پیکره اعلام: یک پیکره استاندارد واحدهای اسمی برای زبان فارسی

شادی حسین‌نژاد [۱]، یاسر شکفته[۲]* و طاهره امامی آزادی [۳]
[۱و۳] پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی، تهران، ایران
[۲] دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران

چکیده

تشخیص واحدهای اسمی یکی از مسائل مطرح در پردازش زبان طبیعی است. کاربرد عمده شناسایی واحدهای اسمی در سامانه‌های خلاصه‌ساز متون، استخراج اطلاعات، پرسش و پاسخ، ترجمه ماشینی و دسته‌بندی اسناد است. یکی از روش‌های تهیه سامانهٔ تشخیص واحدهای اسمی، استفاده از روش‌های مبتنی بر پیکره است. این مقاله نحوه و مراحل تهیه پیکره آعلام – یک پیکره استاندارد با برچسب واحدهای اسمی برای زبان فارسی- را شرح می‌دهد. مجموعه تهیه‌شده با داشتن سیزده برچسب واحدهای اسمی و حجم ۲۵۰ هزار کلمه نیاز سامانه‌های برچسب‌گذاری خودکار در حوزه پردازش زبان طبیعی فارسی را برآورده می‌کند. با استفاده از این پیکره و به‌کارگیری روش یادگیری ماشین میدان تصادفی شرطی، سامانه‌ای برای شناسایی واحدهای اسمی جملات فارسی تهیه شده که دارای دقت ۹۲/۹۴ درصد و فراخوانی ۷۸/۴۸ درصد است.

واژگان کلیدی: پردازش زبان طبیعی، تشخیص واحدهای اسمی، پیکره واحدهای اسمی، یادگیری ماشین، میدان تصادفی شرطی.

A’laam Corpus: A Standard Corpus of Named Entity for Persian Language

Shadi Hosseinnejad [1], Yasser Shekofteh[2]*&Tahereh Emami Azadi [3]
[1,3] Voice and Natural Language Processing Group, Department of Data Processing, RCDAT, Tehran, Iran
[2] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

Abstract

Named entity recognition (NER) is a natural language processing (NLP) problem that is mainly used for text summarization, data mining, data retrieval, question and answering, machine translation, and document classification systems. A NER system is tasked with determining the border of each named entity, recognizing its type and classifying it into predefined categories. The categories of named entities include the names of persons, organizations, locations (e.g. city and country), expressions of times, quantities, monetary expressions, and percentages. In general, corpus-based NER approaches have been proved to be well suited for NER problem. Using a NER corpus, recognition of named entities can be done through ruled-based or machine-learning methods.

Corpus-based NER systems need standard and appropriate annotated corpora. However, such corpora mainly exist in languages such as English, and are rarely found in Persian/Farsi or limited in volume. So, this paper is dedicated to describe the producing procedure of a standard named entity (NE) corpus - A’laam corpus - for Persian language. A’laam corpus contains about 250,000 tokens tagged with 13 NE tags. This corpus has been developed in the Research Center for Development of Advanced Technologies

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

[DOI: 10.29252/jsdp.14.3.127]

(RCDAT). Tokens of A'laam corpus are a part of Farsi Text Corpus. The Farsi Text Corpus is a standard Farsi corpus. This corpus, containing more than 100 million Farsi words, has been developed by the Research Center of Intelligent Signal Processing (changed to the Research Center for Development of Advanced Technologies in 2013). The words of this corpus, selected from diverse written and spoken sources, was tokenized and corrected manually. In addition, a part of the Farsi Text Corpus with 8 million words has part-of-speech (POS) tags at word level. Totally, about 8,400 sentences of the Farsi Text Corpus have been randomly selected to obtain about 250,000 tokens of A'laam Corpus. This corpus included words, POS tags, and named entity tags.

To evaluate A'laam corpus, a Persian NER system was trained based on this corpus. This corpus was so divided into the train and test sections. The train section accounted for 90% of the corpus and the remaining 10% belonged to the test section. Using Conditional Random Fields (CRF) method, the Persian NER system resulted in a 92.94% Precision and 78.48% Recall.

Keywords: **Natural language Processing, Named Entity Recognition, Named Entity Corpus, Machine learning, Conditional Random Field.**

۱- مقدمه

تشخیص واحدهای اسمی[1] یکی از مسائل پردازش زبان طبیعی[2] است که کاربردهای عمده آن در سامانه‌های خلاصه‌ساز متون، استخراج اطلاعات، بازیابی اطلاعات، پرسش و پاسخ، مترجم ماشینی و دسته‌بندی اسناد است [10]. واحد اسمی به یک یا چند کلمه‌ای در متن گفته می‌شود که به‌طور خاص به یک موجود در جهان خارج اشاره کند و بتواند آن را از نمونه‌های مشابه تمیز دهد [36].

مسأله شناسایی واحدهای اسمی شامل دو بخش کلی «شناسایی» و «طبقه‌بندی» است. گاه از اصطلاح NERC[3] یا شناسایی و طبقه‌بندی واحدهای اسمی به جای NER استفاده می‌شود [24]. وظیفه یک سامانه تشخیص واحدهای اسمی، علاوه‌بر تعیین مرز هر یک از واحدهای اسمی، تشخیص نوع آن واحد و قراردادن آن در دسته‌های از پیش تعیین شده است. این دسته‌ها شامل اسامی خاص افراد، مکان‌ها (شهر، کشور و غیره)، اسامی سازمان‌ها، عبارت‌های زمانی، کمیت‌ها، عبارت‌های پولی، درصد و غیره می‌شوند. اگرچه در کاربردهای مختلف ممکن است دسته‌های دیگری نیز تعریف شوند. برای مثال در حوزه پزشکی نام بیماری‌ها و داروها؛ در حوزه زیست‌شناسی نام انواع پروتئین‌ها، دی‌ان‌ای[4]ها و نوع سلول می‌توانند از دسته‌های واحدهای اسمی باشند [2]. در حوزه شیمی، نام انواع فرمول‌ها و مواد شیمیایی می‌توانند از دسته‌های واحدهای اسمی باشند [29]. برای تولید سامانه تشخیص واحدهای اسمی روش‌های مختلفی وجود دارد. به‌طور کلی تشخیص واحدهای اسمی با استفاده از روش‌های مبتنی بر قانون و یا روش‌های یادگیری ماشین انجام می‌شود. برای تولید سامانه‌های تشخیص واحدهای اسمی با روش یادگیری ماشین نیاز به یک پیکره استاندارد و مناسب وجود دارد. اگرچه چنین پیکره‌ای در زبان‌های مختلف از جمله انگلیسی تهیه شده ، اما در زبان فارسی تاکنون گزارش جامعی پیرامون تهیه یک پیکره استاندارد در جهت برچسب‌گذاری کامل واحدهای اسمی ارائه نشده است.

در دو دهه گذشته در زبان‌های مختلف از جمله زبان انگلیسی پژوهش‌های زیادی در این زمینه انجام گرفته است. تشخیص واحدهای اسمی در زبان انگلیسی در رقابت‌های CoNLL-2003[5]، MUC-6[6]، MUC-7 موضوع پژوهش و بررسی بوده است. مطالعات اولیه مساله تشخیص واحدهای اسمی در رقابت CoNLL-2002 برای زبان‌های هلندی و اسپانیایی انجام گرفته است [31]. زبان آلمانی نخستین‌بار در رقابت CoNLL-2003 مورد توجه قرار گرفت [32]. زبان ژاپنی در رقابت MUC-6 و IREX [34] برای نخستین‌بار مطالعه شده است. همچنین برای زبان‌های فرانسه [27]، ایتالیایی [14]، پرتغالی [26]،[33]، عربی [6] و برخی زبان‌های دیگر نیز مطالعاتی صورت گرفته است. سابقه مطالعات تشخیص واحدهای اسمی در زبان فارسی زیاد نیست و تنها چند پژوهش محدود برای آن صورت گرفته است [2].

