تشخیص اسامی اشخاص با استفاده از افزایش کلمه‌های نامزد اسم در میدان‌های تصادفی شرطی برای زبان عربی

مجید عسگری بیدهندی و بهروز مینایی بیدگلی
دانشکدهٔ کامپیوتر، دانشگاه علم و صنعت ایران،تهران، ایران

چکیده

تشخیص و استخراج واحدهای اسمی مانند نام اشخاص، مکان ها، تاریخ و ساعت، در داده‌کاوی از یک منبع الکترونیکی یا متنی بسیار مفید است. تشخیص درست واحدهای اسمی، یک نیاز مهم در حل مسائلی در حوزه های جدید مانند پاسخ‌گویی به سؤال ها، سیستم های خلاصه سازی، بازیابی اطلاعات، استخراج اطلاعات، ترجمهٔ ماشینی، تفسیر ویدئویی و جستجوی معنایی در وب است. در سال های گذشته تلاش دانشمندان برای انجام عملیات تشخیص واحدهای اسمی برای زبان انگلیسی و دیگر زبان‌های اروپایی به نتایج بسیار خوبی منجر شده است، اما برای زبان‌هایی مانند فارسی و عربی، نتایج مناسب حاصل نشده است. یکی از اصلی‌ترین اهداف عملیات تشخیص واحدهای اسمی، تشخیص اسامی اشخاص است. در این مقاله سامانه‌ای برای تشخیص اسامی با به‌کارگیری مفهوم «کلمه های نامزد اسم» در مراحل آموزش و پیش‌بینیِ مدلی مبتنی بر میدان‌های تصادفی شرطی معرفی شده است. به‌طورخاص، همراه با توسعهٔ این سامانه، پیکره‌های متنی استانداردی از روی متون دینی کهن به زبان عربی ساخته شده است. همچنین حاصل کار سامانه بر روی داده‌های روزنامه‌ای که توسط محققان دیگر ایجاد شده، بررسی شده است و نتایج به‌دست آمده در مقایسه با نتایج سامانه‌های دیگر روی همان داده‌ها، نشان می‌دهد با استفاده از این روش، دقت تشخیص اسامی در متون عربی به مقدار قابل توجهی بالا رفته است.

واژگان کلیدی: تشخیص واحدهای اسمی، یادگیری ماشین، میدان‌های تصادفی شرطی، زبان فارسی، زبان عربی.

۱- مقدّمه

تشخیص واحدهای اسمی[1] که به آن شناسایی واحدهای اسمی[2] و نیز استخراج واحدهای اسمی[3] اطلاق می‌شود، یک زیروظیفه از استخراج اطلاعات و به معنی پردازش مستنداتی است که در آن به دنبال مکان‌یابی عناصر اسمی در متن و دسته‌بندی آنها به رده‌های از پیش تعیین‌شده مانند اسامی اشخاص، سازمان‌ها (شرکت‌ها، سازمان‌های دولتی و غیره)، مکان‌ها (شهرها، کشورها، رودخانه‌ها و غیره)، عبارت‌های زمانی، کمیت‌ها، مقدارهای پولی، درصدها و غیره هستیم.

در دههٔ گذشته، عملیات تشخیص واحدهای اسمی در زبان‌های اروپایی به نتایج بسیار خوبی منجر شده (Nadeau & Sekine, 2007)(Chinchor, 1995) اما در مورد زبان‌های عربی و زبان‌های مشابه آن هنوز نتایج قابل قبولی حاصل نشده است. یکی از مهم‌ترین و چالش‌برانگیزترین زیرمجموعه‌های عملیات تشخیص واحدهای اسمی، تشخیص اسامی اشخاص است(Elsebai & Meziane, 2011).

در این مقاله، تشخیص نام اشخاص در چهار مورد مختلف از متون زبان عربی مد نظر قرار گرفته است. متون روزنامه‌ای، متون تاریخی، روایی و فقهی. همان‌طور که اشاره شد، این عملیات، کاربردهای بسیار مفیدی در پردازش زبان

[1] Named Entity Identification
[2] Named Entity Recognition
[3] Named Entity Extraction

طبیعی به‌خصوص برای متون دینی دارد. از جمله، یافتن ارجاعات میان متون (Nadeau & Sekine, 2007)، تشخیص شباهت میان متون(Alhelbawy & Gaizauskas, 2012)، بررسی ارتباط تاریخی دو متن، طبقه‌بندی متون دینی و غیره.

لازم است در این بخش اشاره‌ای به برخی از ویژگی‌های چالش‌برانگیز زبان عربی در عملیات پردازش متن نیز داشته باشیم. یکی از مهم‌ترین تفاوت‌های زبان عربی و زبان‌های مشابه آن، رسم‌الخط این زبان است. زبان فارسی نیز از لحاظ رسم‌الخط تشابهات زیادی به زبان عربی دارد. یکی از اصلی‌ترین ویژگی‌های رسم‌الخط زبان عربی در پردازش متن که البته به‌طور خاص در تشخیص واحدهای اسمی چالش بزرگی محسوب نمی‌شود (آن‌طور که در بخش (۴-۲-۲) نیز توضیح داده شده است)، این است که شکل یک نویسه با توجه به مکان قرارگیری آن در کلمه، تغییر می‌کند. استراتژی ملحق‌کردن اجزا برای فرم‌دادن به کلمات باعث جداافتادگی[1] داده‌ها می‌شود (Benajiba, 2009).

ویژگی دیگر این رسم‌الخط این است که بسیاری از حروف (از جمله اِعراب) به‌خصوص در متون مدرن به‌هیچ‌وجه نوشته نمی‌شوند و در عین حال وجود اعراب باعث تغییر معنای کلمه به‌طور کلی می‌شود. برای مثال ملک به صورت‌های مُلک، مَلَک و مِلک قابل نوشتن است. همچنین اعراب حرف آخر کلمه می‌تواند گاهی تنها عامل مشخص‌کنندهٔ نقش کلمه در جمله باشد، اما در عین حال نوشته نشود.

زبان عربی حساس به حروف کوچک و بزرگ نیست. این خصوصیت یکی از بزرگترین کمبودها در عملیات تشخیص واحدهای اسمی در این زبان است؛ زیرا وجود این خصیصه در زبان‌هایی مانند زبان انگلیسی کمک زیادی به عملیات تشخیص واحدهای اسمی می‌کند(Benajiba, 2009).

در نهایت زبان عربی دارای نظام پیچیدهٔ ساخت‌واژی است. این ویژگی عملیات نرمال‌سازی و نشانه‌گذاری را به‌شدت سخت می‌کند(Benajiba, 2009). زبان عربی چسبندگی زیادی دارد، زیرا فرم اصلی یک کلمه به‌صورت پیشوند(ها) +ریشه + پسوند(ها) است و تعداد پیشوند یا پسوندها ممکن است هیچ یا بیشتر باشد. پسوندها به ریشه متصل می‌شوند تا اصطلاح مورد نیاز را تولید کنند. برای مثال کلمهٔ «المنزل» در انگلیسی به معنای The House است. این مثال نشان می‌دهد که یک کلمهٔ عربی می‌تواند به دو کلمهٔ انگلیسی ترجمه شود. یک مثال پیچیده‌تر در بخش (۴-۲-۱) بیان شده است.

در ادامهٔ این مقاله ابتدا در بخش دو، به بررسی کارهای محقّقان دیگر در مسألهٔ تشخیص واحدهای اسمی در زبان عربی پرداخته‌ایم. در بخش (۳) نیز، چگونگی ساخت پیکرهٔ متنی NoorCorp براساس استانداردهای ذکرشده در کنفرانس CoNLL[2] برای پیکره‌های متنی عملیات تشخیص واحدهای اسمی، بررسی شده است.

در بخش (۴)، سامانهٔ Noor ANER برای تشخیص اسامی خاص در متون روزنامه‌ای و دینی زبان عربی بر اساس مدل میدان‌های تصادفی شرطی و مفهوم افزایش کلمه‌های نامزد اسم به‌عنوان یک راه حل پیشنهادی برای بهبود نتایج، به‌طور کامل معرفی شده است.

در بخش (۵) نیز به ارزیابی سامانه‌ی ذکر شده پرداخته‌ایم. در پایان، در بخش (۶) نتیجه‌گیری و چند پیشنهاد برای بهبود عملکرد سامانه در آینده ارائه شده است.

