روشی جدید جهت استخراج موجودیت‌های اسمی در عربی کلاسیک

سید محمد باقر سجادی[1*]، حسن رشیدی[2] و بهروز مینایی بیدگلی[3]
[1]دانشکده کامپیوتر، دانشگاه آزاد تهران مرکز، تهران، ایران
[2]دانشکده ریاضی و علوم کامپیوتر، دانشگاه علامه طباطبایی، تهران، ایران
[3]دانشکده کامپیوتر، دانشگاه علم و صنعت، تهران، ایران

چکیده

تشخیص واحدهای اسمی به‌عنوان یکی از سامانه‌های پردازش زبان طبیعی عبارت از تشخیص اسامی خاص و طبقه‌بندی آن‌ها به یکی از گروه‌های شخص، مکان، سازمان و زمان است. این عملیات به دلیل تأثیر قابل توجه در بهبود کارایی دیگر حوزه‌های پردازش زبان طبیعی مانند ترجمهٔ ماشین، بازیابی اطلاعات، خوشه‌بندی نتایج جستجو و پرسش و پاسخ، در سال‌های اخیر مورد توجه پژوهش‌گران در زبان عربی نیز قرار گرفته است. گرچه بیشتر پژوهش‌ها در این حوزه روی عربی استاندارد امروزی انجام شده است، اما در این مطالعه عربی کلاسیک مورد توجه است. در همین راستا، روشی جدید جهت تشخیص واحدهای اسمی در زبان عربی ارائه می‌شود. در این پژوهش یک پیکره متنی عربی کلاسیک به نام نورکورپ، متشکل از ۱۳۰ هزار کلمه برچسب‌گذاری‌شده توسط متخصصان، معرفی می‌شود؛ همچنین از یک فرهنگ لغات شامل ۱۸۰۰۰ اسامی اشخاص که از کتب حدیثی استخراج شده است، به‌عنوان منابع خارجی استفاده می‌شود. مدل پیش‌بینی، بر اساس مجمع رده‌بندها و یک روش دومرحله‌ای پیشنهاد شده است؛ به‌طوری‌که در مرحله نخست تشخیص واحدهای اسمی از طریق الگوریتم آدابوست M1 و در مرحله دوم طبقه‌بندی آن‌ها به گروه‌های ازپیش‌تعیین‌شده توسط الگوریتم آدابوست M2 انجام می‌شود. به‌منظور غلبه بر چالش‌های زبان عربی عملیات نشانه‌گذاری، برچسب‌گذاری ادات سخن و قطعه‌کردن عبارت پایه به کار گرفته‌شده است. با استفاده از یک روش آماری، برخی از کلمات پر کاربرد در واحدهای اسمی به‌عنوان کلمات کلیدی استخراج شدند. نتیجهٔ به‌دست‌آمده از مدل پیشنهادی در ارزیابی F-measure معادل ۸۶/۸۵ درصد است که بیان‌گر عملکرد مطلوب مدل است. در آخر، روش پیشنهادی روی یک پیکره استاندارد امروزی به نام انرکورپ اعمال و نتایج با پیکره نورکورپ مقایسه شده‌اند.

واژگان کلیدی: تشخیص واحدهای اسمی، مجمع رده‌بندها، روش بوستینگ، زبان عربی کلاسیک

A New Approach for Extracting Named Entity in Classical Arabic

Seyed Mohamad Bagher Sajadi[1*], Hassan Rashidi[2] & Behrouz Minaei[3]
[1]Computer Engineering, Islamic Azad University Central Tehran Branch, Tehran, Iran
[2]Mathematics and Computer Science, University of Allameh Tabataba'i, Tehran, Iran
[3]Computer Engineering, University of Science and Technology, Tehran, Iran

Abstract

In Natural Language Processing (NLP) studies, developing resources and tools makes a contribution to extension and effectiveness of researches in each language. In recent years, Arabic Named Entity Recognition (ANER) has been considered by NLP researchers due to a significant impact on improving other NLP tasks such as Machine translation, Information retrieval, question answering, query result clustering, etc. While most of these researches are based on Modern Standard Arabic (MSA), in this paper, we focus on Classical Arabic (CA) literature. We propose a corpus called NoorCorp with 130k labeled words for research purposes

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

which is annotated by expert human resources manually. This corpus is based on a Historic-Islamic book of 1200 years ago including 1843 sentences and 127550 words. We also collected about 18k proper names from old Hadith books as a gazetteer which is called NoorGazet used as a future. In this paper, we propose a new approach to extract named entities (NEs) including person, location, organization and time. We use hybrid approach benefiting from advantages of Rule based approach and Machine learning approach. We divided the NoorCorp into two parts of training and test sets containing 80% and 20% of the data set respectively. Prediction model, based on Boosting method, was developed in two steps which Adaboost.M1 is employed to identify NEs and Adaboost.M2 is employed to classify NEs. There are many methods using multiple classifiers as voters and summing up their results, among which, ensemble methods are those which generate multiple hypotheses using the same base learner. We developed an ensemble consisting of 50 members (classifiers) based on decision stump to implement the weak learner. Since only 17% of the text data is composed of name entity labels, we had to deepen the tree while restricting pruning. We exploited tokenizing, part of speech (POS) tagging, and base phrase chunking (BPC) to overcome linguistic obstacles in Arabic including Meaning ambiguity, Optional diacritics, Complex morphology and Nonstandard written text. Moreover, using a statistical technique, the most frequently used words extracted as key words. Results show that performance of the method is better than decision tree as the base classifier. An overall F-measure value of 86.85 obtained which is better than base line about 20% and CART decision tree about 12%. Since CA corpus consists of simpler linguistic patterns compared to MSA, we applied the proposed approach on ANERCorp as Modern Standard Arabic corpus. Results show that the proposed model outcome on CA corpus is about 19% better than MSA. This result is due to the fact that there are plenty of NEs entered to MSA from other languages. These proper names do not have specific patterns and do not exist in the gazetteer. In addition, many NE's are not distributed uniformly in ANERcorp which considerably reduces the results accuracy.

Keywords: Named entity recognition (NER), Ensemble learning, Boosting method, Classical Arabic Language.

۱- مقدمه

تشخیص واحدهای اسمی[1] یا موجودیت‌های اسمی به‌عنوان یکی از زیرعملیات پردازش زبان طبیعی[2] عبارت از استخراج اسامی خاص و طبقه‌بندی آن‌ها به گروه‌های مختلفی از قبیل شخص، مکان، سازمان، زمان و... است [1]. کارآیی این عملیات روی دیگر عملیات پردازش زبان طبیعی مانند ترجمه ماشین، بازیابی اطلاعات، پرسش و پاسخ و خوشه‌بندی متن تأثیر به‌سزایی دارد [2]. ابزاری جهت استخراج اسامی خاص در میان متن از مهم‌ترین منابع مورد نیاز در هر زبان است. مطالعه‌ای که روی روزنامه‌های انگلیسی و فرانسوی انجام شده است، نشان می‌دهد ۱۰ درصد متون را موجودیت‌های اسمی تشکیل می‌دهند [3].

به‌عنوان مثال با توجه به جمله زیر:

"John left Washington D.C. at 4 a.m. and reached New York City at 7h30 a.m."

«جان، واشنگتن را در ساعت ۴ صبح ترک کرد و در ساعت ۷:۳۰ صبح به شهر نیویورک رسید.»

یک سامانه تشخیص واحد اسمی باید قادر باشد جان، واشنگتن و شهر نیویورک را به‌عنوان واحد اسمی شناسایی کند و سپس جان را در گروه شخص و دو موجودیت اسمی دیگر را در گروه مکان طبقه‌بندی کند؛ به‌طور کلی سه رویکرد برای مواجهه با این عملیات وجود دارد [4]:

- رویکرد بر اساس قوانین: این رویکرد وابسته به زبان است و از قوانین زبان‌شناسی که به‌صورت دستی استخراج شده استفاده می‌کند.
- رویکرد بر اساس یادگیری ماشین: این رویکرد مستقل از زبان است و روش‌های یادگیری باناظر را به کار می‌گیرد.
- رویکرد ترکیبی: الگوهای قوانین زبان‌شناسی و روش‌های یادگیری باناظر را با یکدیگر ترکیب می‌کند.

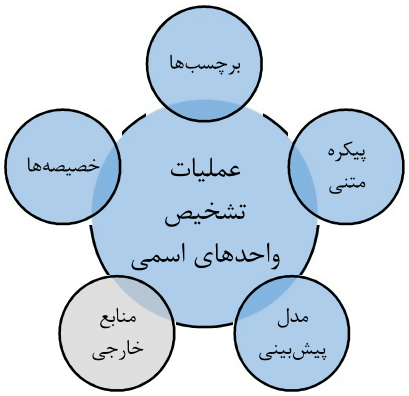
مهم‌ترین عواملی که در یک مسئله تشخیص واحد اسمی باید مورد مطالعه قرار گیرد، عبارتند از: استفاده از پیکره متنی مناسب، تعیین برچسب‌ها و رده‌ها[3]، انتخاب خصیصه‌ها و به‌کارگیری روش‌های مختلف رده‌بندی[4]. استفاده از منابع خارجی مانند فرهنگ لغات نیز می‌تواند برای بهبود نتایج و یا دیگر اهداف پژوهشاتی به کار گرفته شود. در شکل (۱) مهم‌ترین عناصر تشکیل‌دهنده عملیات تشخیص واحدهای اسمی در رویکرد یادگیری ماشین، آورده شده است.