این مقاله به بررسی روند تولید یک پیکره استاندارد واحدهای اسمی برای زبان فارسی می‌پردازد. در بخش دوم ابتدا کارهای پیشین صورت گرفته در تشخیص واحدهای اسمی بررسی و در بخش سوم نحوه تهیه و تولید پیکره استاندارد واحدهای اسمی زبان فارسی (پیکره اعلام) بیان می‌شود. بخش چهارم اختصاص به معرفی سامانه تشخیص

[1] Named Entity Recognition (NER)
[2] Natural Language Processing
[3] Named Entity Recognition and Classification
[4] DNA

[5] Conference on computational natural language learning
[6] Message Understanding Conference

[DOI: 10.29252/jsdp.14.3.127]

واحدهای اسمی دارد. این سامانه با استفاده از پیکره اعلام تهیه شده است. همچنین در این بخش به بیان نتایج حاصله از ارزیابی سامانه پرداخته می‌شود و درنهایت بخش پنجم به نتیجه‌گیری و ارائه پیشنهاداتی برای کارهای آتی تخصیص یافته است.

۲- کارهای پیشین

در این بخش به بررسی پژوهش‌های مهم انجام‌شده در زمینه تشخیص واحدهای اسمی می‌پردازیم. بررسی این پژوهش‌ها به دو بخش تقسیم شده است. در بخش نخست به بررسی رقابت‌های معروف برگزارشده و پیکره‌های تولیدشده در این زمینه پرداخته می‌شود. این رقابت‌ها شامل پژوهش‌های بنیادی عملیات تشخیص واحدهای اسمی هستند که در آنها پیکره‌ای برای این منظور تهیه شده است و شرکت‌کنندگان در این رقابت‌ها با استفاده از داده تهیه شده، به تولید سامانه‌های تشخیص واحدهای اسمی پرداخته‌اند. در بخش دوم سامانه‌های تشخیص واحدهای اسمی بر اساس روش به‌کار رفته در تولید آنها بررسی می‌شوند.

۲-۱- رقابت‌های تشخیص واحدهای اسمی

در سال ۱۹۹۶ میلادی در رقابت MUC-6 [17] بود که برای نخستین‌بار اصطلاح تشخیص واحد اسمی به‌کار رفت. در این رقابت وظیفه شرکت‌کنندگان تشخیص و دسته‌بندی[1] سه نوع اصلی واحد اسمی بود. این سه نوع واحد اسمی عبارتند از: واحدهای عددی، واحدهای زمانی و واحدهای اسامی. در این رقابت واحدهای اسمی به صورت زیر تعریف شده‌اند:

ENAMEX: اسامی اشخاص، سازمان‌ها و مکان‌ها

TIMEX: تاریخ و زمان

NUMEX: عبارات پولی، درصد و کمیت

داده‌های مورد استفاده در این رقابت شامل ۳۱۸ عنوان مقاله از WSJ[2] است که به‌صورت دستی برچسب‌گذاری شده است. پس از رقابت MUC-6 رخدادهای علمی زیادی با موضوع تشخیص واحدهای اسمی شکل گرفت.

در سال ۱۹۹۸ میلادی در رقابت MUC-7 [11] نیز تشخیص واحدهای اسمی مد نظر قرار داشت. در این رقابت واحدهای معرفی‌شده شامل اسامی اشخاص، سازمان، مکان، تاریخ، زمان، درصد و عبارات پولی بودند. پیکره مورد استفاده در این رقابت بخشی از پیکره New York Times News Service است.

طبق تعاریف **دارپا** [12] واحدهای اسمی به سه دسته اصلی واحدهای اسمی، زمانی و شماره‌ای[3] تقسیم‌بندی و هر یک از این سه دسته شامل زیردسته‌هایی می‌شوند. این زیردسته‌ها برای واحدهای اسمی شامل، اسم شخص، مکان و سازمان است. واحدهای زمانی شامل تاریخ، زمان و بازه زمانی می‌شوند و واحدهای شماره‌ای، عبارات پولی، درصد، اعداد اصلی و اندازه‌های استاندارد مانند سن، حجم، وزن و غیره را دربر می‌گیرند.

IREX [34] رقابتی است که برای زبان ژاپنی در سال‌های ۱۹۹۸ تا ۱۹۹۹ در حال انجام بوده است. در این رقابت بخش تشخیص واحدهای اسمی، شامل تشخیص اسامی اشخاص، مکان‌ها، مصنوعات بشر، زمان و عبارات عددی است.

در سال ۲۰۰۲ و ۲۰۰۳ میلادی رقابت‌های CoNLL با هدف تشخیص واحدهای اسمی شکل گرفت. در سال ۲۰۰۲ میلادی این رقابت‌ها برای زبان‌های اسپانیایی و هلندی برگزار شد و رقابت سال ۲۰۰۳ میلادی به تشخیص واحدهای اسمی در زبان‌های انگلیسی و آلمانی اختصاص داشت. پیکره مورد استفاده در این رقابت پیکره رویترز [30] برای زبان انگلیسی و پیکره چندزبانه ECI [5] برای زبان آلمانی است. واحدهای اسمی به‌کار رفته در این رقابت‌ها چهار واحد اسمی شخص، مکان، سازمان و متفرقه است. گروه متفرقه شامل موارد: تاریخ، زمان، عدد، عبارات پولی، عبارات اندازه‌گیری و درصد است. برچسب‌گذاری این داده‌ها به‌صورت دستی و با استفاده از روش IOB صورت گرفته است. روش IOB یک روش برچسب‌گذاری است که در آن هر کلمه یک برچسب مخصوص به خود دارد. کلمه آغازین واحد اسمی برچسب B، کلمات داخل واحد اسمی برچسب I و کلمات خارج آن برچسب O می‌گیرند [6].

برنامه ACE[4] [15] از سال ۱۹۹۹ میلادی آغاز شد و تا سال ۲۰۰۴ میلادی ادامه یافت. این برنامه در آغاز به استخراج واحدهای اسمی اختصاص داشت اما از سال ۲۰۰۰

[1] classify
[2] Wall Street Journal
[3] number
[4] The Automatic Content Extraction

[DOI: 10.29252/jsdp.14.3.127]

میلادی به بعد علاوه‌بر واحدهای اسمی، روابط و رخدادها نیز در دستور کارش قرار گرفت. همچنین زبان‌های عربی و چینی نیز علاوه‌بر زبان انگلیسی اضافه شدند. واحدهای این برنامه تا سال ۲۰۰۰ شامل پنج واحد اسامی اشخاص، سازمان‌ها، مکان‌های سیاسی[1]، مکان و امکانات[2] بود که از سال ۲۰۰۴ دو واحد اسمی، وسایل نقلیه و سلاح نیز به آنها اضافه شد. مکان سیاسی به اسامی‌ای مانند نام شهرها و کشورها اطلاق می‌شود و مکان به اسامی‌ای مانند نام کوه‌ها و رودخانه‌ها و مانند آن گفته می‌شود.

در سال ۲۰۰۶ میلادی رقابت HAREM[3] [33] برای تشخیص واحدهای اسمی در زبان پرتغالی شکل گرفت. در این رقابت پیکره‌ای حاوی بیش از ۴۶۶ هزار کلمه در هشت دسته برچسب‌گذاری شده است که هر دسته شامل چند زیر دسته نیز می‌شود. دسته‌های مشخص‌شده عبارتند از: شخص، زمان، مکان، سازمان، کمیت، رخداد، انتزاعی و اشیاء.

