پیما: پیکرهٔ برچسب‌خوردهٔ موجودیت‌های اسمی زبان فارسی

مهساسادات شهشهانی، مهدی محسنی، آزاده شاکری* و هشام فیلی

دانشکده مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی دانشگاه تهران، تهران، ایران

چکیده

هدف در مسألهٔ تشخیص موجودیت‌های اسمی، رده‌بندی اسامی خاص متن با برچسب‌هایی همچون شخص، مکان، و سازمان است. این مسأله به‌عنوان یکی از گام‌های پیش‌پردازشی بسیاری از مسائل پردازش زبان طبیعی مطرح است. اگر چه در زبان انگلیسی پژوهش‌های زیادی در این حوزه انجام شده و سامانه‌ها به کیفیت *F1* بالای نود درصد دست یافته‌اند، در زبان فارسی به‌دلیل نبود یک مجموعه دادهٔ استاندارد، پژوهش‌های کمی در این زمینه انجام شده است. در این پژوهش به ساخت چنین مجموعه‌داده‌ای می‌پردازیم و آن را به‌صورت آزاد در اختیار پژوهش‌گران قرار می‌دهیم؛ سپس با استفاده از این مجموعه‌داده به طراحی سامانه آماری با استفاده از مدل میدان‌های تصادفی شرطی و نیز سامانه‌ای مبتنی بر شبکه‌های عصبی بازگشتی از نوع LSTM برای تشخیص موجودیت‌های اسمی می‌پردازیم. در پیکرهٔ ایجادشده هفت نوع موجودیت شخص، مکان، سازمان، زمان، تاریخ، درصد، و مقادیر پولی برچسب خورده‌اند و درنتیجه تمام ارزیابی‌های سامانهٔ طراحی‌شده بر روی این هفت برچسب انجام می‌گیرد. برای طراحی این سامانه، پس از آموزش یک سامانهٔ آماری مبتنی بر الگوریتم CRF، از خروجی این سامانه به‌عنوان یک ویژگی برای آموزش یک شبکهٔ عصبی بازگشتی LSTM دوطرفه استفاده می‌کنیم. علاوه‌بر این ویژگی، از خوشه‌بندی واژگان به روش k- means نیز بهره می‌بریم. برای این کار، شمارهٔ خوشهٔ واژگان را به‌عنوان یک ویژگی در اختیار شبکهٔ عصبی LSTM قرار می‌دهیم و به این ترتیب سامانهٔ ترکیبی نهایی ساخته می‌شود. این شیوهٔ ترکیب مدل CRF با مدل شبکهٔ عصبی و نیز استفاده از شمارهٔ خوشه برای هر واژه در روش خوشه‌بندی k-means نوآوری این پژوهش محسوب می‌شود. نتایج آزمایش‌ها نشان می‌دهد که با استفاده از مدل نهایی به *F1* برابر با ۸۷ درصد در سطح واژه و هشتاد درصد در سطح عبارت موجودیت اسمی می‌رسیم. همچنین آزمایش‌ها نشان می‌دهد که روش پیشنهادی برای استفاده از خروجی مدل CRF به‌عنوان یک ویژگی در ورودی مدل شبکهٔ عصبی باعث می‌شود که با دراختیارداشتن حجم کمتری از دادهٔ برچسب‌خورده به کیفیت قابل قبولی در تشخیص موجودیت‌های اسمی برسیم که این مسأله می‌تواند در زبان‌هایی که حجم دادهٔ برچسب‌خوردهٔ آن‌ها محدود است، مفید باشد.

واژگان کلیدی: پیکرهٔ موجودیت‌های اسمی، تشخیص موجودیت‌های اسمی، روش قاعده‌محور، روش مبتنی بر یادگیری عمیق، روش میدان‌های تصادفی شرطی

PAYMA: A Tagged Corpus of Persian Named Entities

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery* & Heshaam Faili

Department of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

Abstract

The goal in the named entity recognition task is to classify proper nouns of a piece of text into classes such as person, location, and organization. Named entity recognition is an important preprocessing step in many natural language processing tasks such as question-answering and summarization. Although many research studies have been conducted in this area in English and the state-of-the-art NER systems have reached performances of higher than 90 percent in terms of F1 measure, there are very few research studies on this

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال: ۹۶/۱۱/۳۰ ● تاریخ پذیرش: ۹۷/۱۰/۱۹ ● نوع مطالعه: کاربردی

Archive of SID

www.SID.ir

task in Persian. One of the main important reasons for this may be the lack of a standard Persian NER dataset to train and test the NER systems. In this research we create a standard tagged Persian NER dataset which will be distributed freely for research purposes. In order to construct this standard dataset, we studied the existing standard NER datasets in English and came to the conclusion that almost all of these datasets are constructed using news data. Thus we collected documents from ten news websites in Persian. In the next step, in order to provide the annotators with guidelines to tag these documents, we studied the guidelines used for constructing CoNLL and MUC English datasets and created our own guidelines considering the Persian linguistic rules. Using these guidelines, all words in documents can be labeled as person, location, organization, time, date, percent, currency, or other (words that are not in any of these 7 classes). We use IOB encoding for annotating named entities in documents, like most of the existing English NER datasets. Using this encoding, the first token of a named entity is labeled with B, and the next tokens (if exist) are labeled with I. The words that are not part of any named entity are labeled with O. The constructed corpus, named PAYMA, consists of 709 documents and includes 302530 tokens. 41148 tokens out of these tokens are labeled as named entities and the others are labeled as O. In order to determine the inter-annotator agreement, 160 documents were labeled by a second annotator. Kappa statistic was estimated as 95% using words that are labeled as named entities. After creating the dataset, we used the dataset to design a hybrid system for named entity recognition. We trained a statistical system based on the CRF algorithm, and used its output as a feature to train a bidirectional LSTM recurrent neural network. Moreover, we used the k-means word clustering method to cluster the words and fed the cluster number of each word to the LSTM neural network. This form of combining CRF with neural networks and using the cluster number for each word is the novelty of this research work. Experimental results show that the final model can reach an *F*1 score of 87% at word-level and 80% at phrase level.

Keywords: Persian named entity corpus, named entity recognition, rule-based model, deep-learning based model, conditional random field's method

۱- مقدّمه

موجودیت‌های اسمی[1]، واحدهای اسمی معنادار متن هستند که مفهوم و منظور اصلی یک متن را مشخص می‌کنند. از مهم‌ترین انواع موجودیت‌های اسمی می‌توان به اسامی افراد، سازمان‌ها، مکان‌ها، پول، درصد، تاریخ، و زمان اشاره کرد.[23] سامانه‌های تشخیص موجودیت اسمی، به شناسایی موجودیت‌های اسمی یک متن می‌پردازند و آن‌ها را در یکی از انواع مشخص رده‌بندی می‌کنند. تشخیص موجودیت‌های اسمی کاربردهای فراوانی در سامانه‌های استخراج اطلاعات، سامانه‌های پرسش و پاسخ [22]، رده‌بندی متون [16] و بهینه‌سازی جستجو [24] دارد.

اگرچه پژوهش‌هایی که در زمینهٔ تشخیص موجودیت‌های اسمی انجام‌شده در زبان‌های مختلف وسیع بوده و تاحدودی نتایج قابل قبولی حاصل شده است، فعالیت‌های انجام‌شده در این زمینه در زبان فارسی چندان گسترده نبوده و نتایج ارایه‌شده پاسخ‌گوی نیازها نیست. در زبان فارسی، مجموعه دادهٔ استانداردی برای آموزش و ارزیابی سامانه‌های تشخیص موجودیت اسمی وجود ندارد که همین مسأله انجام پژوهش در این زمینه را بسیار دشوار کرده است. تنها مجموعه دادهٔ موجود برای زبان فارسی در [26] ارائه شده است که برچسب‌های آن شامل شخص، مکان، سازمان، امکانات، رویداد، و سایر موجودیت‌هاست. این مجموعه داده با استفاده از بخشی از پیکرهٔ بی‌جن‌خان ساخته شده است و تاریخ اسناد خبری مورد استفادهٔ آن به‌روز و جدید نیست. در این پژوهش به ساخت یک مجموعه دادهٔ استاندارد با حجم بیشتر و متون جدیدتر می‌پردازیم؛ سپس از این مجموعه‌داده برای آموزش و ارزیابی سامانه‌ای جهت شناسایی و رده‌بندی موجودیت‌های اسمی در متون فارسی استفاده می‌کنیم.

در ادامهٔ مقاله در بخش دوم به بررسی و مرور اجمالی سامانه‌های تشخیص موجودیت اسمی موجود برای زبان‌های انگلیسی و فارسی می‌پردازیم؛ سپس در بخش سوم چگونگی ساخت مجموعه دادهٔ استاندارد موجودیت‌های اسمی را به‌طور مشروح توضیح می‌دهیم. در بخش چهارم به شرح سامانهٔ طراحی‌شده برای زبان فارسی در این پژوهش می‌پردازیم و در بخش پنجم آزمایش‌های انجام‌شده با استفاده از سامانهٔ طراحی‌شده و مجموعه دادهٔ فراهم‌شده را شرح می‌دهیم و درنهایت در بخش ششم به جمع‌بندی و بیان کارهای آینده می‌پردازیم.

۲- کارهای پیشین

با توجه به آن که تلاش می‌کنیم مجموعه‌دادهٔ استانداردی برای زبان فارسی ایجاد کنیم، ابتدا مروری گذرا بر روی مجموعه‌داده‌های موجود زبان انگلیسی خواهیم داشت و از آن‌جا که می‌خواهیم با بررسی کارهای پیشین انجام‌شده در

[1] named entities

Archive of SID
www.SID.ir

زبان فارسی و با استفاده از روش‌های موجود در زبان انگلیسی به طراحی سامانه‌ای کارآمد برای تشخیص موجودیت‌های اسمی زبان فارسی بپردازیم، در ادامه به مرور اجمالی پژوهش‌های انجام‌شده در زبان انگلیسی و سپس به بررسی پژوهش‌های زبان فارسی می‌پردازیم. همچنین از آن جا که انتخاب ویژگی‌ها به‌اندازهٔ انتخاب روش در این مسأله اهمیت دارد، هر بخش را به دو زیربخش تقسیم می‌کنیم. در زیربخش نخست به مرور روش‌های ارائه‌شده و در زیربخش دوم به معرفی ویژگی‌های ارائه‌شده برای حل مسئلهٔ تشخیص موجودیت‌های اسمی می‌پردازیم.

۲-۱- مجموعه‌داده‌ها

در [25] نشان داده شده که سامانهٔ تشخیص موجودیت‌های اسمی به حوزه و سیاق متن حساس است و سامانه‌ای که برای یک حوزه (مانند حوزهٔ پزشکی) و یا یک سیاق (مانند متون رسمی) آموزش داده شده، در سایر حوزه‌ها (مانند اجتماعی یا سیاسی) و یا سایر سیاق‌ها (مانند وبلاگ یا میکروبلاگ) بیست تا چهل درصد کاهش کیفیت خواهد داشت. بنابراین به نظر می‌رسد اینکه سامانهٔ جامعی آموزش داده شود که روی تمام حوزه‌ها و یا سیاق‌ها عملکرد خوبی داشته باشد، عملی نیست. با این وجود انتخاب متون خبری از این جهت که دربرگیرندهٔ حوزه‌های متعددی از متون در سیاق رسمی است، به‌نظر انتخاب خوبی می‌رسد و بررسی مجموعه داده‌های استاندارد موجود در زبان انگلیسی هم بر این ادعا صحه می‌گذارد؛ چون که در تمامی مجموعه‌داده‌های استاندارد بررسی‌شده، مجموعهٔ متون از متون خبری انتخاب شده‌اند.

در مجموعه‌دادهٔ CoNLL2003[1] [30] که برای دو زبان انگلیسی و آلمانی ایجاد شده است، از مجموعه متون خبرگزاری رویترز برای زبان انگلیسی و مجموعه متون خبری روزنامه‌ی Frankfurter Rundschau برای زبان آلمانی استفاده شده است. این مجموعه‌داده از ۱۴۹۸۷ جمله که معادل ۲۰۳۶۲۱ واژه است تشکیل شده است.

در مجموعه‌دادهٔ MUC6[2] [29] از مجموعهٔ متون خبری وال استریت ژورنال و در مجموعه‌دادهٔ MUC7[3] [9] از مجموعه خبرهای موجود در نیویورک‌تایمز استفاده شده است.