[1] Named entity recognition

[2] Natural language processing (NLP)

[3] Class

[4] Classification



(شکل-۱): عناصر عملیات تشخیص واحدهای اسمی در رویکرد مبتنی بر یادگیری ماشین

(Figure-1): The Most Important Elements in the NER Task in Machine learning approach

هر پژوهش موفق در حوزه پردازش زبان طبیعی نیاز به منابع و ابزارهای زبانی دارد که در این میان وجود یک پیکره متنی غنی ضروری است. در حال حاضر چهار پیکره متنی عربی جهت استفاده در عملیات تشخیص واحد اسمی در دسترس است [5]:

- ACE 2003
- ACE 2004
- ACE 2005
- ANERCorp

پیکره‌های متنی بالا بر اساس عربی استاندارد امروزی تهیه شده‌اند و در بیشتر پژوهش‌ها از این پیکره‌ها استفاده شده است. متون عربی به سه دسته تقسیم می‌شوند: عربی کلاسیک[1]، عربی استاندارد امروزی[2] و عربی محاوره‌ای[3] [6]. عربی کلاسیک، ویرایش رسمی ادبیات اسلامی است که در حدود ۱۵۰۰ سال قبل مورد استفاده قرار گرفته است و شامل بیش‌تر متون مذهبی عربی می‌شود [7]. عربی استاندارد امروزی مربوط به زبان رسمی حاضر است که در روزنامه‌ها، مجلات، آموزش و غیره به کار می‌رود. عربی محاوره‌ای، مکالمه غیر رسمی است که روزانه میان عرب‌زبانان به‌کار گرفته می‌شود [8].

در این پژوهش تمرکز اصلی روی پیکرهٔ متنی عربی کلاسیک جهت آموزش و آزمون است. ما از رویکرد یادگیری ماشین به‌منظور توسعه مدل تشخیص و از رویکرد بر اساس قوانین به‌منظور انتخاب خصیصه‌هایی مانند برچسب ادات سخن[4]، قطعه عبارت پایه[5] استفاده می‌کنیم.

در ادامه، در بخش دوم مشخصات زبان عربی که در تشخیص واحدهای اسمی حائز اهمیت است معرفی می‌شود. در بخش سوم کارهای انجام‌شده در زبان عربی مورد بررسی قرار و در بخش چهارم روش پیشنهادی شامل منابع زبانی، پیش‌پردازش، انتخاب خصیصه، مدل پیشنهادی و پس‌پردازش به تفصیل بحث می‌شود. بخش پنجم به آزمایش‌های انجام‌شده و ارزیابی آن‌ها اختصاص دارد.

۲- مشخصات زبان عربی در عملیات تشخیص واحدهای اسمی

عملیات تشخیص واحدهای اسمی در زبان‌هایی مانند زبان عربی که دارای ریخت‌شناسی[6] غنی هست، نسبت به دیگر زبان‌ها پیچیدگی بیشتری دارند [9]؛ به همین جهت قبل از پیاده‌سازی هر مسئله زبان طبیعی، ویژگی‌های زبان مربوطه می‌بایست مورد مطالعه قرار گیرد. مشخصات زبان عربی که به‌عنوان موانعی برای استخراج اسامی خاص شناخته می‌شوند، عبارت‌اند از:

- ابهام معنایی: برخی از اسامی خاص در زبان عربی وجود دارند که می‌توانند به‌صورت کلمات معمولی نیز به‌کار گرفته شوند. برای مثال کلمات جمیله، فرید، کریم و غیره هم به‌عنوان اسامی شخص و هم به‌عنوان صفت مورد استفاده قرار می‌گیرند.
- عدم وجود قوانین املایی: زبان عربی برخلاف بسیاری از زبان‌های دیگر، دارای حروف بزرگ نوشتاری[7] نیست [5]. به عبارت دیگر این زبان به حروف بزرگ و کوچک در متن حساس نیست. این مشخصه مانع بزرگی در عملیات تشخیص واحدهای اسمی در زبان عربی است درحالی‌که این ویژگی در دیگر زبان‌ها یکی از خصیصه‌های مهم محسوب می‌شود. به‌طور کلی در زبان عربی هیچ قانون املایی برای تشخیص واحدهای اسمی وجود ندارد.

[1] Classical Arabic (CA)

[2] Modern Standard Arabic (MSA)

[3] Colloquial Arabic Dialects

[4] Part Of Speech (POS)

[5] Base Phrase Chunk (BPC)

[6] Morphology

[7] Capital Letters

- مشکل ابهام‌زدایی کلمات: زبان عربی برای ابهام‌زدایی کلمات از اعراب استفاده می‌کند. چهار صدای کوتاه به‌عنوان اعراب در این زبان وجود دارد: فتحه، ضمه، کسره و سکون. در زبان عربی تعدادی از کلمات می‌توانند به‌لحاظ حروف به‌طور کامل شبیه به هم باشند؛ درحالی‌که تلفظ و معنی آن‌ها متفاوت است. در چنین حالتی برای ابهام‌زدایی لغوی از اعراب استفاده می‌شود. بنابراین معنی کلمه وابسته به اعرابی است که در آن استفاده می‌شود. برای مثال " دَخَلَ " به معنی «او وارد شد است» درحالی‌که "دَخْلْ " به معنی «درآمد» است. در عربی استاندارد امروزی، اعراب حذف شده و فقط گاهی اوقات اعراب استفاده‌شده در آخرین بخش کلمه نگه داشته می‌شود. این موضوع افزایش ابهامات لغوی را در پی دارد و خواننده در این حالت می‌بایست از طریق اطلاعات متنی (یعنی از طریق موضوع، معنا و گرامر جمله) ابهام‌زدایی لغوی کلمات را انجام دهد [10]. در جـــدول(۱) مثالی از پیچیدگی معنایی که به‌دلیل حذف اعراب ایجاد می‌شود آورده شده است. برخی از این ابهام‌ها در زبان فارسی نیز صادق است [11].

(جدول-۱): ابهام معنایی در زبان عربی بدون اعراب
(Table-1): Lexical Ambiguity in Arabic without diacritics

کلمه	نویسه	اعراب	نقش
حسن	Hsn	حَسَن	نام
حسن	Hsn	حُسْن	صفت
حسن	Hsn	حَسُنَ	فعل

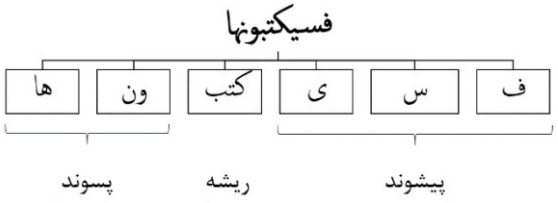
- مشکل جداافتادگی داده‌هـا[1]: زبان عربی دارای چسبندگی بالایی[2] است؛ زیرا هر کلمه از ترکیب وندها تشکیل می‌شود [12]. کلیه اسامی، صفت‌ها، قیدها و فعل‌های عربی از هزار ریشه سه و چهار حرفی و به‌ندرت پنج حرفی به‌دست می‌آید [13]. شکل عمومی یک کلمه در عربی بدین‌صورت است:

پیشوندها + ریشه + پسوندها

پیشوندها حروف تعریف، اضافه و یا ربط و پسوندها مفعول یا مراجع ملکی می‌توانند باشند؛ که تعداد این وندها ممکن است، هیچ یا بیشتر باشد. برای به‌دست‌آوردن عبارت مورد نظر پسوندها به ریشه متصل می‌شوند. برای مثال: کلمه "فسیکتبونها" به معنی "then they will write it" است. این مثال نشان می‌دهد که چگونه یک کلمه در زبان عربی به چند کلمه در زبان انگلیسی ترجمه می‌شود. این کلمه بر اساس شکل عمومی توضیح داده شده این‌گونه نمایش داده می‌شود:

ف + س + ی + کتب + ون + ها

کلمه کتب، ریشه و حروف قبل از آن پیشوند و حروف بعد از آن پسوند هستند. شـــــکل(۲) استراتژی الحاق جهت تشکیل کلمات را در زبان عربی نشان می‌دهد. از نگاه تشخیص واحدهای اسمی این خصوصیت زبان عربی مانع بزرگی محسوب می‌شود؛ زیرا باعث جداافتادگی داده‌ها می‌شود. بنابراین برای آموزش خوب نیاز به یک پیکره متنی خیلی بزرگ است.



(شکل-۲): مثالی از تشکیل یک کلمه در زبان عربی
(Figure-2): A Simple Example of the Composition of an Arabic Word

۳- کارهای مرتبط

اگرچه در سال‌های اخیر زبان عربی به‌طور فزاینده‌ای در بین پژوهش‌گران پردازش زبان طبیعی مورد توجه قرار گرفته است؛ اما به‌طورکلی کارهای انجام‌شده روی تشخیص واحدهای اسمی در زبان عربی محدود هستند. پژوهش‌ها نشان می‌دهد تا سال ۲۰۰۵ هیچ مطالعه‌ای بر اساس رویکرد یادگیری ماشین در این حوزه انجام نشده است و تنها تعداد کمی از پژوهش‌ها به قوانین دستی و قواعد ریخت‌شناسی اختصاص دارند. پژوهش‌ها [14] جزو نخستین کارهایی است که رویکرد یادگیری ماشین در آن ظاهر می‌شود. تأکید اصلی نویسندگان بر موضوع پسوند و پیشوند کلمات است و راه‌کارهایی برای قطعه‌سازی و جداکردن ریشه از وندهایش ارائه می‌شود. آن‌ها رده‌بـــند[3] آنتروپی بیشینه را با مجموعه داده ACE 2004 آموزش داده و نشان داده‌اند، استفاده از

[1] Data sparseness

[2] Highly Inflectional

[3] Classifier

خصیصه‌های مربوط به وندها و قطعه‌سازی آن‌ها باعث دو درصد بهبود در نتایج ACE می‌شود.

بن‌عجیبا و همکاران با استفاده از مدل آنتروپی بیشینه به نتیجه‌ای معادل ۵۵/۲۳ درصد در معیار F-measure دست یافتند [15]. به‌دلیل عدم وجود پیکره متنی استاندارد و قابل دسترس، آن‌ها پیکره متنی خود را برگرفته از ۳۱۶ مقاله از میان روزنامه‌های مختلف با موضوعات متنوع و براساس استاندارد اجلاس یادگیری زبان طبیعی[1]، با عنوان انرکورپ[2] به‌منظور اهداف پژوهشی توسعه دادند. همچنین فرهنگ‌نامه انرگزت[3] را متشکل از سه فرهنگ اشخاص، مکان‌ها و سازمان‌ها به‌عنوان یک منبع خارجی جهت بهبود نتایج تولید کردند؛ سپس در جهت توسعه مطالعه قبل، از یک رویافت دومرحله‌ای شامل کشف مرزهای واحدهای اسمی و رده‌بندی آن‌ها از طریق روش آنتروپی بیشینه استفاده کرده و به نتیجه‌ای معادل ۶۵/۹۱ درصد دست یافتند [16]. در کارهای بعدی [17]،[18]، روی یافتن خصیصه‌های بهینه در تشخیص واحدهای اسمی تمرکز شده است. آن‌ها ۲۲ خصیصه را روی ۹ پیکره متنی مختلف و با استفاده از رده‌بندهای آنتروپی بیشینه، ماشین بردار پشتیبان و میدان‌های تصادفی شرطی به کار گرفته و هر خصیصه را بر اساس میزان تأثیرشان رتبه‌بندی کردند. بهترین نتیجه معادل ۸۳/۳۴ درصد با استفاده از پانزده خصیصه بهینه، روی داده روزنامه‌ای ACE 2003 به‌دست آمده است [19].