در سال ۲۰۰۵ میلادی پیکره هم‌مرجع و واحدهای **اسمی**BBN [37] منتشر شد. این پیکره برروی یک میلیون قطعه[4] پیکره PennTreeBank [21] تهیه شده است که مطالب آن برگرفته از WSJ[5] است. پیکره از دو بخش تشکیل شده و بخش نخست شامل برچسب‌گذاری هم‌مرجع ضمایر است و بخش دوم به برچسب‌گذاری واحدهای اسمی اختصاص دارد. واحدهای اسمی این پیکره در سه دسته اصلی قرار دارند که هر دسته شامل زیردسته‌های متعددی می‌شود و در مجموع ۶۴ دسته واحد اسمی را شامل می‌شوند. این پیکره برای استفاده در برنامه ACE تولید شده است.

در پیکره Ontonotes نیز برچسب‌گذاری واحدهای اسمی صورت گرفته است [28]. طبق تعریف استفاده شده در تهیه این پیکره، ۱۱ واحد اسمی و ۷ واحد رقمی برچسب‌گذاری می‌شوند.

در زبان عربی، علاوه‌بر پیکره‌های تولیدشده در برنامه ACE، پیکره مهم ANERCorp [6] در سال ۲۰۰۷ میلادی تهیه شده است. این پیکره به‌صورت رایگان قابل دسترس است. واحدهای اسمی به کار رفته در این پیکره شامل شخص، مکان، سازمان و متفرقه (طبق تعریف CoNLL2003) است. حجم این پیکره ۱۵۰ هزار کلمه است. همچنین به همراه این پیکره فرهنگ اسامی اشخاص، مکان‌ها و سازمان‌ها نیز تهیه شده است.

پیکره کلمات[6] [16] با حجم بیش از ۱۸ میلیون کلمه عربی، شامل ۲۰۲۹۱ سند متنی است که هر سند برچسب‌های متعددی دارد. یکی از ویژگی‌های این پیکره داشتن برچسب واحدهای اسمی است. پیکره کلمات به صورت خودکار با سامانه تشخیص واحدهای اسمی ANER برچسب‌گذاری شده است. از آنجایی‌که ANER بر روی پیکره ANERCorp آموزش داده شده، برچسب‌های پیکره کلمات همانند برچسب‌های پیکره ANER است.

در همین اواخر پیکره‌ای برای زبان فارسی تولید و گزارش شده که شامل چهارصد هزار قطعه از پیکره متنی بیجن خان است که به‌صورت دستی برچسب‌گذاری شده است. این پیکره تنها شامل سه برچسب شخص، مکان و سازمان است و برچسب‌های دیگر را در بر نمی‌گیرد. تعداد کل واحدهای اسمی این پیکره در حدود ۱۵ هزار واحد اسمی است [3].

همچنین شرکت تجاری appen[7] در سال ۲۰۱۵ میلادی برای هر یک از زبان‌های عربی، انگلیسی، فارسی، ژاپنی، کره‌ای، چینی، روسی و اردو پیکره‌های واحدهای اسمی جداگانه با حدود پانصد هزار کلمه برای هر زبان منتشر کرده است. این پیکره‌ها تنها حاوی هشت برچسب واحد اسمی شامل شخص، سازمان، مکان، ملیت، دین، عنوان، مکان‌های سیاسی و امکانات می‌باشند. اطلاعات بیشتری از این پیکره در دسترس نیست.

۲-۲- سامانه‌های تشخیص واحدهای اسمی

تشخیص واحدهای اسمی، با روش‌های متفاوتی انجام می‌گیرد. از جمله این روش‌ها می‌توان به استخراج با واژگان، روش‌های مبتنی بر یادگیری باسرپرست و یا نیمه‌سرپرستی شده و یا بدون سرپرستی اشاره نمود.

روش‌های مبتنی بر واژگان با دو رویکرد کلی انجام می‌شوند: ۱) تهیه اسامی به‌صورت دستی؛ ۲) استخراج اسامی از منابعی مانند ویکی‌پدیا [38]. علی‌رغم سهولت پیاده‌سازی، روش‌هایی که تنها از فرهنگ لغات[8] در تعیین واحدهای اسمی استفاده می‌کنند، به‌دلیل درنظرنگرفتن

[1] Geo-Political Entities (GPE)
[2] facility
[3] HAREM - Avaliac¸˜ao de sistemas de Reconhecimento de Entidades Mencionadas
[4] token
[5] Wall Street Journal
[6] Kalimat
[7] www.appen.com
[8] dictionary

[DOI: 10.29252/jsdp.14.3.127]

محتوای متن[1]، دقت شناسایی و طبقه‌بندی بالایی ندارند و به‌طورمعمول به‌عنوان زیربرنامه جانبی یا درون بلوک روش‌های مبتنی بر یادگیری قرار می‌گیرند و به‌تنهایی استفاده نمی‌شوند.

در برخی از روش‌ها از قوانین برای تعیین واحدهای اسمی استفاده می‌شود [13]. مجموعه دیگری از روش‌ها از کنار هم قرار گرفتن کلمات در پیکره در کنار اطلاعاتی مانند شبکه واژگانی بهره می‌جویند [20]. در این روش‌ها که از آن به روش‌های بدون سرپرست یاد می‌شود، نوع واحدهای اسمی از برچسب‌های پیشنهادشده در شبکه واژگان تعیین می‌شوند. این روش، امکان شناسایی واحدهای اسمی کم‌رخداد را نیز فراهم می‌سازد که در کنار سایر روش‌های NERC و در ترکیب با آنها قابل استفاده خواهد بود. این قبیل روش‌ها، نیازی به پیکره متنی برچسب خورده جهت تعلیم ندارند، با این وجود برای ارزیابی آنها وجود داده متنی با برچسب واحدهای اسمی ضروری است. بهترین عملکرد بین سامانه‌های تشخیص واحدهای اسمی به روش‌های باسرپرست مربوط می‌شود. در این روش‌ها یادگیری روی داده برچسب‌خورده انجام می‌شود. نمونه‌هایی از الگوریتم‌های به‌کاررفته در یادگیری باسرپرست عبارتند از: مدل مخفی مارکوف[2] [8]، روش مبتنی بر آنتروپی بیشینه[3] [9]، درخت تصمیم[4] [35] و روش میدان‌های تصادفی شرطی[5] [22]. البته در نبود مجموعه دادگان برچسب‌خورده کافی از روش‌های نیمه‌سرپرستی‌شده یا سامانه‌های چندمرحله‌ای استفاده می‌شود [18]. دلیل موفقیت این روش‌ها، توجه به محتوای متن است که در تشخیص عبارات مبهم، بهتر از سایر روش‌ها عمل می‌کنند. در سال‌های اخیر استفاده از روش‌های ترکیبی مبتنی بر قاعده و روش یادگیری ماشینی رواج یافته و پژوهش‌هایی در این زمینه در زبان‌های مختلف صورت گرفته است [25].

در پژوهشی برای زبان عربی [4] پس از تهیه پیکره‌ای استاندارد از متون عربی کهن، با استفاده از روش میدان تصادفی شرطی سامانه تشخیص اسامی اشخاص تولید شده است. پیکره تهیه‌شده در این پژوهش، NoorCorp نام دارد. این پیکره برمبنای دستورالعمل CoNLL-2003 تهیه شده و چهار برچسب شخص، مکان، سازمان و متفرقه را دارد. قالب برچسب‌گذاری این پیکره IOB است. در زبان فارسی نیز چند پژوهش در زمینه تولید سامانه برچسب‌گذار واحدهای اسمی انجام گرفته است. با توجه به در دسترس‌نبودن داده مناسب و کافی در زبان فارسی، تمرکز اغلب این پژوهش‌ها بر روش‌های مبتنی بر قانون است [2]. سادات‌مرتضوی در پژوهشی در زبان فارسی سامانه‌ای برای تشخیص واحدهای اسمی و دسته‌بندی آنها در زبان فارسی تهیه کرده است. این سامانه با به‌کارگیری ساختار واژگانی اسامی خاص و نیز الگوهای متنی ممکن برای اسم‌های خاص متعلق به یک دسته، سعی در شناسایی واحدهای اسمی می‌کند [2].