همچنین در مجموعه‌دادهٔ ACE2[4] از متون خبری واشنگتن‌پست، VOA و ...، در مجموعه‌دادهٔ ACE2003[5] از اخبار پخش‌شده، newswireها و... و در مجموعه‌دادهٔ ACE2004[6] از متون اخبار پخش‌شدهٔ رادیویی استفاده شده است.

۲-۲- پژوهش‌ها در زبان انگلیسی

۲-۲-۱- روش‌ها

مسأله تشخیص موجودیت‌های اسمی ذاتاً از جنس مسائل برچسب‌گذاری توالی[7] است. منظور از برچسب‌گذاری توالی، دراختیارداشتن دنباله‌ای از اشیاست که با توجه به ترتیب آن‌ها قرار است به تک‌تک اشیا برچسبی تعلق گیرد. در مسألهٔ تشخیص موجودیت‌های اسمی این اشیا، واژگان و برچسب‌ها، انواع موجودیت‌های اسمی هستند. به همین دلیل از ابتدای پژوهش در این مسأله، از روش‌های معروف و فراگیر برچسب‌گذاری توالی استفاده شده است. در یک تلاش ابتدایی در [6] سامانه‌ای با استفاده از زنجیرهٔ پنهان مارکف[8] ارائه شد. در این سامانه از روش زنجیرهٔ پنهان مارکُف مرتبه نخست که از معروف‌ترین و متداول‌ترین روش‌های معرفی‌شده برای برچسب‌گذاری دنباله‌هاست، استفاده شده است. در مدل زنجیرهٔ پنهان مارکف، چند متغیر مشاهده‌شده وجود دارد که با توجه به آن‌ها و با دانستن احتمال انتقال بین حالت‌ها به‌دنبال دنباله‌ای از حالت‌های پنهان با بیشترین احتمال هستیم. در مسألهٔ تشخیص موجودیت‌های اسمی، متغیرهای مشاهده‌شده، واژگان و حالت‌های پنهان برچسب‌ها هستند.

در [7] از مدل بیشینهٔ بی‌نظمی مارکف[9] استفاده شده است. در این روش (MEMM) بر خلاف زنجیرهٔ پنهان مارکف که تنها از برچسب متغیر قبلی برای تعیین برچسب متغیر کنونی استفاده می‌کند، امکان استفاده از ویژگی‌های دیگر نیز وجود دارد.

در [19] از روش میدان‌های تصادفی شرطی برای تشخیص موجودیت‌های اسمی استفاده شده است. در این روش، تمامی برچسب‌های خروجی با درنظرگرفتن کل ورودی تولید می‌شود. در بیش‌تر مقالاتی که پس از این مقاله برای حل مسألهٔ تشخیص موجودیت‌های اسمی ارائه شده‌اند و مورد مطالعه قرار گرفته‌اند، از روش CRF[10] به‌عنوان روش اصلی و پایه‌ای استفاده شده است.

در هر سه روش مطرح‌شده از فرض ساده‌سازی مارکف برای تصمیم‌گیری در مورد برچسب‌ها استفاده شده است.

[1] http://www.clips.uantwerpen.be/conll2003/ner/
[2] https://catalog.ldc.upenn.edu/LDC2003T13
[3] https://catalog.ldc.upenn.edu/LDC2001T02
[4] https://catalog.ldc.upenn.edu/LDC2003T11
[5] https://catalog.ldc.upenn.edu/LDC2004T09
[6] https://catalog.ldc.upenn.edu/LDC2005T09
[7] sequence
[8] Hidden Markov Model (HMM)
[9] Maximum Entropy Markov Model
[10] Conditional Random Fields

Archive of SID
www.SID.ir

فرض مارکف بیان می‌کند که برچسب هر متغیر تنها به چند متغیر در یک پنجرهٔ محدود اطراف آن بستگی دارد. اگر از این فرض استفاده نشود، محاسبات و تعداد متغیرها به‌قدری زیاد می‌شود که راه بهینه‌ای برای حل مسأله وجود نخواهد داشت. در [12] از روش CRF برای حل مسألهٔ تشخیص موجودیت‌های اسمی استفاده شده است؛ ولی برای آن که بتوان از درجه‌های بالاتر وابستگی بین متغیرها استفاده کرد، از روش نمونه‌برداری گیبس استفاده کرده است.

بر خلاف بیش‌تر مقالات که از روش میدان‌های تصادفی شرطی برای تشخیص موجودیت‌های اسمی استفاده می‌کنند، در [28] از روش رده‌بندی بردار پشتیبان (SVC[1]) خطی به این منظور استفاده شده است.

در سال‌های اخیر با ظهور شبکه‌های عصبی عمیق، تحول شگرفی در روش‌های ارائه‌شده برای مسائل پردازش زبان طبیعی رخ داده است. برای مسألهٔ تشخیص موجودیت‌های اسمی نیز پژوهش‌هایی بر این اساس صورت گرفته است. در [11] از نمایش طیفی نویسه‌ها[2] که از طریق اعمال یک شبکهٔ عصبی هم‌گشتی[3] به‌دست‌آمده، استفاده شده است. این سامانه مستقل از زبان است و نیازی به تعریف ویژگی ندارد و ویژگی‌ها را به‌صورت خودکار یادگیری می‌کند.

در [10] از یک سامانهٔ ترکیبی از شبکه‌های عصبی هم‌گشتی برای یادگیری نمایش توزیع‌شدهٔ نویسه‌ها و شبکه‌های عصبی بازگشتی[4] از نوع LSTM برای حفظ وابستگی‌های طولانی بین واژگان استفاده شده است.

در [18] روشی ترکیبی از شبکه‌های عصبی بازگشتی و میدان‌های تصادفی شرطی استفاده شده است. با استفاده از یک شبکهٔ عصبی بازگشتی از نوع LSTM، زمینهٔ سمت چپ و راست واژگان مدل می‌شود؛ سپس از ترکیب آن‌ها زمینهٔ کلی واژه به‌دست می‌آید. درنهایت به جای نمایش اولیهٔ عبارات، الگوریتم میدان‌های تصادفی شرطی روی نمایش حاصل از این ترکیب اعمال می‌شود.

۲-۲-۲- ویژگی‌ها

در [15] مجموعهٔ خوب و جامعی از ویژگی‌ها پیشنهاد شده است. ویژگی‌های اصلی پیشنهادشده را می‌توان به دو دستهٔ پایه و ریشه‌یابی‌شده تقسیم کرد که ویژگی‌های ریشه‌یابی‌شده همان ویژگی‌های پایه هستند که به جای خود واژگان به ریشهٔ آن‌ها اعمال می‌شود.

این مقاله ویژگی‌های پایه را به شش دستهٔ واژه (نمایش برداری واژهٔ جاری و دو واژه قبل و بعد از آن در پنج بردار جداگانه)، کیف واژگان (نمایش برداری واژهٔ جاری و دو واژه قبل و بعد از آن در یک بردار بدون نگه‌داشتن اطلاعات مربوط به مکان قرارگیری)، ان-گرام (مشابه ویژگی واژه ولی به‌جای یونیگرام، بایگرام جایگزین می‌شود)، ویژگی‌های املایی (بزرگ‌بودن تمام حروف واژه، بزرگ بودن تنها حرف نخست واژه، ترکیبی از حروف بزرگ و کوچک، شامل‌بودن عدد، شامل‌بودن علائمی مانند '، _، . و ... ، مخففبودن و ...)، الگوهای املایی (مشابه ویژگی واژه، ولی با این تفاوت که واژگان یکسان‌سازی می‌شوند. به این ترتیب که تمام حروف بزرگ به 'A'، حروف کوچک به 'a'، ارقام به '1'، فاصله‌ها به یک ' ' و سایر علائم به '-' تغییر یابند.) و وندها (پیشوند و پسوند واژه شامل ۲-۴ حرف نخست یا آخر واژه) تقسیم کرده است.

در [23] ویژگی‌ها به سه دسته تقسیم شده‌اند که در ادامه به توضیح آن‌ها می‌پردازیم:

۱- ویژگی‌های سطح واژه: بر اساس نویسه‌های واژه تعریف می‌شوند و شامل ویژگی‌های املایی که گفته شد نیز می‌شوند. از جملهٔ این ویژگی‌ها می‌تواند به بزرگ و کوچکی حروف، علائم نگارشی، ساخت‌واژه (پیشوند، پسوند و ریشهٔ واژگان)، برچسب ادات سخن، و طول واژه اشاره کرد.

۲- ویژگی‌های مبتنی بر دیکشنری یا فهرست‌ها: یکی از کارهایی که به سامانهٔ تشخیص موجودیت‌های اسمی کمک می‌کند، استفاده از فهرست‌هایی از اسامی خاص اشخاص مشهور، سازمان‌ها، مکان‌ها و مخففهای آن‌هاست. ویژگی‌های مبتنی بر فهرست بر همین اساس تعریف می‌شوند؛ به این صورت که حضور یا عدم حضور آن‌ها در فهرست‌های هر یک از موجودیت‌ها به‌عنوان یک ویژگی در نظرگرفته می‌شود.

۳- ویژگی‌های سطح سند و پیکره: این ویژگی‌ها بر اساس محتوا و ساختار سند تعریف می‌شوند. از جملهٔ این دسته از ویژگی‌ها می‌توان به تعداد رخدادهای واژه در سند، تعداد رخدادهای واژه با حروف کوچک و بزرگ در سند، و محل قرارگیری در جمله اشاره کرد.

در [31] تأثیر روش‌های نمایش واژگان به‌عنوان ویژگی بررسی شده است. در این مقاله از روش خوشه‌بندی براون[5] و روش‌های نمایش طیفی واژگان HLBL و Collobert and

[1] Support Vector Classification
[2] character embedding
[3] Convolutional Neural Network (CNN)
[4] Recurrent Neural Network (RNN)
[5] Brown clustering

Archive of SID
www.SID.ir

Weston استفاده شده است. با بررسی روش‌های مختلف نمایش واژگان و اضافه‌کردن آن‌ها به‌عنوان ویژگی به یک مدل حاضر، به این نتیجه رسیده است که استفاده از خوشه‌بندی براون در مسألهٔ تشخیص موجودیت‌های اسمی، موثرتر از استفاده از روش‌های نمایش توزیع‌شدهٔ یادشده است. در این روش واژگان به‌صورت سلسله‌مراتبی خوشه‌بندی می‌شوند.

در [21] از این روش برای رده‌بندی واژگان در طبقات مختلف موجودیت‌های اسمی استفاده شده است. به این ترتیب با دانستن اینکه به‌عنوان مثال Microsoft برچسب سازمان می‌گیرد، از این واقعیت که NIKE با Microsoft در یک خوشه قرار گرفته استفاده کرده و به NIKE هم برچسب سازمان تخصیص می‌دهد.

در [27] نیز روش‌های مختلف نمایش توزیع‌شده برای افزودن به‌عنوان ویژگی با هم مقایسه شده‌اند که روش word2vec ارائه‌شده در [20] مناسب‌تر از سایرین بوده است. در [28] هم از word2vec استفاده شده است؛ ولی به جای آن که نمایش توزیع‌شده به‌صورت مستقیم به‌عنوان ویژگی به کار رود، نمایش‌های توزیع‌شده خوشه‌بندی شده و شمارهٔ خوشهٔ هر واژه به‌عنوان ویژگی به آن اضافه شده است. به این ترتیب مقدار این ویژگی برای واژگان با معانی یا کاربردهای نزدیک به هم مشابه خواهد بود.

۳-۲- پژوهش‌ها در زبان فارسی

۱-۳-۲- روش‌ها

در [4] به معرفی سامانه‌ای جهت تشخیص و رده‌بندی موجودیت‌های اسمی در زبان فارسی پرداخته شده است . این سامانه با به‌کارگیری الگوهای متنی ممکن برای اسم‌های خاص متعلق به یک دسته، سعی در شناسایی موجودیت‌های اسمی دارد و از برچسب‌های نحوی و معنایی برای رفع ابهام استفاده می‌کند. این سامانه به‌طور کامل قاعده‌محور است و از روش‌های یادگیری استفاده نمی‌کند.

در [1] با استفاده از مجموعه دادهٔ پژوهشگاه توسعه فناوری‌های پیشرفته خواجه‌نصیرالدین طوسی، روش‌های گوناگونی شامل روش شبکه عصبی، رده‌بند K نزدیک‌ترین همسایه، رده‌بند خطی، و رده‌بند بیزین مورد آزمون قرار گرفته و نشان داده شده است که رده‌بند خطی بهترین و رده‌بند مبتنی بر شبکه عصبی بدترین نتیجه را از نظر معیار F1 دارند.