نویسندگان در [4]، با استفاده از روش ماشین بردار پشتیبان و همچنین مجموعه‌ای از خصیصه‌های مستقل و وابسته به زبان به نتیجه‌ای معادل با ۸۳/۲۰ درصد دست یافتند. [20]، یک روش ساده را جهت ترکیب دو رویکرد یادگیری ماشین و بر اساس قوانین دستی به کار گرفتند که نتیجه ۸۷/۷۷ در معیار F-measure را به دنبال داشت. آن‌ها جهت پیاده‌سازی مدل، از درخت تصمیم J48[4] و جهت آموزش و ارزیابی مدل، از پیکره متنی ACE2003 و انرکورپ استفاده کردند. [21] یک روش ترکیبی را به‌منظور بهره‌مندی از مزایای رویکرد یادگیری ماشین و قوانین دستی استفاده کرده است. نتایج به‌دست‌آمده در معیار Fmeasure در سه رده شخص، مکان و سازمان، به‌ترتیب برابر است با ۹۴، ۹۰ و ۸۸ درصد. [22] پژوهش قبلی را با یک روشی شامل سه مؤلفه زیر توسعه می‌دهند:

- مؤلفه یادگیری ماشین: از ماشین بردار پشتیبان استفاده می‌کند و به مجموعه‌ای از خصیصه‌های استخراج‌شده از پیکره متنی وابسته است.
- مؤلفه بر اساس قوانین: وابسته به مجموعه‌ای از قوانین دستوری است.
- مؤلفه انتخاب/اصلاح برچسب: شناسایی خطاهای رخ داده و اصلاح آن‌ها

خروجی مؤلفه بر اساس قوانین به‌عنوان خصیصه وارد مؤلفه یادگیری ماشین می‌شود. مؤلفه انتخاب/اصلاح نتایج دو مؤلفه قبل را با یکدیگر مقایسه و برخی از خطاهای تشخیص را شناسایی و اصلاح می‌کند. نتایج به‌دست‌آمده در این پژوهش بالاترین درصد را در معیار FMeasure در میان همه پژوهش‌های مربوطه به خود اختصاص داده است [23].

پژوهش‌هایی که با استفاده از مجمع رده‌بندها در عملیات تشخیص موجودیت‌های اسمی انجام شده بسیار محدود است. در [24]، همه مطالعاتی که به‌صورت موفق با استفاده از روش مجمع[5] درحوزه پردازش زبان طبیعی انجام شده جمع‌آوری شده است. در این میان سه پژوهش مربوط به عملیات پردازش زبان طبیعی است که در هیچ‌کدام روش Boosting به کار گرفته نشده است. [25] با استفاده از آدابوست دودویی، دو ماژول را جهت تشخیص و رده‌بندی موجودیت‌های اسمی در دو زبان آلمانی و انگلیسی توسعه داده‌اند. الگوریتم بوستینگ به‌وسیله چند درخت تصمیم کوچک با عمق ثابت پیاده‌سازی شده است. در کار بعدی، آن‌ها از Adaboost.MH به‌عنوان آدابوست چند رده‌ای استفاده کردند [26]. در این دو پژوهش هیچ توضیحی راجع به پیاده‌سازی الگوریتم و همچنین تحلیل عملکرد آن ارائه نشده است. [27] ترکیبی از آدابوست M1 و درخت تصمیم C4.5 را روی مجموعه بزرگی از خصیصه‌ها در زبان انگلیسی و مجارستانی به کار گرفتند و تعداد تکرار را برابر سی قرار دادند. تأکید آن‌ها روی الگوریتم بوستینگ نیست بلکه تمرکز روی رأی‌گیری میان چند رده‌بند است. بنابراین، پژوهش‌ها نشان می‌دهد، تاکنون، هیچ مطالعه‌ای جهت تشخیص واحدهای اسمی در زبان عربی با استفاده از مجمع رده‌بندها انجام نگرفته است.

[1] Conference on Natural Language Learning (CoNLL)

[2] ANERCorp

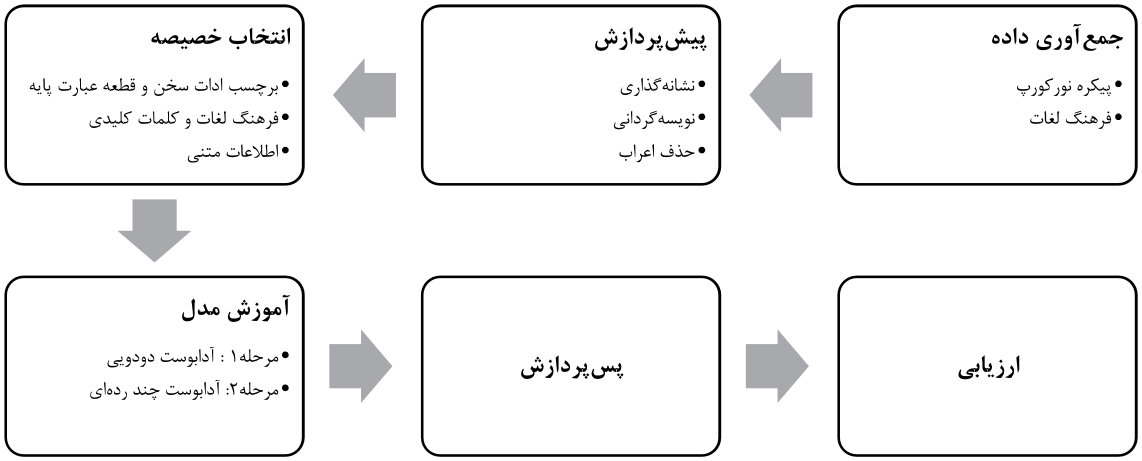
[3] ANERGazet

[4] یک پیاده‌سازی از الگوریتم C4.5 برای درخت تصمیم

[5] Ensemble method

۴- روش پیشنهادی

در این بخش روشی جدید جهت شناسایی و استخراج موجودیت‌های اسمی در عربی کلاسیک ارائه می‌شود. شکل (۳) فرآیند تشخیص واحدهای اسمی در روش پیشنهادیفرآیند تشخیص واحدهای اسمی را در روش پیشنهادی به‌طور اجمال نشان می‌دهد. در ادامه هر یک از این بخش‌ها بررسی خواهد شد.



(شکل-۳): فرآیند تشخیص واحدهای اسمی در روش پیشنهادی
(Figure-3): NER process steps in our approach

۴-۱- منابع زبانی عربی

۴-۱-۱- پیکره متنی

عملیات تشخیص واحدهای اسمی، نیازمند یک پیکره متنی غنی جهت آموزش و ارزیابی مدل است. در این پژوهش، ما یک پیکره متنی عربی کلاسیک را که ویرایش‌شدهٔ پیکره نورکورپ[1] [28] است، ارائه می‌کنیم. این پیکره براساس کتاب «وقعه صفین» نوشته «محمد بن محمد مفید» معروف به شیخ مفید نگاشته‌شده در سال ۲۳۰ هجری قمری به‌عنوان یک کتاب تاریخی تهیه شده و شامل ۱۸۴۳ جمله و ۱۲۷۵۵۰ کلمه است. مشخصات مربوط به نرخ حضور اسامی خاص در این پیکره متنی بر اساس چهار طبقه شخص، مکان و سازمان و زمان در جدول (۲) آمده است.

(جدول-۲): نرخ حضور واحدهای اسمی در پیکره نورکورپ
(Table-2): The frequency rate of different name entities in Noorcorp

طبقه	تعداد	نسبت به پیکره متنی	نسبت به واحدهای اسمی
شخص	۱۹۲۸۵	٪۱۵	٪۸۷
مکان	۱۰۱۷	۰/۸٪	٪۵
سازمان	۹۹۱	۰/۸٪	٪۵
زمان	۶۷۹	۰/۵٪	٪۳
واحد اسمی	۲۱۹۷۲	۱۷/۲٪	۱۰۰٪

جدول (۳) نرخ حضور اسامی خاص را در پیکره متنی انرکورپ نشان می‌دهد. انرکورپ یک پیکره عربی استاندارد امروزی است که بر اساس روزنامه‌ها و ویکی‌پدیای عربی جمع‌آوری شده و دارای ۵۰۹۶ جمله و ۱۵۰۲۸۶ کلمه است.

(جدول-۳): نرخ حضور واحدهای اسامی در پیکره انرکورپ
(Table-3): The frequency rate of different name entities in ANERCorp

طبقه	تعداد	نسبت به پیکره متنی	نسبت به واحدهای اسمی
شخص	۶۴۳۶	۴/۳٪	٪۴۳
مکان	۵۰۳۴	۳/۳۵٪	٪۳۴
سازمان	۳۴۰۳	۲/۲۶٪	٪۲۳
واحد اسمی	۱۴۸۷۱	٪۱۰	۱۰۰٪

از مقایسه جدول(۲) و جدول (۳) نکات جالبی به‌دست می‌آید. نرخ حضور اسامی اشخاص در پیکره کلاسیک حدود چهار برابر، بیشتر از پیکره امروزی است، درحالی‌که نرخ حضور اسامی مکان و سازمان در پیکره امروزی بیش از چهار برابر پیکره کلاسیک است. همچنین میانگین طول جمله در پیکره کلاسیک ۱۳۲ کلمه است؛ درحالی‌که در پیکره امروزی سی کلمه است که نمایان‌گر طولانی‌تر بودن جملات در عربی کلاسیک است.