۲- کارهای مشابه

در بیشتر مقالات ارائه‌شده در زمینهٔ تشخیص واحدهای اسمی، مهم‌ترین معیارهای ارزیابی نتایج، سه معیار دقت، بازخوانی و F هستند که معیار آخر میانگین هارمونیک دو معیار اول است. به همین دلیل مهم‌تر و جامع‌تر از دو معیار دیگر محسوب می‌شود (Chinchor, 1995). در اولین تلاش‌های صورت گرفته در زمینهٔ تشخیص واحدهای اسمی در زبان عربی، موفق‌ترین سامانه‌ها (بر اساس معیار F)، سه نرم‌افزار ANERsys (Benajiba,2007a) مبتنی بر مدل بیشترین آنتروپی، ANERsys 2 مبتنی بر ماشین بردار پشتیبان و Siraj بودند(Benajiba,2007b). Benajiba و همکارش در سال ۲۰۰۸ نرم‌افزاری مبتنی بر میدان‌های تصادفی شرطی[3] ارائه داده که نتایج بهتری را نسبت به نرم‌افزارهای دیگر با استفاده از ترکیب چند مدل میدان‌های تصادفی شرطی تولید کرده است (Benajiba,2008). اما بهترین نتایج حاصل شده بر روی تشخیص اسامی اشخاص تا به امروز (بر اساس معیار F)، بر روی داده‌های روزنامه‌ای (با موضوع عام) در مقالهٔ حریص *الجُمَیلی* و همکارانش (Al-Jumaily,2011) معرفی شده است که این گروه نیز

[1] Sparseness

[2] Conference On Computational Natural Language Learning

[3] Conditional Random Fields

نتایج روش مورد نظر خود را مانند موارد ذکرشدهٔ قبلی بر روی پیکرهٔ متنی ANERcorp اجرا کرده‌اند. البته روش ارائه‌شده در مقالهٔ مزبور از روش‌های یادگیری ماشین استفاده نمی‌کند و روش معرفی‌شده شامل تعدادی از قوانین است که به‌صورت دستی توسط افراد خبره گردآوری شده‌اند. به همین دلیل نتایج حاصل‌شده از نتایج دیگران بهتر است.

همچنین کوشش‌هایی برای انجام عملیات تشخیص واحدهای اسمی در متون عربی با زمینه‌های خاص توسط محققان انجام شده‌اند. برای مثال *فهری* و همکارانش در (Fehri,2011)، روشی را روی متون ورزشی ارائه کرده‌اند که ادعا شده است به معیار f برابر ۹۴ درصد دست یافته است. همچنین در (Elsebai,2011) نویسندگان از روشی بر مبنای مجموعه‌ای از کلمه‌های کلیدی به جای استفاده از تکنیک‌های گرامری، آماری یا یادگیری ماشین پیچیده استفاده کرده‌اند که به‌طور طبیعی به اندازهٔ آن روش‌ها قابل اطمینان و انتقال نیست. در مقالهٔ ذکرشده، معیار f برابر ۸۵ حاصل شده است.

۳- آماده‌سازی پیکرهٔ متنی NoorCorp و فرهنگ لغات NoorGaz

همان‌طور که در کنفرانس آموزش محاسباتی زبان طبیعی در سال ۲۰۰۳ گزارش شد (Sang,2003)، پیشنهاد شده است که یک پیکرهٔ متنی برچسب‌گذاری‌شده استاندارد شامل کلمه‌های متن به همراه برچسب مربوط به آن کلمه باشد. طبقات یکسان که در کنفرانس ششم ادراک زبان معین شده بودند(Chinchor, 1995)، عبارت از نام سازمان‌ها، مکان‌ها و اشخاص بودند. بقیهٔ مواردی که در تعریف واحدهای اسمی گنجانده می‌شوند و جزء سه مورد ذکر شده نیستند نیز، با برچسب MISC در پیکره مشخص می‌شوند. بنابراین هر کلمه از متن باید با یکی از برچسب‌های زیر مشخص شود:

- B-PERS: شروع یک نام خاص از اشخاص؛
- I-PERS: میانه یا ادامهٔ یک نام خاص از اشخاص؛
- B-LOC: شروع نام یک مکان؛
- I-LOC: میانه یا ادامهٔ نام یک مکان؛
- B-ORG: شروع نام یک سازمان؛
- I-ORG: میانه یا ادامهٔ نام یک سازمان؛
- B-MISC: شروع یک نام که جزء هیچ یک از این موارد نیست: نام خاص از اشخاص، نام مکان‌ها و نام سازمان‌ها؛
- I-MISC: میانه یا ادامهٔ نامی که جز هیچ یک از این موارد نیست: نام خاص از اشخاص، نام مکان‌ها و نام سازمان‌ها؛
- O: کلمه‌ای که یک واحد اسمی نباشد.

در کنفرانس آموزش محاسباتی زبان طبیعی، همچنین تصمیم گرفته شد که فرمت یکسانی برای داده‌های آموزش و آزمودن برای همهٔ زبان‌ها استفاده شود (Sang,2003) که شامل دو ستون است: ستون اول برای کلمه‌ها و ستون دوم برای برچسب‌های هر کلمه. شکل (۱)، دو نمونه از پیکره‌های متنی استاندارد را که حاوی برچسب‌های واحدهای اسمی هستند، نمایش می‌دهد. در سمت چپ، یک متن برچسب‌خوردهٔ انگلیسی نمایش داده شده است. بین کلمهٔ مورد نظر و برچسب آن، به‌طور عمومی یکی از نویسه‌های Tab و یا Space قرار می‌گیرند. هر کلمه نیز به همراه برچسب یا برچسب‌هایش در یک خط نوشته می‌شود. تک تک علائم نگارشی نیز به‌طور خاص در یک خط نگاشته می‌شوند و به‌عنوان عامل بزرگی به فرآیند یادگیری کمک می‌کنند. (Benajiba,2007a)

```
Arsenal     B-ORG
captain     0
Robin       B-PER
van         I-PER
Persie      I-PER
has         0
spoken      0
of          0
his         0
love        0
for         0
London      B-LOC
,           0
     ...
```

```
و          0
جلس        0
سليمان     B-PER
قليلا      0
ثم         0
نهض        0
ف          0
خرج        0
إلى        0
حسن        B-PE
بن         I-PER
على        I-PER
و          0
هو         0
قاعد       0
في         0
المسجد     B-LOC
الكوفة     B-LOC
```

(شکل-۱): پیکره‌های متنی استاندارد برای تشخیص واحدهای اسمی

به‌منظور ساخت یک پیکرهٔ متنی مناسب برای سنجیدن عملیات تشخیص واحدهای اسمی بر روی متون دینی، سه پیکرهٔ متنی برای زبان عربی با استفاده از داده‌های موجود در «مرکز تحقیقات کامپیوتری علوم اسلامی[1]» آماده شده است.

[1] برای آشنایی با فعالیت‌های این مجموعه می‌توانید به نشانی زیر مراجعه کنید:

Http://Www.Noorsoft.Org/

این سه پیکره براساس داده‌های برچسب‌خورده از متون سه کتاب زیر فراهم آمده‌اند:

- کتاب «وقعهٔ صفین» نوشتهٔ «نصربن مزاحم منقری» نگاشته شده در سال ۲۱۲ هجری قمری به عنوان یک کتاب تاریخی؛
- کتاب «الارشاد فی معرفهٔ حجج اللّه علی العباد» نوشتهٔ «محمّد بن محمّد مفید» معروف به شیخ مفید نگاشته شده در سال ۲۳۰ هجری قمری به‌عنوان یک کتاب روایی؛
- کتاب «شرایع الاسلام فی مسائل الحلال و الحرام» نوشتهٔ «جعفر بن حسن» (محقق حلّی) نگاشته شده در سال ۶۷۶ هجری قمری به‌عنوان یک کتاب فقهی.

این پیکره‌های متنی از لحاظ تعداد کلمه‌ها (بعد از نشانه‌گذاری)، درصد اسامی اشخاص، اسامی جغرافیایی، اسامی گروه‌ها و سازمان‌ها و دیگر اسامی در جدول (۱) با یکدیگر مقایسه شده است. همچنین برای اینکه مقایسه حجمی میان این پیکره‌های متنی با پیکرهٔ متنی ANERcorp به‌راحتی برای خواننده امکان‌پذیر باشد، اطلاعات مربوط به آن پیکره نیز در ردیف آخر جدول آورده شده است.