در [5] استفاده از ترکیب مدل مخفی مارکف و قواعد تعیین‌شده، سامانه‌ای برای تشخیص موجودیت‌های اسمی پیشنهاد شده است. درواقع این سامانه از نوع ترکیبی است (ترکیب قاعده‌محور و یادگیری ماشین). آن‌ها سامانهٔ خود را روی مجموعه‌دادهٔ برچسب‌خورده‌ای که از متون خبرگزاری مهر ساخته‌اند، مورد آزمون قرار داده‌اند.

در [2] روشی ترکیبی از روش‌های قاعده‌محور و آماری ارائه شده است. برای تهیهٔ فهرست واژگان از ویکی‌پدیای فارسی استفاده شده است. همچنین به جز سامانه با سه نوع برچسب موجودیت اسمی (شخص، مکان و سازمان) از سامانه با شش نوع برچسب نیز استفاده شده است (شخص، مکان، سازمان، امکانات، محصول و رویداد). داده‌های آزمون به صورت دستی و با استفاده از پیکرهٔ بی‌جن‌خان تهیه شده است. همچنین برای فراهم‌کردن یک سامانهٔ تشخیص موجودیت‌های اسمی مبتنی بر یادگیری ماشین، به‌صورت دستی داده‌های آموزش فراهم شده است و روش‌های یادگیری ماشین مورد آزمون قرار گرفته که بهترین مدل ارائه‌شده مدل مبتنی بر میدان‌های تصادفی شرطی بوده است. در [3] تأثیر استفاده از ویژگی کسرهٔ اضافه در تشخیص موجودیت‌های اسمی مورد بررسی قرار گرفته است.

۲-۳-۲- ویژگی‌ها

در زبان انگلیسی ویژگی‌ای که بسیار مورد استفاده قرار گرفته و تأثیر به‌سزایی بر دقت سامانه‌های تشخیص موجودیت اسمی داشته، ویژگی بزرگ و کوچکی حروف است که در خط فارسی استفاده نمی‌شود.

در زبان فارسی ویژگی کسرهٔ اضافه وجود دارد که می‌تواند مشخص‌کنندهٔ خوبی برای مرز عبارت‌های موجودیت اسمی باشد. به‌عنوان مثال پایان عبارت اسمی «سازمانِ شهرداریِ استانِ تهران» که باید به‌عنوان موجودیت اسمی برچسب سازمان بگیرد، از این طریق قابل تشخیص است. البته دقت تشخیص کسرهٔ اضافه صددرصد نیست و می‌تواند با خطا همراه باشد. با این وجود، تأثیر ویژگی کسرهٔ اضافه در [2] مورد بررسی قرار گرفته است و بهبود چهار درصدی در حالت سه‌برچسبی شخص، مکان و سازمان نسبت به حالت عدم استفاده از این ویژگی گزارش شده است.

در [14] از ویژگی‌های مبتنی بر سند (جایگاه واژه در جمله و سند) و مبتنی بر فهرست (حضور یا عدم حضور در فهرست‌های خاص انواع مختلف موجودیت‌های اسمی) استفاده شده است. همچنین خروجی سامانه قاعده‌محور به‌عنوان یک ویژگی به سامانه مبتنی بر یادگیری ماشین داده شده است.

Archive of SID

www.SID.ir

در [1] از نقش دستوری کلمهٔ قبل و بعد، طول واژه، جمع یا مفردبودن اسم، وجود پسوندهای خاص برخی موجودیت‌های اسمی، وجود «ی» نسبی در آخر واژه، و درصد حضور واژه در کل متون آموزشی به‌صورت اسم مکان و اسم خاص شخص استفاده شده است.

۳- پیکره

در این بخش فرآیند تولید پیکرهٔ موجودیت‌های اسمی ارائه می‌شود. شرح مختصر شیوه‌نامه برچسب‌زنی، انتخاب اسناد پیکره، کیفیت‌سنجی برچسب‌زنی و آمار پیکره نهایی در ادامه آورده می‌شود.

۳-۱- شیوه‌نامه

تهیهٔ شیوه‌نامه برچسب‌زنی موجودیت‌های اسمی که حاوی قواعد تشخیص و برچسب‌زنی موجودیت‌های مختلف سازمان، شخص، مکان، تاریخ، پول، و درصد است با رجوع به دو شیوه‌نامهٔ استاندارد MUC[1] و CoNLL[2] و با تطبیق با قواعد زبان فارسی تهیه شده است. برچسب‌زن با توجه به بافت، واژگان یا عباراتی را که جزء موجودیت‌های هفت‌گانهٔ پیش‌گفته دسته‌بندی می‌شوند، تعیین می‌کند. شیوه رمزگذاری برچسب موجودیت‌ها IOB است، یعنی برچسب نخستین قطعه[3] (واژه) با B، و برچسب قطعه‌های (واژگان) دیگر، در صورت وجود، با I شروع می‌شود. واژگانی نیز که موجودیت نیستند، برچسب O خواهند داشت.

قواعد کلی برچسب‌زنی تأیید می‌کند که موجودیت‌هایی که در پی هم می‌آیند تا حد امکان برچسب جداگانه می‌گیرند. موارد استثنایی برای این قاعده کلی وجود دارد. برای مثال توالی عباراتی که جهت تعیین یک موجودیت واحد آورده می‌شوند، مانند «دانشکده مهندسی برق و کامپیوتر دانشگاه تهران» همگی یک برچسب موجودیت می‌گیرند. در دلالت‌های التزامی که به‌طورمعمول بین موجودیت مکان و سازمان رخ می‌دهد و نیاز به مرجع‌یابی وجود دارد، مرجع‌یابی انجام نمی‌شود. به‌عنوان مثال در «ایران ادعای آمریکا را رد کرد» واژگان «ایران» و «آمریکا» برچسب مکان می‌گیرند.

در عبارات هم‌پایه، همهٔ عبارت یک برچسب می‌گیرد. به‌عنوان نمونه «آمریکای شمالی و جنوبی» درکل یک برچسب مکان می‌گیرد. هم‌نامی‌ها نیز جزء موجودیت لحاظ می‌شوند و برچسب می‌گیرند. به‌عنوان مثال «فرودگاه بین‌المللی تهران» که هم‌نام «فرودگاه بین‌المللی امام خمینی» است برچسب موجودیت می‌گیرد. اگر صفت درون عبارت موجودیت قرار گیرد برچسب می‌گیرد، ولی اگر خارج از عبارت موجودیت باشد برچسب نمی‌گیرد. به‌عنوان نمونه «خلیج همیشگی فارس» درکل یک برچسب مکان می‌گیرد، ولی در «سید حسین محلاتی جلیل‌القدر» تنها بخش «سید حسین محلاتی» برچسب شخص می‌گیرد.

تفاوت برچسب‌های زمان و تاریخ، مدت متفاوت طول زمانی آن‌هاست. واحدهای زمانی کمتر از یک روز با برچسب «زمان» و واحدهای زمانی بیش از یک روز با برچسب «تاریخ» مشخص می‌شوند. برای مثال عبارت «روز ۳ خرداد ۱۳۶۴» با برچسب «تاریخ» و عبارت «ساعت ۱۴ روز ۳ خرداد ۱۳۶۴» با برچسب «زمان» مشخص می‌شود.

آنچه گفته شد، مختصری از شیوه‌نامه برچسب‌زنی موجودیت‌های اسمی بود. به جهت رعایت اختصار، ذکر جزئیات شیوه‌نامه برچسب‌زنی در همه موجودیت‌ها مقدور نیست. به همین دلیل به ذکر کلیات بسنده می‌کنیم و متن کامل شیوه‌نامه را با پیکره موجودیت‌های اسمی منتشر خواهیم کرد.

برچسب‌زن‌های این پیکره دارای تحصیلات کارشناسی ارشد و آشنا به موضوع بودند. درنهایت برای ارزیابی کیفیت برچسب‌زنی بخشی از داده‌ها برای کیفیت‌سنجی به یک برچسب‌زن با تحصیلات دکترای زبان‌شناسی داده شد تا دوباره برچسب‌زنی شود.

۳-۲- انتخاب اسناد

با هدف افزایش تنوع اسناد پیکره، منابع دریافت اخبار فهرستی از خبرگزاری‌ها و سایت‌های خبری مختلف در نظر گرفته می‌شود: خبرگزاری فارس، خبرگزاری مهر، ایرنا (خبرگزاری جمهوری اسلامی)، ایسنا (خبرگزاری دانشجویان ایران)، همشهری آنلاین، تابناک، فرارو، ورزش سه، انتخاب، و باشگاه خبرنگاران جوان.

انتخاب اسناد مجموعه‌داده از بین تعداد بسیار زیاد اخبار باید به‌گونه‌ای انجام شود که توزیع موضوعی اسناد پیکره نهایی با توزیع موضوعی اخبار دنیای واقعی هم‌خوانی داشته باشد و میانگین طول اسناد مجموعه‌داده نیز به میانگین طول اسناد دنیای واقعی نزدیک باشد. همچنین برای این که تنوع بیشتری در اسناد مجموعه‌داده نهایی وجود داشته باشد، باید اسناد از زمان‌های مختلف انتخاب شوند. با این ترتیب رفتار اسناد پیکره به اخبار دنیای واقعی نزدیک‌تر می‌شود.

[1] wwwnlpir.nist.gov/related_projects/muc/proceedings/ne_task.html
[2] www.cnts.ua.ac.be/conll2003/ner/
[3] token

Archive of SID
www.SID.ir

برای انتخاب اسناد مجموعه داده ابتدا نیاز است، اسناد خبری از منابع یادشده با به‌کارگیری یک خزش‌گر جمع‌آوری شوند. بازهٔ زمانی دریافت اخبار ده ماه نخست سال ۱۳۹۵ در نظر گرفته شده است. پس از حذف اسناد نامناسب که تنها شامل تصاویر یا متن‌های بسیار کوتاه هستند (زیر هفتاد نویسه) در مجموع حدود هفتصدهزار سند خبری به‌دست آمد. پیش‌پردازش اسناد که شامل هنجارسازی و واحدسازی متن است با ابزار Persianp[1] صورت گرفته است.

برای تعیین موضوع اسناد از سامانهٔ رده‌بند موضوعی اسناد خبری پروژه هشتگ فارسی[2] استفاده شده و اسناد در شش دستهٔ موضوعیِ «سیاست»، «اقتصاد»، «ورزش»، «فرهنگ و هنر»، «دانش و فناوری» و «جامعه» دسته‌بندی می‌شوند.

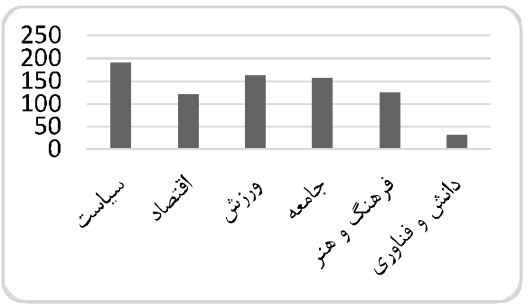
در هر موضوع میانگین طول و انحراف معیار اسناد محاسبه و با توجه به این دو مشخصه از اسناد طبق توزیع نرمال نمونه‌گیری انجام و با این روش میانگین طول اسناد با دنیای واقعی یکسان و پراکندگی اسناد نسبت به پارامتر طول نیز حفظ می‌شود؛ سپس در بازه‌های زمانی مختلف به‌صورت میانگین از موضوعات مختلف به‌گونه‌ای سند انتخاب می‌شود که توزیع موضوعی اسناد انتخابی با توزیع موضوعی همه اسناد برابر باشد.

برای اینکه پیکره درنهایت شامل اسنادی از منابع مختلف خبری باشد، از هر منبع خبری به نسبت اخباری که هر منبع خبری در یک موضوع منتشر کرده است از مجموع اسناد نمونه‌گیری‌شده سند انتخاب می‌شود؛ چون برخی از خبرگزاری‌ها اخبار بسیار بیشتری نسبت به دیگر خبرگزاری‌ها و پایگاه‌های خبری منتشر می‌کنند و جهت رعایت توازن، حداکثر بیست درصد اسناد مجموعه داده نهایی می‌تواند از یک منبع خبری باشد. تنها مورد استثنا، سایت ورزش سه است که چون اختصاص به ورزش دارد سی درصد اسناد ورزشی را به خود اختصاص می‌دهد.

