سیستم شناسایی و طبقه‌بندی اسامی در متون فارسی

سید عبدالحمید اصفهانی[1]، سعید راحتی قوچانی[2] و نادر جهانگیری[3]
[1] گروه کامپیوتر، دانشگاه آزاد اسلامی واحد قاینات،
[2]دانشکده فنی و مهندسی، گروه هوش مصنوعی، دانشگاه آزاد اسلامی مشهد،
[3] دانشکده ادبیات و علوم انسانی، گروه زبان‌شناسی همگانی، دانشگاه فردوسی مشهد

چکیده:

یک سیستم شناسایی و طبقه‌بندی اسامی، سیستمی است که می‌تواند یک یا چند نوع از اسامی را در متن شناسایی و طبقه‌بندی کند این اسامی می‌توانند اسامی اشخاص، ارگان‌ها، شرکت‌ها، اسامی مکان‌ها (کشور، شهر، خیابان و مانند آن) اسامی زمان (تاریخ و ساعت) مقادیر مالی، درصدها و مانند آن باشد. هر چند که در دهۀ اخیر کارهای زیادی بر روی سیستم‌های شناسایی و طبقه‌بندی اسامی در زبان‌های مختلف و دامنه‌های مختلف انجام شده است، امّا در زبان فارسی، با توجّه به عدم وجود یک مجموعه داده کامل به همراه برچسب‌های غنی، تاکنون سیستمی برای طبقه‌بندی اسامی ایجاد نشده است. در این پژوهش از مجموعه دادۀ پژوهشکدۀ پردازش هوشمند علائم استفاده شده است. روش کار بدین صورت است که در ابتدا الگوریتم پیش‌پردازش اسامی را با استفاده از برچسب دستوری کلمات از داده‌ها جدا شده و سپس مصدرها، اسامی زمان، اسامی شمارشی، اعداد را هم از مجموعۀ داده حذف می‌کند. این کار باعث می‌شود تا حجم طبقات در داده‌های آموزشی متوازن‌تر گردد؛ در استخراج ویژگی از تابع N-gram استفاده شده است. پس از استخراج ویژگی، سیستم را با چهار طبقه‌بندی کنندۀ خطی، بیزین، نزدیک‌ترین همسایگی و شبکۀ عصبی آموزش می‌دهیم. عدم تنوع در اسامی زمان و هم‌چنین عدم اختلاط و یا اختلاط کم این اسامی با اسامی طبقات دیگر، این امکان را فراهم می‌کند تا بتوان با استفاده از یک سیستم مبتنی بر حافظه، اسامی زمان را در یک متن شناسایی کرد. با استفاده از شبکۀ عصبی نتایج بسیار مناسبی در جداسازی اسامی مکان و افراد از بقیۀ اسامی به‌دست آمده است (۹۹٪) و طبقه‌بندی کنندۀ KNN و طبقه‌بندی کنندۀ خطّی به‌طور میانگین اسامی مکان و افراد و اسامی عمومی طبقه‌بندی مقدار ۹۱ ٪ بر اساس معیار F-measure به‌دست آمده است. در طبقه‌بندی اسامی زمان با استفاده از یک فهرست کمکی مقدار ۹۶ ٪ بر اساس معیار F-measure به‌دست آمده است.

واژگان کلیدی: پردازش زبان طبیعی- شناسایی و طبقه‌بندی اسامی- انتخاب ویژگی- تابع N-gram

۱- مقدمه

سیستم شناسایی و طبقه‌بندی اسامی[1]، سیستمی است که وظیفۀ آن شناسایی و طبقه‌بندی اسامی موجودیت‌ها در یک متن است. این سیستم یکی از پرکاربردترین سیستم‌هایی است که در پردازش زبان به‌کار می‌رود. یک سیستم NER می‌تواند یک یا چند نوع از اسامی را در متن شناسایی و طبقه‌بندی کند. این اسامی می‌توانند اسامی اشخاص، ارگان‌ها، شرکت‌ها، اسامی مکان‌ها (کشور، شهر، خیابان و مانند آن) اسامی زمان (تاریخ و ساعت) مقادیر مالی، درصدها و مانند آن باشد. در سال‌های اخیر اغلب تحقیقات انجام شده مربوط به اسامی اشخاص، مکان‌ها و ارگان‌ها بوده است.*(Nadeau and Sekine, 2007)*

کاربرد اولیۀ NER برای استخراج اطلاعات از متن[2] (IE) بوده است *(Babych and Hartley, 2003)* و اصولاً پیدایش این شاخه در NLP جهت استخراج اطلاعات از متون بوده است؛ امّا به مرور، کاربردهای مهمّ دیگری برای NER تعریف گردیده است. در ادامه چند کاربرد مهم‌تر را به‌طور خلاصه مرور می‌کنیم:

طبقه‌بندی متون[3] : یک سیستم طبقه‌بندی کننده می‌تواند به جای این‌که ویژگی‌های خود را از کلّ متن

[1] Named Entity Recognition

[2] Information Extraction

[3] Text categorization

استخراج کند، ابتدا توسط NER اسامی را از متن استخراج کرده و سپس طبقه‌بندی را بر اساس این اسامی انجام دهد. استفاده از NER دقّت طبقه‌بندی‌کننده را تا حدّ قابل توجّهی افزایش می‌دهد (*Kumaran and Allan*).

سیستم پرسش و پاسخ[1]: پاسخ بسیاری از سؤالات یک متن مانند کی؟ کجا؟ چه زمانی؟ چه مقدار؟ چه کسی؟ در اسامی حاضر در متن قرار دارد که می‌توان آن‌ها را توسط یک سیستم NER از متن استخراج کرد (*Moll´a, et al., 2006*).

خلاصه‌سازی متون[2]: داده‌های به‌دست آمده از یک سیستم NER شامل اطّلاعات کلی از یک متن می‌باشند که در خلاصه‌کردن متون بسیار مفید است(*Hassel, 2003*).

بهینه‌کردن جستجو: زمانی که ما یک کلمه را جستجو می‌کنیم با توجّه به فرم و نقش آن نتایج متفاوتی به‌دست می‌آید. به‌طور مثال اگر "استاد یوسفی" را جستجو کنیم منظورمان اسم شخص است و نتایجی مانند "خیابان استاد یوسفی" که اسم مکان است، غیرمفید هستند(*Pasca, 2004*). از سیستم NER در بخش‌های دیگر پردازش زبان مانند ترجمهٔ ماشینی[3]، شناسایی مرجع ضمیر هم استفاده می‌شود(*Babych and Hartley, 2003*).

یکی از اوّلین تحقیقات که در زمینهٔ NER انجام گرفته است مربوط به کار" *لیزا راو*" می‌باشد که در سال ۱۹۹۱ در هفتمین کنفرانس هوش مصنوعی IEEE ارایه گردیده است. (*Rau, 1991*) این مقاله یک سیستم جهت استخراج و شناسایی اسامی کمپانی‌ها معرّفی گردید. این سیستم برای شناسایی اسامی و طبقه‌بندی آن‌ها از توابع اکتشافی[4] و یک سری قوانین دستی بهره می‌گرفت.

پس از این مقاله تا سال ۱۹۹۵ چندین مقالهٔ دیگر در زمینهٔ NER ارایه شد. در اکثر این مقالات از یک‌سری فهرست‌های مربوط به اسامی ارگان‌ها، افراد، مکان‌ها و مانند آن در شناسایی و طبقه‌بندی اسامی استفاده می‌کردند و در این سیستم‌ها قدرت سیستم، رابطهٔ مستقیم با فهرست‌های مورد استفاده داشت. در سال ۱۹۹۹ "*آندری میخو*" سیستمی را معرّفی کرد که در آن از فهرست اسامی استفاده نشده بود. (*Mikheev, et al., 1999*) "*دیمیترا*" از سیستم‌های NER در شناسایی و طبقه‌بندی اسامی در متون اقتصادی یونان استفاده کرده است. این سیستم اسامی را بر اساس گرامر شناسایی و طبقه‌بندی می‌کند. (*Farmakiotou, et al. 2000*) "*های‌لونگ*" در شناسایی و طبقه‌بندی اسامی، سیستمی را معرّفی کرد که از ویژگی‌های عمومی در طبقه‌بندی استفاده می‌کرد. (*Leong and Ng, 2002*). "*دانگ و ژیان سو*" از یک طبقه‌بندی‌کننده HMM بر پایهٔ برچسب‌زنندهٔ قطعه در سیستم NER خود استفاده کردند (*Zhou and Su, 2002*). "*ژانگ ولی*" سیستمی را برای شناسایی و طبقه‌بندی اسامی در زبان چینی ارایه دادند (*Zhang, et al. 2003*). "*فلوریا*" برای بالابردن دقّت در سیستم NER از ترکیب طبقه‌بندی کننده‌ها استفاده کرد. برای این کار از چهار طبقه‌بندی کنندهٔ متفاوت استفاده کردند. (طبقه‌بندی کنندهٔ خطّی، Max-Entropy، یادگیری تقویتی، HMM (*Florian, et al., 2003*) "*گرور*" کاربرد سیستم NER را در بازیابی اطلاعات از متون تاریخی مورد بررسی قرار داد (*Grover, et al., 2007*). "*بنجامین فاربر*" با همکاری "*نزار حبشه*" از یک برچسب‌زنندهٔ ریخت‌ساختی برای بهبود سیستم‌های NER در زبان عربی استفاده کردند که باعث کاهش ۱۴ ٪ خطا در این سیستم‌ها گردید.(*Farber, et al., 2008*) "*علی‌الصبا*" برای شناسایی اسامی درست در زبان عربی از یک تابع اکتشافی استفاده کرده است که اسامی افراد را در متون عربی شناسایی و استخراج می‌کند (*Elsebai, 2008*). "*خالد شالان*" هم یک سیستم را جهت شناسایی اسامی افراد در زبان عربی ارایه کرد که بر پایهٔ گرامر زبان عربی عمل می‌کند (*Shaalan and Raza, 2009*). "*ابراهیم الخراشی*" یک روش جدید در سیستم‌های NER را معرفی کرد که به جای شناسایی و طبقه‌بندی اسامی، اسامی افراد را بر اساس ریشهٔ اسم و گرامر زبان عربی، تولید و شناسایی می‌کرد (*Alkharashi, 2009*).