در پژوهش دیگری [1] با استفاده از چهار طبقه‌بندی‌کننده خطی، بیزین، نزدیکترین همسایگی[6] و شبکه عصبی سامانه تشخیص واحدهای اسمی را آموزش داده‌اند. داده مورد استفاده در این سامانه با اعمال قوانین بر روی پیکره متنی فارسی تهیه شده است. با استفاده از برچسب اجزای کلام پیکره متنی فارسی سه دسته از اسامی (اسامی شخص، مکان و اسامی عمومی) از پیکره استخراج شده‌اند و سپس طبقه‌بندی‌کننده‌های مختلفی برای تشخیص واحدهای اسمی آموزش داده شده است. نتیجه این سامانه با استفاده از طبقه‌بندی‌کننده خطی و نزدیکترین همسایگی مقدار f-measure حدود ۹۱٪ گزارش شده است. نتیجه این پژوهش برای تشخیص اسامی زمان با استفاده از یک فهرست کمکی f-measure حدود ۹۶٪ به دست آمده است.

با استفاده از فرهنگ لغات و اعمال فیلترها [19] سامانه‌ای برای تشخیص واحدهای اسمی تولید و در تولید این سامانه از فرهنگ‌های متعددی استفاده شده و این سامانه قادر به تشخیص چهار واحد اسمی اسم شخص، مکان، سازمان و متفرقه است. دقت گزارش‌شده ۸۲/۷۳٪ بر مبنای معیار f-measure است.

در پژوهش دیگری در زمینه تولید سامانه تشخیص واحدهای اسمی برای زبان فارسی با استفاده از ترکیب روش یادگیری ماشین مدل مخفی مارکف و روش مبتنی بر قواعد سامانه‌ای تولید شده که قادر به تشخیص سه واحد اسمی اسامی اشخاص، مکان و سازمان است. این سامانه بر روی حجم داده ۳۲ هزار کلمه‌ای آزمایش شده و نتیجه آن ۸۵٪/۹۳ برای معیار f-measure گزارش شده است [23].

[1] context
[2] Hidden markov model (HMM)
[3] Maximum Entropy (ME)
[4] Decesion tree (DT)
[5] Conditional random field (CRF)
[6] KNN

[DOI: 10.29252/jsdp.14.3.127]

پیش‌نیاز روش‌های یادگیری سرپرستی‌شده، داشتن دادگان برچسب‌خورده است. چنین دادگانی شامل مجموعه‌های اولیه برچسب‌خورده به‌صورت دستی هستند. در ادامه این مقاله به توصیف پیکره واحدهای اسمی تهیه‌شده برای زبان فارسی پرداخته خواهد شد که این پیکره استاندارد «پیکره اَعلام»[1] نام‌گذاری شده است. همچنین ابزار تشخیص واحدهای اسمی تولیدشده برمبنای این پیکره نیز در ادامه شرح داده شده است.

۳- فرآیند تولید پیکره اعلام

تولید پیکره استاندارد واحدهای اسمی زبان فارسی در چهار مرحله صورت گرفته است:

۱- انتخاب برچسب‌های واحدهای اسمی (تهیه دستورالعمل برچسب‌گذاری)
۲- انتخاب دادگان خام مناسب
۳- برچسب‌گذاری دستی دادگان برمبنای دستورالعمل برچسب‌گذاری
۴- تبدیل دادگان برچسب‌گذاری‌شده به فرمت استاندارد

۳-۱- انتخاب برچسب‌های واحدهای اسمی

در پیکره تولیدشده که در این مقاله به شرح آن می‌پردازیم، اساس کار در تولید دستورالعمل برچسب‌گذاری، تعریف ارائه‌شده از واحدهای اسمی در پیکره Ontonotes است. در مواردی نیز از تعریف ارائه‌شده توسط دارپا[2] در سال ۱۹۹۹ میلادی استفاده شده است. در دستورالعمل تهیه پیکره، بنا به ویژگی‌های زبان فارسی و نیازهای پردازشی، برخی واحدهای اسمی پیکره Ontonotes حذف یا اضافه و برخی دیگر با یکدیگر درآمیخته شده‌اند. در جدول (۱) واحدهای اسمی تعریف‌شده در پیکره Ontonotes مشاهده می‌شود.

(جدول-۱): واحدهای اسمی پیکره Ontonotes.
(Table-1): Different named entities in Ontonotes Corpus.

ردیف	واحد اسمی	تعریف
۱	شخص	اسامی افراد و شخصیت‌های غیرحقیقی (افسانه‌ای)
۲	نورپ[3]	مخفف سه کلمه ملیت، دین یا گروه سیاسی
۳	امکانات	ساختمان‌ها، پل‌ها، فرودگاه‌ها و ...
۴	سازمان	شرکت‌ها، ادارات دولتی، نشریه‌ها و ...
۵	مکان‌های سیاسی[4]	کشور، شهر، استان، ایالت
۶	مکان	مکان‌هایی که در دسته‌بندی ردیف ۵ نمی‌گنجند، مانند اسامی کوه‌ها، رودخانه‌ها و ...
۷	محصول	وسایل نقلیه، اسلحه، مواد غذایی و ...
۸	رخداد	نام جنگ‌ها، انقلاب‌ها، رخدادهای ورزشی و ...
۹	اثر هنری[5]	نام کتاب‌ها، آثار هنری، موسیقی و ...
۱۰	قانون	اسناد قانونی (که نام داشته باشند)
۱۱	زبان	اسامی زبان‌ها
۱۲	تاریخ	تاریخ یا بازه تاریخی
۱۳	زمان	زمان‌های کوتاهتر از یک روز
۱۴	درصد	درصد شامل %
۱۵	پول	عبارات پولی
۱۶	کمیت	مقادیر
۱۷	عدد اصلی	اعداد اصلی
۱۸	عدد ترتیبی	اعداد ترتیبی

تفاوت‌های عمده برچسب‌گذاری پیکره Ontonotes و «پیکره اَعلام» به شرح زیر است:

- سه واحد اسمی مکان و مکان سیاسی و امکانات، در پیکره اعلام یک واحد در نظر گرفته می‌شوند و تمایزی میان آنها وجود ندارد. علت این تصمیم دشواربودن تمایز میان این سه واحد بالاخص واحدهای اسمی مکان و مکان سیاسی است.
- دو واحد اسمی زبان و نورپ در یکدیگر ادغام شده و واحد ال‌نورپ[6] را تشکیل داده‌اند. به‌دلیل اینکه زبان‌ها در بسیاری موارد با نام ملیت‌ها یکی هستند و این مسأله تمایز میان این دو را سخت می‌کند، برای مثال «زبان هلندی» و «ملیت

[1] نحوه دسترسی به پیکره از طریق ارتباط با speech@rcdat.ir امکان‌پذیر است.
[2] DARPA
[3] NORP
[4] GPE
[5] Work Of Art
[6] LNORP

[DOI: 10.29252/jsdp.14.3.127]

هلندی». لذا نیاز بود که زبان و ملیت در یک دسته‌بندی قرار بگیرند.

- واحد اسمی بازه اضافه شده است. درواقع تاریخ به دو واحد اسمی تاریخ و بازه زمانی تفکیک شده است.
- واحدهای قانون، اثر هنری و محصولات حذف شده‌اند. به‌دلیل اینکه رخداد این واحدهای اسمی کم است.