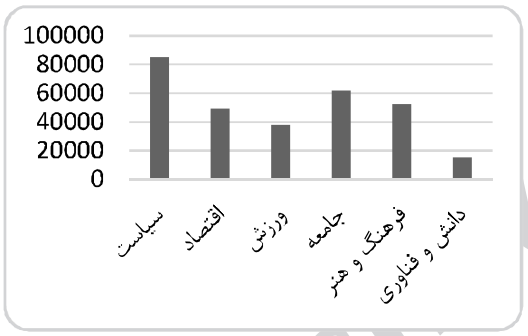
به این ترتیب پیکرهٔ نهایی شامل اسنادی است که از نظر موضوع، میانگین طول، و زمان با اخبار دنیای واقع هم‌خوانی دارد. شکل (۱) نمودار تعداد اسناد را در موضوعات مختلف نشان می‌دهد و شکل (۲) تعداد واژگان در هر موضوع را مشخص می‌کند.

۳-۳- کیفیت برچسب‌زنی

برای اینکه برچسب‌زنی کیفیت‌سنجی شود، بخشی از پیکره توسط دو شخص مختلف برچسب‌گذاری و میزان توافق بین برچسب‌ها محاسبه می‌شود. از اسناد مجموعه‌دادهٔ ۱۶۰ سند برچسب‌خورده که در مجموعه حدود ۲۰۳۰ جمله دارد، برای برچسب‌زنی مجدد انتخاب می‌شود. این اسناد شامل ۵۳۸۰۰ قطعه هستند. اگر اختلافی بین برچسب‌ها مشاهده شود، یا معلول اختلاف برداشت از بافت واژه و تشخیص متفاوت در نوع موجودیت است یا اشتباهی توسط یکی از برچسب‌زن‌ها رخ داده و نوع برچسب اشتباه وارد شده است. در این ۱۶۰ سند تعداد اختلاف بین دو برچسب‌زن در همه قطعه‌ها ۳۸۶ مورد و میزان هماهنگی برچسب‌زن‌ها ۹۹ درصد است.



(شکل-۱): تعداد اسناد انتخابی در موضوعات مختلف
(Figure-1): Number of selected documents per topic



(شکل-۲): تعداد واژه‌ها در اسناد انتخابی در موضوعات مختلف
(Figure-2): Number of words in selected documents per topic

با توجه به اینکه درصد بالایی از قطعه‌ها موجودیت اسمی نیستند و برچسب «O» دارند، می‌توان تنها با لحاظ‌کردن قطعه‌هایی که برچسب موجودیت دارند نیز درصد اختلاف را بررسی کرد؛ چون هیچ مرجع مطلقی برای تعیین اینکه کدام قطعه‌ها موجودیت اسمی هستند، وجود ندارد، برای محاسبه تعداد قطعه‌هایی که موجودیت اسمی هستند، اجتماع تعداد قطعه‌هایی که توسط برچسب‌زن‌ها موجودیت تشخیص داده شده‌اند، ملاک قرار می‌گیرد. درمجموع ۷۶۸۳ قطعه موجودیت اسمی دارند. با درنظرگرفتن تعداد اختلافات (۳۸۶) در این حالات میزان توافق برچسب‌زن‌ها ۹۵ درصد است.

همچنین معیار کاپا برای محاسبه میزان توافق بین برچسب‌زن‌ها را محاسبه می‌کنیم. معیار کاپا توافق را بر اساس

[1] persianp.ir
[2] http://pclp.itrc.ac.ir/?q=products&page=8(ردیف ۲۵۶)

Archive of SID
www.SID.ir

میزان فاصله رأی برچسب‌زن‌ها نسبت به حالتی که برچسب‌زنی را تصادفی انجام دهند، محاسبه می‌کند. نتایج در حالتی که همه برچسب‌ها ازجمله برچسب «O» وجود دارد به شرح زیر است:

Kappa = 0.9723; 95% CI: 0.9695 to 0.9750
p-value < 2.2e-16

نتایج در حالتی که برچسب «O» در نظر گرفته نمی‌شود به شرح زیر است:

Kappa = 0.9409; 95% CI: 0.9352 to 0.9467
p-value < 2.2e-16;

همان‌طور که مشاهده می‌شود، میزان p-value در هر دو حالت بسیار پایین است و نتیجه حکایت از توافق بالا بین برچسب‌زن‌ها دارد.

۴-۳- آمار پیکره

پیکرهٔ نهایی شامل ۷۰۹ سند خبری است که درمجموع شامل ۷۱۴۵ جمله است. در این مجموعه داده ۳۰۲۵۳۰ قطعه (واژه) وجود دارد که ۴۱۱۴۸ قطعه برچسب موجودیت دارند و مابقی قطعه‌ها برچسب موجودیت ندارند (برچسب O). آمار تعداد موجودیت‌های مختلف (تک‌واژگان یا عبارات نشان‌دهنده یک موجودیت) و تعداد قطعه‌های برچسب‌خورده در هر نوع موجودیت در جدول (۱) آمده است.

برای محاسبهٔ تعداد موجودیت‌های یکتا هر موجودیت تنها یک بار شمرده و تکرارهای مختلف آن نادیده گرفته می‌شود.

(جدول-۱) تعداد قطعه‌های برچسب‌خورده در هر نوع موجودیت
(Table-1) Number of labeled tokens per entity type

نوع موجودیت	تعداد موجودیت	تعداد قطعه‌ها	تعداد موجودیت‌های یکتا
شخص	4447	7675	2507
سازمان	6360	16964	2413
مکان	6223	8782	1288
زمان	281	732	801
تاریخ	1858	4259	157
مبالغ مالی	527	2037	436
درصد	326	699	179
تعداد کل	20022	41148	7781

۴- روش پیشنهادی

برای پیاده‌سازی سامانهٔ تشخیص موجودیت‌های اسمی از روش‌های قاعده‌محور، آماری، و یادگیری عمیق استفاده شده است که در ادامه به توضیح جزئیات هر یک می‌پردازیم.

۴-۱- روش قاعده‌محور

در روش‌های قاعده‌محور، قواعدی به‌صورت دستی بر اساس شمّ زبانی خبره یا به‌صورت نیمه‌خودکار و با مشاهده مصادیق متعدد تدوین شد که با استفاده از آن قواعد موجودیت‌های اسمی متن تشخیص داده شود. قواعد به دو گونه قواعد منظم روی توالی نویسه‌های هر قطعه و قواعد منظم روی توالی قطعه‌های متن است. به‌عنوان مثال موجودیت تاریخ (مانند ۱۳۹۸/۱/۱۰) را می‌توان با استفاده از عبارات منظم روی توالی نویسه‌های قطعه تشخیص داد. قواعد منظم روی توالی قطعه‌ها نیز قابل تعریف هستند. این قواعد را می‌توان به‌صورت قواعد ساده پیش‌توالی و فهرست (مانند قاعده «"دانشگاه آزاد اسلامی" + نام شهر» برای تشخیص موجودیت سازمان) یا قواعدی مشابه عبارات منظم ولی برای توالی قطعه‌ها تهیه و برای تشخیص موجودیت‌ها استفاده کرد.

۴-۲- روش آماری

با توجه به آنچه در بخش قبل گفته شد، در بیشتر مقالات از روش CRF به‌دلیل تناسب کامل با نوع مسئله به‌عنوان روش اصلی و پایه در مسألهٔ تشخیص موجودیت‌های اسمی استفاده شده است. حتی در روش‌های جدیدتر که در سال‌های اخیر با استفاده از یادگیری عمیق ارائه شده، هم‌چنان از CRF در ترکیب با یادگیری عمیق استفاده شده است. به‌عنوان مثال در [18] از ترکیب CRF و شبکه‌های LSTM استفاده شده است. به همین دلیل تصمیم گرفتیم از این روش به‌عنوان روش آماری منتخب استفاده کنیم و تمرکز را در ادامه بر تعریف ویژگی‌های مناسب قرار دهیم. در ادامه ابتدا به توضیح مختصر روش میدان‌های تصادفی شرطی می‌پردازیم و سپس ویژگی‌های مورد استفاده را معرفی می‌کنیم.

۴-۲-۱- میدان‌های تصادفی شرطی

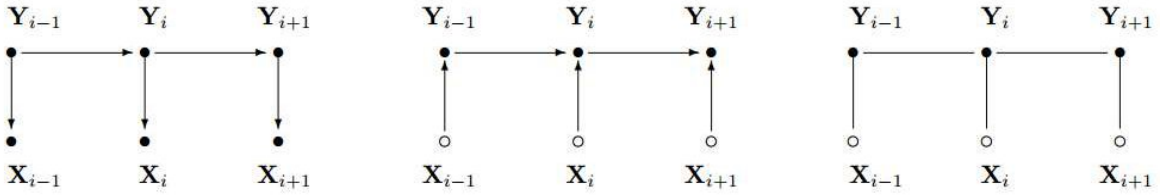
روش بیشینهٔ بی‌نظمی مارکف اگرچه قوت و دقت زیادی دارد و برای مدل‌کردن مسائل برچسب‌گذاری توالی بسیار مناسب است، یک ضعف جدّی دارد که از آن با عنوان اریب برچسب یاد می‌شود. این مشکل را می‌توان این‌طور توضیح داد که حالت‌های با بی‌نظمی کم روی توزیع انتقال‌ها، درواقع مشاهدهٔ خود را نادیده می‌گیرند. برای حل این مشکل، با حفظ تمام مزایای روش MEMM، روش میدان‌های تصادفی شرطی (CRF) ارائه شده است. در شکل (۳) تفاوت روش MEMM،HMM و CRF در قالب تصویر نشان داده شده است. همان‌طور که در این شکل مشخص است، در روش

Archive of SID
www.SID.ir

HMM هر حالت، تنها با توجه به حالت قبلی تولید و مشاهده‌ها از حالت‌ها تولید و در MEMM، هر حالت با توجه به حالت قبلی و مشاهدهٔ کنونی تعیین می‌شود. در CRF، کل توالی حالت‌ها از کل توالی مشاهدات تولید می‌شود و به همین علت مانند روش MEMM دچار سوگیری روی تعداد حالت‌های خروجی از هر حالت نمی‌شود. گروه پردازش زبان طبیعی دانشگاه استنفورد[1]، مجموعه‌ای از ابزارهای پردازش زبان را به‌صورت متن‌باز در اختیار پژوهش‌گران قرار داده‌اند. یکی از این ابزارها تشخیص موجودیت‌های اسمی[2] است. این ابزار به زبان جاوا نوشته شده و مبتنی بر روش CRF است.

همچنین مجموعه‌ای غنی از ویژگی‌ها به‌صورت پیش‌فرض برای آن تعریف شده‌اند و به‌سادگی قابل استفاده هستند. علاوه‌براین در بسیاری از مقالات از همین سامانه به‌عنوان سامانهٔ پایه برای پیاده‌سازی یک ابزار تشخیص موجودیت اسمی استفاده شده است و مقالهٔ متناظر آن ارجاعات زیادی دارد.

برای استخراج ویژگی‌های بن‌واژه، برچسب عبارت اسمی، و برچسب ادات سخن از ابزار Persianp استفاده شده است.



(شکل-۳): تفاوت روش‌های HMM،MEMM و CRF. برگرفته از [17] . در شکل سمت چپ HMM، در وسط MEMM و در سمت راست CRF نمایش داده شده است.

(Figure-3): Difference between HMM (left), MEMM (middle), and CRF (right) methods [17]

۲-۲-۴- ویژگی‌ها

ویژگی‌هایی که ما برای آموزش سامانهٔ CRF استفاده کرده‌ایم، ویژگی‌های ابتدایی و پایه‌ای هستند. ویژگی‌های املایی تعریف‌شده در مقالات انگلیسی به‌صورت مستقیم برای فارسی قابل استفاده نیست و باید برای زبان فارسی بازتعریف شوند. برای مثال می‌توان به ویژگی دارابودن ارقام به‌عنوان یک پیش‌فرض برای تعلق واژه به دستهٔ موجودیت‌های اسمی درصد، تاریخ، و زمان اشاره کرد. ما از ویژگی‌های املایی استفاده نکرده‌ایم. همچنین به جای استفادهٔ مستقیم از ویژگی کسرهٔ اضافه، از ویژگی برچسب عبارت اسمی[3] استفاده کردیم؛ چون موجودیت‌های اسمی به‌طورمعمول یک عبارت اسمی کامل هستند؛ علاوه‌براین، از ویژگی برچسب ادات سخن استفاده کردیم. درضمن اگرچه استفاده از ریشهٔ[4] واژگان در پژوهش‌های بازیابی اطلاعات مفید است، در سایر کارهای مربوط به پردازش زبان استفاده از بن‌واژه[5] نتیجهٔ بهتری دارد [13]، زیرا در تعیین ریشهٔ واژگان تنها به ظاهر واژگان توجه می‌شود در حالی که در تعیین بن‌واژه به سطح معنایی واژگان و ارتباط واژگان مجاور با هم نیز توجه می‌شود. به همین دلیل از ویژگی بن‌واژه نیز علاوه‌بر خود واژه استفاده کردیم. علاوه‌بر این‌ها استفاده از ویژگی‌های مبتنی بر دیکشنری نیز کاربردی است؛ بدین‌ترتیب فهرستی از موجودیت‌های اسمی بدون ابهام فراهم کردیم و حضور یا عدم حضور واژه در این فهرست را به‌عنوان ویژگی در نظر گرفتیم. برای تهیهٔ این فهرست‌ها ابتدا باید فهرست اولیه‌ای از موجودیت‌های اولیه تهیه و سپس با بررسی خبره موارد نامطلوب حذف شود. فهرست موجودیت‌های اسمی مورد استفاده در این پژوهش، از ویکی‌پدیای فارسی استخراج شده و شامل سه برچسب اصلی شخص، مکان و سازمان است. در این فهرست، تعداد ۴۲۴۴ موجودیت شخص، ۴۸۴۷ موجودیت مکان و ۵۲۷۳۱ موجودیت سازمان وجود دارد.