بسیاری از کارهای انجام شده، مخصوص یک دامنهٔ خاص از متون می‌باشد. این تحقیقات هم در متون رسمی مانند روزنامه‌ها، متون علمی، کتاب‌ها و مانند آن و هم در متون غیررسمی مانند نامه‌های الکترونیکی که نسبت به متون رسمی بی‌قاعده‌تر است، انجام شده است. آقای "*چانگ و سانگ*" برای استخراج اسامی از متون غیر رسمی از الگوریتم‌های بیشترین شباهت و حوزه‌های تصادفی شرطی (CRF) استفاده کردند (*Chang and Sung, 2005*). "*مینکو و وانگ*" هم سیستمی جهت استخراج نام‌ها از پست الکترونیکی طراحی کردند (*Minkov, et al., 2005*). یکی از

[1] Question Answering
[2] text summarization
[3] Machine Translation
[4] heuristic

حوزه‌هایی که در آن تحقیقات زیادی انجام شده است، متون پزشکی است. در این متون با استفاده از NER اطّلاعات پزشکی افراد استخراج و طبقه‌بندی می‌شود. آقای *"بورستل"* برای سیستم شناسایی و طبقه‌بندی اسامی در متون پزشکی از طبقه‌بندی کنندهٔ CRF جهت طبقه‌بندی کلمات استفاده کرده است *(Settles, 2004)*. *"رومان کلینگر و کریستف فردریک"* در مقالهٔ خود از سیستم شناسایی و طبقه‌بندی ژن‌ها و فرآورده‌های ژنی در متون استفاده کردند آن‌ها با استفاده از دو طبقه‌بندی کنندهٔ CRF (یکی برای عبارات کوتاه و یکی برای عبارات بلند) داده‌ها را طبقه‌بندی کردند *(Klinger, et al., 2007)*. *"کیم و سامگ"* یک روش نو بر پایهٔ آموزش فعال بر اساس بیشترین ارتباط مرزی (MMR)Maximal marginal relevenc برای شناسایی و طبقه‌بندی اسامی پزشکی ارایه کردند *(Kim, et al., 2006)*. هر چند که برای سیستم‌های شناسایی وطبقه‌بندی اسامی در زبان‌های مختلف و دامنه‌های مختلف کارهای زیادی انجام شده است. امّا با توجّه به مشکلات و محدودیت‌هایی که در زبان فارسی وجود دارد و در بخش سوم این پژوهش مورد بررسی قرار گرفته است هنوز سیستمی که بتواند اسامی را از متون فارسی استخراج و طبقه‌بندی کند وجود ندارد.

ما در این پژوهش در بخش بعدی ویژگی‌هایی را که در سیستم‌های NER استفاده شده است، مرور کرده و سپس در بخش سوم تفاوت‌های یک سیستم NER فارسی با یک سیستم NER انگلیسی را بررسی کرده و با توجّه به این تفاوت‌ها یک بُردار ویژگی برای سیستم NER فارسی استخراج می‌کنیم. NER فارسی که بتواند اسامی افراد و مکان‌ها را تشخیص دهد. ما ابتدا اسامی را با استفاده از برچسب دستوری[1]، کلمات را جدا کرده سپس با استفاده از الگوریتم‌هایی مصدرها، اسامی زمان، اسامی شمارشی، اعداد را هم از مجموعه دادهٔ خود حذف می‌کنیم. سپس سیستم را با طبقه‌بندی کننده‌های خطّی، بیزین، نزدیک‌ترین همسایگی و شبکهٔ عصبی آموزش می‌دهیم. در بخش ششم سیستم شناسایی اسامی زمان را مورد بررسی قرار می‌دهیم.

۲- استخراج ویژگی برای سیستم شناسایی و طبقه‌بندی اسامی

ما در این بخش ویژگی‌هایی را که در اغلب سیستم‌های NER استفاده می‌شود مرور می‌کنیم. این ویژگی‌ها را می‌توان به سه قسمت تقسیم کرد: ۱- ویژگی‌های کلمه‌ای ۲- ویژگی‌های فهرستی ۳- ویژگی‌های سندی.

۲-۱- ویژگی‌های کلمه‌ای

این ویژگی‌ها مربوط به نویسه‌های سازندهٔ کلمه است و از روی شکل و ظاهر خود کلمه استخراج می‌گردد. این ویژگی‌ها توصیف‌گر حالت کلمه، نقش کلمه، حالت عددی و مانند آن است. ویژگی‌های زیر از این دسته‌اند: *(Nadeau and Sekine, 2007)*

ویژگی‌هایی که مربوط به بزرگ یا کوچک بودن حروف در کلمه است، مانند: ۱- شروع شدن کلمه با حروف بزرگ ۲- بزرگ بودن تمام حروف در کلمه ۳- وجود حروف بزرگ در وسط کلمه مانند eBay

ویژگی‌های عددی کلمات مانند ۱- الگوی اعداد به کار رفته (استفاده از ۲ یا ۴ عدد جهت نمایش تاریخ و مانند آن)۲- استفاده از اعداد رومی در کلمه ۳- اعداد شمارشی و یا ترتیبی ۴- استفاده از عدد در وسط کلمه مانند W3C

ویژگی‌های نویسهٔ کلمات مانند: ۱- ضمیر اوّل شخص ۲- ضمایر ملکی ۳- حروف یونانی

ویژگی‌های صرفی کلمات مانند: ۱- پسوند ۲-پیشوند ۳- ریشهٔ کلمات

ویژگی‌های تابعی مانند: ۱- تابع آلفا ۲- تابع N-gram ۳- استفاده از الگو برای کلمات

۴- طول کلمه استفاده از نقش کلمهٔ دستوری در جمله

استفاده از نویسه‌های غیر حرفی مانند: کاما یا نقطه در کلمات

۲-۲- ویژگی‌های فهرستی

یک‌سری دیگر از ویژگی‌هایی که برای کلمات یک جمله در سیستم‌های NER در نظر گرفته می‌شود، ویژگی‌هایی است که از یک‌سری فهرست‌ها استخراج می‌گردد؛ برای استخراج این ویژگی‌ها از فهرست‌هایی که شامل اسامی افراد، مکان‌ها، شهرها، کشورها، افراد و مانند آن است، استفاده می‌شود. وجود کلمه در هر یک از این فهرست‌ها، نشان‌دهندهٔ یکی از ویژگی‌های کلمه است. در بعضی موارد وجود یک کلمه در یک فهرست می‌تواند با احتمال، طبقهٔ آن کلمه را مشخّص کند، ولی به‌دلیل ابهام در نقش و جایگاه اسامی، این امکان وجود دارد که اسمی در یک فهرست قرار داشته باشد؛ ولی متعلّق به طبقهٔ آن فهرست نباشد. بزرگ بودن این فهرست‌ها باعث می‌شود تا ویژگی‌هایی با دقّت بالاتری استخراج گردد؛

[1] Part Of Speech (POS)

امّا در بسیاری موارد، محدودیت وجود دارد کـه نمـی‌تـوان از فهرست‌هایی بزرگ استفاده کرد. به‌همین دلیل از یک سـری روش‌ها در جستجو استفاده می‌شـود تـا فهرسـت‌هـا بتواننـد تعـداد کلمـات بیشـتری را تحـت پوشـش قـرار دهنـد *(Nadeau and Sekine, 2007)*.