درنتیجه این تغییرات در پیکره اَعلام، سیزده واحد اسمی معرفی شده است. واحدهای اسمی و تعریف هر یک از آنها در جدول (۲) توضیح داده شده است. نکته مهم در برچسب‌گذاری این پیکره این است که واحدهای اسمی قابل تجزیه نیستند و همچنین واحدهای اسمی تودرتو در این پیکره وجود ندارند. در برچسب‌گذاری شخص، شاخص‌های آن یعنی کلماتی مانند: «آقا»، «خانم»، «دکتر»، «جناب»، «شهر»، «کشور» و امثال اینها داخل واحد اسمی در نظر گرفته نمی‌شوند. همچنین صفاتی که به یک واحد اسمی اضافه می‌شوند داخل واحد اسمی محسوب نمی‌شوند. در واقع یک واحد اسمی به‌صورت نام متعارفی که به آن معروف است در نظر گرفته می‌شود. به‌عنوان مثال در عبارت «کلانشهر تهران»، تهران برچسب مکان می‌گیرد و کلمه کلانشهر داخل واحد اسمی نیست.

تفاوت مهمی میان سازمان‌ها و گروه‌های سیاسی و مذهبی وجود دارد. مواردی که برچسب سازمان می‌گیرند، سازمان‌ها و شرکت‌های رسمی هستند؛ اما گروه‌های سیاسی و مذهبی برچسب LNORP می‌گیرند.

(جدول-۲): برچسب‌های موجود در پیکره فارسی اعلام
(Table-2): Different named entities in A'laam corpus.

ردیف	واحدهای اسمی در پیکره اعلام	معادل در پیکره Ontonotes	تعریف
۱	شخص	شخص	نام و نام خانوادگی اشخاص و شخصیت‌های غیرحقیقی (افسانه‌ای) مثال: «رستم دستان»، «عبید زاکانی»
۲	ال‌نورپ	ترکیب زبان[1] و نورپ	مخفف چهار کلمه: ملیت، دین، گروه سیاسی و زبان
۳	سازمان	سازمان	شرکت‌ها، ادارات دولتی، نشریه‌ها و ... مثال: «روزنامه اطلاعات»، «اداره آموزش و پرورش»
۴	مکان	ترکیب مکان و مکان سیاسی و امکانات	نام مکان‌های سیاسی و جغرافیایی مثال: «زعفرانیه»،«تبریز»، «کارون»
۵	رخداد	رخداد	رخدادهای مهم تاریخی: نام جنگ‌ها، انقلاب‌ها، رخدادهای ورزشی و ... مثال: «انقلاب اسلامی ایران»، «جام جهانی فوتبال»
۶	تاریخ	تاریخ	بیان کلی و جزئی از تاریخ: مثال:« ۲۲ بهمن ۱۳۵۷»
۷	بازه	بخشی از تعریف تاریخ	زمان سپری شده یا بازه زمانی مثال: «ده سال گذشته»
۸	زمان	زمان	زمان‌های کوتاهتر از یک روز مثال: «ساعت ۶ صبح»
۹	درصد	درصد	درصد شامل ٪ مثال: «۱۷٪»، «سی و هشت درصد»
۱۰	پول	پول	عبارات پولی مثال: «۲۰۰ تومان»
۱۱	اندازه	کمیت	اندازه‌گیری‌های استاندارد مانند: سن، مساحت، فاصله، انرژی، سرعت، دما، حجم، وزن و سایر عبارت و ساختارهای بیانگر اندازه مثال: «۴۷ متر»، «صفر درجه سانتیگراد»
۱۲	عدد اصلی	عدد اصلی	شمارش عددی یا کمیت برخی از اشیاء: (مانند کسری از چیزی، اعداد اعشاری و صحیح) مثال:«سه»، «۷/۱»
۱۳	عدد ترتیبی	عدد ترتیبی	اعداد ترتیبی مثال: «نخستین»، «دوم»
			مثال: «ایرانی»، «جمهوری‌خواه»

[1] language

[DOI: 10.29252/jsdp.14.3.127]

۲-۳- انتخاب دادگان خام مناسب

پیکره متنی فارسی[1] یک پیکره استاندارد در زبان فارسی است که در بیش‌تر پژوهش‌های زبانی این حوزه مورد استفاده قرار گرفته است. این پیکره شامل بیش از یکصد میلیون کلمه فارسی است که توسط پژوهشگاه توسعه فناوری‌های پیشرفته خواجه‌نصیرالدین طوسی[2] تهیه و جمع‌آوری شده است. تمام این داده‌ها که از منابع متنوع نوشتاری و گفتاری انتخاب شده‌اند، تقطیع و قطعات به‌صورت دستی تصحیح شده‌اند. بخشی از این پیکره که شامل هشت میلیون کلمه است علاوه بر تقطیع و تصحیح دارای برچسب اجزای کلام (POS) در سطح کلمه است [7].

یکی از مراحل مهم در تولید پیکره‌های متنی انتخاب دادگان خام اولیه است. این دادگان باید با اهدافی که از تهیه پیکره درنظر گرفته شده است، هم‌خوانی داشته باشد. در تولید پیکره واحدهای اسمی اعلام، از داده‌های موجود در بخش برچسب‌گذاری‌شده پیکره متنی فارسی (پیکره هشت میلیون کلمه‌ای) استفاده شده است تا از اطلاعات اجزای کلام[3] دقیق کلمات نیز استفاده شود. تعداد ۸۴۰۰ جمله به‌صورت تصادفی از این دادگان انتخاب شده است. انتخاب تصادفی موجب این می‌شود که در حجم محدود داده، کلمات یکتا و واحدهای اسمی متنوع‌تر و بیشتری موجود باشند. همچنین مدل تعلیم‌یافته روی پیکره‌ای با این ویژگی، از قدرت تعمیم‌پذیری بالاتری برخوردار خواهد بود. جملات انتخاب‌شده شامل بیش از ۲۵۰ هزار کلمه هستند؛ سپس جملات انتخاب‌شده مجدد مورد بازبینی قرار گرفته و یکسان‌سازی‌هایی در آنها صورت گرفته است.

۳-۳- برچسب‌گذاری دادگان و تبدیل به قالب استاندارد

در مرحله سوم تهیه پیکره، جملات انتخاب‌شده بر مبنای دستورالعمل شرح داده‌شده در بخش ۳ به‌صورت دستی توسط افراد آموزش‌دیده برچسب‌گذاری و پس از تهیه دادگان و برچسب‌گذاری دستی آنها، دادگان به قالب استانداردِ IOB تبدیل شدند. در این قالب به هر کلمه موجود در متن یکی از سه برچسب B[4]، I[5] و یا O[6] اختصاص داده می‌شود. این برچسب‌ها نمایش‌دهنده جایگاه کلمه در واحدهای اسمی هستند. کلمه آغازگر یک واحد اسمی، برچسب B، سایر کلمات واحدهای اسمی برچسب I و در نهایت باقی‌مانده کلمات برچسب O دارند. ۱۳ برچسب واحدهای اسمی در پیکره به‌کار رفته که پس از تبدیل به قالب IOB تعداد این برچسب‌ها به ۲۷ برچسب رسیده است. در جدول (۳) برچسب‌های ریز واحدهای اسمی پیکره اعلام (همراه با برچسب‌های I و B) نمایش داده شده است.

در جدول (۴)، مشخصات مربوط به هر برچسب و تعداد رخداد برچسب در کل پیکره اعلام مشاهده می‌شود. ستون نخست این جدول نام واحدهاست. در ستون دوم تعداد رخداد هر برچسب مشاهده می‌شود. این تعداد برابر با تعداد برچسب‌های B است. در ستون سوم تعداد رخداد برچسب‌های I مشاهده می‌شود. در ستون چهارم مجموع ستون دوم و سوم محاسبه شده و این مقدار برابر است با تعداد قطعاتی که هر برچسب به آنها اختصاص داده شده است. تفاوت میان مقدار ستون دوم و چهارم در مثال زیر مشخص شده است:

در عبارت «آقای دکتر حسن روحانی»، تعداد واحد شخص یک است، ولی تعداد قطعاتی که واحد اسمی شخص به آنها اختصاص داده شده دو است. با توجه به این مقادیر مشاهده می‌شود که از کل حدود ۲۵۰ هزار کلمه انتخاب شده، ۳۰۲۸۱ قطعه برچسب واحدهای اسمی دارند که این تعداد شامل ۱۸۶۶۴ واحد اسمی می‌شود.