به این ترتیب ویژگی‌های مورد استفاده از این قرار هستند:

- واژهٔ جاری و یک واژهٔ قبل و یک واژهٔ بعد از آن
- برچسب ادات سخن واژهٔ جاری و واژهٔ قبل و واژهٔ بعد از آن
- برچسب عبارت اسمی واژهٔ جاری و واژهٔ قبل و واژهٔ بعد از آن
- بن‌واژهٔ واژهٔ جاری و واژهٔ قبل و واژه بعد
- ان-گرام‌های سطح نویسه واژه تا سقف شش نویسه (فقط از ابتدا یا انتهای واژه): برای بررسی پیشوندها و پسوندها
- ویژگی‌های مبتنی بر دیکشنری به‌صورت تطابق جزئی

[1] https://nlp.stanford.edu
[2] https://nlp.stanford.edu/software/CRF-NER.html
[3] NP-Chunk
[4] stem
[5] lemma

Archive of SID
www.SID.ir

۳-۴- یادگیری عمیق

همان‌طور که در بخش قبل گفته شد، در اغلب سامانه‌های طراحی‌شده با استفاده از یادگیری عمیق از شبکه‌های عصبی بازگشتی با استفاده از LSTM استفاده شده است. این شبکه‌های عصبی برای مدل‌سازی مسائل از جنس برچسب‌گذاری دنباله‌ها مناسب هستند و استفاده از LSTM امکان استفاده از اطلاعات دور از واژهٔ جاری را برای برچسب‌گذاری فراهم می‌کند. ما نیز در این جا از شبکهٔ عصبی با استفاده از LSTM استفاده کرده‌ایم.

برای اجرای شبکهٔ عصبی از ابزار openNMT[1] استفاده شده است. این ابزار یک چارچوب عمومی برای پیاده‌سازی انواع مدل‌های یادگیری عمیق است که به‌ویژه تمرکزش روی مدل‌های توالی به توالی است. در این مدل‌ها، ورودی یک توالی و خروجی نیز یک توالی است که به‌الزام طول دنبالهٔ خروجی با دنبالهٔ ورودی یکسان نیست؛ ولی جز این، حالتی هم برای مدل‌های برچسب‌زنی توالی دارد که در آن طول دنبالهٔ خروجی و ورودی به‌الزام یکسان است. ما در این جا از این حالت استفاده کرده‌ایم. در این حالت شبکه عصبی از یک رمزگذار[2] تشکیل شده است و به‌دلیل نظیربودن هر ورودی با یک خروجی نیازی به رمزگشا[3] ندارد. شبکه عصبی مورد استفاده از دو لایه LSTM تشکیل شده و طول بردارهای طیفی واژگان پانصد و تعداد گره‌های شبکه نیز پانصد است.

برای آموزش شبکهٔ عصبی، ویژگی‌ها به بردارهای نمایش طیفی واژگان ملحق شده و یک بار شبکهٔ عصبی با سلول‌های LSTM در جهت رفت (از ابتدا به انتهای جمله) و بار دیگر در جهت برگشت (از انتها به ابتدای جمله) آموزش داده می‌شود.

برای خوشه‌بندی بردارهای نمایش طیفی واژگان جهت استفاده از شمارهٔ خوشه به‌عنوان ویژگی برای مدل شبکهٔ عصبی، از ابزار word2vec استفاده شد[4]. این ابزار بردارهای نمایش طیفی واژگان را روی پیکرهٔ ورودی یادگیری و این بردارها را بر اساس الگوریتم k-means خوشه‌بندی می‌کند. تعداد خوشه‌ها قابل تنظیم در ورودی است. در این پژوهش تعداد خوشه‌ها برابر ۱۵۰۰ قرار داده شده است.

۱-۳-۴- ویژگی‌ها

به‌طورعمومی در روش‌های مبتنی بر شبکهٔ عصبی عمیق نیازی به تعریف و استخراج ویژگی‌ها نیست و شبکهٔ عصبی عمیق خود به یادگیری ویژگی‌های مورد نیاز در داده‌ها می‌پردازد. برای این که بدون تعریف ویژگی، مدل، خود قادر به استخراج ضمنی ویژگی‌های داده‌ها باشد، حجم زیادی دادهٔ برچسب‌خورده نیاز است. به همین دلیل، در غیاب حجم بالای داده‌های برچسب‌خورده انتظار کیفیت بالایی از مدل آموزش‌داده‌شده نداریم. در [18] روشی برای افزودن ویژگی‌ها پیشنهاد و نشان داده شده است که با افزودن ویژگی بزرگ و کوچکی حرف نخست واژگان (برای زبان انگلیسی) می‌توان به بهبود سه درصدی در معیار F1 رسید؛ یعنی افزودن ویژگی‌ها می‌تواند به بهبود کیفیت سامانه حاصل کمک کند. در این پژوهش، از همین روش برای اضافه‌کردن ویژگی‌ها به شبکهٔ عصبی استفاده می‌کنیم. مطابق این روش، بردار ویژگی‌ها به بردار نمایش طیفی واژگان پیوست می‌شود.

با استفاده از این قابلیت به‌منظور ترکیب سامانه آماری مبتنی بر الگوریتم CRF و شبکهٔ عصبی، از خروجی CRF به‌عنوان یک ویژگی در آموزش مدل شبکهٔ عصبی استفاده می‌کنیم. شهود ما برای انجام این کار این است که خروجی CRF می‌تواند پیش‌فرضی برای تولید برچسب صحیح در شبکهٔ عصبی ایجاد کند. همچنین، پیش‌تر از خروجی مدل قاعده‌محور به‌عنوان یک ویژگی در ورودی CRF استفاده شده است [14] و ما تصمیم گرفتیم همین روش را برای ترکیب CRF و LSTM به کار گیریم.

علاوه‌بر استفاده از خروجی سامانه CRF، استفاده از یک ویژگی دیگر در اینجا پیشنهاد داده می‌شود. در [31] پیشنهاد شده است که از کد خوشه‌ها در روش خوشه‌بندی براون به‌عنوان یک ویژگی در سامانه‌های تشخیص موجودیت‌های اسمی استفاده شود. در این روش خوشه‌بندی، واژگان با معنای نزدیک به هم در خوشه‌های یکسان قرار می‌گیرند و هرچه طول کد خوشهٔ دو واژه، اشتراک بیشتری داشته باشد، آن دو واژه از نظر معنایی به هم نزدیک‌ترند. به این ترتیب می‌توان متوجه نزدیکی واژگان «تهران» و «اصفهان» شد که هر دو باید برچسب مکان بگیرند و همچنین می‌توان متوجه دوری معنایی واژگان «تهران» و «علی» شد که باید برچسب‌های متفاوت بگیرند. آموزش خوشه‌بندی براون بسیار زمان‌بر است. به جای آن، از شمارهٔ خوشهٔ هر واژه در خوشه‌بندی نمایش طیفی واژگان به روش k-means با معیار فاصلهٔ کسینوسی استفاده می‌کنیم. به این ترتیب، برای مثال واژگان «شنبه»، «یکشنبه»، «دوشنبه» و سایر روزهای هفته که با احتمال بالایی باید برچسب «تاریخ» بگیرند، در یک خوشه قرار می‌گیرند. همچنین، واژگان «اصفهان»، «تهران» و «تبریز» که با احتمال بالایی برچسب

[1] http://opennmt.net
[2] encoder
[3] decoder
[4] https://github.com/dav/word2vec

Archive of SID
www.SID.ir

«مکان» می‌گیرند، در یک خوشه قرار می‌گیرند. فایدهٔ این ویژگی این است که اگر برای مثال اسم شهری در هیچ یک از فهرست‌ها و یا داده‌های آموزشی نباشد، هم‌خوشه‌بودن آن با واژگانی که می‌دانیم برچسب مکان می‌گیرند به شناسایی آن اسم به‌عنوان مکان کمک شایانی می‌کند؛ به این ترتیب سامانهٔ طراحی‌شده پیشنهادی از چندین قسمت شامل بخش قاعده‌محور (شامل فهرست‌محور و مبتنی بر عبارات منظم)، آماری مبتنی بر CRF، یادگیری عمیق مبتنی بر LSTM، و ترکیب آن‌ها تشکیل می‌شود. در ادامه با انجام آزمایش‌های متعدد با استفاده از مجموعه‌دادهٔ طراحی‌شده به ارزیابی هر بخش می‌پردازیم.

۵- ارزیابی

در این بخش ابتدا به معرفی معیارهای ارزیابی مورد استفاده می‌پردازیم؛ سپس نتایج ارزیابی سامانه با استفاده از سامانهٔ قاعده‌محور، سامانهٔ آماری و شبکه عصبی و ترکیب آن‌ها را به همراه تحلیلی بر خطاهای سامانهٔ نهایی ارائه می‌دهیم. درنهایت عملکرد سامانهٔ طراحی‌شده را با سایر سامانه‌های موجود که امکان استفاده از آن‌ها فراهم بوده مقایسه می‌کنیم.

تمام ارزیابی‌ها به روش k-fold cross-validation انجام شده که برای بخش CRF، مقدار k برابر پنج و برای شبکهٔ عصبی به‌دلیل نیاز به حجم بیشتری از دادهٔ آموزشی این مقدار برابر با ده در نظر گرفته شده است. هر بار عمل تقسیم‌بندی داده‌ها به k بخش مساوی، پنج بار تکرار و میانگین مقدارهای دقت، فراخوانی، و F1 گزارش شده است. برای مشخص‌کردن معناداربودن یا نبودن بهبودها، از آزمون آماری Student's t-test جفتی با سطح اطمینان ۹۵ درصد استفاده شده است.

۵-۱- معیارهای ارزیابی

برای آن که مشخص شود عملکرد یک سامانهٔ طراحی‌شده تا چه اندازه خوب است، لازم است ارزیابی صورت گیرد. در این پژوهش از روش ارزیابی CONLL استفاده می‌کنیم. در این روش، سه معیار دقت[1]، فراخوانی[2]، و معیار F1 گزارش می‌شود. رابطهٔ محاسبهٔ این سه معیار در زیر آمده است:

$$\text{دقت} = \frac{\text{تعداد تشخیص‌های صحیح}}{\text{تعداد کل تشخیص‌ها}}$$

$$\text{فراخوانی} = \frac{\text{تعداد تشخیص‌های صحیح}}{\text{تعداد موجودیت‌های واقعی}}$$

$$F1 = \frac{2 \times \text{دقت} \times \text{فراخوانی}}{\text{دقت} + \text{فراخوانی}}$$

دقت، مشخص‌کنندهٔ درصد موجودیت‌های اسمی است که توسط سامانه به‌درستی تشخیص داده شده‌اند و فراخوانی، درصد موجودیت‌های اسمی واقعی است که سامانه توانسته آن‌ها را تشخیص دهد.

وقتی دقت در سطح عبارت محاسبه می‌شود، مرز دقیق عبارت موجودیت اسمی مهم است و حتی اگر یک واژه از آن عبارت به‌عنوان جزئی از آن شناسایی نشده باشد، جزء تشخیص‌های درست لحاظ نمی‌شود. به‌عنوان مثال در عبارت «دانشکدهٔ مهندسی برق و کامپیوتر دانشکدهٔ فنی دانشگاه تهران» کافی است که یک واژه (برای مثال دانشکده) به‌عنوان جزئی از موجودیت، برچسب نخورده باشد. در این صورت حتی اگر تمام واژگان دیگر به‌درستی برچسب‌گذاری شده باشند، در محاسبهٔ دقت، تأثیر منفی می‌گذارد. علت این است که تعداد تشخیص‌های درست ثابت می‌ماند؛ اما تعداد کل تشخیص‌ها افزایش و به این ترتیب دقت کاهش می‌یابد. بنابراین تشخیص‌ندادن یک موجودیت اسمی طولانی از نظر محاسبهٔ دقت در سطح عبارت به تشخیص ناقص آن ارجحیت دارد. از آن جا که در کاربردهای واقعی ممکن است تشخیص ناقص موجودیت اسمی هم کاربرد داشته باشد و نیازی به تشخیص صددرصدی آن نباشد، در این پژوهش دقت در سطح واژه را هم گزارش کرده‌ایم.