۲-۳- ویژگی‌های سندی

این ویژگی‌ها مربوط به اطّلاعاتی از کلمـه اسـت کـه در کـلّ سند وجود دارد. این اطّلاعات از طریـق پـردازش کـلّ اسـناد به‌دست می آید. واضح است که اگر حجم اسناد و داده‌های ما زیـاد باشـد، ویژگـی‌هـای قـوی‌تـری اسـتخراج مـی‌گـردد. ویژگی‌های سندی، ماورای ساختار کلمه است و درواقع از یک بررسی آماری از کل سند به‌دست می آید. در ادامه چند نـوع از این ویژگی‌ها را بررسی می کنیم.

وجود چند شکل از کلمه در سند: اگر یک کلمه در یک سند به چند شکل مختلف ظـاهر شـود، ماننـد: boy و Boy آن‌گاه می‌توان نتیجه گرفت که شکل‌هایی دیگر کلمه مربوط به مسائل دیگری هم‌چون شروع جمله است با استفاده از این نوع ویژگی می‌توان تا حدّی از این ابهامات جلوگیری کرد.

ویژگـی محلـی: تعـداد تکـرار در مـتن یـا پـاراگراف، موقعیت کلمه در متن یا پاراگراف

۳- تفاوت‌های یک سیستم NER فارسی با سیستم NER انگلیسی

قبل از این‌که به بررسی و انتخاب ویژگی‌های لازم بـرای یـک سیستم شناسایی و طبقه‌بندی اسـامی فارسـی بپـردازیم، بـه بررسی مشکلات و تفاوت‌های ایـن سیسـتم بـا سیسـتم‌هـای NER انگلیسی زبان می‌پردازیم.

در اکثر زبان‌های لاتین و غربی کلمـاتی کـه بـا حـروف بزرگ نوشته می‌شوند، نامزدهای مناسبی جهت اسامی خاص هستند. در این زبان‌ها تفـاوت در شـکل حـروف (بـزرگ یـا کوچک بودن) علاوه‌بر این‌که در شناسایی اسامی به سیستم کمک می‌کند همـان‌طـور کـه در قسـمت قبـل گفتـه شـده بسیاری از ویژگی‌های مربوط به دسته‌بندی اسامی نیز بر این اساس استخراج شده است.

دومین مشکل عمده‌ای که در شناسـایی و طبقـه‌بنـدی اسامی وجود دارد، مربوط به رسم‌الخط زبان فارسی است، در رسم‌الخطی که توسّط فرهنگستان ادب و زبان فارسـی مـورد تأییـد و تصـویب قـرار گرفتـه اسـت، اغلـب پسـوندهـا و پیشوندهای یک کلمه حـدّ الامکـان بـه صـورت جـدا از هـم نوشته می‌شود، این امر باعث می‌شود که حدّ و مرز یک کلمه را نتوان بـه‌درسـتی تشـخیص داد و هـم‌چنـین پسـوندهـا و پیشوندهای یک کلمه را به‌درستی شناسایی کرد. برای مثـال جملهٔ " آقای حسین‌زاده گفت" از سه کلمـه تشـکیل شـده است و واژهٔ "زاده " نقش پسوند را در کلمه " حسـین‌زاده " دارد اما در جمله "حسین زاده زهراست " واژه " حسین " و " زاده " دو کلمه جدا و مستقل هستند، این مسئله علاوه بر این که شناسایی حدود کلمه را مشکل مـی‌کنـد در مـواردی که ما پیشوندها و پسوندها را به‌عنـوان یـک ویژگـی در نظـر می‌گیریم، مشکل‌ساز خواهد بود *(خ.فرشیدورد ۱۳۸۶)*.

هر چند که فرهنگستان ادب و زبان فارسـی اسـتفاده از نیم فاصله را برای جلوگیری از ایـن مشـکل پیشـنهاد کـرده است، ولی با توجّه به این‌که در بسیاری از متون فارسی ایـن مسئله کم‌تر رعایت می‌شود، هنوز ایـن شناسـایی پسـوندهـا برای سیستم وجود دارد (www.persianucudery.org).

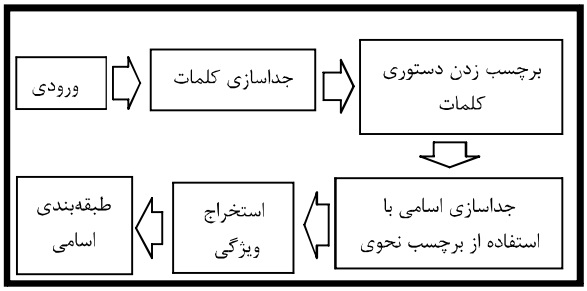
تفاوت عمدهٔ دیگری که بین زبـان فارسـی و انگلیسـی وجود دارد مربوط به تعداد صداهاسـت. در زبـان انگلیسـی ۹ واکه وجود دارد، درصـورتی‌کـه در زبـان فارسـی مـا ۶ واکـه بیشتر نداریم. این مسئله شاید به‌اندازهٔ دو مسـئلهٔ قبلـی کـه بیان شد در شناسایی و طبقه‌بندی اسامی مشکل ایجاد نکند. اما زمانی که بخواهیم از بعضی ویژگی‌های نویسه‌ای کلمه که در آن از صدای فونتیک کلمات اسـتفاده شـده اسـت و یـا از الگوریتم ساندکس کـه بـرای جسـتجو در فهرسـت اسـتفاده مـی‌شـود، اسـتفاده کنـیم بـا مشـکل مواجـه مـی‌شـویم. *(مشکوة الدینی ۱۳۸۵)*.

مشکل عمدهٔ دیگری که برای یک سیستم NER فارسی وجود دارد، عدم وجود یک مجموعه دادهٔ قوی جهت آمـوزش سیستم اسـت. همـان‌طـور کـه گفتـه شـد در سیسـتم‌هـای شناسـایی و طبقـه‌بنـدی کـه از روش‌هـای یـادگیری بـا سرپرستی، جهت آموزش سیستم اسـتفاده مـی‌شـود، وجـود یک مجموعه دادهٔ قوی می‌توانـد کمـک بزرگـی در بـالابردن صحّت و دقّت سیستم داشته باشد.

۴- سیستم شناسایی طبقه‌بندی اسـامی فارسی

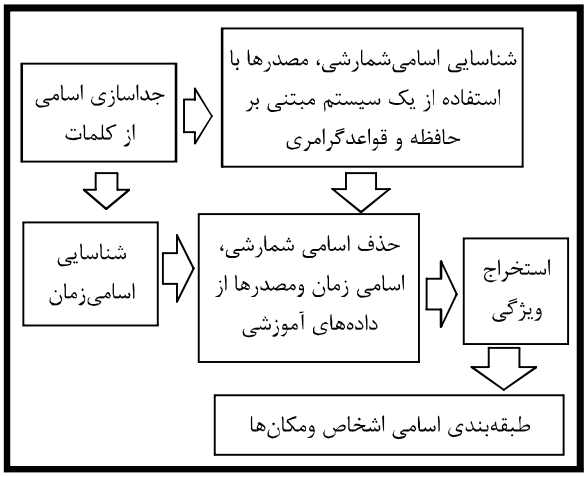
(شـکل ۱) مراحـل کـار را بـرای یـک سیسـتم NER نشـان می‌دهد. در این سیستم یک متن، به‌عنوان ورودی به سیستم ارایه می‌شود و اسامی به تفکیک طبقهٔ مربوط به آن به‌عنـوان خروجی سیستم به‌دست می‌آید. مراحل لازم جهت شناسـایی و طبقه‌بندی اسامی یک متن عبارتند از :

۱- یک متن با ورودی به سیستم ارایه می‌شود، مانند جملهٔ: "حسن شنبه به تهران می‌رود".
۲- با استفاده از یک نشان‌گذار کلمات در متن ورودی از هم‌دیگر جدا می‌شوند. (حسن/ شنبه/ به/ تهران/ می‌رود)
۳- با استفاده از برچسب‌زن دستوری نقش دستوری هر کلمه مشخّص می‌گردد (حسن/اسم | شنبه/اسم | به/حرف | تهران/اسم | می‌رود/فعل)
۴- با استفاده از برچسب دستوری کلمات، اسامی شناسایی شده و از بقیهٔ کلمات متن جدا می‌شوند. (حسن / شنبه / تهران)
۵- برای هر یک از اسامی بُردار ویژگی لازم استخراج می‌شود.
۶- با استفاده از یک دسته‌بند آماری، اسامی در طبقات مربوطه طبقه‌بندی می‌شوند.