ستون پنجم نشانگر درصد هر یک از واحدهای اسمی نسبت به کل واحدهاست. برای مثال چهارده درصد از کل واحدهای اسمی برچسب‌گذاری‌شده «سازمان» بوده است. همان‌طورکه مشاهده می‌شود، بیشترین درصد واحدهای اسمی اختصاص به سه واحد اصلی دارد که در بیش‌تر رقابت‌ها و پیکره‌های واحدهای اسمی تعریف شده‌اند: واحدهای «مکان»، «شخص» و «سازمان». همچنین «عدد اصلی» نیز درصد بالایی را به خود اختصاص داده است. این مساله با توجه به تعریف آن طبیعی به نظر می رسد.

ستون ششم نشان‌دهنده درصد قطعاتی است که یک برچسب خاص به آنها اطلاق شده است. در این پیکره ۲/۶۴ درصد از کل قطعات برچسب «سازمان» داشته‌اند. مشاهده

[1] Farsi Text Corpus

[2] این پژوهشکده از سال ۱۳۹۲ با مجوز رسمی وزارت علوم، تحقیقات و فناوری به پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی ارتقا یافته است.

[3] Part of speech (POS)

[4] Begin

[5] Inside

[6] Outside

[DOI: 10.29252/jsdp.14.3.127]

می‌شود که تنها دوازده درصد از پیکره شامل قطعات است که برچسب واحدهای اسمی گرفته‌اند. درنهایت در ستون آخر میانگین تعداد قطعات در هر واحد اسمی مشاهده می‌شود.

بیشترین عدد این ستون عدد ۳/۴۵ کلمه به‌ازای هر برچسب «رخداد» است و کمترین آن ۱/۰۳ کلمه به‌ازای هر برچسب «عدد ترتیبی» است. از روی این مقادیر می‌توان به‌طورتقریبی ماهیت تعریف هر واحد اسمی را بررسی کرد.

در ادامه و در جدول (۵) نمونه‌ای از یک عبارت برچسب‌خورده موجود در پیکره تهیه‌شده مشاهده می‌شود.

(جدول-۳): برچسب‌های ریز واحدهای اسمی و شمارش هریک در پیکره فارسی اَعلام.
(Table-3): The frequency for different fine-grained named entities in A'laam Corpus.

ردیف	برچسب	شمارش در پیکره	ردیف	برچسب	شمارش در پیکره
۱	B-سازمان	2632	۱٤	I-سازمان	4019
۲	B-مکان	4904	۱٥	I-مکان	782
۳	B-شخص	2863	۱٦	I-شخص	1468
٤	B-اصلی	3046	۱۷	I-اصلی	367
٥	B-تاریخ	1328	۱۸	I-تاریخ	1884
٦	B-LNORP	1413	۱۹	I-LNORP	190
۷	B-رخداد	431	۲۰	I-رخداد	1060
۸	B-بازه	345	۲۱	I-بازه	458
۹	B-اندازه	388	۲۲	I-اندازه	388
۱۰	B-پول	234	۲۳	I-پول	518
۱۱	B-ترتیبی	677	۲٤	I-ترتیبی	25
۱۲	B-درصد	200	۲٥	I-درصد	262
۱۳	B-زمان	183	۲٦	I-زمان	216

(جدول-۴): مشخصات مربوط به برچسب‌های موجود در پیکره فارسی اَعلام
(Table-4): The frequency rate of different named entities in A'laam Corpus.

ردیف	برچسب واحد اسمی	تعداد B	تعداد I	مجموع	درصد نسبت کلمات دارای برچسب بر کل کلمات پیکره	درصد نسبت تعداد واحدها به کل واحدها	میانگین تعداد کلمه در هر واحد اسمی
۱	سازمان	2632	4019	6651	2.64	14.1	2.52
۲	مکان	4904	782	5686	2.25	26.3	1.15
۳	شخص	2863	1468	4331	1.72	15.4	1.51
۴	اصلی	3046	367	3413	1.35	16.3	1.12
۵	تاریخ	1328	1884	3212	1.27	7.1	2.41
۶	ال‌نورپ	1413	190	1603	0.63	7.6	1.13
۷	رخداد	431	1060	1491	0.59	2.3	3.45
۸	بازه	345	458	803	0.31	1.9	2.32
۹	اندازه	388	388	776	0.31	2.1	2.00
۱۰	پول	234	518	752	0.29	1.2	3.21
۱۱	ترتیبی	677	25	702	0.27	3.6	1.03
۱۲	درصد	200	262	462	0.18	1.1	2.31
۱۳	زمان	183	216	399	0.15	1.0	2.18
---	مجموع	18664	11637	30281	12.03	100	1.62

[DOI: 10.29252/jsdp.14.3.127]

(جدول-۵): نمونه‌ای از یک جمله در پیکره اعلام.
(table-5): A sample sentence of A'laam Corpus.

شماره قطعه	کلمه	برچسب POS	برچسب NE
1	محمد	N	B-شخص
2	نشاسته‌ریز	AJ	I-شخص
3	،	PUNC	O
4	مدیرکل	N	O
5	فنی	AJ	B-سازمان
6	و	CONJ	I-سازمان
7	حرفه‌ای	AJ	I-سازمان

۴- سامانه تشخیص واحدهای اسمی

در این بخش از پیکره متنی تولیدشده (پیکره اَعلام) در ساخت یک سامانه تشخیص واحدهای اسمی فارسی استفاده و برای تعلیم مدل از روش‌های مبتنی بر یادگیری استفاده شده است. در این حوزه یکی از بهترین عملکردها در بین روش‌های یادگیری ماشینی به روش سی‌آراف[1] اختصاص دارد [22]. الگوریتم میدان تصادفی شرطی مبتنی بر نظریه احتمال شرطی و نظریه گراف است. در سی‌آراف طبقه‌بندی با توجه به بافت انجام می‌شود و اطلاعات داده‌های همسایه مد نظر قرار می‌گیرد.

اگر O={o1,o2,...ot} دنباله مجموعه مشاهدات (یعنی در اینجا مجموعه کلمه‌های موجود در جمله) باشند و s={s1,s2,...,st} مجموعه برچسب‌های منتسب به جمله ورودی باشد؛ احتمال شرطی دنباله برچسب‌ها به شرط دنباله ورودی، به‌وسیله رابطه (۱) تعریف می‌شود:

$$P(s \mid o) = \frac{1}{Z_0} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k (s_{t-1}, s_t, o, t)\right) \qquad (۱)$$

صورت این کسر مجموع مقادیر توابع ویژگی است که برای متغیر تصادفی s برحسب مشاهدات o به‌دست می‌آید و Z0 برای نرمال‌سازی به‌کار می‌رود. f_k یک تابع ویژگی[2] است که متغیرهای آن یک مشاهده، جایگاه آن در دنباله مشاهدات و حالت متناظر با آن مشاهده است. خروجی هر تابع ویژگی یک عدد است و به هر تابع ویژگی k در هنگام آموزش یک وزن (λk) اختصاص داده می‌شود.

برخلاف بسیاری از روش‌ها، مدل تصادفی شرطی تنها به یک نمونه بسنده نمی‌کند و برای برچسب‌گذاری، بافتی را که نمونه در آن قرار دارد، مورد توجه قرار می‌دهد. در تولید این سامانه از روش سی.آر.اِف و ابزار متن‌باز[3] PocketCRF[4] استفاده شده است.