۵-۲- ارزیابی بخش‌های مختلف سامانه

۵-۲-۱- روش قاعده‌محور

برای تشخیص قاعده‌محور موجودیت‌های اسمی فارسی از ابزار TokenRegex استنفورد [8] استفاده و قواعد را مختص ابزار آماده‌سازی می‌کنیم.

نتیجهٔ ارزیابی سامانهٔ قاعده‌محور طراحی‌شده بر روی کل مجموعه داده، به‌تفکیک نوع برچسب در سطح واژه و عبارت در جدول (۲) و جدول (۳) گزارش شده است. همان طور که انتظار می‌رفت، دقت این سامانه تاحدودی بالا ولی فراخوانی آن پایین است.

۵-۲-۲- روش آماری

نتیجهٔ ارزیابی سامانهٔ آماری طراحی‌شده بر روی کل مجموعه داده، به‌تفکیک نوع برچسب در سطح واژه و عبارت در جدول‌های (۴ و ۵) گزارش شده است. همان‌طور که از این

[1] precision
[2] recall

Archive of SID
www.SID.ir

جدول‌ها مشخص است، با استفاده از سامانهٔ آماری به دقت، فراخوانی و درنتیجه F1 بالاتری دست پیدا کرده‌ایم. این مسأله نشان می‌دهد که به‌دلیل درحدکفایت بزرگ‌بودنِ حجم مجموعه‌داده، سامانهٔ آماری موفق به یادگیری قواعد اصلی تشخیص موجودیت‌های اسمی شده است. تعداد قطعه‌هایی که با برچسب زمان در مجموعه‌داده مشخص شده‌اند، کم‌تر از سایر انواع موجودیت‌های اسمی بوده، درحالی‌که تنوع قوانین برای این موجودیت زیاد است. به همین دلیل فراخوانی برای این برچسب کم‌تر از سایر برچسب‌ها بوده است. در مورد درصد و واحد پول با آن که تعداد قطعه‌های مشخص‌شده با این برچسب‌ها در مجموعه داده کم‌تر از سایر موجودیت‌ها (به جز زمان) بوده، دقت و فراخوانی بالایی به‌دست آمده است. علت این امر به کم‌تربودن تعداد و پیچیدگی قواعد تشخیصی برای این دو برچسب برمی‌گردد.

(جدول-۲):نتایج سامانهٔ قاعده‌محور- در سطح واژه
(Table-2):Rule-based system results – word level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.86	0.44	0.58
مکان	0.80	0.69	0.74
سازمان	0.75	0.54	0.62
زمان	0.92	0.52	0.66
تاریخ	0.86	0.33	0.48
مبالغ مالی	0.99	0.75	0.85
درصد	0.98	0.92	0.95
کل	0.81	0.55	0.65

(جدول-۳):نتایج سامانهٔ قاعده‌محور- در سطح عبارت
(Table-3):Rule-based system results – phrase level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.86	0.44	0.58
مکان	0.80	0.69	0.74
سازمان	0.75	0.54	0.63
زمان	0.92	0.52	0.66
تاریخ	0.86	0.33	0.48
مبالغ مالی	0.99	0.75	0.85
درصد	0.98	0.92	0.95
کل	0.81	0.55	0.65

(جدول-۴): نتایج سامانهٔ آماری – سطح واژه
(Table-4):Statistical system results – word level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.93	0.73	0.82
مکان	0.90	0.76	0.83
سازمان	0.93	0.75	0.83
زمان	0.91	0.62	0.74
تاریخ	0.89	0.79	0.84
مبالغ مالی	0.98	0.84	0.90
درصد	0.98	0.91	0.95
کل	0.92	0.76	0.83

(جدول-۵):نتایج سامانهٔ آماری – سطح عبارت
(Table-5):Statistical system results – phrase level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.80	0.65	0.72
مکان	0.82	0.81	0.82
سازمان	0.76	0.67	0.70
زمان	0.72	0.61	0.66
تاریخ	0.77	0.71	0.74
مبالغ مالی	0.83	0.75	0.79
درصد	0.92	0.91	0.91
کل	0.80	0.72	0.75

۵-۲-۳- ترکیب روش‌های قاعده‌محور و آماری

مطابق [14] از خروجی مدل قاعده‌محور به‌عنوان یک ویژگی در مدل CRF استفاده کردیم و برای بررسی تأثیر استفاده از آن، مدل را یک بار بدون استفاده از این ویژگی و بار دیگر با استفاده از آن ساختیم. نتایج ارزیابی نشان می‌دهد که این ویژگی تأثیری بر کیفیت مدل CRF نداشته و کمکی به بهبود آن نکرده است. دلیل این مسأله این است که مدل CRF قواعد تعریف‌شده در بخش قاعده‌محور را در فرآیند یادگیری آموزش دیده است. همچنین از حضور یا عدم حضور واژه در فهرست موجودیت‌های اسمی پرتکرار کم‌ابهام به‌عنوان یک ویژگی دودویی استفاده می‌کند و به همین دلیل از بخش فهرست‌محور هم بی‌نیاز است.

۵-۲-۴- یادگیری عمیق

از آن جا که الگوریتم‌های یادگیری عمیق برای آن که به‌خوبی عمل یادگیری را انجام دهند به حجم زیادتری از دادهٔ آموزشی نسبت به حالت یادگیری ماشین معمولی دارند، آزمایش‌های این حالت را با استفاده از k-fold و با k=10 انجام دادیم. نتایج آزمایش‌ها بدون استفاده از ویژگی‌ها، با استفاده از ویژگی خروجی مدل CRF و با افزودن ویژگی شمارهٔ خوشه در جدول‌های (۶ تا ۱۱) گزارش شده است.

نتایج ارزیابی بخش شبکهٔ عصبی LSTM دوطرفه در دو سطح واژه و عبارت در جدول (۶) و جدول (۷) ارائه شده است. در این حالت از هیچ ویژگی از پیش تعریف‌شده‌ای استفاده نشده است. شبکهٔ عصبی به حجم دادهٔ زیادی نیاز

Archive of SID

دارد که بتواند ساختارها را یاد بگیرد. با توجه به کم‌بودن حجم داده با برچسب‌های زمان، تاریخ، درصد و مقادیر پولی همان‌طور که انتظار می‌رفت، کیفیت سامانهٔ یادگیری‌شده برای این برچسب‌ها بسیار پایین ولی در مورد سه برچسب اصلی، کیفیت قابل قبول است.

(جدول-۶):نتایج سامانهٔ مبتنی‌بریادگیری عمیق بدون ویژگی- سطح واژه

(Table-6):Deep-learning based system without features – word level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.86	0.77	0.81
مکان	0.82	0.77	0.79
سازمان	0.83	0.79	0.81
زمان	0.60	0.35	0.44
تاریخ	0.72	0.67	0.70
مبالغ مالی	0.82	0.45	0.58
درصد	0.78	0.42	0.55
کل	0.82	0.74	0.78

(جدول-۷):نتایج سامانهٔ مبتنی‌بر یادگیری عمیق بدون ویژگی- سطح عبارت

(Table-7):Deep-learning based system without features – phrase level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.71	0.67	0.69
مکان	0.70	0.76	0.73
سازمان	0.55	0.63	0.59
زمان	0.14	0.13	0.13
تاریخ	0.39	0.50	0.44
مبالغ مالی	0.03	0.03	0.03
درصد	0.04	0.05	0.04
کل	0.58	0.63	0.61

همان‌طور که مشخص است، مدل شبکهٔ عصبی طبق ارزیابی در سطح عبارت، موفق به یادگیری برچسب‌های درصد، مقادیر پولی و زمان نشده ولی نتایج ارزیابی‌های سطح واژه برای این سه برچسب بسیار متفاوت است. برای درک این مسأله عبارت پولی «هزار و ۳۳۶ دلار» را در نظر می‌گیریم. این عبارت باید به‌طور کامل برچسب مقادیر پولی بگیرد؛ ولی مدل LSTM فقط «۳۳۶ دلار» را برچسب مقادیر پولی زده و موفق به تشخیص «هزار و» نشده است. همچنین در مورد زمان، عبارت «پیش از ظهر» را در نظر می‌گیریم. مدل LSTM ظهر را به‌عنوان «زمان» شناسایی می‌کند، ولی واژگان «پیش» و «از» واژگان عمومی هستند که در بسیاری از رخدادهای خود جزئی از هیچ موجودیت اسمی نیستند. همین مسأله باعث شده است که در عبارات با طول بیش‌تر، بخشی از عبارت موجودیت اسمی تشخیص داده نشود که این مسأله منجر به کیفیت بسیار پایین در حالت ارزیابی سطح عبارت شده است. به‌ویژه توجه به این نکته نیز ضروری است که میانگین طول موجودیت‌های این سه برچسب بیشتر از چهار برچسب دیگر بوده است که همین مسأله احتمال آن را که عبارت به‌طور کامل شناسایی شود، کاهش می‌دهد.

(جدول-۸):نتایج سامانهٔ مبتنی‌بریادگیری عمیق با استفاده از ویژگی خروجی CRF- سطح واژه

(Table-8):Deep-learning based system results with CRF output as a feature – word level

نوع موجودیت	دقت	فراخوانی	کل
شخص	0.82	0.91	0.86
مکان	0.82	0.88	0.85
سازمان	0.81	0.89	0.85
زمان	0.67	0.83	0.74
تاریخ	0.81	0.85	0.83
مبالغ مالی	0.85	0.94	0.89
درصد	0.90	0.93	0.92
کل	0.81	0.89	0.85

(جدول-۹):نتایج سامانهٔ مبتنی‌بریادگیری عمیق با استفاده از ویژگی خروجی CRF- سطح عبارت

(Table-9):Deep-learning based system results with CRF output as a feature – phrase level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.74	0.79	0.77
مکان	0.84	0.81	0.83
سازمان	0.71	0.70	0.71
زمان	0.54	0.51	0.52
تاریخ	0.72	0.68	0.70
مبالغ مالی	0.57	0.59	0.58
درصد	0.88	0.82	0.85
کل	0.79	0.77	0.78

(جدول-۱۰):نتایج سامانهٔ ترکیبی با استفاده از ویژگی شمارهٔ خوشه- سطح واژه

(Table-10):Hybrid system results with cluster number as a feature – word level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.89	0.90	0.89
مکان	0.87	0.85	0.86
سازمان	0.87	0.87	0.87

Archive of SID
www.SID.ir

زمان	0.91	0.66	0.76
تاریخ	0.89	0.80	0.87
مبالغ مالی	0.98	0.84	0.90
درصد	0.98	0.91	0.95
کل	0.88	0.86	0.87

(جدول-۱۱):نتایج سامانهٔ ترکیبی با استفاده از ویژگی شمارهٔ خوشه- سطح عبارت

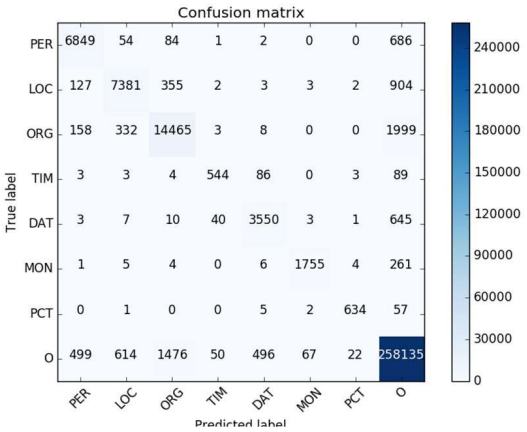
(Table-11):Hybrid system results with cluster number as a feature – phrase level

نوع موجودیت	دقت	فراخوانی	F1
شخص	0.82	0.84	0.83
مکان	0.82	0.85	0.84
سازمان	0.75	0.77	0.76
زمان	0.74	0.63	0.68
تاریخ	0.79	0.72	0.77
مبالغ مالی	0.83	0.74	0.79
درصد	0.92	0.90	0.91
کل	0.80	0.81	0.80

همان‌طور که مشخص است، استفاده از خروجی **CRF** به‌عنوان ویژگی در شبکهٔ عصبی باعث شده است که حتی در مورد برچسب‌هایی که دادهٔ آموزشی کمی داشتند به کیفیت بالای پنجاه در سطح عبارت و بالای هشتاد در سطح واژه برسیم. همچنین افزودن ویژگی شمارهٔ خوشه تأثیر قابل توجهی بر کیفیت سامانه در حالت ارزیابی در سطح واژه داشته است. برای نشان‌دادن نحوهٔ کارکرد این ویژگی به ذکر دو مثال می‌پردازیم. در عبارت «از قول موگرینی»، کلمهٔ «موگرینی» با سامانهٔ شبکهٔ عصبی برچسب **O**، ولی بعد از استفاده از ویژگی شمارهٔ خوشه، به‌درستی برچسب **B_PER** گرفته است. برای درک این که چه اتفاقی افتاده کافی است به واژگانی که در خوشهٔ «موگرینی» قرار گرفته‌اند توجه کنیم. تعدادی از این واژگان از این قرار هستند: مایر، ارنست، روانچی، کمالوندی، استفان و ترودو. از آن جا که واژگان این دسته همگی برچسب شخص دارند، به‌درستی در اینجا «موگرینی» برچسب شخص گرفته است.