(شکل ۱) مراحل شناسایی و طبقه‌بندی اسامی یک متن

با توجّه به این مطلب که در داده‌های آموزشی مورد استفاده در این تحقیق، کلمات یک متن شناسایی شده و دارای برچسب دستوری می‌باشند؛ بنابراین مراحل مربوط به این الگوریتم‌ها در اینجا مورد بحث قرار نمی‌گیرد (مراحل ۲ و ۳) درواقع تحقیق انجام شده در این پژوهش مربوط به مراحل بعد از برچسب‌زنی دستوری کلمات می‌باشد (مرحلهٔ ۴ به بعد) این مراحل در (شکل ۲) با جزییات بیشتر نشان داده شده است.



(شکل ۲) سیستم ارایه شده جهت شناسایی و طبقه‌بندی اسامی

در سیستم ارایه شده در این پژوهش، متون برچسب خورده به‌عنوان ورودی به سیستم داده می‌شوند. در اوّلین قدم، با استفاده از برچسب دستوری اسامی از بقیهٔ کلمات جدا می‌شوند. پس از جداسازی اسامی با استفاده از یک سیستم مبتنی بر حافظه، اسامی زمان از مجموعه اسامی استخراج می‌گردد. مصدرها و اسامی شمارشی هم با استفاده از یک سیستم مبتنی بر حافظه و قواعد گرامری در مجموعه اسامی شناسایی می‌شوند. پس از شناسایی اسامی زمان، شمارشی و مصدرها، این اسامی، از مجموعهٔ اسامی جدا می‌شوند. این کار جهت متوازن‌تر کردن حجم طبقات اسامی می‌باشد. پس از این کار ویژگی‌های لازم جهت آموزش سیستم از اسامی باقیمانده استخراج می‌گردد. در این مرحله علاوه‌بر ویژگی‌های نویسه‌ای، ویژگی‌های دیگری هم که از کلّ متون آموزشی به‌دست می‌آید، استخراج گردیده است. در استخراج ویژگی‌ها از تابع N-gram استفاده شده است. ویژگی‌های استخراج شده به یک دسته‌بند آماری داده می‌شود تا اسامی مکان و افراد از اسامی عمومی استخراج گردند.

۴-۱- فضای ویژگی برای سیستم شناسایی طبقه‌بندی اسامی فارسی

برای آموزش سیستم خود هر اسم را به‌عنوان یک دادهٔ ورودی در نظر گرفته و برای هر اسم با توجّه به برچسب‌هایی که در این مجموعه داده برای آن در نظر گرفته شده یک سری ویژگی‌هایی را مشخّص می‌کنیم. بعضی از ویژگی‌ها را هم که مربوط به خود کلمه است، می‌توان با استفاده از یک سری از الگوریتم‌های پردازش زبان به‌دست آورد.

۱) نقش دستوری کلمهٔ قبل: برای هر اسم، نوع دستوری کلمهٔ قبلی آن می‌تواند به‌عنوان یک ویژگی که در نوع طبقهٔ آن اسم نقش دارد، در نظر گرفته شود.
۲) نقش دستوری کلمه بعد: برای هر اسم، نوع دستوری کلمه از آن می‌تواند به‌عنوان یک ویژگی که در نوع طبقهٔ آن اسم نقش دارد، در نظر گرفته شود.
۳) طول کلمه
۴) جمع یا مفرد بودن اسم
۵) وجود پسوندهایی که در اغلب موارد باعث تولید اسم مکان می‌شوند مانند گاه، خانه، زار، آباد و مانند آن *(خ.فرشیدورد ۱۳۸۶)*

۶) وجود پسوندهایی که در اغلب موارد باعث تولید اسم خاص افراد می‌شوند؛ مانند پور، زاده، نژاد و مانند آن. *(خ.فرشیدورد ۱۳۸۶)*

۷) وجود "ی" نسبی در آخر کلمه

۸) درصد حضور کلمه در کلّ متون آموزشی به‌صورت اسم مکان

۹) درصد حضور کلمه در کلّ متون آموزشی به‌صورت اسم خاص افراد

دو ویژگی‌ای که در آخر ذکر گردید، با جدول یونیگرامی که از کلیهٔ متون به‌دست آمده، محاسبه می‌شود؛ به‌صورتی‌که ما برای محاسبهٔ هر کدام ابتدا تعداد دفعاتی را که این کلمه در متون آمده است، به‌دست آورده و سپس درصدی را که این کلمه در متون به‌صورت اسم مکان و یا اسم خاص آمده است، محاسبه می‌کنیم.

۴-۲- معرّفی مجموعه داده

همان‌طور که در قبل نیز بیان شد، اگر بخواهیم سیستم شناسایی و طبقه‌بندی متون را به روش با سرپرستی آموزش دهیم، نیاز به یک مجموعه داده برچسب خورده داریم؛ بنابراین ما داده‌های آموزشی خود را جهت آموزش سیستم شناسایی و طبقه‌بندی اسامی فارسی از متون که توسط پژوهشکدهٔ پردازش هوشمند علائم[1] برچسب خورده است استخراج می‌کنیم. در این متون کلمات یک جمله با توجّه به جایگاه نحوی آن کلمه به چند گروه تقسیم شده و برای هر گروه یک‌سری ویژگی‌هایی مختص آن گروه بیان شده است. برای مثال در این متون برای افعال، ویژگی‌هایی مانند زمان فعل، مرکب یا ساده بودن فعل، شخص فعل و مانند آن ذکر شده است و یا برای اسامی، ویژگی‌هایی مانند خاص یا عام بودن اسم، جمع یا مفرد بودن، اسم مکان ذکر گردیده است.

حجم متون برچسب خورده نزدیک به ده میلیون کلمه است. که از این ده میلیون کلمه، بیش از چهار میلیون را اسامی تشکیل می‌دهند. دراین متون ۳۸۰ هزارکلمه در نقش اسم مکان و ۲۳۴ هزار کلمه، هم در نقش اسم خاص وجود دارد. بنابراین در کل متون موجود تنها ده درصد اسامی اسم مکان هستند و هم‌چنین هفت درصد اسامی، اسم خاص هستند.

اما اگر بخواهیم کلمات منحصر به‌فرد را در این متون بشماریم، در ده میلیون کلمه تنها ۱۴۶ هزار کلمهٔ منحصر به‌فرد وجود دارد که از این رقم ۹۶ هزار تا را اسامی به خود اختصاص داده‌اند. ده هزار اسم از این مجموعه حدّاقل یک‌بار به‌عنوان اسم مکان و ۲۹ هزار اسم هم حدّاقل یک‌بار به‌عنوان اسم خاص در متون به‌کار رفته‌اند. ۱۶۰۰ اسم در این متون وجود دارد که در بعضی جملات به‌عنوان اسم مکان و در بعضی جملات، به‌عنوان اسم خاص استفاده شده‌اند.

مجموعه داده به سه طبقه تقسیم می شود: ۱- اسامی مکان ۲- اسامی خاص ۳- اسامی‌ای که نه اسم مکان هستند و نه اسم خاص .

یکی از مشکلات که در این داده‌ها دیده می‌شود، عدم توازن بین حجم طبقات در داده‌هاست. در بسیاری از متون، حجم اسامی خاص و اسامی مکان، درصد کمی از کلّ اسامی را تشکیل می‌دهند. هر چند که این مشکل تا حدّی مربوط به نوع خود داده‌هاست، ولی می‌توان با استفاده از الگوریتم‌هایی، حجم طبقات را متوازن‌تر کرد.

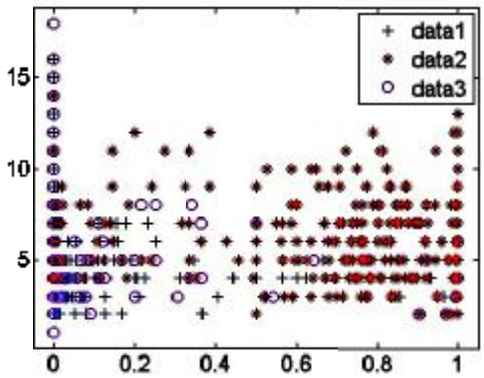
ما برای متوازن‌تر کردن حجم طبقات، اسامی‌ای را که به‌طور صددرصد اطمینان داریم که در متون، نه به‌صورت اسم مکان و نه به‌صورت اسم خاص افراد ظاهر خواهند شد از داده‌های خود حذف می‌کنیم. به‌عنوان مثال مصدرها، اسامی زمان، اسامی شمارشی و عددی و مانند آن اسامی هستند که در هیچ یک از متون به‌صورت اسم مکان و یا اسم زمان ظاهر نشده‌اند. در این پژوهش این کار به‌عنوان یک پیش‌پردازش بر روی داده‌ها انجام شده است، داده‌هایی که در این پژوهش از مجموعه اسامی حذف شده است، شامل مصدرها، اسامی زمان (اسامی ماه‌ها، روزهای هفته و ..)، اسامی شمارشی و عددی است. این کار باعث می شود تا علاوه‌بر متوازن‌تر شدن حجم طبقات، پانزده درصد از حجم کلّ داده‌ها کم شود. این اسامی را می‌توان به روش‌های زیر در یک متن شناسایی کرد: ۱- با استفاده از نوع دستوری کلمه ۲- با استفاده از فهرست‌های کمکی۳- با استفاده از گرامر زبان (جدول شماره ۱) داده‌ها را به تفکیک طبقهٔ آن‌ها مورد بررسی قرار داده است؛ در این جدول داده‌ها بر اساس ویژگی‌هایی که در بخش قبلی انتخاب شده، مورد بررسی قرار گرفته است.