در این پژوهش از استاندارد برچسب‌گذاری IOB استفاده شده است. بر اساس این استاندارد قسمت نخست هر

[1] NoorCorp

واحد اسمی با حرف B برچسب‌گذاری می‌شود و دیگر بخش‌های واحد اسمی با حرف I. دیگر کلماتی که شامل موجودیت‌های اسمی نیستند با حرف O مشخص می‌شوند. جدول (۴) نرخ حضور برچسب‌ها را بر اساس بخش نخست موجودیت اسمی و بخش‌های بعدی آن نشان می‌دهد. با توجه به جدول(۴)، برچسب‌های B در اسامی مکان و سازمان بیش از برچسب‌های I است درحالی‌که در موجودیت‌های اسمی اشخاص، برچسب I بیشتر است. این موضوع بدان معناست که بیشتر اسامی اشخاص و همچنین اسامی زمان حاوی بیش از دو کلمه هستند. در جدول(۵) نرخ برچسب‌گذاری IOB در پیکره انرکورپ در مقایسه با پیکره نورکورپ آمده است.

(جدول-۴): نرخ برچسب‌گذاری IOB در پیکره نورکورپ

(Table-4): The frequency rate of IOB annotation in NOORCorp

برچسب	تعداد	درصد
B-PERS	۷۲۹۱	٪۳۳
I-PERS	۱۱۹۹۴	٪۵۴
B-LOC	۸۷۲	٪۴
I-LOC	۱۴۵	۰/۲٪
B-ORG	۶۲۴	٪۳
I-ORG	۳۶۷	۱/۷٪
B-TIME	۱۷۰	۰/۸٪
I-TIME	۵۰۹	۲/۳٪

(جدول-۵): نرخ برچسب‌گذاری IOB در پیکره انرکورپ

(Table-5): The frequency rate of IOB annotation in ANERCorp

برچسب	تعداد	درصد
B-PERS	۳۶۰۰	٪۲۴
I-PERS	۲۸۳۲	٪۱۹
B-LOC	۴۴۳۰	٪۳۰
I-LOC	۶۰۴	٪۴
B-ORG	۲۰۲۳	٪۱۴
I-ORG	۱۳۸۰	٪۹

۴-۱-۲- فرهنگ لغات

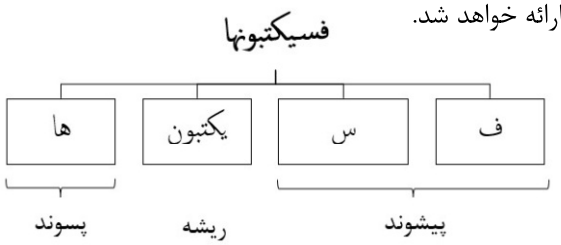
فرهنگ لغات به‌عنوان یک منبع خارجی یکی از منابع مهمی است که در افزایش کارایی عملیات استخراج موجودیت‌های اسمی به کار گرفته می‌شود. ما حدود ۸۸۰۰۰ هزار اسم شخص را از میان کتب مختلف حدیثی استخراج و آن‌ها را به بخش‌های کوچک‌تر تقسیم کرده و سپس نرخ حضور هر بخش را محاسبه کردیم. با افزودن اسامی اشخاص از فرهنگ انرگزت، فرهنگ اسامی اشخاص ما با ۱۸۲۳۸ ردیف تولید گردید.

۴-۲- پیش‌پردازش

به‌منظور آماده‌سازی و افزایش کیفیت پیکره متنی سه عملیات نشانه‌گذاری، نویسه‌گردانی و حذف اعراب انجام شده است.

۴-۲-۱- نشانه‌گذاری

با توجه به این‌که پیچیدگی ریخت‌شناسی عربی موجب جداافتادگی داده‌ها می‌شود عملیات نشانه‌گذاری[1] روی پیکره متنی، به‌منظور کاهش این پیچیدگی به کار گرفته می‌شود. در اینجا منظور از نشانه‌گذاری جداسازی اجزای کلمه تا اندازه‌ای است که هر جزء آن دربرگیرنده معنا باشد؛ یعنی تنها پسوندها و پیشوندهای حاوی معنا از ریشه جدا شده‌اند. بنابراین ما برای کاهش پیچیدگی ریخت‌شناسی زبان عربی، کلمات پیکره متنی را به نشانه‌ها شکسته و نشانه‌ها را برچسب‌گذاری می‌کنیم. در شکل (۴) مثالی از نشانه‌گذاری پیشنهادی آمده است. به‌وسیله این عملیات، ۱۲۷۵۵۰ کلمه در پیکره متنی به ۱۴۶۶۲۴ نشانه تقسیم شدند. در ادامه تأثیر این پیش‌پردازش روی نتیجه نهایی ارائه خواهد شد.



(شکل-۴): مثالی از نشانه‌گذاری در روش پیشنهادی

(Figure-4): An example of tokenizing in our approach

۴-۲-۲- نویسه‌گردانی

منظور از نویسه‌گردانی، نگاشتی از حروف مبدأ (در اینجا عربی) به حروف مقصد (در اینجا انگلیسی) است. این عملیات تأثیری روی نتایج ندارد و تنها به‌منظور سهولت استفاده پیکره متنی در الگوریتم‌ها و ابزارهای زبانی به کار می‌رود. در این پژوهش، تبدیل نویسه‌های UTF8 به لاتین توسط نویسه‌گردانی باک‌والتر[2] انجام شده است.

[1] Tokenizing

[2] Buckwalter

۴-۲-۳- حذف اعراب

گرچه پیکره متنی ما به‌صورت پیش‌فرض دارای اعراب است و عدم وجود اعراب، ابهام‌زدایی را دشوار می‌کند؛ اما با توجه به این‌که تمامی پژوهش‌ها در پردازش زبان عربی، بدون اعراب در نظر گرفته شده است؛ ما نیز جهت مقایسه دقیق‌تر کار خود با دیگران مجبور به حذف آن شده‌ایم.

در جـدول (۶) ساختار پیکره متنی ارائه‌شده مثالی از ساختار پیکره متنی نورکوپ پس از فرآیند پیش‌پردازش آورده شده است.

(جدول-۶): ساختار پیکره متنی ارائه‌شده

(Table-6): NoorCorp structure

نشانه	نویسه‌گردانی	برچسب واحد اسمی
قالوا	VBD_MP3	O
ل	l	O
ما	mA	O
قدم	qdm	O
علی	Ely	B-PERS
بن	bn	I-PERS
أبی	<by	I-PERS
طالب	TAlb	I-PERS
من	mn	O
البصره	AlbSrp	B-LOC
إلی	<lY	O
الکوفة	Alkwfp	B-LOC

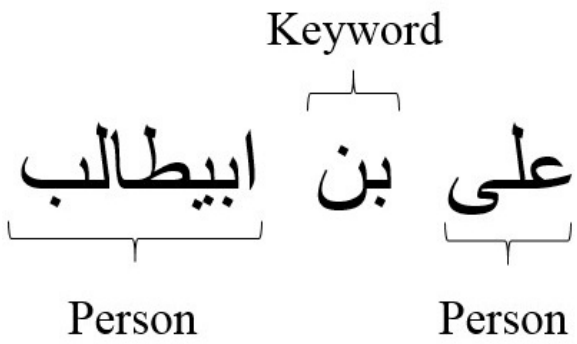
۴-۳- انتخاب خصیصه

از مهم‌ترین و چالش‌برانگیزترین مباحث در تشخیص واحدهای اسمی انتخاب خصیصه‌های کارا جهت تصمیم‌گیری دقیق‌تر است. در برخی از پژوهش‌ها فقط به بررسی انواع خصیصه‌ها و میزان کارایی آن‌ها در یک مسئله تشخیص واحد اسمی پرداخته شده است که این امر اهمیت استخراج و انتخاب خصیصه‌ها را می‌رساند. خصیصه‌هایی که در پیکره متنی ما آمده است عبارت‌اند از:

- برچسب ادات سـخن[1]: فرآیند برچسب‌گذاری ادات سخن عبارت از تعیین نقش نحوی برای هر کلمه در یک جمله است [29]. این عملیات به حل مشکل ابهام معنایی و همچنین ابهام لغوی کمک می‌کند [30].
- برچسب قطعه عبارت پایـه[2]: منظور از قطعه‌کردن عبارت پایه، گروه‌بندی ترتیبی از کلمات همسایه است که تشکیل عبارت‌های نحوی، مانند گروه اسمی و یا گروه فعلی می‌دهند. تشخیص مرز و نوع عبارت نحوی می‌تواند پیش‌پردازش سامانه‌های مهمی در پردازش زبان طبیعی باشد [31].

کارآیی عملیات قطعه‌کردن عبارت پایه و همچنین برچسب‌گذاری ادات سخن تأثیر زیادی روی بهره‌وری سامانه تشخیص واحد اسمی دارد [13]. به‌منظور پیاده‌سازی این دو عملیات از ابزار امیره استفاده شده است [32]. این ابزار بر اساس یادگیری باناظر توسعه داده شده و سه عملیات نویسه‌گردانی، برچسب‌گذاری ادات سخن و قطعه‌کردن عبارت پایه را به‌ترتیب با ۹۹/۲ درصد، ۹۶/۱۵ و ۹۶/۳۳ در معیار Fmeasure فراهم آورده است.

- فرهنگ لغات: ما از فرهنگ لغات گردآوری‌شده به‌عنوان یک خصیصه استفاده کردیم؛ به‌طوری‌که مقادیر ۱،۲،۳،۰ به‌ترتیب به هر یک از گروه‌های شخص، مکان، سازمان و غیره نگاشت می‌شوند. البته قابل ذکر است که فرهنگ لغات ما، در محدوده اسامی اشخاص بسیار غنی اما داده‌های خوبی در دیگر طبقات واحد اسمی ندارد.
- کلمات کلیدی: برخی از کلمات در عربی کلاسیک وجود دارند که به‌طور مشخص قبل یا بعد از موجودیت اسمی ظاهر می‌شوند. ما برخی از این کلمات را با استفاده از یک کار آماری برای گروه‌های مختلف واحد اسمی استخراج کردیم. در شـــکل (۵) یکی از واژه‌های کلیدی پرکاربرد در گروه اسامی اشخاص نشان داده شده است.