همان طور که مشاهده می‌شود این چهار نوع پیکرهٔ متنی که با توجه به تاریخ تألیف خود، متون عربی مدرن و عربی کهن را پوشش می‌دهند، از لحاظ ساختاری تفاوت‌های آشکاری با هم دارند. در دادهٔ روزنامه‌ای همان‌طورکه انتظار می‌رود، هم تعداد اسامیِ اشخاص و هم تعداد مکان‌ها در متن به تعداد نسبتاً بالایی استفاده می‌شود؛ اما نسبت تعداد اسامی به متن در مقایسه با کتاب‌های تاریخی کمتر است. در کتب روایی تعداد اسامی اشخاص در متن زیادتر است؛ اما نام مکان‌ها نسبتاً کمتر دیده می‌شود. در کتب فقهی به‌طور اساسی نام‌های خاص کمتر یافت می‌شوند و همان طور که دیده می‌شود این نسبت کمتر از یک درصد است (درمجموع ۲۳۳ اسامی خاص اشخاص در کل ۴۸۵۸۲ کلمه). نسبت تعداد انواع واحدهای اسمی به کل آنها به تفکیک پیکرهٔ متنی در جدول (۲) آورده شده است.

یکی از ابزارهایی که کمک زیادی به‌بالابردن دقت الگوریتم‌های ما برای تشخیص اسامی خاص می‌کند، فرهنگ لغتی[1] از اسامی است. بدین منظور نزدیک به ۸۸۰۰۰ نام از داده‌های برچسب‌خوردهٔ نرم‌افزار «جامع‌الاحادیث» با همکاری «مرکز تحقیقات کامپیوتری علوم اسلامی» جمع‌آوری شد. سپس این اسامی به اجزای خود شکسته شدند. برای مثال نام «حسن بن علی بن عبد الله بن المغیره» به شش جزء غیر تکرارِ «حسن»، «بن»، «علی»، «عبد»، «الله» و «المغیره» شکسته شد که در آن یکی از اجزا («بن») سه بار تکرار شده است. این اجزا برای تمامی اسامی به دست آمده و به مجموعه به‌صورت جداجدا به همراه تعداد تکرار خود اضافه شدند. همچنین اسامی ذکرشده در فرهنگ ANERgazet (Benajiba,2007a) به آنها اضافه شد و درنهایت یک پایگاه داده با ۱۸۲۳۸ ردیف به‌دست آمد.

۴- سامانهٔ Noor ANER

سامانهٔ Noor ANER، یک سامانهٔ مبتنی بر میدان‌های تصادفی شرطی است که متن ورودی را بعد از انجام پیش‌پردازش‌های لازم تحلیل کرده و اسامی اشخاص را در آن نشانه‌گذاری می‌کند. در ادامه در مورد ساختار این سامانه توضیح داده می‌شود.

۴-۱- میدان‌های تصادفی شرطی

مدل میدان‌های تصادفی شرطی یک روش مدل‌سازی آماری است که اغلب در شناسایی الگوها مورد استفاده قرار می‌گیرد. به‌طور دقیق‌تر، این روش نوعی از مدل گرافیکی احتمالاتی غیرمستقیم تشخیصی[2] است.

۴-۱-۱- مدل‌های خطی لاجیستیک

فرض کنید x یک نمونه باشد، و y یک برچسب ممکن برای آن. یک مدل خطی لاجیستیک فرض می‌کند که

$$p(y|x; w) = \frac{e^{\sum_j w_j F_j(x,y)}}{Z(x, w)}$$

که در آن Z تابع افراز نامیده می‌شود و برابر است با:

$$Z(x, w) = \sum_{y'} e^{\sum_j w_j F_j(x,y')}$$

هر کدام از عبارات $F_j(x, y)$ یک تابع خصیصه نامیده می‌شود و هر مقدار w_j که در واقع یکی از عناصر بردار وزن یا w است، وزن مرتبط با آن تابع خصیصه است که براساس اهمیت تابع مقدرا آن بیشتر می‌شود.

[1] Gazetteer

[2] Discriminative Undirected Probabilistic Graphical Model

بنابراین، با داشتن ورودی x، برچسب پیش‌بینی شده توسط مدل

$$\hat{y} = \text{argmax}_y p(y|x; w) = \text{argmax}_y \sum_j w_j F_j(x, y)$$

خواهد بود.

(جدول- ۱): مقایسهٔ حجم پیکره‌های متنی موجود در NoorCorp

نام پیکره	تعداد کلمه‌ها	اشخاص	مکان‌ها	گروه‌ها	دیگر اسامی	موضوع
وقعهٔ صفین	۲۳۵۸۴۲	۶/۴۷٪	۱۰/۰۶٪	۶/۵۹٪	۰٪	تاریخی
الارشاد	۱۳۴۳۱۶	۱۴/۳۱٪	۱/۰۷٪	۲/۳۶٪	۰٪	روایی
شرایع	۴۸۵۸۲	۰/۴۸٪	۱/۱۱٪	۰٪	۰٪	فقهی
ANERcorp	۱۵۰۲۸۵	۴/۲۸٪	۳/۳۴٪	۲/۲۶٪	۱/۱۰٪	روزنامه‌ای

(جدول- ۲): نسبت واحدهای اسمی در پیکره‌های متنی مورد استفاده در مقاله

پیکره	اشخاص	مکان‌ها	گروه‌ها	دیگر
وقعهٔ صفین	۲۷/۹۸٪	۴۳/۵۲٪	۲۸.۵٪	۰٪
ارشاد	۸۰/۶۶٪	۶/۰۳٪	۱۳/۰۳٪	۰٪
شرایع	۳۰/۱۹٪	۶۹.۸۱٪	۰٪	۰٪
ANERcorp	۳۸/۹۸٪	۳۰/۴۲٪	۲۰/۵۸٪	۱۰/۰۱٪

۴-۱-۲- میدان‌های تصادفی شرطی[1]

مدل میدان‌های تصادفی شرطی درحقیقت نوع خاصی از مدل‌های خطی لاجیستیک است. منظور ما از میدان‌های تصادفی شرطی در این مقاله، به‌طورکلی میدان‌های تصادفی شرطی با حلقه‌ی خطی[2] است.

۴-۱-۳- استنتاج و یادگیری در میدان‌های تصادفی شرطی

آموزش یک مدل CRF به معنای پیدا کردن بردار وزن‌های w است به گونه‌ای که بهترین پیش‌بینی ممکن را برای هر نمونه‌ی آموزشی $\bar{x}$ ارائه دهد:

$$\bar{y}^* = \text{argmax}_{\bar{y}} p(\bar{y}|\bar{x}; w)$$

در هر حال، قبل از اینکه بخواهیم در مورد فرآیند آموزش سخن بگوییم بایستی، باید متوجه این باشیم که دو مشکل اساسی در مرحلهٔ استنتاج برای ما وجود دارد: نخست، چگونه می‌توانیم معادله را برای هر $\bar{x}$ (نماد بالای x به معنای برداری از تمام xها است) و هر مجموعه‌ای از وزن‌های w به‌صورت کارا محاسبه کنیم. این محاسبه برای برچسب‌های $\bar{y}$ از مرتبهٔ نمایی است.

دوم، با داشتن هر $\bar{x}$ و $\bar{y}$ ما باید مقدار زیر را ارزیابی کنیم:

$$p(\bar{y}|\bar{x}; w) = \frac{1}{Z(\bar{x}, w)} e^{\sum_j w_j F_j(\bar{x}, \bar{y})}$$

مشکل در اینجا مخرج کسر است که باز روی تمام دنباله‌های $\bar{y}$ عمل می‌کند:

$$Z(\bar{x}, w) = \sum_{\bar{y}'} e^{\sum_j w_j F_j(\bar{x}, \bar{y}')}$$

برای هر دوی این موارد به روش‌های ابتکاری و میان‌بر نیاز داریم که بدون پردازش لحظه‌ای روی همهٔ $\bar{y}$ها، تمامی آنها به‌گونه‌ای کارا پردازش شوند. این حقیقت که در میدان‌های تصادفی شرطی، هر تابع خصیصه تنها به دو برچسب که کنار هم قرار دارند، وابسته است، به ما برای پیداکردن چنین راه حلی کمک خواهد کرد (Elkan,2008)، (Lafferty,2001) و (Sutton,2007).