۳-۵- تحلیل خطا

برای تحلیل نتایج و شناخت منابع خطای سامانه می‌توان به ماتریس درهم‌ریختگی[1] برچسب‌زنی سامانه مراجعه کرد. این ماتریس در شکل (۴) نمایش داده شده است (برای ترسیم این ماتریس از ابزار scikit-learn[2] در زبان پایتون استفاده شده است).



(شکل-۴): ماتریس درهم‌ریختگی سامانهٔ تشخیص موجودیت اسمی (ستون‌ها برچسب‌های صحیح و ردیف‌ها برچسب‌های تشخیص‌داده شده است)

(Figure-4): Confusion matrix of the named entity recognition system (columns are the true labels and rows are the predicted labels)

در شکل (۴) نتایج سامانهٔ ترکیبی نهایی نشان داده شده است. با توجه به این ماتریس و خروجی‌های سامانه، به شناسایی منابع اصلی خطا و درصورت امکان وضع قوانین پس‌پردازشی برای رفع برخی خطاها می‌پردازیم.

منابع اصلی خطا از این قرار هستند:

۱. خطای نوع نخست: واژگانی که جزئی از یک موجودیت بوده‌اند، ولی به‌اشتباه برچسب O گرفته‌اند.

یکی از دلایل اصلی بروز خطای نوع نخست، وجود واژگانی است که واژگان عمومی حساب می‌شوند و خاص موجودیت نیستند. در این صورت در تعداد زیادی از تکرارهای خود در دادهٔ آموزش با برچسب O ظاهر شده‌اند و همین امر باعث بروز خطا می‌شود. برای مثال کلمهٔ «انتقال» یک کلمهٔ عمومی، ولی در «انتقال خون» جزئی از یک موجودیت سازمان است.

در موارد دیگر، برخی ساختارهای پیچیده توسط سامانه تشخیص داده نشده‌اند. برای مثال، ساختار زمان یا تاریخ «از ... تا ...» (مثل عبارت موجودیت اسمی زمان «از ساعت ۲۰ تا ۲۱» و موجودت اسمی تاریخ «از هفتم تا سیزدهم بهمن») به طور کامل شناسایی نشده است و واژه‌های «از» و «تا» در همه جا برچسب O دارند. برای تشخیص این ساختار، یک قاعدهٔ پس‌پردازشی به سامانه اضافه کرده‌ایم تا عبارات با الگوی «از + برچسب زمان + تا + برچسب زمان» را به‌طور کامل با برچسب زمان و عبارات با الگوی «از + برچسب تاریخ + تا + برچسب تاریخ» را با برچسب تاریخ مشخص کند.

[1] Confusion matrix

[2] http://scikit-learn.org/

Archive of SID
www.SID.ir

۲. خطای نوع دوم: واژگانی که جزئی از هیچ موجودیتی نبوده‌اند، ولی به‌اشتباه برچسب موجودیت گرفته‌اند.

یکی از دلایل اصلی بروز خطای نوع دوم، وقوع واژگانی است که در تعداد زیادی از وقوع‌های خود در متن برچسب موجودیت داشته‌اند. به‌عنوان مثال، کلمهٔ «دفتر» در بسیاری از وقوع‌های خود درواقع جزئی از یک موجودیت سازمان بوده است، ولی در جملهٔ «این برنامه در دفتر بیروت تولید می‌شود.»، دفتر برچسب سازمان نباید بگیرد. واژگانی از این دست مثل «شهر»، «استان» و «کشور» در وقوع‌های زیادی به‌اشتباه برچسب موجودیت گرفته‌اند. در یک قانون پس‌پردازشی به حذف این گونه واژگان در شرایطی که به‌تنهایی آمده باشند و کلمهٔ بعد از آن‌ها برچسب موجودیت نگرفته باشد، می‌پردازیم.

۳. خطای نوع سوم: خطا در تشخیص بین برچسب‌های مکان و سازمان.

خطای نوع سوم درواقع یک خطای مفهومی است. به‌عنوان مثال، «دانشگاه تهران» در عبارت «او به دانشگاه تهران رفت.»،موجودیت «مکان» و در عبارت «او در دانشگاه تهران کار می‌کند.» موجودیت سازمان است که تشخیص این مسأله برای سامانه نیاز به درک سطح بالاتری از معنای جمله دارد. در سامانهٔ ما این گونه موجودیت‌ها برچسب «مکان» گرفته‌اند.

نوع دیگر این خطا زمانی رخ می‌دهد که نام یک مکان جزئی از نام سازمان باشد مانند عبارت‌های موجودیت سازمان «تیم ملی کشتی فرنگی ایران» و «وزارت کشور عربستان» که در آن‌ها «ایران» و «عربستان» باید برچسب سازمان می‌گرفته‌اند، اما به‌اشتباه به‌عنوان «مکان» برچسب خورده‌اند. برای رفع این خطا، برچسب‌های مکانی را که بلافاصله بعد از یک برچسب سازمان آمده‌اند و با آن بخش در یک عبارت اسمی قرار داشته‌اند، به برچسب «سازمان» تغییر می‌دهیم.

۴. خطای نوع چهارم: خطا در تشخیص بین برچسب‌های زمان و تاریخ.

تشخیص موجودیت‌های زمان و تاریخ در بعضی موارد برای کاربر انسانی نیز دشوار است. به‌عنوان مثال «از شامگاه سه‌شنبه تا صبح چهارشنبه» به‌دلیل آن که دربرگیرندهٔ واحد زمانی بیش از یک روز است باید به‌عنوان یک موجودیت «تاریخ» تشخیص داده شود؛ ولی سامانه به اعتبار کلمهٔ «شامگاه» که در بیش‌تر عبارات بیان‌گر زمان است، این عبارت را به‌عنوان زمان تشخیص داده است. در برخی کاربردهای دیگر مانند «صبح روز شنبه ۲ تیر ۱۳۹۷»، مدل نهایی به بخش «صبح روز شنبه» برچسب زمان و به بخش «۲ تیر ۱۳۹۷» برچسب تاریخ داده است که با یک قاعدهٔ پس‌پردازشی، برچسب تاریخ بلافاصله بعد از یک برچسب زمان را به زمان تبدیل می‌کنیم.

۵. خطای نوع پنجم و ششم: تشخیص نادرست مکان یا سازمان به‌عنوان شخص.

از موارد بروز اصلی این خطا زمانی است که نام شخص به‌عنوان اسم یک سازمان یا مکان آمده باشد. به‌عنوان مثال «مدرسهٔ شهید مطهری» باید برچسب مکان یا سازمان (وابسته به جمله) بگیرد که برچسب شخص گرفته است یا در عبارت «دفتر نماینده ویژه بان کی مون در امور سوریه»، کل عبارت باید برچسب سازمان بگیرد که سامانه به‌اشتباه «بان کی مون» را به‌عنوان شخص در نظر گرفته است. برای حل این مشکل، با یک قاعدهٔ پس‌پردازشی، بخشی از یک عبارت موجودیت که برچسب شخص گرفته است، اگر در میان برچسب‌های مکان (یا سازمان) قرار گرفته باشد یا بلافاصله بعد از یک برچسب مکان (یا سازمان) آمده باشد، تبدیل به برچسب مکان (یا سازمان) می‌شود.

۴-۵- مقایسه با سایر سامانه‌ها

برای مقایسهٔ نتایج ارزیابی روش پیشنهادی این پژوهش با کارهای موجود، دو روش پایه انتخاب کردیم:

الف) CRF:

مدل ترکیبی از قاعده‌محور و CRF که بر اساس پایان‌نامه [2] در شرکت آرمان‌رایان شریف[1] پیاده‌سازی شده و کیفیت آن از ابزارهای تشخیص موجودیت‌های اسمی پیشین زبان فارسی بهتر بوده است. داده‌ای که این ابزار روی آن آموزش داده شده است، با دادهٔ استفاده‌شده در پژوهش حاضر به‌طورکامل متفاوت بوده است (از نظر نوع، زمان، طول سندها و توزیع موضوعی). سامانهٔ پیاده‌سازی‌شده در شرکت آرمان‌رایان شریف، ترکیبی از روش‌های قاعده‌محور و CRF است و روی زیرمجموعهٔ برچسب‌خورده‌ای از پیکرهٔ بی‌جن‌خان آموزش داده شده است.

[1] http://armansoft.ir

Archive of SID
www.SID.ir

ب) ترکیب LSTM و CRF:

یک روش مستقل از زبان مبتنی بر ترکیب LSTM و CRF که در [18] ارائه شده است. این روش برای زبان‌های آلمانی، اسپانیایی، انگلیسی و هلندی مورد ارزیابی قرار گرفته است که طبق نتایج اعلام‌شده، برای دو زبان نخست از بهترین روش موجود بهتر بوده و برای دو زبان دوم هم فاصلهٔ چندانی با بهترین روش موجود نداشته است، با این تفاوت که بر خلاف بهترین روش‌های موجود قبلی، نیازی به تعریف و استخراج ویژگی‌ها ندارد. این روش در بسیاری مقالات زبان‌های دیگر به‌عنوان روش پایه در ارزیابی‌ها مورد استفاده قرار گرفته است. در این روش، مدل CRF به جای دنبالهٔ واژگان ورودی روی دنبالهٔ تولیدشده توسط مدل LSTM اعمال می‌شود.

در جدول (۱۲) نتایج ارزیابی سامانهٔ پیشنهادی با دو روش پایه روی سه برچسب اصلی مقایسه شده است. روش ترکیبی LSTM و CRF ارائه‌شده در [18] با آن که روی زبان‌های دیگر عملکرد بسیار خوبی نشان داده، روی مجموعه‌دادهٔ پیما برای زبان فارسی به کیفیت بسیار پایینی رسیده است. به‌نظر می‌رسد حجم مجموعه‌دادهٔ پیما برای آموزش شبکهٔ عصبی آن کافی نبوده و برای اجرای این روش، از کد نوشته‌شده توسط نویسندگان مقاله استفاده شده است[1].

(جدول-۱۲):مقایسهٔ روش پیشنهادی با کارهای موجود
(Table-12):Comparison of the proposed method with the previous works

روش		دقت	فراخوانی	F1
آرمان‌رایان شریف (CRF) [2]	سطح واژه	0.74	0.63	0.68
ترکیب LSTM و CRF در [18]	سطح واژه	0.81	0.38	0.52
	سطح عبارت	0.48	0.44	0.46
روش پیشنهادی	سطح واژه	0.87	0.87	0.87
	سطح عبارت	0.79	0.82	081

همان‌طور که از نتایج مندرج در جدول (۱۲) مشخص است، روش پیشنهادی در این پژوهش روی سه برچسب اصلی در هر دو سطح واژه و عبارت، با معیارهای دقت، فراخوانی و F1 بهتر از روش‌های پایه عمل کرده است.