(شکل-۵): یک مثال از به‌کارگیری کلمات کلیدی به عنوان خصیصه

(Figure-5): A sample for using keyword as a feature

[1] Part Of Speech (POS) tag

[2] Base Phrase Chunk (BPC) tag

- اطلاعات متنی: برخی از واحدهای اسمی بر اساس کلمات قبلی و یا بعدی خود مشخص می‌شوند. به‌عنوان مثال «کوفه» نام مکان است درحالی‌که «اهل الکوفه» نام سازمان است و یا «تمیم» نام شخص است درحالی‌که «بنی تمیم» نام سازمان است. بنابراین با استفاده از ویژگی توالی داده‌ها در پیکره متنی می‌توان به همسایه‌های کلمات توجه نمود. ما پنجره‌ای به طول $[-n,+n]$ نشانه تعریف کرده و $n=2$ در نظر گرفتیم. به عبارت دیگر از دو نشانه قبل از نشانه جاری تا دو نشانه بعد به همراه برچسب ادات سخن، محدوده‌ای است که خصیصه‌های متنی انتخاب شده‌اند.

۴-۴- مدل پیشنهادی

در میان روش‌های متعدد یادگیری ماشین، ما روش‌های ترکیب رده‌بندها یا مجمع رده‌بندها را به کار گرفتیم. شیوه‌های مختلفی برای به‌کارگیری چند رده‌بند و ترکیب نتایج آن‌ها وجود دارد؛ اما منظور از مجمع رده‌بندها شیوه‌هایی است که چند رده‌بند با استفاده از یادگیرنده پایه یکسان[1] تولید می‌شوند. این روش‌ها به‌دلیل قدرتی که در تصمیم‌گیری و پایین‌آوردن نرخ خطای تشخیص دارند، در پردازش زبان‌های طبیعی نیز مورد استفاده می‌توانند قرار گیرند [24]. درواقع این روش‌ها با استفاده از تعداد زیادی رده‌بند، سعی می‌کنند از زوایای مختلف به مسئله نگاه کنند و نقاط ضعف را با یک نظرسنجی ساده میان یکدیگر پوشش دهند. از میان روش‌های موجود در مجمع رده‌بندها از روش بوستینگ استفاده کرده‌ایم. این روش با تمرکز روی نمونه‌هایی که سخت‌تر رده‌بندی می‌شوند، به‌صورت افزایشی سعی در تقویت رده‌بندها و نتیجه پیش‌بینی دارد. یکی از مهم‌ترین و گسترده‌ترین شکل‌های بوستینگ، الگوریتم آدابوست[2] توسعه‌یافته در سال ۱۹۹۶ است [33]. ایده اصلی پشت این الگوریتم توسعه افزایشی مجمع D به‌منظور بهبود پیش‌بینی، به‌وسیله تمرکز روی بخش‌های سخت‌تر داده است. در مرحله آموزش، مجموعه داده Z به N نمونه تقسیم می‌شود. به هر یک از نمونه‌های مجموعه آموزشی یک وزن w اختصاص می‌یابد که در ابتدا وزن همه نمونه‌ها یکسان است. رده‌بند k در مرحله k به مجمع اضافه شده و با استفاده از زیرمجموعه‌ای از نمونه‌ها و با توجه به تابع توزیع وزن w که در آخرین مرحله یعنی $k-1$ به‌دست‌آمده آموزش می‌بینند. محسابه نرخ خطای رده‌بند که نمایان‌گر دقت رده‌بند است در فرمول (۱) آمده است.

$$\epsilon_k = \sum_{j=1}^{N} w_j^k l_k^j , \qquad (۱)$$
$$l_k^j = 1 \; if \; D_k misclassifies \; Z_j and \; l_k^j = 0 \; otherwise$$

بر اساس دقت رده‌بندی یک ضریب رأی‌گیری به‌عنوان وزن به هر رده‌بنداعطا می‌شود؛ بنابراین پس از ساخت مجمع، رأی‌گیری میان رده‌بندها به‌صورت وزن‌دار خواهد بود. وزن α_k برای رده‌بند k مطابق فرمول (۳) محاسبه می‌شود.

$$\beta_k = \frac{\epsilon_k}{1-\epsilon_k} \qquad \epsilon_k \in (0,0.5) \qquad (۲)$$

$$\alpha_k = \frac{1}{2} ln \frac{1}{\beta_k} \qquad (۳)$$

بر اساس میزان خطای رده‌بند k تابع توزیع مجموعه آموزشی مرحله بعد تولید یا به عبارت دیگر نمونه‌ها برای رده‌بند $k+1$ دوباره وزن‌دهی می‌شوند. در هر بار وزن‌دهی، نمونه‌هایی که به‌درستی رده‌بندی شده‌اند با کاهش و نمونه‌هایی که به‌غلط رده‌بندی شده‌اند با افزایش وزن روبه‌رو می‌شوند؛ یعنی ممکن است وزن یک نمونه در چند مرحله افزایش یا کاهش یابد. رده‌بند جدید بر اساس تابع توزیع جدید آموزش داده شده و با اطلاع از این‌که رده‌بندهای قبلی در تشخیص کدام نمونه‌ها دچار مشکل هستند، روی نمونه‌های سخت‌تر تمرکز می‌کند. تابع توزیع $w+1$ در مرحله k بر اساس فرمول (۴) تولید می‌شود.

$$w_j^{k+1} = \frac{w_j^k \beta_k^{(1-l_j^k)}}{\sum_{i=1}^{N} w_j^k \beta_k^{(1-l_j^k)}} , \qquad j = 1, \cdots, N \qquad (۴)$$

در مرحله رده‌بندی، برای پیش‌بینی نمونه x یک پشتیبان به نام μ برای هر طبقه t در نظر گرفته می‌شود. وزن رده‌بندهایی که به طبقه t رأی می‌دهند با یکدیگر جمع شده و به‌عنوان پشتیبان طبقه t در نظر گرفته می‌شود. رأی‌گیری وزن‌دار به‌منظور پیش‌بینی برچسب برای هر نمونه x مطابق فرمول (۵) انجام می‌گیرد.

[1] Same base learner

[2] Adaptive Boosting (Adaboost)

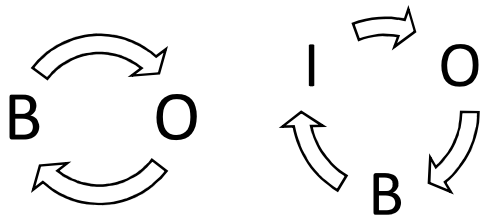
$$\mu_t(x) = \sum_{D_t(x)=\,\omega_t} \alpha_t \qquad (۵)$$

استفاده از لگاریتم برای محاسبه وزن هر رده‌بند (α_t)، به‌دلیل کاهش تأثیر زیاد رده‌بندهای با دقت بالا رأی‌گیری و قراردادن همه وزن‌ها در یک محدوده مشخص است. هر طبقه‌ای که حداکثر بیشینهٔ را داشته باشد، به‌عنوان طبقهٔ نمونه x انتخاب می‌شود.

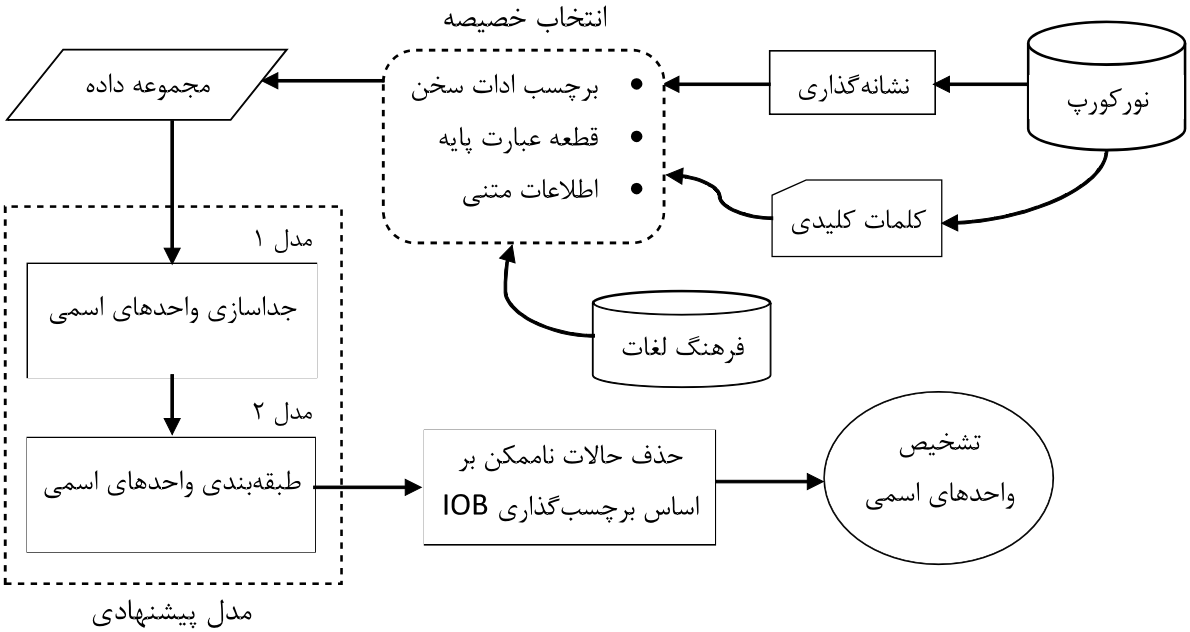
ما با استفاده از الگوریتم آدابوست و یک روش دو-مرحله‌ای، مدل خود را پیاده‌سازی کردیم. در مرحله نخست، مدل ما با استفاده از آدابوست M1 که یک الگوریتم بوستینگ برای تشخیص دو رده است، موجودیت‌های اسمی را جدا می‌کند. به عبارت دیگر مدل ۱ یکی از دو برچسب موجودیت اسمی و یا غیره را به هر نشانه اختصاص می‌دهد. در مرحله دوم با استفاده از الگوریتم آدابوست M2 که یک الگوریتم رده‌بندی چند رده[1] است مدل ۲ ایجاد شده و موجودیت‌های اسمی طبقه‌بندی می‌شوند. در این مرحله تمامی نمونه‌های مجموعه آموزشی، موجودیت‌های اسمی هستند که از مدل ۱ به‌دست آمده‌اند. بنابراین باید مشخص گردد جزو کدام طبقه از موجودیت‌های اسمی قرار می‌گیرند. شکل (۷) مراحل اصلی روش پیشنهادی را به‌صورت جزئی‌تر نشان می‌دهد.