وقتی که مجموعه‌ای از نمونه‌های آموزشی را در اختیار داریم، فرض می‌کنیم هدف ما پیداکردن پارامترهای w_j ای است که احتمال شرطی نمونه‌های آموزشی بیشینه شود. برای این منظور همان‌طور که در بالا شرح داده شد، می‌توانیم از روش شیب‌دار استفاده کنیم. پس نیاز است که مشتق جزئی معیار شباهت شرطی را برای یک نمونهٔ آموزشی برای هر w_j محاسبه کنیم. بیشینه‌کردن p همان بیشینه‌کردن ln p است:

[1] Conditional Random Fields (Crf)
[2] Linear-Chain Crf

$$\frac{\partial}{\partial_{w_j}} \ln\ p(y|x;w) = F_j(x,y) - \frac{\partial}{\partial_{w_j}} \log Z(x,w)$$
$$= F_j(x,y) - E_{y\prime:p(y\prime|x;w)}[F_j(x,y')].$$

به بیان دیگر مشتقات جزئی نسبت به وزن i ام مقدار تابع خصیصهٔ i ام برای برچسب درست y منهای مقدار متوسط تابع خصیصه برای همهٔ برچسب‌های ممکن y' است. توجه کنید که این مشتق‌گیری مقدار حقیقی را برای توابع خصیصه مجاز می‌سازد، نه فقط مقادیر صفر و یک را.

شیب‌دهی معیار شباهت شرطی وقتی که تمامی مجموعهٔ آموزشی T در دست باشد، مجموع شیب‌دهی‌ها برای هر نمونهٔ آموزشی است. بیشینهٔ مطلق کل این شیب‌دهی‌ها برابر صفر است، پس داریم:

$$\sum_{\langle x,y\rangle \in T} F_j(x,y) = \sum_{\langle x,,\rangle \in T} E_{y:p(y\prime|x;w)}[F_j(x,y)]$$

این معادله برای همهٔ نمونه‌های آموزشی درست است نه برای تک‌تک نمونه‌ها.

سمت چپ معادلهٔ بالا مقدار مجموع تابع خصیصهٔ j روی همهٔ مجموعهٔ آموزشی است و سمت راست، مقدار مجموع تابع خصیصهٔ j که توسط مدل پیش‌بینی شده. درنهایت در هنگام بیشینه‌کردن معیار شباهت شرطی با روش شیب‌دهی برخط، دست‌کاری وزن w_j به‌صورت زیر خواهد بود:

$$w_j := w_j + \alpha(F_j(x,y) - E_{y\prime:p(y\prime|x;w)}[F_j(x,y')])$$

۴-۲- پیش‌پردازش‌های انجام‌شده روی متون

در طول فرآیند یادگیری و آزمودن و نیز برچسب‌گذاری روی داده‌های نادیده، پیش‌پردازش‌های مختلفی روی داده‌های آموزشی انجام می‌شود که در این بخش به آنها اشاره می‌کنیم.

۴-۲-۱- نشانه‌گذاری

یکی از عملیات‌هایی که به دقتِ پردازش متون کمک می‌کند و به‌عنوان یکی از پیش‌پردازش‌های مهم همواره مورد استفاده قرار می‌گیرد، نشانه‌گذاری[1] است. برای کلمه «نشانه» در عبارت «نشانه‌گذاری»، تعاریف مختلفی وجود دارد(Fehri,2011). اما نشانه‌گذاری در اینجا به معنی جداکردن تمامی اجزای یک کلمه به ساده‌ترین حد ممکن است طوری که هر جزء به‌تنهایی متضمن معنایی در زبان مورد نظر باشد. برای مثال کلمهٔ «وسیکتبون‌ها» در زبان عربی (به معنای «و آنها آن را خواهند نوشت») را می‌توان به‌صورت «و + س + یکتبون + ها» و طبق تعریف دیگری به‌صورت «و + س + ی + کتب + ون + ها» نشانه‌گذاری کرد. در اینجا «ی» و «ون» متضمن معنای دستوری هستند نه واژگانی، مانند جزء اول، کلمه‌ی «و».

ما برای عملیات نشانه‌گذاری متن، از نسخهٔ ۱/۲ نرم‌افزار «امیرهٔ»[2] که توسط *مونا دیاب* توسعه داده شده است، بهره گرفته‌ایم. توضیحات بیشتر در مورد این نرم‌افزار در بخش ۴-۲-۳ آورده شده است.

۴-۲-۲- نویسه‌گردانی

یکی از اصلی‌ترین پیش‌پردازش‌ها که اغلب آخرین پیش‌پردازش نیز می‌باشد، نویسه‌گردانی[3] است. نویسه‌گردانی به تغییر نویسه‌های کلمه‌ها از یک زبان به نویسه‌های یک زبان دیگر گفته می‌شود که اغلب زبان مقصد، زبان انگلیسی است. در این فرآیند هر نویسه به‌طور دقیق به یک نویسه در زبان مقصد تبدیل می‌شود و هر نویسهٔ زبان مقصد نیز به‌طور دقیق معادل یک نویسه یا گروه نویسهٔ یکسان از زبان مبدأ است.

در سامانهٔ Noor NER از نویسه‌گردانی باک‌والتر Buckwalter استفاده شده است (Habash,2007). وقتی این نویسه‌گردانی روی یک کلمه انجام می‌شود، بدون توجه به شمایل نویسهٔ عربی (یعنی بدون توجه به اینکه نویسه در کجای کلمه واقع شده است)، این نویسه به یک نویسهٔ انگلیسی تبدیل می‌شود. چند نمونه از نویسه‌گردانی در شکل (۳) نشان داده شده است.

ستون دوم از سمت چپ، دادهٔ واقعی به زبان عربی و ستون اول از سمت چپ، کلمهٔ نویسه‌گردانی شدهٔ دادهٔ واقعی است. ورودی بسیاری از برنامه‌های پردازشگر متون زبان طبیعی، داده‌ها را در ستون اول ورودی خود می‌پذیرند. به همین دلیل دادهٔ نویسه‌گردانی شده، که درحقیقت ورودی مورد انتظار این برنامه‌هاست، در ستون اول نوشته شده است.

[1] Tokenizing

[2] نوشتار انگلیسی نام این نرم‌افزار به صورت Amira 2.1 است. بعد از وارسی مقالات مشخص شد که کلمهٔ امیرهٔ (کلمهٔ امیر با تای تأنیث) درست است، با اینکه تلفظ نام انگلیسی برای یک فارسی زبان این شائبه را ایجاد می‌کند که نام نرم‌افزار امیرا باشد.

[3] Transliteration

۴-۳- آماده‌سازی پیکرهٔ متنی و آموزش مدل میدان‌های تصادفی شرطی

برای پیاده‌سازی و آموزش میدان‌های تصادفی شرطی از نرم‌افزار flexCRF استفاده شده است. این نرم‌افزار برای انجام عملیات آموزش به فرمت خاصی از ورودی نیاز دارد. این ورودی از سه ستون اصلی تشکیل شده است:

در ستون اول، دادهٔ اصلی به‌صورت نویسه‌گردانی‌شده قرار می‌گیرد. دادهٔ اصلی در اینجا منظور جملات بدون برچسب هستند. در این نرم‌افزار و نرم‌افزارهای مشابه، همان‌طور که در بخش ۴-۱ هم توضیح داده شد، پردازش توالی‌ها مهم است؛ و توالی‌ها در مسألهٔ ما جملات هستند. هر جمله با یک نویسهٔ نقطه به پایان می‌رسد؛ و طبق قرارداد بعد از هر جمله یک خط خالی قرار داده می‌شود.

ستون دوم، توابع خصیصه هستند که طبق قرارداد خاصی که در مستندات برنامه flexCRF معین شده‌اند[1]، تولید می‌شوند. برای هر کلمه، ممکن است انواع مختلفی از توابع خصیصه تولید شوند؛ اما محدودیت میدان‌های تصادفی شرطی نیز باید مد نظر قرار گیرند. به این معنا که هر تابع خصیصه می‌تواند تنها بر اساس کلمهٔ جاری، یک یا دو کلمهٔ قبل یا بعد از آن، مسند[2] جاری و یک یا دو مسند قبل یا بعد از آن تعریف شود. نرم‌افزار ما توابع خصیصهٔ زیر را به‌صورت ورودی در قالب‌های زیر به برنامهٔ flexCRF می‌دهد:

- یک کلمه
- دو کلمهٔ متوالی
- یک مسند
- دو مسند متوالی
- سه مسند متوالی
- یک کلمه و یک مسند
- دو مسند و یک کلمه
- دو کلمه و یک مسند
- برچسب واحد اسمی کلمهٔ مورد نظر برای آموزش و آزمودن.