(جدول ۱) ویژگی‌های استخراج شده به تفکیک هر طبقه

اسامی خاص افراد	اسامی مکان	اسامی عمومی	
٪۱۲	٪۱۰	٪ ۷۸	حجم داده‌ها
۵/۳۷	۵/۴۸	۵	میانگین طول هر کلمه
٪۹۶	٪۹۱	٪۸۵	درصد مفرد بودن کلمه
٪۲	٪۸۷	٪۰/۳	میانگین درصد حضور هر کلمه در متون به‌صورت اسم مکان
٪۷۰	٪۱	٪۱/۴	میانگین درصد حضور هر کلمه در متون به‌صورت اسم خاص
٪۱۰	٪۶/۷	٪۱۵	درصد اسامی که به "ی" ختم می‌شوند

[1] http:// www.rcisp.com

همان‌طور که در جدول بالا مشاهده می‌شود، با آن‌که در پیش‌پردازش داده‌ها سعی شد تا در حجم طبقات توازنی به‌وجود آید، امّا هنوز این مسئله به‌طور کامل حل نشده است؛ که البتّه همان‌طور که بیان شد این مشکل تا حدّی مربوط به نوع خود داده‌هاست و در بسیاری از متون، حجم اسامی خاص و اسامی مکان درصد کمی از کلّ اسامی را تشکیل می‌دهند. میانگین طول کلمات در بین اسامی خاص و اسامی مکان نسبت به کل اسامی بیشتر است. این ویژگی می‌تواند نقش خوبی در طبقه‌بندی اسامی داشته باشد. اسامی مکان به‌طور میانگین در ۸۷ درصد مواقعی که در بقیهٔ متون استفاده شده‌اند در نقش اسم مکان بوده‌اند. این عدد برای اسامی خاص هفتاد درصد است. این دو ویژگی (درصد حضور کلمه به‌صورت اسم مکان در کل متون و درصد حضور کلمه به‌صورت اسم خاص) می‌تواند نقش خوبی در تشخیص طبقه اسامی داشته باشد. در (شکل ۳) این مسئله نمایش داده شده است؛ امّا همان‌طور که در (جدول ۱) آمده است، در اسامی مکان کم‌تر از پسوندهای مکان‌ساز استفاده شده است که این به‌دلیل تنوع این اسامی در متون است (اسامی شهرها، محله‌ها، کشورها و مانند آن) این عدد در اسامی خاص، کم‌تر است. دلیل این مسئله این است که در اسامی خاص تنها پیشوندها و پسوندهایی که در بعضی نام‌های خانوادگی افراد وجود دارد، استفاده شده است (مانند پور، نژاد، مقدم و مانند آن) و پیشوند و پسوند که در دیگر کلمات خاص استفاده شده باشد وجود ندارد.



(شکل ۳) توزیع داده‌ها بر اساس ویژگی درصد حضور کلمه به صورت مکان و طول کلمه

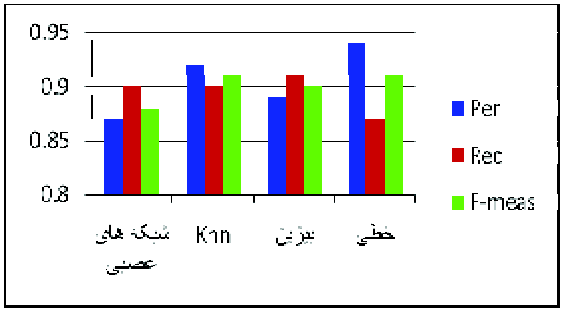
۵- طبقه‌بندی اسامی

پس از این‌که ویژگی‌های لازم برای آموزش سیستم استخراج گردید، در این مرحله با استفاده از طبقه‌بندی‌کننده، داده‌های خود را طبقه‌بندی می‌کنیم. در این مرحله از چهار طبقه‌بندی‌کنندهٔ متفاوت استفاده شده است.

طبقه‌بندی‌کنندهٔ خطّی با توجّه به داده‌های آموزشی، فضای ویژگی را به چند قسمت تقسیم می‌کند. مرز بین این فضاها تابعی خطّی می‌باشد. طبقه‌بندی با الگوریتم، نزدیک‌ترین همسایگی با ویژگی ورودی پنج طبقه هر داده را از روی پنج همسایهٔ نزدیک آن به‌دست می‌آورد. سومین الگوریتم مورد استفاده، طبقه‌بندی کننده بیزین است که داده‌ها را با استفاده از رابطهٔ بیزین طبقه‌بندی می‌کند. آخرین الگوریتم، طبقه‌بندی با استفاده از یک شبکهٔ عصبی سه‌لایه است که لایهٔ میانی آن دارای بیست گره می‌باشد. (جدول ۲) نتایج حاصل از طبقه‌بندی اسامی با طبقه‌بندی‌کننده‌های خطّی و غیر خطّی را نشان می‌دهد. همان‌طور که در (جدول ۲) می‌بینید طبقه‌بندی‌کننده‌های خطّی نتایج ضعیف‌تری نسبت به طبقه‌بندی‌کننده‌های غیر خطی دارند. لازم به ذکر است در هر الگوریتم ۷۵٪ از داده‌ها برای آموزش و ۲۵٪ از داده‌ها به‌عنوان داده‌های آزمون در نظر گرفته می‌شود.

(جدول ۲) آموزش سیستم با طبقه‌بندی‌کننده‌های مختلف

	شبکه عصبی	Knn	بیزین	خطی
Precision	٪۸۷	٪۹۲	٪۸۹	٪۹۴
Recall	٪۹۰	٪۹۰	٪۹۱	٪۸۷
F-measure	٪۸۸	٪۹۱	٪۹۰	٪۹۱



(شکل ۴) مقایسهٔ پارامترهای فراخوانی و دقّت در طبقه‌بندی‌کننده‌ها

(جداول ۳، ۴، ۵) میزان خطای هر طبقه‌بندی‌کننده را به تفکیک هر طبقه نشان می‌دهد. در تمامی روش‌ها میزان خطا در طبقهٔ سوم که مربوط به اسامی خاص افراد است نسبت به دو طبقهٔ دیگر بیشتر است. این مسئله را از چند جهت می‌توان مورد بررسی قرار داد: ۱- اسامی افراد دارای تنوع بیشتری نسبت به دو طبقهٔ دیگر هستند. ۲- بخش عمده‌ای از اسامی که در این طبقه وجود دارد، مربوط به زبان‌های دیگر است که وارد زبان فارسی شده و کم‌تر می‌توان برای این اسامی ساختار مشخصی تعریف کرد. این نکات باعث شده تا شناسایی این اسامی برای طبقه‌بندی‌کننده سخت‌تر شود.

(جدول۳) طبقه‌بندی داده‌ها با طبقه‌بندی‌کنندهٔ خطّی

	Precision	Recall	F-measure
اسامی عمومی	٪۹۵	٪۹۹	٪۹۷
اسامی مکان	٪۹۵	٪۸۸	٪۹۲
اسامی افراد	٪۹۰	٪۷۶	٪۸۴
کل اسامی	٪۹۴	٪۸۸	٪۹۱

(جدول ۴) طبقه‌بندی داده‌ها با طبقه‌بندی‌کنندهٔ بیزین

	Precision	Recall	F-measure
اسامی عمومی	٪۹۷	٪۹۶	٪۹۷
اسامی مکان	٪۹۱	٪۸۹	٪۹۰
اسامی افراد	٪۸۰	٪۸۸	٪۸۳
کل اسامی	٪۸۹	٪۹۱	٪۹۰

(جدول ۵) طبقه‌بندی داده‌ها با طبقه‌بندی‌کنندهٔ K-means با K=5

	Precision	Recall	F-measure
اسامی عمومی	٪۹۶/۵	٪۹۸	٪۹۷
اسامی مکان	٪۹۰	٪۹۱	٪۹۱
اسامی افراد	٪۹۰/۵	٪۸۰	٪۸۵
کل اسامی	٪۹۲	٪۹۰	٪۹۱

در تمامی روش‌ها میزان خطا در طبقهٔ اوّل نسبت به دو طبقهٔ دیگر کم‌تر است. این مسئله به این دلیل اتّفاق افتاده است که میزان داده‌ها در این طبقه نسبت به دو طبقهٔ دیگر بیشتر است و داده‌های این طبقه نسبت به دو طبقهٔ دیگر بیشتر آموزش دیده‌اند. این مسئله در نتایج طبقه‌بندی با شبکهٔ عصبی نمایان‌تر است. این به‌دلیل حساس‌بودن شبکهٔ عصبی به مسئلهٔ over training می‌باشد.