۴-۵- پس‌پردازش

در برچسب‌گذاری IOB توالی برچسب‌ها به‌ترتیب و به‌صورت گردشی مطابق شکل (۶) توالی‌های ممکن در برچسب‌گذاری IOB خواهد بود. در این ترتیب هیچ‌گاه I پس از O قرار نمی‌گیرد؛ بنابراین، زمانی که در فرآیند پیش‌بینی چنین ترتیبی رخ می‌دهد ما برچسب I را به B تبدیل می‌کنیم. این عملیات نتایج را سه درصد بهبود داده است.



(شکل-۶): توالی‌های ممکن در برچسب‌گذاری IOB
(Figure-6): A sample for using keyword as a feature



(شکل-۷): روش پیشنهادی
(Figure-7): The proposal approach in detail

۵-آزمایش و نتایج

پیکره متنی نورکورپ را به دو بخش مجموعه آموزش و مجموعه آزمون تقسیم کردیم که به‌ترتیب ۸۰ و ۲۰ درصد از پیکره را شامل می‌شوند. مدل پیشنهادی در دو مرحله و به‌وسیله ابزار MATLAB پیاده‌سازی شده است.

۱-۵- ساخت مدل

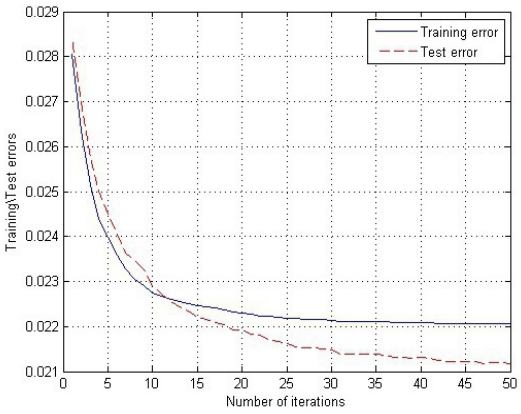
با استفاده از رده‌بند پایه درخت تصمیم، مجمعی شامل ۵۰

[1] Multi class

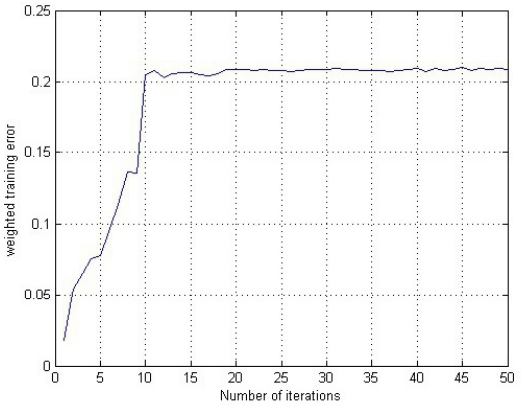
عضو تولید کردیم و از ریشه تصمیم[1] که مدلی یک‌سطحی از درخت تصمیم است، به‌منظور پیاده‌سازی یادگیرنده ضعیف[2] بهره‌گرفتیم. به‌دلیل این‌که تنها هفده درصد از پیکره متنی را واحدهای اسمی تشکیل می‌دهند، مجبور شدیم عمق درخت را بیشتر و حرص‌کردن را محدود کنیم. کمینه تعداد نمونه‌ها در هر برگ، یک در نظر گرفته شده و این درحالی است که از بیش‌برازش[3] جلوگیری شده است.

در شکل(۸) میزان خطای رده‌بندی پس از فرآیند آموزش و آزمون مدل به‌صورت مقایسه‌ای به نمایش در آمده است. این منحنی نشان می‌دهد که رده‌بندها در هر مرحله از تکرار به چه میزان در یادگیری الگوی داده موفق عمل کرده‌اند. محور افقی تعداد اعضای مجمع در هر مرحله از الگوریتم و محور عمودی نرخ خطای مجمع را نشان می‌دهد. مجمع ما در پنجاه مرحله و مطابق الگوریتم آدابوست شکل گرفته که در هر مرحله یک رده‌بند به آن اضافه شده است. نکته قابل توجه در این شکل، سرعت هم‌گرایی[4] است؛ در تکرارهای اولیه آموزش مدل، خطای رده‌بندی بالاست؛ اما در هر مرحله دقت رده‌بندی با تمرکز روی داده‌های سخت‌تر تقویت می‌شود. به این ترتیب خطای رده‌بندی بسیار کاهش یافته و در تکرار بیستم الی سی‌ام نتیجه نهایی حاصل می‌شود بنابراین تعداد اعضای بهینه مجمع برای مسئله ما درحدود بیست است. نکته جالب دیگر این است که در تکرارهای بالاتر، دقت رده‌بندی مجموعه آزمون بیش از مجموعه آموزش است. این موضوع طبق یکی از خصوصیت‌های بوستینگ رخ داده است، زمانی که نرخ خطای آموزش ثابت می‌شود، نرخ خطای آزمون همچنان می‌تواند کاهش یابد [34].

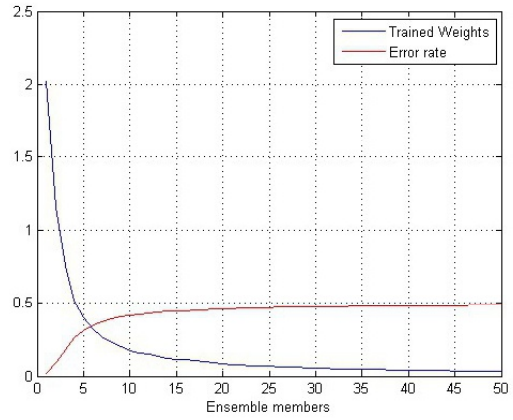
در شکل(۹) نرخ خطای رده‌بندی به‌صورت مجزا برای هر عضو مجمع آورده شده است. با توجه به این نمودار رده‌بندهای اولیه دارای نرخ خطای کمی هستند؛ اما با سخت‌ترشدن رده‌بندی در ادامهٔ روند آموزش، نرخ خطای رده‌بندها نیز افزایش می‌یابد تا جایی که نرخ خطای چهل عضو آخر بیش از بیست درصد است.



(شکل-۸): مقایسه نرخ خطای رده‌بندی در مرحله آموزش و آزمون
(Figure-8): Comparing training and test error



(شکل-۹): نرخ خطای رده‌بندی برای هر عضو مجمع
(figure-9): Training error for each learner individually



(شکل-۱۰): وزن اختصاص یافته به هر رده‌بند بر اساس نرخ خطا
(figure-10): Assigned weight to each learner based on error rate

هر چه نرخ خطای یک عضو بیشتر می‌شود، وزن کمتری به آن اختصاص می‌یابد؛ درنتیجه تأثیر اعضای با دقت پایین، در رأی‌گیری وزن‌دار کاهش می‌یابد. درشکل(۱۰) نرخ خطای هر رده‌بند با منحنی قرمز و وزن

[1] Decision stump
[2] Weak learner
[3] Overfitting
[4] Converge

اختصاص داده شده بر اساس نرخ خطا با منحنی آبی نشان داده شده است. با توجه به الگوریتم آدابوست، میزان خطای هر رده‌بند (متغیر ϵ) در بازه ۰ تا ۰/۵ است؛ یعنی دقت هر عضو از مجمع باید بیش از پنجاه درصد باشد.

۵-۲- ارزیابی مدل

مدل ایجادشده روی مجموعه آزمون به کار گرفته شد که نمودار تشخیص مجمع درشکل(۸) به‌صورت هاشور خورده آمده است. منحنی حاکی از نتیجه بسیار خوب مدل است. جدول (۷) نتایج خط پایه را نشان می‌دهد که از اعمال مدل پیشنهادی روی مجموعه آزمون به‌همراه دو خصیصه برچسب ادات سخن و قطعه عبارت پایه به‌دست آمده است. برای ارزیابی‌ها از معیار F-measure استفاده شده است [35].

(جدول-۷): نتایج خط پایه

(Table-7): Base line results

گروه	Precision	Recall	F-Measure
شخص	۷۳/۸۸	۶۶/۳	۶۹/۸۹
مکان	۸۲/۶۱	۲۴/۰۵	۳۷/۲۵
سازمان	۴۹/۵۶	۳۴/۵۷	۴۰/۷۳
زمان	۶۶/۱۵	۱۹/۲۸	۲۹/۸۶
همه	**۷۳/۳۵**	**۶۱/۰۷**	**۶۶/۶۵**

در جدول(۸) نتیجه نهایی به‌دست‌آمده از پیش‌بینی مدل ارائه‌شده روی مجموعه آزمون به تفکیک واحدهای اسمی آمده است. مدل ما در تشخیص اسامی اشخاص نسبت به سایر طبقات اسمی با نتیجه‌ای معادل ۹۱/۲۵ درصد، عملکرد بهتری دارد. این نتیجه مطلوب دو عامل اصلی دارد:

- اسامی اشخاص دارای الگوی ساده‌تری هستند.
- توزیع این اسامی در پیکره متنی تاحدودی یکپارچه است. گرچه تنوع در این اسامی زیاد است اما تعداد قابل توجهی از هر اسم در مجموعه آموزشی وجود دارد.
- فرهنگ لغات مورد استفاده در مورد اسامی اشخاص از غنای مناسبی برخوردار است. درحالی‌که در دیگر گروه‌ها، اسامی قابل توجهی در فرهنگ لغات وجود ندارد.

به‌طورکلی، نتیجه به‌دست‌آمده از روش پیشنهادی بیش از بیست درصد نسبت به خط پایه بهبود پیدا کرده است. در این میان بالاترین سهم به دو گروه شخص و زمان با افزایش ۲۲ و ۱۸ درصدی اختصاص دارد. این موضوع نشان می‌دهد، خصیصه‌های ارائه‌شده در این پژوهش تأثیر به‌سزایی روی موجودیت‌های اشخاص و زمان دارند. اما اسامی مکان و سازمان نسبت به خط پایه بهبود قابل توجهی نیافته‌اند و به‌ترتیب یک و سه درصد افزایش داشته‌اند.