در اینجا منظور از مسند، اطلاعات کمکی است که به الگوریتم آموزش داده می‌شود. این اطلاعاتِ کمکی، در حقیقت اطلاعاتِ خصیصهٔ ما هستند و بنابراین باید در مرحلهٔ پیش‌پردازش، بدون دخالت انسان قابل تولید باشند؛ برای مثال، به‌صورتی‌که ما با استفاده از نرم‌افزار امیره،

```
'  ء | x  خ | -  ـ | K  ٍ
|  آ | d  د | f  ف | a  َ
>  أ | *  ذ | q  ق | u  ُ
&  ؤ | r  ر | k  ك | i  ِ
<  إ | z  ز | l  ل | ~  ّ
}  ئ | s  س | m  م | o  ْ
A  ا | $  ش | n  ن | P  پ
b  ب | S  ص | h  ه | J  چ
p  ة | D  ض | w  و | V  ژ
t  ت | T  ط | Y  ى | G  گ
v  ث | z  ظ | y  ي |
j  ج | E  ع | F  ً |
H  ح | g  غ | R  ٌ |
```

(شکل- ۲): نویسه‌گردانی باک‌والتر برای زبان عربی

```
w        و       CC        CC     O
jls      جلس     VBD_MS3   VP     O
slymAn   سليمان  NNP       NP     B-PERS
qlylA    قليلا   NN        ADJP   O
vm       ثم      CC        CC     O
nhD      نهض     VBD_MS3   VP     O
f        ف       RP        RP     O
xrj      خرج     VBD_MS3   VP     O
<ly      إلى     IN        PP     O
AlHsn    الحسن   DET_NNP   NP     B-PERS
bn       بن      NNP       NP     I-PERS
Ely      علي     NNP       NP     I-PERS
w        و       CC        CC     O
hw       هو      PRP_MS3   NP     O
qAEd     قاعد    NN        NP     O
fy       في      IN        PP     O
Almsjd   المسجد  DET_NN    NP     B-LOC
Alkwfh   الكوفه  DET_JJ    NP     I-LOC
```

(شکل- ۳): پیکرهٔ متنی بعد از نویسه‌گردانی و افزودن برچسب‌های ادات سخن

۴-۳-۲- برچسب‌گذاری ادات سخن

مجموعهٔ نرم‌افزاری امیره برای زبان عربی استاندارد مدرن توسط مونا دیاب در دانشگاه کلمبیا توسعه یافته است. امیره یک مجموعهٔ جانشین برای نرم‌افزار ASVMTools است. ابزار امیره شامل یک نشانه‌گذار (TOK)، یک برچسب‌گذار ادات سخن(POS) و یک جداساز عبارات پایه (BPC) است.

طبق گزارش‌های مندرج در (Diab,2009) این سامانه، بسیار سریع و قابل اطمینان ساخته شده است. همچنین امکان دست‌کاری پارامترهای مختلف زمان اجرا در این نرم‌افزار تعبیه شده است. امیره برای کاربردها و عملیات‌های مختلف پردازش زبان طبیعی در مقالات متعددی مورد استفاده قرار گرفته است. (Diab,2009) ما از این نرم‌افزار برای کمک به آموزش مدل میدان‌های تصادفی شرطی و نیز نشانه‌گذاری متون بهره جسته‌ایم.

[1] مستندات مربوطه از این آدرس قابل دسترسی هستند:
Http://Flexcrfs.Sourceforge.Net/Documents.Html

[2] Predicate

برچسب‌های ادات سخن هر کلمه را تولید کرده‌ایم. یک قالب تابع خصیصه، به‌عنوان مثال قالب «یک کلمه و یک مسند» می‌تواند به «یک کلمه و یک برچسب ادات سخن» تبدیل شود. مشخص است که این روش دریافت ورودی، برنامهٔ flexCRF را برای کارکردن در فضاهای مختلف بسیار انعطاف پذیر کرده است.

همان‌طور که در شکل (۳) می‌بینید ما برای هر کلمه، اطلاعات گوناگونی را در اختیار داریم:

- دادهٔ نویسه‌گردانی‌شده (که توسط نرم‌افزار از روی دادهٔ اصلی تولید شده است)
- دادهٔ اصلی (که توسط تایپیست نوشته شده است)
- برچسب ادات سخن (که توسط نرم‌افزار امیره از روی دادهٔ اصلی به‌دست آمده است)
- برچسب جداسازی عبارات پایه (که توسط نرم‌افزار امیره از روی دادهٔ اصلی به‌دست آمده است)
- برچسب واحدهای اسمی (از بعد از پرسش از زبان‌شناسان و افراد خبره این حوزه نوشته شده‌اند)

مشخص است که مورد آخر فقط برای آموزش و آزمودن مورد استفاده قرار می‌گیرد و در مرحله پیش‌بینی[1] به آن دسترسی نداریم.

۴-۴- برچسب‌های غنی شده با کلمه‌های نامزد اسمی

همان‌طور که در بخش ۴-۳ اشاره شد، یادگیری مدل میدان‌های تصادفی شرطی در سامانهٔ Noor ANER از روی برچسب‌های ادات سخن صورت می‌گیرد که درحقیقت گزاره‌های ما هستند. اما این گزاره‌ها، برچسب‌های سادهٔ ادات سخن نیستند و برای آموزش به شیوهٔ خاصی با اطّلاعات موجود در فرهنگ لغت غنی شده‌اند.

ایدهٔ نهفته در این غنی‌سازی، درحقیقت راه حل ابتکاری ما برای بهبود عملیات تشخیص واحدهای اسمی است. در اینجا ما برچسب‌های ادات سخن را از طریق اجرای برنامهٔ امیره به‌دست می‌آوریم. سپس با استفاده از روش ارائه‌شده در زیر، آنها را با اسامی موجود در فرهنگ لغت غنی‌تر می‌کنیم:

- اگر کلمهٔ مورد نظر در فرهنگ لغات موجود بود، عبارت _NAME را به ابتدای برچسب ادات سخن آن کلمه اضافه می‌کنیم. به چنین کلمه‌ای یک کلمهٔ «نامزد اسم» می‌گوییم.
- اگر دو یا چند نامزد اسم پشت سر هم حاضر شوند، برچسب تمام آنها به NAME2 تغییر می‌کند.

یعنی درحقیقت درصورتی‌که تعداد برچسب‌های ادات سخن ما n باشد، تعداد آن‌ها به 2n+1 افزایش پیدا می‌کند. اما چرا انتظار داریم این روش، نتیجهٔ حاصله را بهبود دهد؟ اهمیت قانون دوم، به نظر واضح می‌رسد. اسامی در زبان‌های عربی و نیز در بسیاری از زبان‌های دیگر، صفات هستند و این یکی از چالش‌های اصلی برای تشخیص اسامی در این زبان‌هاست. وقتی دو یا چند صفت پشت سر هم در این زبان‌ها ظاهر می‌شوند، احتمال اسم‌بودن این کلمه‌ها افزایش می‌یابد. به‌طور خاص در زبان عربی کهن و در متون روایی، وجود کلمه‌های نسبتی مانند «بن» نیز این احتمال را افزایش می‌دهند. به‌دلیل این‌که این احتمال در داده‌های ما (اسم‌بودن تمامی اعضای توالی چند کلمه‌ی نامزد اسم) برابر ۹۴ درصد بوده است، به‌طورکلی برچسب ادات سخن را حذف کرده و برچسب جدید NAME2 را جایگزین می‌کنیم. اما قانون اول نیز در بسیاری از موارد مفید واقع می‌شود. درحقیقت ما گزاره‌ای را تولید می‌کنیم که هم حاوی اطلاعات ادات سخن و هم حاوی پیشنهاد نامزد اسم‌بودن است. در عین حال تصمیم‌گیری در مورد اسم‌بودن این کلمه را به خود مدل میدان‌های تصادفی شرطی می‌سپاریم. درحقیقت با دوبرابرکردن برچسب‌های مورد نظر، برچسب‌هایی تولید می‌شوند که احتمال اسم‌بودن کلمه‌های مرتبط به آنها بیشتر از برچسب‌های مشابه آنهاست. در عین حال الگوریتم یادگیری ماشین با استفاده از آنها توابع خصیصهٔ خود را می‌سازد و در هنگام برچسب‌گذاری تصمیم می‌گیرد چگونه از آنها بهره ببرد.

با این توضیحات این موضوع قابل استنباط است که روش پیشنهادشده در اینجا پایداری بالایی دارد؛ زیرا درصورتی‌که برچسب‌گذار ادات سخن که یک برنامهٔ رایانه‌ای است و نیز افزونندهٔ کلمه‌های نامزد اسم، اشتباه کنند، مدل میدان‌های تصادفی شرطی، توابع خصیصهٔ ضعیف‌تری از آنها خواهد ساخت، زیرا تعداد این برچسب‌های اشتباه نسبت به کل برچسب‌ها کم است. نتایج ارائه‌شده در بخش پنج درست بودن ایدهٔ ما را ثابت می‌کنند.