۶- طبقه‌بندی اسامی زمان

در شناسایی اسامی زمان از یک متن، از یک روش جدید و ساده‌تری استفاده می‌کنیم. در شناسایی اسامی زمان به چند دلیل استفاده از روشی که در شناسایی اسامی افراد و اسامی مکان‌ها به‌کار بردیم توصیه نمی‌شود؛ که مهم‌ترین آن‌ها را در ادامه توضیح می‌دهیم:

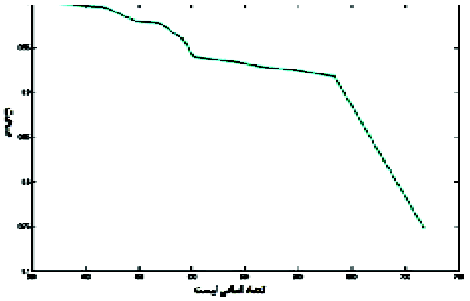
اوّلین مسئله که در مورد اسامی زمان با آن مواجه می‌شویم، حجم این اسامی در متون است. حجم این اسامی نسبت به کلّ اسامی موجود در متون بسیار اندک است. از چهار میلیون اسم که در کلّ مجموعه داده‌های ما وجود دارد، تنها ۱۵۶ هزار تا از این اسامی، اسم زمان هستند. به عبارتی کم‌تر از سه درصد از اسامی موجود در مجموعه داده‌ها، اسم زمان هستند. این مسئله باعث می‌شود تا یکی از طبقات موجود در داده‌های آموزشی حجم بسیار اندکی نسبت به کلّ داده‌های آموزشی داشته باشد و همان‌طور که در قبل هم بیان شد، این مسئله باعث انحراف در آموزش سیستم می‌شود. دومین مسئله که در مورد این اسامی وجود دارد، عدم تنوّع در این اسامی می‌باشد. اگر از ۱۵۶ هزار اسم زمان موجود در متون، اسامی تکراری را حذف کنیم، تنها ۷۱۷ اسم باقی می‌ماند؛ به‌عبارتی دیگر کلّ اسامی زمان را که در متون آموزشی ما وجود دارند، می‌توان در یک فهرست با طول ۷۱۷ اسم قرار داد. ویژگی دیگری که در مورد اسامی زمان وجود دارد، اختلاط کم اسامی این طبقه با اسامی مربوط به طبقات دیگر می‌باشد. به‌عبارت دیگر اگر کلمه در جمله به‌صورت اسم زمان بیاید آن‌گاه این کلمه کم‌تر به صورت‌های دیگر (اسم مکان، اسم فرد و مانند آن) در جملات دیگر متن ظاهر می‌شود. در فهرست بالا (کلّ اسامی زمان) از ۷۱۷ اسم موجود ۳۶۳ اسم وجود دارد که در صددرصد مواقع که در جملات حضور داشته‌اند، غالباً اسم زمان بوده‌اند؛ کلماتی مانند دیروز، امروز، قرن و مانند آن.

این سه مسئله باعث می‌شود تا ما در شناسایی اسامی زمان از روشی متفاوت با روشی که در شناسایی اسامی افراد و مکان‌ها داشتیم استفاده کنیم. دو ویژگی آخری که در مورد اسامی زمان بیان شد، یعنی عدم تنوع در اسامی زمان و هم‌چنین عدم اختلاط و یا اختلاط کم این اسامی با اسامی طبقات دیگر به ما کمک می‌کند تا بتوانیم در شناسایی این اسامی از یک سیستم مبتنی بر حافظه استفاده کنیم. یکی از مسائلی که در سیستم‌های مبتنی بر حافظه وجود دارد میزان حافظهٔ مصرفی و یا به‌طور دقیق‌تر در این سیستم انتخاب حجم مناسبی از اسامی به عنوان اسم زمان است.

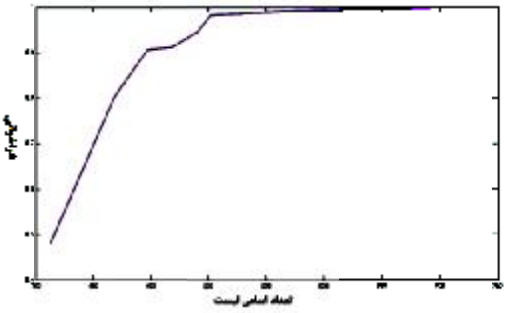
اگر ما تمام اسامی موجود در فهرست را در جملات به عنوان اسم زمان شناسایی کنیم، آن‌گاه تمامی اسامی زمان که در متون وجود دارد به‌درستی شناسایی می‌شود (Recall =٪۱۰۰) ولی مشکلی که به‌وجود خواهد آمد، این است که یک اسم موجود در فهرست در تمام دفعات حضورش در متن به صورت اسم زمان شناسایی می‌شود. برای مثال کلمه‌ای مانند "شهر" که در دفعات معدودی به‌صورت اسم زمان بوده است (از ۹۲۸۴ حضورش در متون آموزشی تنها ۲۷ بار غالباً اسم زمان بوده است) در کل متون به‌صورت اسم زمان شناسایی می‌شود، این امر باعث می‌شود تا سیستم، اسامی نامربوطی را هم به‌عنوان اسم زمان در نظر بگیرد در این حالت تنها ۷۴ درصد از اسامی که سیستم شناسایی کرده اسم زمان هستند (Precision=٪۷۴). از طرفی ما اگر فهرست خود را تنها به ۳۶۳ اسمی که در صددرصد مواقع به‌صورت اسم زمان در متن آمده‌اند محدود

کنیم، آن‌گاه سیستم تمام اسامی‌ای که شناسایی می‌کند، اسم زمان خواهند بود (Precision=٪۱۰۰). اما بسیاری از اسامی که در متن وجود دارند با سیستم شناسایی نمی‌شوند. این سیستم تنها ۴۸ درصد از اسامی زمان موجود در متن را شناسایی می‌کند (Recall=٪۴۸).

بنابراین هر چه فهرست اسامی محدودتر شود، معیار صحّت در سیستم بالا می‌رود و معیار فراخوانی کاهش می‌یابد و از طرف دیگر هر چه فهرست را بزرگ‌تر و کلّی‌تر در نظر بگیریم معیار فراخوانی سیستم بالاتر می‌رود و معیار صحّت سیستم کاهش می‌یابد. این مسئله در (اشکال ۵ و ۶) نشان داده شده است.

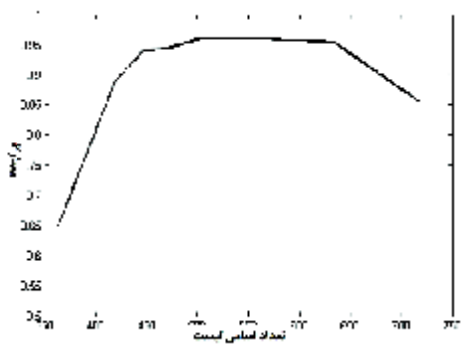


(شکل۵) میزان معیار صحت براساس میزان حجم فهرست



(شکل۶) میزان معیار فراخوانی براساس میزان حجم فهرست

برای این که بتوانیم فهرست مناسبی را جهت شناسایی اسامی استخراج کنیم، از معیار F-measure با ضریب آلفا ۰/۵ استفاده می‌کنیم. همانطور که در فصل قبل بیان شد، این معیار توازنی از مقادیر فراخوانی و صحّت را به ما می‌دهد. شکل زیر مقدار F-measure سیستم را بر اساس حجم فهرست نشان می‌دهد.



(شکل۷) مقدار F-measure براساس میزان حجم فهرست

همان‌طور که در (شکل ۷) مشاهده می‌شود، اگر ما اسامی‌ای را که در بیش از ۶۰ درصد مواقع در متن آمده‌اند و به‌صورت اسم زمان بوده‌اند، در فهرست خود قرار دهیم، آن‌گاه فهرستی با حجم ۴۹۱ کلمه خواهیم داشت و از طرفی سیستم با استفاده از این فهرست می‌تواند ۹۵ درصد اسامی زمان در متن آموزشی را استخراج کند و هم‌چنین ۹۶درصد از اسامی استخراج شده، اسم زمان هستند. معیار F-measure برای این سیستم ۹۶ درصد خواهد بود.