۴-۱- آموزش مدل تشخیص واحدهای اسمی

برای آموزش مدل تشخیص واحدهای اسمی از دو ویژگی اصلی کلمه و برچسب اجزای کلام آن استفاده شده است. برچسب اجزای کلام مورد استفاده، برچسب‌های دقیق اجزای کلام کلمات است (به‌دلیل اینکه دادگان پیکره اعلام از پیکره متنی زبان فارسی انتخاب شده‌اند، در اینجا استفاده از برچسب دقیق اجزای کلام میسر بود).

این ویژگی‌ها به‌صورت توابع ویژگی در مرحلهٔ آموزش به ابزار PocketCRF داده می‌شوند. این توابع ویژگی شامل ترکیب برچسب گروه‌های نحوی با پنجره پنج‌تایی از کلمات، پنجره پنج‌تایی از برچسب اجزای کلمات، بایگرم کلمات، بایگرم برچسب اجزای کلمات و سه‌تایی‌های برچسب اجزای کلام است. در جدول (۶) توابع ویژگی مورد استفاده ذکر شده‌اند.

(جدول-۶): ویژگی‌های به‌کاررفته در سامانه برچسب‌گذار
(Table-6): Features used in NER system.

ردیف	ویژگی‌ها
1	پنجره پنج‌تایی از کلمات
2	پنجره پنج‌تایی از برچسب اجزای کلمات
3	پنجره سه‌تایی از بایگرم (دوتایی) کلمات
4	پنجره پنج‌تایی از بایگرم (دوتایی) برچسب اجزای کلمات
5	پنجره پنج‌تایی از سه‌تایی‌های برچسب اجزای کلام

۴-۲- آزمایش مدل تشخیص واحدهای اسمی

آزمایش‌ها در دو مرحله صورت گرفت. مرحله نخست شامل مجموعه ۸۴ هزار کلمه‌ای (یک سوم کل پیکره) و مرحله دوم شامل مجموعه ۲۲۵ هزار کلمه‌ای (۹۰ درصد کل پیکره) بود. ۱۰ درصد از پیکره نیز برای ارزیابی انتخاب شد که این مجموعه برای هر دو مرحله آزمایش یکسان بود. در هر مرحله یک مدل تعلیم داده و بر روی دادگان آزمایش ارزیابی شد. نتایج مقایسه‌ای این دو مدل در جدول (۷) برحسب معیارهای دقت[5] و فراخوانی[6] نمایش داده شده است.

[1] Conditional Random Fields (CRF)
[2] Feature function
[3] Open source
[4] Sourceforge.net/projects/pocket-crf-1
[5] precision
[6] recall

[DOI: 10.29252/jsdp.14.3.127]

(جدول-۷): دقت و فراخوانی آزمایش دو مدل تولیدشده با حجم دادگان آموزش متفاوت

(Table-7): Results obtained from our model with different data size.

	درصد دقت	درصد فراخوانی
دادگان مرحله اول (۸۴ هزار کلمه)	87.51	55.58
دادگان مرحله دوم (۲۲۵ هزار کلمه)	92.94	78.48

نتایج جدول (۷) نشان‌دهنده آن است که افزایش حجم دادگان تعلیم توانسته است که کارایی سامانه تشخیص واحدهای اسمی مبتنی بر یادگیری ماشین را به‌صورت قابل‌ملاحظه‌ای افزایش دهد. همچنین با مقایسه با سایر کارهای انجام‌گرفته در زبان فارسی مشاهده می‌شود که استفاده از روش‌های یادگیری ماشین نتایج تشخیص واحدهای اسمی را نسبت به روش‌های قاعده‌مند بهتر می‌کند. در پژوهش‌های پیشین در زبان فارسی [2] با استفاده از روش‌های مبتنی بر قانون به دقت حدود ۷۲ درصد و فراخوانی حدود ۷۶ درصد دست یافته‌اند. همان‌طور که انتظار می‌رود، دستیابی به مقادیر فراخوانی بالا در روش‌های مبتنی بر قانون و فرهنگ لغات دور از دسترس نیست؛ اما در روش‌های مبتنی بر یادگیری ماشینی و مبتنی بر پیکره به دقت بالاتری در کارایی سامانه NER می‌توان دست یافت.

۵- جمع‌بندی و کارهای آینده

همان‌طورکه اشاره شد، بخش قابل توجهی از روش‌های موفق در برچسب‌دهی خودکار واحدهای اسمی، از روش‌های یادگیری باسرپرست استفاده می‌کنند. این امر ضرورتِ داشتن مجموعه داده با برچسب‌های NER را نشان می‌دهد.

در این مقاله، نحوه ساخت یک پیکره استاندارد فارسی حاوی برچسب واحدهای اسمی (پیکره آعلام) بررسی شده است. جملات این پیکره از مجموعه پیکره متنی زبان فارسی انتخاب شده‌اند. این مجموعه پالایش‌شده تنوعات مختلف زبانی را در بر دارد و شامل برچسب اجزای کلام است. از پیکره تهیه‌شده در تولید سامانه تشخیص واحدهای اسمی استفاده شد و این سامانه مورد ارزیابی قرار گرفت.

در کارهای آینده با استفاده از ابزار برچسب‌گذار اجزای کلام – که بر روی بخش برچسب دار پیکره متنی فارسی آموزش یافته است - تأثیر برچسب‌گذاری دقیق بر نتایج تشخیص واحدهای اسمی بررسی می‌شود. همچنین می‌توان با روش‌های نیمه‌خودکار حجم داده‌های برچسب‌خورده را افزایش داد. انتظار می‌رود با افزایش حجم دادگان دقت سامانه تعلیم‌یافته، افزایش یابد. به‌علاوه عملکرد مدل‌های تولیدشده با این پیکره را در سامانه‌های استخراج اطلاعات و پرسش و پاسخ مورد ارزیابی می‌توان قرار داد.

۶- مراجع 6-References

[۱] س. ع. اصفهانی، س. راحتی قوچانی و ن. جهانگیری. «سیستم شناسایی و طبقه‌بندی اسامی در متون فارسی»، *پردازش علائم و داده‌ها*، دوره ۱۳، شماره ۱ (پیاپی ۱۳) ; صص. ۷۷-۸۸. ۱۳۸۹.

[1] S. A. Esfahani, S. R. Ghuchani, and N. Jahangirim, "Recognition system of names in Persian texts," *JSDP*, vol. 13, no. 1, pp. 77-88, 1389.

[۲] پ. سادات‌مرتضوی و م. شمس‌فرد. «شناسایی واحدهای اسمی در متون فارسی.» پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر، تهران. ۱۳۸۸.

[2] P. S. Mortazavi and M. Shamsfard, "Recognition of named entities in Persian texts," in *15-th annual conference of computer society of Iran*, Tehran, 1388.

[۳] م. عبدوس؛ ب. مینایی بیدگلی و ح. قدمنان. «تولید پیکره واحدهای اسمی فارسی.» اولین همایش ملی زبان‌شناسی پیکره‌ای، تهران. ۱۳۹۴.

[3] M. Abdoos, B. M. Bidgoli, and H. Ghadmanan, "Production of persian named entity corpus NaExtractiing person names using name candidate injection in a coditional random filed model for Arabic language," in *the first national conference on corpus linguistics*, Tehran, 1394.

[۴] م. عسگری بیدهندی و ب. مینایی بیدگلی. «تشخیص اسامی اشخاص با استفاده از افزایش کلمه‌های نامزد اسم در میدان‌های تصادفی شرطی برای زبان عربی»، پردازش

[DOI: 10.29252/jsdp.14.3.127]

علائم و داده‌ها، دوره ۱۱ شماره ۲۱، صص ۷۳ -۸۵ . ۱۳۹۳.

[4] M. A. Bidhendi and B. M. Bidgoli, "Extracting person names using name candidate injection in a conditional random filed model for Arabic language," *JSDP*, vol. 11, no. 21, pp. 73-85, 2014.