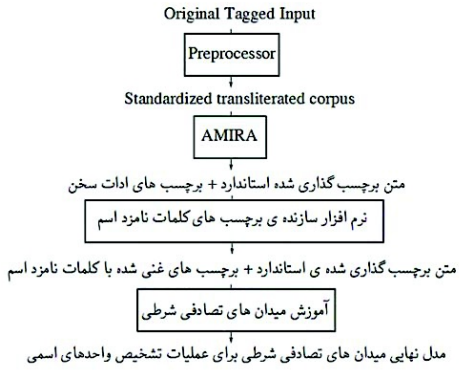
۴-۵- طرح‌وارهٔ کلی کارکرد سامانه

همان‌طور که در بالا ذکر شد، ابتدا متن برچسب‌گذاری اولیهٔ ما با یک نرم‌افزار استانداردسازی و پیش‌پردازش‌های لازم،

[1] Prediction

تبدیل به یک متن «برچسب‌گذاری‌شده»، «استانداردشده» و «نویسه‌گردانی‌شده» می‌شود. سپس با استفاده از نرم‌افزار امیره، متن برچسب‌گذاری‌شده به همراه اطلاعات برچسب‌های ادات سخن به‌دست می‌آید. سپس نرم‌افزار دیگری، برچسب‌های غنی‌شده با کلمه‌های نامزد اسمی را از روی این اطلاعات می‌سازد. آنچه در اینجا حاصل می‌شود، اطلاعات نهایی برای آموزش مدل میدان‌های تصادفی شرطی است. این پردازش به‌طور خلاصه در شکل (۴) نشان داده شده است.

در مرحلهٔ آزمون و یا برچسب‌گذاری نیز همین عملیات تکرار شده و کلمه‌های برچسب‌گذاری‌شده و غنی‌شده به مدل داده می‌شود تا برچسب‌گذاری صورت پذیرد. در بخش بعدی نتایج حاصل از به‌کارگیری این ایده به‌طور کامل بررسی شده‌اند.



(شکل-۴): ساختار سامانهٔ Noor ANER

۵- ارزیابی و تحلیل نتایج

در این بخش به ارزیابی سامانهٔ Noor ANER بر روی متون دینی و همچنین متون عادی روزنامه‌ای خواهیم پرداخت و سامانه را با نرم‌افزارهای مشابه دیگر مقایسه می‌کنیم.

۵-۱- ارزشیابی سامانه بر روی متون دینی

در بخش ۳ اشاره کردیم که سه پیکرهٔ متنی مخصوص متون تاریخی، روایی و فقهی برای آزمایش مدل میدان‌های تصادفی شرطی آماده شدند. با توجه به اینکه سامانهٔ Noor ANER تنها برای برچسب‌گذاری اسامی اشخاص پیاده‌سازی شده است، نتایج این بخش نیز تنها روی اسامی اشخاص ارائه شده است.

همان‌طور که در جدول (۳) می‌بینید، دو معیار دقت و بازخوانی برای داده‌های تاریخی و روایی مقادیر بسیار بالایی را دارند. یکی از اصلی‌ترین دلایل دست‌یابی به این دقت، وجود اسامی کامل (متشکل از نام اصلی، نام پدر، نام اجداد و نیز نام خانوادگی) است که اجزای آنها اغلب با استفاده از کلمهٔ «بن» از یکدیگر جدا شده‌اند. این نوع از اسامی یک الگوی بسیار قدرتمند برای تشخیص به دست می‌دهند. مشخص است که این ویژگیِ داده‌های تاریخی و روایی کار را برای سامانه بسیار ساده‌تر کرده است و دقت این سامانه بر روی دیگر داده‌ها به هیچ وجه به این مقدار نخواهد رسید؛ اما می‌توان ادعا کرد که دقّت سامانه روی داده‌های روایی و تاریخیِ دیگر نیز به همین مقدار خواهد رسید.

داده‌های فقهی اصولاً اسامی کمی را در بر می‌گیرند و تشخیص اسامی در آنها نیز سخت‌تر است؛ زیرا در مرحلهٔ آموزش، الگوهای زیادی را مشخص نمی‌کنند. این مطلب به‌وضوح در نتایج حاصله نیز نشان داده شده است.

(جدول- ۳): نتایج اجرای سامانه بر روی داده‌های تاریخی، روایی و فقهی

پیکرهٔ متنی	موضوع	دقت	بازخوانی	معیارf
وقعهٔ صفین	تاریخی	۹۹/۹۳٪	۹۹/۹۳٪	۹۹/۹۳
الارشاد	روایی	۹۵/۶۲٪	۹۲/۱۶٪	۹۳/۸۶
شرایع	فقهی	۱۰۰/۰۰٪	۶۰/۸۷٪	۷۵/۶۸

۵-۲- ارزیابی سامانه بر روی متون روزنامه‌ای

مدل میدان‌های تصادفی شرطی برای همهٔ پیکره‌های متنی ذکرشده در جدول (۳) آموزش دیده و آزموده شد. در این بخش نتایج حاصل از این روش با پیاده‌سازی‌های دیگر مقایسه است.

موفق‌ترین پیاده‌سازی‌ها برای تشخیص واحدهای اسمی، که امکان مقایسهٔ آنها با یکدیگر برای ما وجود دارد[1]، سه نرم‌افزار ANERsys، ANERsys 2 و Siraj هستند (Benajiba, 2007b). همچنین *Benajiba* و همکارش در سال ۲۰۰۸ نرم‌افزاری مبتنی بر میدان‌های تصادفی شرطی ارائه داده‌اند که نتایج بهتری را نسبت به برنامه‌های دیگر با استفاده از ترکیب چند مدل میدان‌های تصادفی شرطی تولید کرده است(Benajiba, 2008) . اما همان‌طور که در بخش ۲ نیز اشاره شد، تا امروز بهترین نتایج حاصل‌شده

[1] به این معنا که نسخهٔ متن باز یا تجاری با مستندات کافی برای آن موجود باشد

برای عملیات تشخیص اسامی بر روی داده‌های روزنامه‌ای ANERcorp، در مقالهٔ حریص *الجُمَیلی* و همکارانش (Al-Jumaily,2011) ارائه شده است. که البته روش معرفی‌شده توسط آنها از روش‌های یادگیری ماشین استفاده نمی‌کند و قوانین، به‌طور دستی توسط افراد خبره گردآوری شده‌اند.

یک اسکریپت ساده در کنفرانس آموزش محاسباتی زبان طبیعیِ سال ۲۰۰۲ به‌عنوان مبنایی برای مقایسهٔ خروجی عملیات تشخیص واحدهای اسمی معرفی شده است؛ اسکریپت خط مبنا[1] یا baseline در مرحلهٔ آزمودن، به هر کلمه برچسبی را تخصیص می‌دهد که در مرحلهٔ آموزش، آن برچسب به کلمهٔ مورد نظر به نسبت بیشتری تخصیص یافته است. نتایج حاصل از اجرای این اسکریپت در این بخش آورده شده است. همچنین نتایج حاصل از اجرای هر یک از برنامه‌های ذکرشده در سامانهٔ Noor ANER بر روی پیکرهٔ متنی ANERcorp در جداول (۴ تا ۱۰) مقایسه شده‌اند (Benajiba,2007a)، (Benajiba,2007b)، (Benajiba,2008) و (Al-Jumaily,2011). در جدول (۱۱) نتایج به‌دست آمده با استفاده از برچسب‌های ادات سخن و بدون استفاده از برچسب‌های کلمات نامزد اسمی (Benajiba,2008) ، بیان شده است.

۵-۳- نتایج و تحلیل خطاها

همان‌طور که در جدول (۹) مشاهده می‌شود، مقادیر مورد نظر در مورد اسامی متفرقه در مقالهٔ ارائه‌شده، ذکر نشده بودند و درنتیجه این مقادیر در اینجا ارائه نشده است. به‌علاوه با توجه به این که ردیف مجموع در جداول دیگر شامل اسامی متفرقه نیز هست، با این که ما نسبت تعداد هر کدام از انواع اسامی را در اختیار داریم و اعداد این سطر از روی بقیه قابل محاسبه هستند، ذکر آن اعداد نیز صورت نپذیرفته است؛ زیرا به‌طور اساسی حذف اسامی متفرقه، امکان مقایسهٔ صحیح را برای ما ناممکن می‌کند.