۷- نتیجه‌گیری

در این پژوهش ما ابتدا سیستم‌هایی را که برای شناسایی و طبقه‌بندی متون استفاده می‌شود، مرور کردیم و ویژگی‌هایی را که برای این سیستم‌ها از اسامی استخراج می‌شود، به‌اختصار بررسی کردیم. در ادامه مشکلات یک سیستم شناسایی و طبقه‌بندی اسامی فارسی را بررسی کردیم؛ سپس با توجّه به مجموعه داده‌های خود (داده‌های برچسب خوردهٔ پژوهشکدهٔ پردازش هوشمند علائم) برای سیستم خود یک‌سری ویژگی تعریف کردیم؛ در تعریف این ویژگی‌ها ما علاوه‌بر استفاده از نقش نحوی کلمات، از یک ویژگی فهرستی که از کلّ داده‌های برچسب خورده به‌دست می‌آید، هم استفاده کردیم؛ که این ویژگی فهرستی در طبقه‌بندی اسامی، به سیستم کمک بزرگی می‌کند. یکی دیگر از ویژگی‌هایی که در این پژوهش معرفی شده است و از پیشوند و پسوندهای اسم به‌دست می‌آید، وجود پسوندهایی است که بیشتر در اسامی مکان یا اسامی خاص افراد استفاده می‌شوند. این ویژگی در تشخیص اسامی مکان به سیستم کمک خوبی می‌کند؛ ولی چون اسامی خاص افراد تنوع بیشتری دارند، این ویژگی در تشخیص این اسامی کمک کم‌تری می‌کند.

در انتها سیستم به‌وسیلهٔ چند طبقه‌بندی‌کنندهٔ مختلف آموزش داده شده است. طبقه‌بندی‌کنندهٔ خطّی بیش‌ترین میزان صحّت را در بین طبقه‌بندی‌کننده‌ها دارد؛ ولی میزان فراخوانی در این طبقه‌بندی‌کننده پایین می‌باشد. این مسئله می‌تواند به‌دلیل دو ویژگی فهرستی که برای کلمات تعریف شده است (درصد حضور کلمه به‌صورت اسم زمان و مکان) باشد، که اگر طبقه‌بندی‌کننده، اسامی‌ای را که مقدار این دو ویژگی آن‌ها بالاست، انتخاب کند، آن‌گاه همان‌طور که در اسامی زمان نیز بیان شد، مقدار صحّت زیاد و مقدار فراخوانی کم می‌شود. طبقه‌بندی با استفاده از شبکهٔ عصبی،

ضعیف‌ترین نتایج را در بردارد، نامتوازن بودن حجم طبقات باعث این انحراف در طبقه‌بندی شده است و از آن‌جایی‌که شبکهٔ عصبی بیشترین حسّاسیت را به این مسئله دارد؛ این انحراف در شبکهٔ عصبی بیشتر مشاهده می‌شود. دقّت شناسایی اسامی عمومی در شبکهٔ عصبی بسیار بالاست (حجم بالای اسامی عمومی). در بین این چهار طبقه‌بندی‌کننده، الگوریتم نزدیک‌ترین همسایگی، بهترین نتایج طبقه‌بندی اسامی را دارد.

اگر بخواهیم مسئلهٔ طبقه‌بندی را به‌تفکیک طبقات مورد بررسی قرار دهیم، میزان خطا در طبقهٔ سوم که مربوط به اسامی خاص افراد است، نسبت به دو طبقهٔ دیگر بیشتر است؛ این مسئله به‌دلیل این‌که: ۱- اسامی افراد دارای تنوع بیشتری نسبت به بقیهٔ اسامی هستند ۲- بخش عمده‌ای از اسامی که در این طبقه وجود دارد، مربوط به زبان‌های دیگر است که وارد زبان فارسی شده و کم‌تر می‌توان برای این اسامی ساختار مشخّصی تعریف کرد.

هم‌چنین طبقهٔ اوّل نسبت به دو طبقهٔ دیگر درصد خطای کمتری دارد و این مسئله به‌دلیل حجم بالای داده‌های این طبقه در کلّ متون اتّفاق افتاده است. حجم بالای داده‌ها در این طبقه نسبت به دو طبقهٔ دیگر باعث شده است، داده‌های این طبقه نسبت به دو طبقهٔ دیگر بیشتر آموزش ببینند. همان‌طور که بیان شد این مسئله در شبکهٔ عصبی نمایان‌تر است.

این سیستم نسبت به سیستمی که خالد شالان برای استخراج اسامی افراد در متون عربی استفاده کرده بود، در تشخیص اسامی افراد ضعیف‌تر عمل می‌کند. در آن سیستم برای استخراج اسامی افراد از یک متن عربی از قوانینی که در لقب، کنیه و نسب افراد وجود دارد استفاده شده که همان‌طور که گفته شد، اسامی افراد در زبان فارسی ساختار مشخّصی ندارند، امّا در مجموع کلّ اسامی، دقّت این پژوهش بالاترست *(Shaalan and Raza, 2009)*. نتایج این پژوهش در مقایسه با پژوهش بنیجابا که در طبقه‌بندی اسامی عربی از یک مجموعه داده قوی و یک بردار ویژگی بهینه استفاده کرده است*(Benajiba, et al. 2008)*، اسامی خاص افراد با دقّت کمتری شناسایی شده، ولی دقّت شناسایی در اسامی مکان در روش ما بیشتر است. نتایج به‌دست آمده توسط الگوریتم نزدیک‌ترین همسایگی نسبت به نتایج سیستم آقای علی‌الصبا که بر روی متون عربی انجام شده است، دو درصد بهبود را نشان می‌دهد *(Elsebai, 2008)*.

۸- پیشنهادها

همان‌طورکه در بالا بیان شد یکی از مشکلات موجود برای این سیستم‌ها، نامتوازن بودن حجم طبقات در داده‌ها موجب می‌شود تا داده‌های یک طبقه نسبت به بقیهٔ طبقات بیشتر آموزش ببیند و دقّت در طبقه‌بندی داده کاهش یابد؛ هر چند که این مشکل تا حدّی مربوط به نوع خود داده‌هاست؛ یعنی اگر در یک متن عمومی این بررسی صورت گیرد، حجم اسامی مکان و افراد نسبت به کلّ متن بسیار اندک است؛ ولی می‌توان با استفاده از الگوریتم‌هایی حجم طبقات را متوازن‌تر کرد. یکی از این روش‌ها که حذف یک سری از اسامی عمومی از آموزش سیستم است در این پژوهش بررسی گردید.

اگر بتوانیم با استفاده از یک فهرست، اسامی عمومی پرتکرار را از داده‌ها، قبل از طبقه‌بندی حذف کنیم با توجّه به متوازن‌تر شدن حجم طبقات، انتظار می‌رود تا دقّت سیستم نیز افزایش یابد.

با توجّه به همبستگی کلمات در یک جمله و نیز همبستگی جملات اگر بتوان از الگوریتم‌هایی که توالی بین داده‌ها را هم بررسی می‌کند، مانند مدل مخفی مارکوف استفاده کرد، انتظار می‌رود تا دقّت سیستم افزایش یابد. هرچند که در انتخاب ویژگی‌ها کلمات بعد و قبل هر اسم استفاده شده است.

همان‌طور که در نتایج دیدیم روش‌های مختلف طبقه‌بندی، نتایج مختلفی را برای طبقات مختلف به همراه داشت. با توجّه به این نتایج، انتظار می‌رود استفاده از روش‌های مختلف در ترکیب طبقه‌بندی‌ها، دقّت نهایی افزایش یابد.

اگر در استخراج ویژگی‌ها از دیگر توابع موجود در پردازش زبان‌های طبیعی، مانند تابع آلفا استفاده شود، موجب خواهد شد تا بُردار ویژگی تمایز بیشتری بین طبقات ایجاد کند. بنابراین طبقه‌بندی اسامی با دقّت بیشتری انجام شود. همان‌طور که در نتایج نیز مشاهده شد، بیشترین خطا در طبقه‌بندی مربوط به شناسایی اسامی موجود در طبقهٔ اسامی افراد می‌باشد. یکی از دلایلی که باعث بالا رفتن این خطا می‌شود، وجود کلمات غریب از زبان‌های دیگر در زبان فارسی است که این اسامی در طبقات مکان و اسامی افراد خیلی بیشتر از بقیهٔ طبقات وجود دارند. اگر بتوان قبل از طبقه‌بندی با استفاده از یک زیرسیستمی، این اسامی را از متن جدا کنیم (ر.زمردیان ۱۳۸۴) آن‌گاه انتظار می‌رود خطای سیستم NER در شناسایی اسامی افراد و مکان‌ها کمتر شود و دقّت کلّی سیستم بیشتر گردد.

