesi, J.Nivre, "Bootstrapping a Persian Dependency Treebank", Published as a Journal in Special Issue of the Linguistic Issues in Language Technology (LiLT), Heidelberg, Germany, 2012.

[31] S. W. Meng, H. T. Ng, D. Chung, Y. Lim, "A machine learning approach to coreference resolution of noun phrases", *Computational Linguistics*, Vol.27(4), pp. 521–544, 2001.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado and J. Dean, "Distributed representtations of words and phrases and their compositionality", In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pp. 3111:3119. Curran Associates, Inc, 2013.

[33] U. Olga, M. Poesio, C. Giuliano and K. Tymoshenko, "Disambiguation and Filtering Methods in Using Web Knowledge for Coreference Resolution," FLAIRS Conference, 2013.

[34] Ng. Vincent, "Shallow Semantics for Coreference Resolution," 43rd Annual Meeting on Association for Computational Linguistics, 2017.

[35] V. Marc, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, "A model-theoretic coreference scoring scheme,"Proceedings of the 6th conference on Message understanding. Association for Computational Linguistics, 1995.

[36] S. Wiseman, A. M. Rush, S. M. Shieber, and Jason Weston, "Learning anaphoricity and antecedent ranking features for coreference resolution", *ACL-IJCNLP*, pp. 1416–1426, 2015, Beijing, China.

[37] Y. Xiaofeng, G. Zhou, J. Su, and Ch. L. Tan, "Coreference resolution using competition learning approach," 41st Annual Meeting on Association for Computational Linguistics, Volume 1, 2013.

[38] Y. Xiaofeng, J. Su, G. Zhou, and Ch. Lim Tan, "An NP-cluster based approach to coreference resolution," Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, 2014.

[39] X. Luo, "On coreference resolution performance metrics," Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, 2005.

[40] Y. Bishan, C. Cardie, and P. Frazier, "A hierarchical distance-dependent Bayesian model for event coreference resolution," *Transactions of the Association for Computational Linguistics*, Vol.3, pp.517–528, 2015.

[41] Y. Dian, X. Pan, B. Zhang, L. Huang, D. Lu, S. Whitehead, and H. Ji, "RPI BLENDER TAC-KBP2016 system description", In Proceedings of the Text Analysis Conference, 2016.

حسین سهلانی مدرک کارشناسی ارشد خود را در رشته مهندسی کامپیوتر از دانشگاه علم و صنعت در سال ۱۳۹۱ و از سال ۱۳۹۳ در دانشگاه مالک اشتر در مقطع دکترا مشغول به تحصیل است. ایشان در حال حاضر عضو هیأت‌علمی دانشگاه علوم انتظامی امین است. زمینه‌های پژوهشی موردعلاقه ایشان عبارت‌اند از: پردازش تصویر، پردازش زبان طبیعی، تحلیل اطلاعات در شبکه‌های اجتماعی، بازشناسی الگو و شبکه‌های عصبی.
نشانی رایانامهٔ ایشان عبارت است از:
sahlani@mut.ac.ir
sahlani_h@yahoo.com

مریم حورعلی مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک از دانشگاه علم و صنعت ایران در سال ۱۳۸۵ و مدرک دکترای خود را در گرایش مهندسی فناوری اطلاعات از دانشگاه تربیت مدرس اخذ کرده است. ایشان در حال حاضر عضو هیئت‌علمی گروه‌های هوش مصنوعی و فناوری اطلاعات دانشگاه

صنعتی مالک اشتر تهران است. زمینه‌های پژوهشی موردعلاقه ایشان عبارت‌اند از: پردازش متن و زبان طبیعی، تحلیل اطلاعات در شبکه‌های اجتماعی و سامانه‌های فازی.
نشانی رایانامهٔ ایشان عبارت است از:
Mhourali@mut.ac.ir

بهروز مینایی بیدگلی دانش‌آموخته دانشگاه ایالتی میشیگان آمریکا در رشته علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. ایشان در حال حاضر عضو هیأت‌علمی و دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت است. وی سرپرستی گروه پژوهشی فناوری‌های بازی‌های رایانه‌ای و نیز آزمایشگاه داده‌کاوی را به عهده دارد. محاسبات نرم، یادگیری ماشین، بازی‌های رایانه‌ای، داده‌کاوی، متن‌کاوی و پردازش زبان طبیعی، زمینه‌های پژوهشی موردعلاقه ایشان است.
نشانی رایانامهٔ ایشان عبارت است از:
B_minaei@iust.ac.ir