برای اینکه نتایج اعلام‌شده قابل مقایسه و دقیق‌تر باشند، این نتایج فقط بر روی پیکرهٔ متنی ANERcorp اجرا شده‌اند که اولین پیکرهٔ متنی ساخته‌شده برای تشخیص واحدهای اسمی در زبان عربی است و همچنین به‌دلیل در بر گرفتن داده‌های روزنامه‌ای، به نوعی سخت‌ترین نوع از مسألهٔ تشخیص واحدهای اسمی است.

همچنین با اینکه تمرکز اصلی این مقاله بر روی تشخیص اسامی اشخاص است، اما برای مقایسه، حاصل عملکرد مدل میدان‌های تصادفی شرطی بر روی انواع دیگر واحدهای اسمی نیز در اینجا آورده شده است. توجه کنید که برای مثال در تشخیص اسامی مکان‌ها، سازمان‌ها و واحدهای اسمی دیگر از هیچ فرهنگ لغتی استفاده نشده و یکی از دلایل پایین‌تربودن دقت در این زمینه‌ها، همین موضوع است.

در تشخیص اسامی خاص سامانهٔ Noor ANER، دقتی حدود یک واحد بیشتر از بهترین نتایج پیش از خود دارد. همان‌طورکه در بخش ۵-۱ دیدید، بر روی انواع دیگر داده‌ها (داده‌های روایی، تاریخی و فقهی) نیز نتیجهٔ بسیار خوبی تولید شده است.

میزان تأثیر افزودن برچسب‌های نامزد اسمی زمانی بهتر مشخص می‌شوند که توجه نماییم نتایج ذکرشده در بهترین رقیبِ مبتنی بر روش‌های یادگیری ماشین، که توسط Benajiba و همکارانش معرفی شده، حاصل ترکیب چند مدل میدان‌های تصادفی شرطی است که براساس مسندهای زیر بنا شده‌اند:

- برچسب‌های ادات سخن؛
- برچسب‌های جداساز عبارات پایه؛
- اطلاعات فرهنگ لغت؛
- ذکرشدن یک ملیت قبل از کلمه‌ی جاری.

و نتیجهٔ به‌دست آمده برآیند تمام این نتایج است؛ اما سامانهٔ Noor ANER از برچسب‌های جداساز عبارات پایه و ملیت بهره نمی‌برد و با این حال نتیجهٔ حاصل‌شده بالاتر از این سامانه است.

بررسی دقیق‌تر خطاهای سامانه به ما نشان داد که قسمت عمدهٔ اشتباهات سامانه، بیش‌تر بر روی اسامی خارجی وارد شده به زبان عربی است. بیشتر این اسامی حتی یک‌بار هم در فرهنگ لغت ذکر نشده‌اند (برای مثال اسامی افراد انگلیسی که طبیعتاً با رسم‌الخط عربی وارد داده‌های روزنامه‌ای شده‌اند) این چالش یکی از اصلی‌ترین چالش‌های هر سامانه تشخیص واحدهای اسمی محسوب شده و باعث پایین‌تر رفتن خطای بازخوانی می‌شود.

بررسی‌ها نشان دادند یکی از دیگر از دلایل عمدهٔ خطاها که باعث کم‌شدن دقت سامانه شده بود، تشخیص‌دادن صفت به‌عنوان اسامی اشخاص بود. همان‌طور که می‌دانید بیشتر اسامی در عربی درحقیقت می‌توانند نقش صفتی را نیز در جمله اختیار کنند.

[1] این نرم افزار از نشانی Http://Cnts.Ua.Ac.Be/Conll2002/Ner/Bin/Baseline قابل بارگیری است.

(جدول-۴): نتایج اجرای اسکریپت خط مبنا

اسکریپت خط مبنا	دقت	بازخوانی	معیارf
مکان‌ها	۷۵/۷۱٪	۷۶/۹۷٪	۷۶/۳۷
دیگر واحدهای اسمی	۲۲/۹۱٪	۳۴/۶۷٪	۲۷/۵۹
سازمان‌ها	۵۲/۸۰٪	۳۳/۱۴٪	۴۰/۷۲
اشخاص	۳۳/۸۴٪	۱۴/۷۶٪	۲۰/۵۶
مجموع	۵۱/۳۹٪	۳۷/۵۱٪	۴۳/۳۶

(جدول- ۵): نتایج اجرای نرم افزار Siraj

برنامهSiraj	دقت	بازخوانی	معیارf
مکان‌ها	۸۴/۷۹٪	۶۷/۹۱٪	۷۵/۴۲
دیگر واحدهای اسمی	۰٪	۰٪	۰
سازمان‌ها	۰٪	۰٪	۰
اشخاص	۷۴/۶۶٪	۵۵/۸۴٪	۶۳/۸۹
مجموع	۷۸/۹۵٪	۴۶/۶۹٪	۵۸/۵۸

(جدول- ۶): نتایج اجرای نرم افزار ANERsys

برنامهANERsys	دقت	بازخوانی	معیارf
مکان‌ها	۸۲/۱۷٪	۷۸/۴۲٪	۸۰/۲۵
دیگر واحدهای اسمی	۶۱/۵۴٪	۳۲/۶۵٪	۴۲/۶۷
سازمان‌ها	۴۵/۱۶٪	۳۱/۰۴٪	۳۶/۷۹
اشخاص	۵۴/۲۱٪	۴۱/۰۱٪	۴۶/۶۹
مجموع	۶۳/۲۱٪	۴۹/۰۴٪	۵۵/۲۳

(جدول- ۷): نتایج اجرای نرم افزار ANERsys 2

برنامهANERsys 2	دقت	بازخوانی	معیارf
مکان‌ها	۹۱/۶۹٪	۸۲/۲۳٪	۸۶/۷۱
دیگر واحدهای اسمی	۷۲/۳۴٪	۵۵/۷۴٪	۶۲/۹۶
سازمان‌ها	۴۷/۹۵٪	۴۵/۰۲٪	۴۶/۴۳
اشخاص	۵۶/۲۷٪	۴۸/۵۶٪	۵۲/۱۳
مجموع	۷۰/۲۴٪	۶۲/۰۸٪	۶۵/۹۱

(جدول- ۸): نتایج اجرای نرم‌افزار پیشنهادی-Benajiba ۲۰۰۸

نرم‌افزار پیشنهادی ۲۰۰۸Benajiba-	دقت	بازخوانی	معیارf
مکان‌ها	۹۳/۰۳٪	۸۶/۶۷٪	۸۹/۷۴
دیگر واحدهای اسمی	۷۱/۰۰٪	۵۴/۲۰٪	۶۱/۴۷
سازمان‌ها	۸۴/۲۳٪	۵۳/۹۴٪	۶۵/۷۶
اشخاص	۸۰/۴۱٪	۶۷/۴۲٪	۷۳/۳۵
مجموع	۸۶/۹۰٪	۷۲/۷۷٪	۷۹/۲۱

مشاهده می‌شود که نتیجه بیش از دو واحد بهتر شده است که این مقدار بهبود در مقایسه با روش‌های دیگر از نظر آماری بسیار با اهمیت است. به‌علاوه نتایج اشاره‌شده در بخش ۵-۱ هم باید مورد توجه قرار بگیرند.

(جدول- ۹): نتایج اجرای نرم‌افزار پیشنهادی جمیلی-۲۰۱۱

نرم‌افزار پیشنهادی جمیلی- ۲۰۱۱	دقت	بازخوانی	معیارf
مکان‌ها	۸۱/۴۹٪	۶۲/۷۰٪	۷۰/۸۷
دیگر واحدهای اسمی	-	-	-
سازمان‌ها	۶۵/۵۴٪	۵۰/۹۱٪	۵۷/۳۰
اشخاص	۷۷/۶۳٪	۷۴/۹۵٪	۷۶/۲۷
مجموع	-	-	-

(جدول- ۱۰): نتایج اجرای سامانه‌یNoor ANER

سامانهٔ Noor ANER	دقت	بازخوانی	معیارf
مکان‌ها	۷۵/۸۳٪	۷۹/۲۶٪	۷۷/۵۰
دیگر واحدهای اسمی	۶۹/۴۲٪	۳۳/۴۷٪	۴۵/۱۶
سازمان‌ها	۷۵/۹۰٪	۳۸/۳۴٪	۵۰/۹۵
اشخاص	۷۸/۵۱٪	۷۰/۵۷٪	۷۴/۳۳
مجموع	۷۶/۸۵٪	۶۱/۳۸٪	۶۸/۲۵

۶- نتیجه‌گیری و کارهای بیشتر

نتایج بخش‌های قبل نشان دادند که سامانهٔ Noor ANER با استفاده از تزریق کلمات نامزد اسم در تشخیص اسامی اشخاص در حوزه‌های گوناگون (متون روزنامه، تاریخی، روایی و فقهی) نسبت به سامانه‌های قابل مقایسهٔ دیگر بر روی داده‌های یکسان نتایج بهتری را ارائه می‌دهد. بررسی‌های صورت‌گرفته نشان می‌دهد دقت، بازخوانی و معیار f به‌طور خاص روی داده‌های روایی و تاریخی، بسیار بالا هستند.