روش ما در تشخیص همه طبقات اسمی دقت[1] قابل قبولی دارد؛ اما در معیار فراخوانی[2] نتایج ضعیف است. این موضوع بدان معنی است که مدل ما وقتی یک نشانه را واحد اسمی تشخیص می‌دهد با احتمال بسیار بالا به‌درستی عمل کرده است؛ اما در تشخیص بسیاری از واحدهای اسمی دچار مشکل است. به‌عبارت دیگر مدل ما در مرحله دوم مدل پیشنهادی بهتر از مرحله اول عمل می‌کند.

درجدول (۹) جزئیات نتایج به‌دست‌آمده از پیش‌بینی مدل ارائه‌شده روی مجموعه آزمون به‌تفکیک هر برچسب آورده شده است. درواقع این جدول نتایج را بر اساس برچسب‌گذاری IOB نشان می‌دهد. همان‌طور که در جدول مشخص است، مدل ما در تشخیص قسمت اولیه اسامی اشخاص ضعیف‌تر از قسمت‌های بعدی عمل می‌کند. بنابراین به‌منظور بهبود تشخیص اسامی اشخاص می‌بایست تشخیص قسمت اولیه را تقویت کرد؛ اما مدل ما در تشخیص قسمت اولیه اسامی مکان بیش از سه برابر بهتر از تشخیص قسمت‌های بعدی عمل می‌کند. این نتایج به‌طور کامل منطبق با جدول (۴) است به‌طورکلی نرخ برچسب‌های B و I در دو گروه شخص و مکان متضاد است.

به‌منظور بررسی عملکرد مدل دومرحله‌ای پیشنهادی، مدل تک‌مرحله‌ای با استفاده از آدابوست M2 پیاده‌سازی شد که نتایج آن در جدول (۱۰) نمایش داده شده است. مقایسه جدول (۸) و جدول (۱۰) نشان می‌دهد مدل پیشنهادی به‌طور کلی نتیجه را ۱/۷۵ درصد و به‌صورت جزئی واحدهای اسمی شخص، مکان و زمان را به‌ترتیب حدود ۲، ۳ و ۳ درصد بهبود می‌بخشد. گرچه مدل پیشنهادی توسعه قابل توجهی به همراه نداشته اما باعث بهبود نسبی شده است.

(جدول-۸): نتیجه نهایی به‌دست‌آمده از مدل پیشنهادی

(Table-8): Obtained results from our model

گروه	Precision	Recall	F-Measure
شخص	۹۲/۹۲	۸۹/۶۴	۹۱/۲۵
مکان	۸۸/۳	۲۶/۲۷	۴۰/۴۹
سازمان	۶۶/۶۷	۲۹/۶۳	۴۱/۰۳
زمان	۸۸/۸۹	۳۲/۲۹	۴۷/۳۷
همه	**۹۲/۳۸**	**۸۱/۹۵**	**۸۶/۸۵**

[1] Precision

[2] Recall

(جدول-۹): جزییات نتایج به‌دست‌آمده از پیش‌بینی مدل پیشنهادی روی مجموعه آزمون

(Table-9): Obtained results from our model in detail

برچسب	Precision	Recall	F-Measure
B-PERS	۸۷/۰۷	۷۷/۱۴	۸۱/۸
I-PERS	۹۵/۱۹	۹۲/۴۳	۹۳/۷۹
B-LOC	۷۹/۵۷	۲۹/۲۵	۴۲/۷۸
I-LOC	۴۵/۴۵	۷/۴۹	۱۳/۵۲
B-ORG	۶۵/۱۲	۳۱/۴۶	۴۲/۴۲
I-ORG	۶۴/۸۶	۲۳/۸۸	۴۳/۶۴

(جدول-۱۰): نتیجه به‌دست‌آمده از مدل پیشنهادی به‌صورت تک‌مرحله‌ای

(Table-10): Obtained results from our model as a single-step

گروه	Precision	Recall	F-Measure
شخص	۹۲/۴۶	۸۶/۹۷	۸۹/۶۳
مکان	۷۵/۹۶	۲۵/۰	۳۷/۶۲
سازمان	۶۵/۰	۳۲/۱	۴۲/۹۸
زمان	۸۲/۱۴	۳۰/۹۴	۴۴/۹۵
همه	**۹۱/۴۶**	**۷۹/۵۷**	**۸۵/۱**

به‌منظور این‌که عملکرد مدل پیشنهادی بهتر ارزیابی شود؛ مسئله را با الگوریتم کرت پیاده‌سازی کردیم که نتایج آن در جدول (۱۱) آمده است. از مقایسه جدول (۸) و جدول (۱۱) می‌توان دریافت مدل بوستینگ بیش از دوازده درصد دقت درخت تصمیم را افزایش می‌دهد که این موضوع حاکی از برتری مجمع ارائه‌شده نسبت به رده‌بند پایه است.

به‌منظور مقایسه میان پیکره متنی امروزی و کلاسیک، روش پیشنهادی را با اندکی تغییرات روی پیکره انرکورپ به کار گرفتیم که نتایج به‌دست‌آمده در جدول (۱۲) ارائه‌شده است. نتایج نشان می‌دهد که روش پیشنهادی روی پیکره نورکورپ به‌عنوان یک پیکره کلاسیک حدود نوزده درصد بهتر از پیکره انرکورپ به‌عنوان پیکره امروزی عمل می‌کند. نتیجه حاصل‌شده با دلایل ذیل قابل تحلیل است:

- تعدادی از اسامی موجود در پیکره انرکورپ مربوط به زبان‌های دیگر هستند. این اسامی از الگوی خاصی پیروی نمی‌کنند و در فرهنگ لغات نیز وجود ندارند. این موضوع را می‌توان در معیار بازخوانی مدل دنبال کرد. به‌عنوان مثال، با توجه به این‌که بیشتر اسامی سازمان در این پیکره، اسامی لاتین هستند، درنتیجه مدل ما تنها قادر به شناسایی ۳۷ درصد از این اسامی است.
- موجودیت‌های اسمی به‌صورت یکپارچه در پیکره متنی توزیع نشده‌اند، بنابراین موجودیت‌های زیادی در مجموعه آزمون وجود دارند که در مجموعه آموزش حضور ندارند.

یافته‌ها نشان می‌دهد متون عربی کلاسیک از الگوهای ساده‌تری نسبت به متون عربی امروزی برخوردار بوده و دارای تنوع کمتری در موجودیت‌های اسمی هستند.

(جدول-۱۱): نتایج به‌دست‌آمده از درخت تصمیم کرت

(Table-11): Obtained Results from CART decision tree

گروه	Precision	Recall	F-Measure
شخص	۹۳/۴۴	۹۰/۹۶	۹۲/۱۸
مکان	۶۳/۸	۵۹/۶۸	۶۱/۶۷
سازمان	۶۸/۶۱	۴۸/۷۵	۵۷/۰
همه	**۷۸/۹**	**۷۰/۸۲**	**۷۴/۶۴**

(جدول-۱۲): نتایج به‌کارگیری مدل پیشنهادی روی پیکره انرکورپ

(Table-12): Result of our approach on ANERCorp

گروه	Precision	Recall	F-Measure
شخص	۷۲/۶۱	۶۷/۹۱	۷۰/۱۸
مکان	۷۹/۸۶	۷۰/۱۹	۷۴/۷۱
سازمان	۷۵	۳۷/۶۷	۵۰/۱۵
همه	**۷۴/۹۴**	**۶۰/۵۹**	**۶۷/۰۱**

۶- نتیجه‌گیری و کارهای آینده

در این پژوهش روشی جدید جهت استخراج موجودیت‌های اسمی در عربی کلاسیک ارائه شده است. ما یک پیکره متنی عربی کلاسیک شامل ۱۳۰ هزار کلمه برچسب‌گذاری‌شده جهت اهداف پژوهشی توسعه دادیم و از یک فرهنگ لغات شامل ۱۸۰۰۰ اسم شخص، برگرفته از کتب حدیثی، به‌عنوان منبع خارجی استفاده کردیم. به‌منظور غلبه بر چالش‌های زبان عربی عملیات نشانه‌گذاری، برچسب‌گذاری ادات سخن و قطعه‌کردن عبارت پایه به‌کار گرفته شده است؛ همچنین به‌منظور توسعه مدل پیش‌بینی، از یک روش دومرحله‌ای شامل تشخیص اسامی خاص و طبقه‌بندی آن‌ها با استفاده از روش بوستینگ بهره گرفتیم. نتیجهٔ به‌دست‌آمده معادل ۸۶/۸۵ در معیار F-Measure است. روش پیشنهادی روی پیکره انرکورپ به‌عنوان پیکره عربی استاندارد امروزی نیز به کار گرفته شد که نتایج نشان

می‌دهد الگوهای موجود در عربی کلاسیک ساده‌تر است.

پیچیدگی‌هایی زبان عربی باعث شده تا استخراج اسامی خاص در این زبان نسبت به زبان‌های دیگر دقت کمتری داشته باشد؛ لذا همواره نیاز به روش‌هایی برای افزایش دقت وجود دارد. روش بوستینگ برای نخستین بار تا سال ۲۰۱۵ در حوزه تشخیص واحدهای اسمی عربی در این پژوهش مورد بررسی قرار گرفته که نتایج به موفقیت مطلوب این روش اشاره دارد. به‌صورت کلی روش‌های مبتنی بر ترکیب رده‌بندها از قدرت پیش‌بینی بالایی برخوردار هستند؛ زیرا سعی می‌کنند از جهات مختلف مسئله را مورد بررسی قرار دهند؛ بنابراین می‌توان از دیگر روش‌های ترکیب رده‌بندها نیز در حوزه استخراج اسامی خاص بهره برد. ضعف مدل پیش‌بینی ما در مرحله نخست است؛ بنابراین می‌بایست این مرحله بهبود یابد. در کارهای آینده می‌توان فرهنگ لغات را برای اسامی مکان و سازمان توسعه داد. همچنین از قوانین زبان‌شناسی، بیشتر استفاده کرد.

۷- منابع 7-Refrences

[1] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” *Lingvisticae Investig.*, vol. 30, no. 1, pp. 3–26, 2007.

[2] M. Oudah and K. Shaalan, “A Pipeline Arabic Named Entity Recognition using a Hybrid Approach.,” *Coling*, vol. 2, no. December 2012, pp. 2159–2176, 2012.