۹- منابع

خ. فرشیدورد (۱۳۸۶). فرهنگ پیشوندها و پسوندهای زبان فارسی، انتشارات زوار.

(مشکوة الدینی، م. (۱۳۸۵). دستور زبان فارسی بر اساس نظریهٔ گشتاری، انتشارات دانشگاه فردوسی مشهد.

Alkharashi, I.A., 2009. Person Named Entity Generation and Recognition for Arabic Language. Proceedings of the Second International Conference on Arabic Language Resources and Tools. Cairo, Egypt, The MEDAR Consortium: 205-208.

Babych, B. Hartley, A., 2003. Improving machine translation quality with automatic named entity recognition. Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools. Budapest, Hungary:1-8.

Benajiba, Y., Rosso, A.P., et al., 2008. Arabic named entity recognition using optimized feature sets. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii: 284-293.

Chang, Y.-S., Sung, Y.-H., 2005. Applying Name Entity Recognition to Informal Text. CS224N/Ling237.

Elsebai, A., 2008. Arabic Proper Names Recognition Using Heuristics.

Farber, B., Freitag, D., et al., 2008. Improving NER in Arabic Using a Morphological Tagger. The 6th international conference on Language Resources and Evaluation, LREC

Farmakiotou, D., Karkaletsis, V., et al., 2000. Rule-Based Named Entity Recognition for Greek Finncial Texts.

Florian, R., Ittycheriah, b., et al., 2003. Named Entity Recognition through Classifier Combination. Proceedings of the Seventh Conference on Natural Language Learning. HLT-NAACL

Grover, C., Givon, S., et al., 2007. Named Entity Recognition for Digitised Historical Texts. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08) 1343-1346.

Kim, S., Song, Y., et al ., 2006. MMR-based Active Machine Learning for Bio Named Entity Recognition. Human Language Technology. North American New York: 69–72.

Klinger, R., Friedrich, C.M., et al., 2007. Named Entity Recognition with Combinations of Conditional Random Fields. Proceedings of the Second BioCreative Challenge Evaluation Workshop. Madrid, Spain: 89–92.

Kumaran, G., Allan, J.,Text classification and named entities for new event detection, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom: 297-304.

Leong, H., Ng, H.T., 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. Proceedings of the 19th Coling.

Hassel, M., 2003. Exploitation of named entities in automatic text summarization for swedish. Proceedings of NODALIDA.

Mikheev, A., Moens, M., et al., 1999. Named Entity Recognition without Gazetteers. Proceedings of EACL: 1-8.

Minkov, E., Wang, R.C., et al., 2005. Extracting personal names from email: applying named entity recognition to informal text. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada: 443-450.

Moll´a, D., Zaanen, M.v., et al., 2006. Named Entity Recognition for Question Answer. Proceedings of the 2006 Australasian Language Technology Workshop. Sydney: 51-58.

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Linguisticae Investigationes 3(1): 3-26.

Pasca, M., 2004. Acquisition of categorized named entities for web search. Proceedings of the thirteenth ACM international conference on Information and knowledge management. Washington, D.C., USA: 137-145.

Rau, L., 1991. Extracting Company Names from Text. Conference on Artificial Intelligence Applications of IEEE.

Settles, B., 2004. Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets, COLING: 107--110.

Shaalan, K., Raza, H., 2009. Person Name Entity Recognition for Arabic. Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources. Prague, Czech Republic, Association for Computational Linguistics: 17-24.

Zhang, H.P., Liu, Q., et al., 2003. Chiniese name entity recognition using role model. Computational linguistics and Chiniese language processing 8.

Zhou, G., Su, J., 2002. Named Entity Recognition using an HMM-based Chunk Tagger. Proceedings of

the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: 473-480.

Moll´a, D., Zaanen, M.v., et al., 2006. Named Entity Recognition for Question Answer. Proceedings of the 2006 Australasian Language Technology Workshop. Sydney: 51-58.

Babych, B., Hartley, A., 2003. Improving machine translation quality with automatic named entity recognition. Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools. Budapest, Hungary:1-8.

Pasca, M., 2004. Acquisition of categorized named entities for web search. Proceedings of the thirteenth ACM international conference on Information and knowledge management. Washington, D.C., USA: 137-145.

Kumaran, G., Allan, J., 2004. Text classification and named entities for new event detection. In Proceedings of ACM SIGIR2004.

سیدعبدالحمید اصفهانی مدرک کارشناسی خود را در رشتهٔ علوم کامپیوتر در سال ۱۳۸۶ از دانشگاه ولی عصر (عج) رفسنجان و مدرک کارشناسی ارشد را در رشتهٔ مهندسی کامپیوتر – هوش مصنوعی از دانشگاه آزاد اسلامی واحد مشهد اخذ نموده است. از ابتدای سال ۱۳۸۹ در دانشگاه آزاد اسلامی واحد قاینات مشغول به کار می‌باشد و هم اکنون عضو هیأت علمی و مسئول باشگاه پژوهشگران جوان دانشگاه آزاد اسلامی واحد قاینات می‌باشد. زمینه‌های تحقیقاتی مورد علاقهٔ ایشان، پردازش زبان‌های طبیعی، شناسایی الگو و پردازش تصویر می باشد.

نشانی رایانامک ایشان عبارتند از:

hamid_com_81@yahoo.com

سعید راحتی قوچانی متولد ۱۳۴۶ شهرستان قوچان، دانش‌آموخته کارشناسی الکترونیک سال ۱۳۶۹ دانشکده فنی دانشگاه تهران و کارشناسی ارشد مخابرات ۱۳۷۲ دانشگاه آزاد اسلامی تهران جنوب و دکترای مخابرات ۱۳۷۷ دانشگاه آزاد اسلامی واحد علوم و تحقیقات می‌باشد. وی پژوهشگر برتر جشنوارهٔ فردوسی ۱۳۷۹، رتبهٔ اوّل پژوهش سال ۱۳۸۶ و رتبهٔ دوم پژوهش سال ۱۳۸۵ دانشگاه آزاد اسلامی مشهد

شده است. ایشان از سال ۱۳۷۸ به‌عنوان استادیار مخابرات دانشگاه آزاد اسلامی مشهد مشغول به کار می‌باشد. وی از ۱۳۷۸ به مدت سه سال به‌عنوان معاون آموزشی و پژوهشی و بعد از آن به مدت سه سال به‌عنوان رئیس دانشکده مهندسی دانشگاه آزاد اسلامی مشهد و مابین سال‌های ۱۳۸۴ تا ۱۳۸۸ معاون آموزشی و پژوهشی دانشگاه امام رضا (ع) بود. ایشان تاکنون بیش از ۷۵ مقاله در کنفرانس‌های داخلی و خارجی و نشریات به چاپ رسانده است. گرایش‌های تحقیقاتی ایشان پردازش گفتار و آموزش شبکه‌های عصبی و کاربرد آن در مدل‌سازی سیستمهای بیولوژیک می‌باشد.

نشانی رایانامک ایشان عبارتند از:

rahati@mshdiau.ac.ir

نادر جهانگیری در سال۱۳۲۳در رشت متولد شد. تحصیلات کارشناسی را در رشتهٔ زبان و ادبیات انگلیسی و کارشناسی ارشد زبان‌شناسی را در دانشگاه تهران گذراند. در سال ۱۳۵۴ دکترای زبان‌شناسی اجتماعی از دانشگاه لندن اخذ کرد. سوابق علمی وی عبارتاند از: استادیار گروه زبان‌شناسی دانشگاه فردوسی مشهد، تحقیق و تدریس در دانشگاه برکلی آمریکا، تدریس زبان‌شناسی در دانشگاه توکیو، عضویت در انجمن زبان‌شناسی بریتانیا و انجمن زبان‌شناسی امریکا، معرّفی ایشان به‌عنوان یکی از نخبگان جهان از کشور ایران در کتاب ?Who is who سال ۱۹۹۸ میلادی، عضویت در آکادمی علوم نیویورک، تز دکترای وی « بررسی جنبه‌های اجتماعی و فرهنگی زبان‌ها» بوده است. وی همچنین آثاری مانند آواشناسی آکوستیک، فلسفهٔ زبان (ترجمه) "زبان، بازتاب زمان، فرهنگ و اندیشه" را به رشتهٔ تحریر در آورده است و نیز" گویش و لغت‌نامه گیلکی لاهیجانی" در سه جلد که به‌وسیلهٔ مؤسسهٔ تحقیقات آفریقا و آسیا در ژاپن چاپ و منتشر شد، از آثار وی می باشد. زمینه‌های تحقیقی مورد علاقهٔ وی زبان‌شناسی محاسباتی، زبان‌شناسی رایانه‌ای و آفازی می‌باشد.

نشانی رایانامک ایشان عبارتند از:

jahangiri398@yahoo.com