یکی از کارهای ارزشمند صورت‌گرفته در جریان تولید این سامانه، ساخت پیکره‌های متنی مناسبی است که براساس استانداردهای کنفرانس آموزش محاسباتی زبان طبیعی برای زبان عربی ساخته شده است.

استفاده از برچسب‌های ادات سخن در آموزش مدل میدان‌های تصادفی شرطی، به‌نوعی استفاده از ویژگی‌های زبانی است که مدل در آن آموزش می‌بیند؛ اما درصورتی‌که سامانه‌ای برای برچسب‌گذاری ادات سخن در زبان دیگری موجود باشد، اِعمال مدلی مشابه مدل ارائه‌شده در این مقاله بر روی آن زبان ممکن است. این بدان معنی است که روش ارائه‌شده تا حدی قابل انتقال به زبان‌های دیگر است.

(جدول- ۱۱): نتایج اجرای مدل Benajiba مبتنی بر برچسب‌های ادات سخن

مدل Benajiba مبتنی بر برچسب‌های ادات سخن	دقت	بازخوانی	معیارf
مکان‌ها	٪۸۹/۸۸	٪۸۶/۴۹	۸۸/۱۵
دیگر واحدهای اسمی	٪۷۷/۹۱	٪۵۱/۱۵	۶۱/۷۵
سازمان‌ها	٪۸۳/۰۲	٪۵۳/۳۳	۶۴/۹۴
اشخاص	٪۷۹/۲۹	٪۶۵/۴۲	۷۱/۶۹
مجموع	٪۸۵/۲۸	٪۷۱/۸۲	۷۷/۹۷

اما همان‌طور که اشاره شد، در سامانه‌ای Noor ANER از ویژگی‌های منحصر به زبان عربی در بهبود عملکرد این سامانه استفاده نشده است. توسعه‌های آیندهٔ این سامانه می‌تواند بر پایهٔ افزودن توابع خصیصه‌ای که منحصر به ویژگی‌های زبان عربی هستند، بنا شوند.

این مقاله به‌طور خاص روی تشخیص اسامی اشخاص متمرکز شده است. پس یکی از توسعه‌های آیندهٔ آن می‌تواند دربرگرفتن تشخیص انواع دیگر واحدهای اسمی باشد. برای این منظور باید فرهنگ لغات مناسب نیز استخراج شود. همچنین مدل‌هایی که توسط محققان دیگر برای تشخیص واحدهای اسمی ساخته شده‌اند، از توابع خصیصهٔ مختلف، استفاده می‌کنند. برای مثال توابع خصیصهٔ ساخته شده با برچسب‌های ادات سخن، توابع خصیصهٔ ساخته‌شده با برچسب‌های جداساز عبارات پایه و ...؛ اما همان‌طور که شرح داده شد ما تنها از برچسب‌های ادات سخن در این سامانه استفاده کرده‌ایم. بهره‌گرفتن از چنین رهیافت‌هایی می‌تواند نتایج حاصله را بهبود ببخشد.

استفاده از نتایج ارائه‌شده توسط چند سامانهٔ تاحدودی متفاوت و ترکیب نتایج آنها (به‌عنوان مثال از طریق رأی‌گیری، Bagging یا Boosting) از معمول‌ترین روش‌های بهبود نتایجی در چنین سامانه‌هایی است (Romaszko,2012). بنابراین یکی دیگر از کارهای آینده می‌تواند استفاده از روش‌های ترکیبی باشد.

۷- مراجع

Al-Jumaily, H. and Martínez, P. and Martínez-Fernández, J.L. and Van der Goot, E. A real time Named Entity Recognition system for Arabic text mining. Language Resources and Evaluation, :1--21, 2011.

Alhelbawy, A., & Gaizauskas, R. (2012). Named Entity Based Document Similarity with SVM-Based Re-ranking for Entity Linking. Advanced Machine Learning Technologies and Applications, 379–388.

Benajiba, Y. and Rosso, P. ANERsys 2.0: Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information. Proc. of Workshop on Natural Language-Independent Engineering, IICAI-2007, 2007.

Benajiba, Y. and Rosso, P. Arabic named entity recognition using conditional random fields. Proc. of Workshop on HLT & NLP within the Arabic World, LREC, pages 143--153, 2008.

Benajiba, Y. and Rosso, P. and BenedíRuiz, J. Anersys: An arabic named entity recognition system based on maximum entropy. Computational Linguistics and Intelligent Text Processing, :143--153, 2007.

Benajiba, Y., Diab, M., & Rosso, P. (2009). Arabic named entity recognition: A feature-driven study. IEEE Transactions on Audio, Speech and Language Processing, 17(5), 926–934.

Chinchor, N. (1995). Statistical significance of MUC-6 results. Proceedings of the 6th Conference on Message Understanding, 39–43.

Diab, Mona. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagg-ing, and Base Phrase Chunking. In Choukri, Khalid and Maegaard, Bente, editors, Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2009. The MEDAR Consortium.

Elsebai, A., & Meziane, F. (2011). Extracting person names from Arabic newspapers. Innovations in Information Technology 87–89.

Elkan, Charles. Log-linear models and conditional random fields. 2008.

Fehri, H. and Haddar, K. and Hamadou, A.B. Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation

Model. International Workshop Finite State Methods and Natural Language Processing, pages 134, 2011.

Habash, Nizar and Soudi, Abdelhadi and Buckwalter, Timothy. On Arabic Transliteration. In Soudi, Abdelhadi and Bosch, Antal van den and Neumann, Günter and Ide, Nancy, editors, Arabic Computational Morphology in Text, Speech and Language Technology, pages 15--22. Springer Netherlands, 2007.

Lafferty, John and McCallum, Andrew and Pereira, Fernando. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Eighteenth International Conference on Machine Learning, 2001.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, (1991), 1–20.

Romaszko, L.; Fac. of Math., Inf. & Mech., Univ. of Warsaw, Warsaw, P. (2012). An Ensemble-Based Named Entity Recognition Solution for Detecting Consumer Products. In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference (pp. 865 – 868). Brussels.

Sutton, Charles. and McCallum, Andrew. An Introduction to Conditional Random Fields for Relational Learning. 2007.

Sang, Tjong Kim. Erik F. and De Meulder, Fien. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 in CONLL '03, pages 142-147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Webster, Jonathan J. and Kit, Chunyu. Tokenization as the initial phase in NLP. Proceedings of the 14th conference on Computational linguistics - Volume 4 in COLING '92, pages 1106–1110, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

مجید عسگری بیدهندی تحصیلات خود را در مقطع کارشناسی ارشد در دانشگاه علم و صنعت ایران در سال ۱۳۹۰ به پایان رسانده است. ولی در حال حاضر نیز دانشجوی مقطع دکترای هوش مصنوعی و رباتیک در همان دانشگاه است. حوزه‌های تحقیقاتی تخصصی وی عبارتند از: استخراج آزاد اطلاعات، سامانه‌های پرسش و پاسخ، بازیابی اطلاعات و ترجمه ماشینی.

نشانی رایانامهٔ ایشان عبارت است از

majid_asgari@comp.iust.ac.ir

بهروز مینایی بیدگلی: دکترای خود را در رشته علوم و مهندسی کامپیوتر از دانشگاه ایالتی میشیگان آمریکا در سال ۱۳۸۴ گرفت. تخصص او هوش مصنوعی و داده‌کاوی است. او هم اکنون به‌عنوان عضو هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه علم وصنعت به تدریس دروس هوش مصنوعی و نرم‌افزار مشغول می‌باشد. او سرپرستی گروه متن‌کاوی برای متون عربی و فارسی را در پژوهشکده داده‌کاوی نور را نیز به عهده دارد. از سال ۱۳۸۶ ریاست بنیاد ملی بازی‌های رایانه‌ای بر عهده ایشان است.

نشانی رایانامهٔ ایشان عبارت است از

b_minaei@iust.ac.ir