[3] S. Abuleil and M. Evens, “Extracting Names From Arabic Text for Question-Answering Systems.,” *Riao*, pp. 638–647, 2004.

[4] R. Koulali and A. Meziane, “A contribution to arabic named entity recognition,” in *International Conference on ICT and Knowledge Engineering*, 2012, pp. 46–52.

[5] K. Shaalan, “A Survey of Arabic Named Entity Recognition and Classification,” *Comput. Linguist.*, vol. 40, no. July 2013, pp. 469–510, 2014.

[6] N. Y. Habash, “Introduction to Arabic natural language processing,” *Synth. Lect. Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–187, 2010.

[7] H. Al-Jumaily, P. Martínez, J. L. Martínez-Fernández, and E. Van der Goot, “A real time Named Entity Recognition system for Arabic text mining,” *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 543–563, 2012.

[8] M. Korayem, D. Crandall, and M. Abdul-Mageed, “Subjectivity and sentiment analysis of arabic: A survey,” *Adv. Mach. Learn. ...*, 2012.

[9] Y. Maynard, D., Tablan, V., Ursu, C., Cunningham, H. ve Wilks, “Named Entity Recognition from Diverse Text Types,” in *Recent Advances in Natural Language Processing*, Springer, 2001, pp. 440–451.

[10] I. a Alkharashi, “Person Named Entity Generation and Recognition for Arabic Language,” in *the Proceedings of 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt*, 2009, pp. 205–208.

[11] ب. وزیرنژاد، ف. سلطانزاده، م. مهدوی و م. مرادی، "ویرایش‌گر متن شریف: سامانهٔ ویرایش و خطایابی املایی زبان فارسی"، *مجله پردازش علائم و داده‌ها*، شماره۱۲، صفحات ۴۳-۵۲، ۱۳۹۴.

[11] B. Vazirnejad, F. Soltanzadeh, M. Mahdavi, and M. Moradi, “Sharif Text Editor: A Persian Editor and Spell Checker System.,” *JSDP*, vol. 12, no. 4, pp. 43–52, 2016.

[12] I. A. Al-sughaiyer and I. A. Al-kharashi, “Arabic Morphological Analysis Techniques : A Comprehensive Survey,” *J. Am. Soc. Information Science and Technology*, vol. 55, no. 3, pp. 189–213, 2004.

[13] K. Darwish, A. Abdelali, and H. Mubarak, “Using Stem-Templates to improve Arabic POS and Gender/Number Tagging,” in *International Conference on Language Resources and Evaluation (LREC-2014)*, 2014, pp. 2926–2931.

[14] I. Zitouni, J. Sorensen, X. Luo, and R. Florian, “The impact of morphological stemming on Arabic mention detection and coreference resolution,” *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages,* June 29, pp. 63–70, 2005.

[15] Y. Benajiba, P. Rosso, M. Bened, and J. Bened iRuiz, “ANERsys : An Arabic Named Entity Recognition System Based on Maximum Entropy,” *Names*, pp. 143–153, 2007.

[16] Y. Benajiba and P. Rosso, "ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information.," in *3rd Indian International Conference on Artificial Intelligence (IICAI-07)*, 2007, pp. 1814–1823.

[17] Y. Benajiba and P. Rosso, "Arabic named entity recognition using conditional random fields," *Proc. Work. HLT NLP within ...*, 2008.

[18] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition using optimized feature sets," *Proc. Conf. Empir. Methods Nat. Lang. Process. EMNLP 08*, no. October, pp. 284–293, 2008.

[19] D. Valencia, "Arabic Named Entity Recognition," *Audio, Speech, Lang. Process. IEEE Trans.*, vol. 17, no. May, pp. 151–152, 2010.

[20] S. Abdallah, K. Shaalan, and M. Shoaib, "Integrating rule-based system with classification for arabic named entity recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7181 LNCS, no. PART 1, pp. 311–322, 2012.

[21] K. Shaalan and M. Oudah, "A hybrid approach to Arabic named entity recognition," *J. Inf. Sci.*, vol. 40, no. 1, pp. 67–87, 2014.

[22] M. A. Meselhi, H. M. Abo Bakr, I. Ziedan, and K. Shaalan, "Hybrid Named Entity Recognition-Application to Arabic Language," in *Computer Engineering & Systems (ICCES), 2014 9th International Conference on*, 2014, pp. 80–85.

[23] M. A. Meselhi, H. M. A. Bakr, I. Ziedan, and K. Shaalan, "A Novel Hybrid Approach to Arabic Named Entity," in *Machine Translation*, Springer, 2014, pp. 93–103.

[24] F. Enríquez, F. L. Cruz, F. J. Ortega, C. G Vallejo, and J. A. Troyano, "A comparative study of classifier combination applied to NLP tasks," *Inf. Fusion*, vol. 14, no. 3, pp. 255–267, 2013.

[25] X. Carreras, L. Marquez, and L. Padró, "Named entity extraction using adaboost," 2002, pp. 1–4.

[26] X. Carreras, L. Màrquez, and L. Padró, "A simple named entity extractor using AdaBoost," *... seventh Conf. Nat. ...*, 2003.

[27] G. Szarvas, R. Farkas, and A. Kocsor, "A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms," *Structure*, pp. 267–278, 2006.

[28] م. عسگری بیدهندی و ب. مینایی بیدگلی، "تشخیص اسامی اشخاص با استفاده از تزریق کلمه‌های نامزد اسم در میدان‌های تصادفی شرطی برای زبان عربی"، *مجله پردازش علائم و داده‌ها*، شماره ۱۱، صفحات ۷۳-۸۵، ۱۳۹۳.

[28] M. Asgari Bidhendi and B. Minaei Bidgoli, "Extracting person names using name candidate injection in a conditional random field model for Arabic language," *JSDP*, vol. 11, no. 1, pp. 73–85, 2014.

[29] م. رضائی شریف آبادی و پ. خسروی‌زاده، "برچسب‌زنی خودکار نقش‌های معنایی در جملات فارسی به کمک درخت‌های وابستگی"، *مجله پردازش علائم و داده‌ها*، شماره ۱۳، صفحات ۲۷-۳۸، ۱۳۹۵.

[29] M. Rezaei Sharifabadi and P. Khosravizadeh, "Automatic Labeling of Semantic Roles in Persian Sentences using Dependency Trees," *JSDP*, vol. 13, no. 1, pp. 27–38, 2016.

[30] F. Al Shamsi and A. Guessoum, "A hidden Markov model-based POS tagger for Arabic," in *Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France*, 2006, pp. 31–42.

[31] آ. سلیمی بدر و م. همایون‌پور، "تعیین مرز و نوع عبارات نحوی در متون فارسی"، *مجله پردازش علائم و داده‌ها*، شماره ۱۰، صفحات ۶۹-۸۶، ۱۳۹۲.

[31] A. Salimibadr and M. M. Homayounpour, "Phrase chunking in Persian texts," *JSDP*, vol. 10, no. 2, pp. 69–86, 2014.

[32] M. Diab, "Second Generation AMIRA Tools for Arabic Processing : Fast and Robust Tokenization , POS tagging , and Base Phrase Chunking," *Proc. Second Int. Conf. Arab. Lang. Resour. Tools*, pp. 285–288, 2009.

[33] L. Kuncheva, "Combining Pattern Classifiers

methods and algorithms. John Wiley&Sons," *Inc. Publ. Hoboken*, 2004.

[34] C. M. Bishop and others, *Pattern recognition and machine learning*, vol. 1. springer New York, 2006.

[35] ر. طباطبائی، م. فیضی درخشی و س. معصومی، "ارائه یک سیستم هوشمند و معناگرا برای ارزیابی سیستم های خلاصه ساز متون"، *مجله پردازش علائم و داده‌ها*، شماره ۱۲، صفحات ۳-۱۱، ۱۳۹۴.

[35] R. Tabatabaei, M. R. Feizi-Derakhshi, and S. Masoumi, "Proposing an intelligent and semantic-based system for Evaluating Text Summarizers," *JSDP*, vol. 12, no. 2, pp. 3–11, 2015.

بهروز مینایی بیدگلی دانش‌آموختهٔ دانشگاه ایالتی در رشته علوم میشیگان آمریکاو مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. ایشان در حال حاضر عضو هیأت علمی ودانشیار دانشکده مهندسی کامپیوتر دانشگاه علم وصنعت است. وی سرپرستی گروه پژوهشی فناوری‌های بازی‌های رایانه‌ای و نیز آزمایشگاه داده‌کاوی را به عهده دارد. محاسبات نرم، یادگیری ماشین، بازی‌های رایانه‌ای، داده‌کاوی، متن کاوی و پردازش زبان طبیعی، زمینه‌های پژوهشی مورد علاقه ایشان است.

نشانی رایانامهٔ ایشان عبارت است از:

b_minaei@iust.ac.ir

سید محمد باقر سجادی، مدرک کارشناسی ارشد خود را از دانشگاه آزاد قزوین در سال ۱۳۹۳ در رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک اخذ کرد. ایشان هم اکنون دانشجوی مقطع دکتری مهندسی نرم‌افزار در دانشگاه آزاد تهران مرکز می‌باشد. زمینه‌های پژوهشی مورد علاقه وی، پردازش زبان طبیعی، وب معنایی و داده‌های پیوندی است.

نشانی رایانامهٔ ایشان عبارت است از:

moh.sajadi.eng@iauctb.ac.ir

حسن رشیدی، دانشیار دانشکده ریاضی و علوم کامپیوتر در دانشگاه علامه طباطبایی است. او مدرک کارشناسی را در رشته مهندسی کامپیوتر و مدرک کارشناسی ارشد را در رشته مهندسی برنامه ریزی سیستم ها، هر دو از دانشگاه صنعتی اصفهان، و دکترای خود را از دانشگاه اسکس انگلستان دریافت کرد. حوزه کاری ایشان شامل مهندسی نرم‌افزار و الگوریتم‌های بهینه‌سازی است. ایشان مقالات پژوهشی بسیاری در کنفرانس‌ها و نشریات علمی بین المللی منتشر کرده است.

نشانی رایانامهٔ ایشان عبارت است از:

hrashi@atu.ac.ir