[1] https://github.com/glample/tagger

۶- نتیجه‌گیری و کارهای آینده

مسألهٔ تشخیص موجودیت‌های اسمی به‌عنوان یک گام پیش‌پردازشی برای بسیاری از مسائل پردازش زبان طبیعی مطرح است. این مسأله در زبان فارسی به‌دلیل نبود یک مجموعه‌دادهٔ استاندارد، کمتر مورد پژوهش قرار گرفته است. در این پژوهش با بررسی ویژگی‌های مجموعه‌داده‌های استاندارد موجود در سایر زبان‌ها و به‌ویژه زبان انگلیسی، تلاش کردیم که مجموعه دادهٔ استانداردی برای زبان فارسی ایجاد کنیم. با توجه به این که در بسیاری از مجموعه‌داده‌های استاندارد موجود در زبان انگلیسی از متون خبری به‌عنوان منبع ساخت استفاده شده است، با جمع‌آوری متون خبری خبرگزاری‌های متعدد و با استفاده از برچسب‌گذارهای انسانی، به ساخت مجموعه‌دادهٔ استانداردی برای زبان فارسی اقدام کردیم؛ سپس با مطالعهٔ سامانه‌های طراحی‌شده برای تشخیص موجودیت‌های اسمی در زبان انگلیسی و با استفاده از مجموعه‌دادهٔ تهیه‌شده به طراحی سامانه‌ای برای زبان فارسی پرداختیم. سامانهٔ ما از بخش‌های متعدد شامل قاعده‌محور (تشکیل‌شده از فهرست‌محور و قواعد منظم)، آماری و یادگیری عمیق تشکیل شده است. با انجام آزمایش‌هایی روی بخش‌های مختلف سامانه طراحی‌شده، با سامانه ترکیب‌شده از CRF و LSTM و با استفاده از ویژگی شمارهٔ خوشه به بهترین نتیجه بر اساس معیار F1 دست یافتیم.

با توجه به دقت و فراخوانی حاصل از ارزیابی روش‌های مختلف روی مجموعه‌دادهٔ فراهم‌شده و با مقایسهٔ حجم و روش برچسب‌زنی آن با مجموعه‌داده‌های استاندارد موجود در زبان انگلیسی می‌توان به این نتیجه رسید که مجموعه‌داده برای انجام پژوهش‌های بیشتر در زمینهٔ تشخیص موجودیت‌های اسمی مناسب است. در این جا سامانه‌ای اولیه با استفاده از ویژگی‌های ابتدایی برای تشخیص موجودیت‌های اسمی پیشنهاد داده و آن را مورد ارزیابی قرار دادیم. در ادامهٔ این پژوهش می‌توان شبکه‌های یادگیری عمیق پیچیده‌تری را مورد آزمون قرار داد. به‌ویژه می‌توان مانند پژوهش‌های انجام‌گرفته در زبان انگلیسی از شبکه‌های CNN برای آموزش نمایش طیفی نویسه‌ها استفاده کرد.

تشکر و قدردانی

این پژوهش در راستای یکی از پروژه‌های طرح جویش‌گر بومی و تحت حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) انجام شده است. نویسندگان از حمایت‌های این پژوهشگاه قدردانی می‌کنند.

Archive of SID

www.SID.ir

7- References — ۷- مراجع

[۱] س.ع. اصفهانی، س. راحتی قوچانی و ن. جهانگیری، «سیستم شناسایی و طبقه‌بندی اسامی در متون فارسی»، پردازش علایم و داده‌ها، دوره ۷ شمارهٔ ۱، ۱۳۸۹.

[1] S. A. Esfahani, S. Rahati Ghouchani, and N. Jahangiri, "Persian named entity recognition and classification", *Journal of Signal and Data Processing*, vol. 7, no. 1, 2010.

[۲] م. عبدوس، «ارائه روشی جهت تشخیص واحدهای اسمی در زبان فارسی با استفاده از محتوای ویکی‌پدیای فارسی»، پایان‌نامهٔ کارشناسی ارشد، دانشگاه علم و صنعت ایران، تهران، ایران، ۱۳۹۴.

[2] M. Abdous, "Recognizing Persian Named Entities Using Persian Wikipedia Content", M.S Thesis, Iran University of Science and Technology, Tehran, Iran, 2015.

[۳] م. عبدوس و ب. مینایی بیدگلی، «بهبود شناسایی موجودیت‌های نامدار فارسی با استفاده از کسره اضافه»، پردازش علائم و داده‌ها، دورهٔ ۱۴، شمارهٔ ۴، ۱۳۹۶.

[3] M. Abdous and B. Minaei Bidgoli, "Improving Named Entity Recognition Using Izafe in Farsi", *Journal of Signal and Data Processing*, vol. 14, no. 4, 2017.

[۴] پ.س. مرتضوی و م. شمس‌فرد، «شناسایی موجودیت نام‌دار در متون فارسی»، پانزدهمین کنفرانس انجمن کامپیوتر ایران، تهران، ۱۳۸۸.

[4] P. S. Mortazavi, M. Shamsfard, "Named Entity Recognition in Persian Texts", *in 15th National CSI Computer Conference*, Tehran, Iran, 2009.

[5] F. Ahmadi and H. Moradi, "A Hybrid Method for Persian Named Entity Recognition," *in 7th Internatonal Conference on Information Knowledge Technology*, 2015.

[6] D. M. Bikel, S. Miller, R. M. Schwartz, and R. Weischedel, "Nymble: A High-Performance Learning Name-Finder", *in Proceedings of the fifth conference on Applied natural language processing*, pp. 194–201, 1997.

[7] A. Borthwick and J. Sterling, "NYU: Description of the MENE Named Entity System as used in MUC-7," *Proceedings of the 7th Message Understanding Conference (MUC-7)*, 1998.

[8] A. X. Chang and C. D. Manning, "TOKENS REGEX : Defining Cascaded Regular Expressions over Tokens," Stanford University Technical Report, 2004.

[9] A. Chinchor, "OVERVIEW OF MUC-7 / MET-2 Overviews of English and Multilingual Tasks," *in Proceedings of Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax*, Virginia, April 2, 1997.

[10] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *in Transactions of the Association for Computational Linguistics*, vol. 4 pp. 357 370, 2016.

[11] C. dos Santos and V. Guimar, "Boosting Named Entity Recognition with Neural Character Embeddings," *in Fifth Named Entity Recognition Workshop*, joint with 53rd ACL and the 7th IJCNLP, 2015, pp. 25 33.

[12] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," *in Proceedings of the 43rd annual meeting on association for computational linguistics*, 2005.

[13] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 2nd editio. Prentice-Hall, 2009.

[14] M. K. Khormuji and M. Bazrafkan, "Persian Named Entity Recognition based with Local Filters," *International Journal of Computer Applications*, vol. 100, no. 4, pp. 1–6, 2014.

[15] M. Konkol, T. Brychcín, and M. Konopík, "Latent semantics in Named Entity Recognition," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3470–3479, 2015.

[16] G. Kumaran and J. Allan, "Text Classification and Named Entities for New Event Detection," *in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 297–304.

[17] J. Lafferty and A. Mccallum, "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data Conditional Random Fields : Probabilistic Models for Segmenting and," *in Proceedings of the eighteenth international conference on machine learning*, ICML, 2001, vol. 1, no. June, pp. 282–289.

[18] G. Lample, M. Ballesteros, S. Subramaninan, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," *in Proceedings of NAACL-HLT 2016*, 2016, no. July.

[19] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," *Proceedings of the seventh conference on*

Archive of SID

www.SID.ir

Natural language learning at HLT-NAACL 2003, vol. 4, 2003,pp. 188–191.

[20] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *in Proceedings of the International Conference on Learning Representations (ICLR 2013)*, 2013, pp. 1–12.

[21] S. Miller, J. Guinness, and A. Zamanian, "Name Tagging with Word Clusters and Discriminative Training," *in Proceedings of HLT-NAACL*, 2004.

[22] D. Molla, Me. van Zaanen, and D. Smith, "Named Entity Recogntion for Question Answering," *Proceedings of the 2006 Australasian language technology workshop*, vol. 4, 2006, pp. 51–58.

[23] D. Nadeau, "A Survey of Named Entity Recognition and Classification," *Linguisticae Investigationes*, no. 30, p. 3–26., 2007.

[24] M. Pasca, "Acquisition of Categorized Named Entities for Web Search," *Thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 137–145.

[25] T. Poibeau and L. Kosseim, "Proper Name Extraction from Non-Journalistic Texts," *in Proc. Computational Linguistics in the Netherlands*, 2001, pp. 144–157.

[26] H. Poostchi and M. Piccardi, "PersoNER: Persian Named-Entity Recognition," *in Proceedings of Coling 2016, the 26th International Conference on Computational Linguistics*, 2016, pp. 3381–3389.

[27] M. Seok, H. Song, C. Park, J. Kim, and Y. Kim, "Named Entity Recognition using Word Embedding as a Feature 1," *International Journal of Software Engineering and Its Applications*, vol. 10, no. 2, pp. 93–104, 2016.

[28] S. K. Sienˇ, "Adapting word2vec to Named Entity Recognition," *in Proceedings of the 20th Nordic Conference of Computational Linguistics*, NODALIDA 2015, 2015, pp. 239–243.

[29] B. M. Sundheim, "Overview of Results of the MUC-6 Evaluation," *in Proceedings of the 6th conference on Message understanding. Association for Computational Linguistics*, 1996, pp. 13–31.

[30] E. F. Tjong, K. Sang, and F. De Meulder, "Language-Independent Named Entity Recognition," in Proc. CoNLL, 2003.

[31] J. Turian, L. Ratinov, Y. Bengio, and J. Turian, "Word Representations: A Simple and General Method for Semi-supervised Learning," *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, no. July, pp. 384–394, 2010.

مهساسادات شهشهانی پژوهشگر مقطع دکترا در زمینهٔ بازیابی و استخراج اطلاعات در دانشگاه آمستردام (UvA) و فارغ‌التحصیل مقطع کارشناسی ارشد رشتهٔ مهندسی کامپیوتر-نرم‌افزار در دانشکده مهندسی برق و کامپیوتر پردیس دانشکده‌های فنی دانشگاه تهران است. این مقاله حاصل پژوهش وی در زمان کار روی پایان نامهٔ کارشناسی ارشد است.

نشانی رایانامهٔ ایشان عبارت است از:

m.shahshahani@uva.nl

مهدی محسنی دانشجوی دکترای دانشکده مهندسی برق و کامپیوتر دانشگاه تهران در رشته هوش مصنوعی و رباتیک است و به حوزه‌های پژوهشی پردازش زبان طبیعی و تولید پیکره‌های زبانی علاقه‌مند است. پژوهش‌های اخیر وی در زمینه تشخیص موجودیت‌های اسمی، خلاصه‌سازی خودکار متون و شبکه‌های عصبی end-to-end است. از سوابق فعالیت ایشان می‌توان به همکاری با کارگروه خط و زبان فارسی و آزمایشگاه زبان‌شناسی دانشگاه تهران اشاره کرد. توسعه ابزارهای تحلیل زبانی و تولید منابع زبانی برای زبان فارسی از زمینه‌های خاص پژوهشی این پژوهش‌گر است.

نشانی رایانامهٔ ایشان عبارت است از:

mahdi.mohseni@ut.ac.ir

آزاده شاکری دانشیار دانشکده مهندسی برق و کامپیوتر پردیس دانشکده‌های فنی دانشگاه تهران است. ایشان مدرک دکترای خود را در سال ۱۳۸۷ از دانشگاه ایلینویز اوربانا - شمپین در آمریکا دریافت کرد. از زمان پیوستن به دانشکدهٔ مهندسی برق و کامپیوتر، وی سرپرستی آزمایشگاه پژوهشی سامانه‌های هوشمند اطلاعات را بر عهده دارد و در زمینه‌های مدیریت اطلاعات متنی، بازیابی اطلاعات، متن کاوی، و داده کاوی کار می‌کند.

نشانی رایانامهٔ ایشان عبارت است از:

shakery@ut.ac.ir

Archive of SID

www.SID.ir

هشام فیلی کارشناسی در گرایش نرم‌افزار را از دانشگاه صنعتی شریف در سال ۷۶ و مقطع ارشد در همان گرایش و مقطع دکترا در گرایش هوش مصنوعی را در سال‌های ۷۸ و ۸۵ در همان دانشگاه به اتمام رساند. از سال ۱۳۸۶ نیز به عضویت هیأت علمی دانشکده مهندسی برق و کامپیوتر پردیس دانشکده‌های فنی، دانشگاه تهران درآمد. با تأسیس آزمایشگاه پردازش هوشمند متن و زبان طبیعی، فعالیت‌های پژوهشی خود را در راستای هوش مصنوعی و به‌طور خاص پردازش هوشمند متن در دانشکده مهندسی برق و کامپیوتر ادامه داده است.

از سال ۹۳ تا کنون نیز هم‌زمان متصدی معاونت مرکز فناوری اطلاعات و فضای مجازی دانشگاه تهران است.

نشانی رایانامهٔ ایشان عبارت است از:

hfaili@ut.ac.ir

Archive of SID

www.SID.ir