[5] S. Armstrong-Warwick, et al., "Data in your language: the ECI multilingual corpus." In *Proceedings of the International Workshop on Sharable Natural Language Resources,* 1994.

[6] Y. Benajiba, P. Rosso, and J. BenediRuiz, "ANER-sys: An Arabic Named Entity Recognition system based on Maximum Entropy," *Computational Linguistics and Intelligent Text Processing*. pp. 143-153. 2007.

[7] M. Bijankhan, J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi, "Lessons from Building a Persian Written Corpus: Peykare," *Language Resources and Evaluation*, vol. 45, no. 2, pp. 143–164. 2011.

[8] D. M. Bikel, S. Miller, R. Schwartz, R. Weischedel, "Nymble: a High-Performance Learning Name-finder". in *Proceedings of Conference on Applied Natural Language Processing*. 1997.

[9] A. Borthwick, J. Sterling, , E. Agichtein, E, and R. Grishman. "NYU: Description of the MENE Named Entity System as used in MUC-7". in *Proceedings of the Seventh Message Understanding Conference,* 1998.

[10] W. Che, M. Wang, C. D. Manning, and T. Liu, "Named Entity Recognition with Bilingual Constraints," In *HLT-NAACL*, pp. 52-62, 2013.

[11] N. Chinchor and P. Robinson, "MUC-7 named entity task definition." in *Proceedings of the 7th Conference on Message Understanding*, 1997.

[12] N. Chinchor, et al., "1999 Named Entity Recognition Task Definition," *MITRE and SAIC,* 1999.

[13] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan. "Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* pp. 1002–1012. 2010.

[14] A. Cucchiarelli and P. Velardi, "Unsupervised named entity recognition using syntactic and semantic contextual evid-ence," *Computational Linguistics,* vol. 27, no.1, pp. 123-131, 2001.

[15] G. R. Doddington, et al., "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation," in *Proceedings of LREC,* 2004.

[16] M. El-Haj and R. Koulali, "KALIMAT a multipurpose Arabic Corpus," in *Second Workshop on Arabic Corpus Linguistics (WACL-2),* pp. 22-25, 2013.

[17] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History," in *The 16th International Conference on Computational Linguistics COLING*, 1996.

[18] W. Liao and S. Veeramachaneni, "A Simple Semi-supervised Algorithm For Named Entity Recognition". In *Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*, pp. 58–65, 2009.

[19] M. K. Khormuji and M. Bazrafkan, "Persian Named Entity Recognition based with Local Filters". *International Journal of Computer Applications*, vol. 100, no. 4, 2014.

[20] B. Magnini, M. Negri, R. Prevete and H. Tanev, "A WordNet-based approach to Named Entities recognition,". in *Proceeding SEMANET'02 Proceedings of the 2002 workshop on Building and using semantic networks,* vol. 11, pp. 1-7, 2002.

[21] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics,* vol. 19, no. 2, pp. 313-330, 1993.

[22] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature Induction and Web-enhanced Lexicons," in *Proceedings of CONLL*, pp. 188–191, 2003.

[23] H. Moradi and F. Ahmadi, "A hybrud approach for Persian Named Entity Recognition," in *7th conference on information and*

[DOI: 10.29252/jsdp.14.3.127]

knowledge Technology (IKT), Urmia, Iran. 2015.

[24] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes,* vol. 30, no. 1, pp. 3-26, 2007.

[25] M. Oudah and K. F. Shaalan. "A Pipeline Arabic Named Entity Recognition using a Hybrid Approach," in *Proceedings of CoNLL,* 2012.

[26] D. Palmer, and et al., "A statistical profile of the named entity task," in *Proceedings of the fifth conference on Applied natural language processing,* 1997.

[27] T. Poibeau, "The multilingual named entity recognition framework," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics,* 2003.

[28] S. Pradhan, and et al., "OntoNotes: A Unified Relational Semantic Representation," in *International Conference on Semantic Computing,* pp. 405–419, 2007.

[29] T. Rocktschel, M. Weidlich, and U. Leser, "ChemSpot: a hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, pp. 1633-1640, 2012.

[30] T. Rose, M. Stevenson, and M. Whitehead, "The Reuters Corpus Volume 1-from Yesterday's News to Tomorrow's Language Resources," in *Proceedings of LREC*, vol. 2, 2002.

[31] T. J. Sang and F. Erik, "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition," in *Proceedings of CoNLL*, pp. 155–158, Taipei, Taiwan, 2002.

[32] T. K. Sang, F. Erik, and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL* vol. 4, 2003.

[33] D. S. Diana, et al., "Harem: An advanced ner evaluation contest for portuguese," in *Proceedings of LREC,* 2006.

[34] S. Sekine and H. Isahara, "IREX: IR & IE Evaluation Project in Japanese," in *Proceedings of LREC*, 2000.

[35] S. Sekine and C. Nobata, "Definition, Dictionaries and tagger for extended named entity Hierarchy," in *Proceedings of conference on Language Resources and Evaluation*, 2004.

[36] S. Rahul, "Named Entity Recognition: A Literature Survey," 2014.

[37] R. Weischedel and A. Brunstein, "BBN Pronoun Coreference and Entity Type Corpus LDC2005T33," in *Web Download. Philadelphia: Linguistic Data Consortium*, 2005.

[38] F. Wu, and S. Weld, "Open information extraction using wikipedia," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 118–127, 2010.

شادی حسین‌نژاد فارغ التحصیل رشته مهندسی کامپیوتر از دانشگاه تهران در سال ۱۳۹۱ و همچنین مدرک کارشناسی ارشد را در رشته زبان‌شناسی رایانشی در سال ۱۳۹۳ از دانشگاه صنعتی شریف دریافت کرده و از سال ۱۳۹۵ دانشجوی دکترای زبان‌شناسی در دانشگاه الزهرا است. وی از سال ۱۳۹۲ در پژوهشگاه خواجه‌نصیر گروه پردازش صوت و زبان طبیعی مشغول به فعالیت است. زمینه پژوهشی مورد علاقه ایشان پردازش زبان طبیعی، زبان‌شناسی رایانشی، آهنگ و آواشناسی است.
نشانی رایانامه ایشان عبارت است از:
hosseinnejad@rcdat.ir

یاسر شکفته تحصیلات خود را در مقطع کارشناسی در دو رشته مهندسی پزشکی-بیوالکتریک و مهندسی برق-الکترونیک به‌ترتیب در سال‌های ۱۳۸۴ و ۱۳۸۵ در دانشگاه صنعتی امیرکبیر به پایان رساند. ایشان در سال‌های ۱۳۸۷ و ۱۳۹۲ مدارک کارشناسی ارشد و دکترای خود را در رشته مهندسی پزشکی (گرایش بیوالکتریک) از همان دانشگاه اخذ کرد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش سیگنال دیجیتال، پردازش صوت و زبان طبیعی است.

[DOI: 10.29252/jsdp.14.3.127]

نشانی رایانامه ایشان عبارت است از:
y_shekofteh@sbu.ac.ir

طاهره امامی آزادی فارغ التحصیل کارشناسی ارشد رشته مهندسی پزشکی گرایش بیوالکتریک از دانشگاه صنعتی امیرکبیر است. ایشان از سال ۱۳۸۶ در پژوهشگاه توسعه فناوری‌های پیشرفته خواجه‌نصیرالدین طوسی مشغول به کار است. عمده فعالیت‌های پژوهشی ایشان در حوزه‌های پردازش علائم حیاتی، پردازش گفتار و پردازش زبان طبیعی بوده است. موضوعات پژوهشی مورد علاقه ایشان عبارتند از: شناسایی الگو، یادگیری ماشینی و داده‌کاوی.
نشانی رایانامه ایشان عبارت است از:
t.emami@rctad.ir

[DOI: 10.29252/jsdp.14.3.127]