بسم الله الرحمن الرحیم

روش‌های تشخیص موجودیت‌های اسمی در متن

محسن ایمانی

86521062

Imany.mohsen@gmail.com

چکیده

در این پژوهش به بررسی روش‌های تشخیص و استخراج موجودیت‌های اسمی در متن پرداخته شده‌است. این کار را می‌توان با روش‌هایی از قبیل روش‌های مبتنی بر آمار، روش‌های بر مبنای استفاده از واژه‌نامه و روش‌های بر مبنای قاعده و با استفاده از عبارات باقاعده انجام داد. در این متن به معرفی این روش‌ها پرداخته و سپس به شرح برخی از روش‌های آماری تشخیص موجودیت‌های اسمی خواهیم پرداخت.

واژه‌های کلیدی: پردازش زبان‌های طبیعی، موجودیت‌های اسمی، مدل به‌هم‌ریختگی بیشینه، مدل پنهان مارکوف

مقدمه

موجودیت‌های اسمی در متن به عباراتی گفته می‌شود که حاوی اسامی اشخاص، سازمان‌ها و یا مکان‌ها باشند (Erik, Tjong, & Fien, 2003). اولین مرحله در اغلب کاربردهای استخراج اطلاعات تشخیص و دسته‌بندی موجودیت‌های اسمی در یک متن است. به این عملیات تشخیص موجودیت‌های اسمی[1] اطلاق می‌شود (Martin, & Jurafsky 2007). اسامی خاصی که تشخیص داده می‌شوند و همچنین قالبی که برای دسته‌بندی آن‌ها به کار می‌رود وابسته به نوع کاربرد آن خواهد بود. در سامانه‌های تشخیص موجودیت‌های اسمی عمومی بیشتر به سمت پیدا نمودن اسامی اشخاص، مکان‌ها و سازمان‌هایی که در یک متن معمولی خبری ذکر شده‌است تمرکز می‌شود. برنامه‌های عملیاتی نیز بر این مبنا برای تشخیص موارد گوناگونی از اسامی ژن‌ها و پروتئین‌ها *(Settles, 2005)* گرفته تا اسامی دوره‌های دانشگاهی *(McCallum, 2005)* به کار می‌رود.

روش تشخیص موجودیت‌های اسمی بر مبنای قواعد

در روش برمبنای قواعد[2] سعی می‌شود قالب و شکل کلی موجودیت‌های اسمی را به صورت یک عبارت باقاعده نمایش داده شده تا سامانه اسامی خاص را بر مبنای این عبارات تشخیص دهد. در واقع در این روش موجودیت‌های اسمی را به وسیله‌ی مولفه‌هایی که در ظاهر این عبارات ممکن است موجود باشد تشخیص می‌دهد. برای مثال در زبان انگلیسی دو حرف بزرگ مجاور هم احتمالا یک اسم خواهد بود و یا عباراتی که در آن‌ها کلمات و یا حروفی از قبیل Dr، .Mr و ... شروع می‌گردد و یا به حروفی از قبیل MD خاتمه می‌یابد احتمالا اسم یک شخص خواهد بود. شکل ظاهری بعضی از موجودیت‌های اسمی را می‌توان به صورت یک عبارات باقاعده نمایش داد. برای مثال نشانی ایمیل را می‌توان به صورت عبارت باقاعده‌ی (مثال1) نوشت (LingePipe: Named Entity Tutorial, 2003 - 2009).

```
[A-Za-z0-9](([_\.\-]?[a-zA-Z0-9]+)*)@([A-Za-z0-9]+)(([\.\-]?[a-zA-Z0-9]+)*)\.([A-Za-z]{2,})
```

(مثال 1)

برخی دیگر از موجودیت‌های اسمی از قبیل تاریخ، شماره‌ی تلفن، کد پستی و ... را به صورت عبارات باقاعده مشخص نموده، به وسیله‌ی آن‌ها این اسامی را از داخل متن تشخیص داد.

---

[1] Named – Entity Recognition(NER)
[2] Rule – Based

روش‌های مبتنی بر واژه‌نامه

در بسیاری از کاربردها، به نسبت ساده‌تر می‌باشد که یک لیست از موجودیت‌های اسمی و نوع آن‌ها تهیه شده و موجودیت‌های اسمی داخل متن با استفاده از این لیست تشخیص داده شود. برای مثال، اسامی مکان‌ها در فرهنگ‌های جغرافیایی[1] که شامل میلیون‌ها اسم مکان به همراه مشخصات جغرافیایی آن می‌باشد به عنوان یک واژه‌نامه مناسب برای تشخیص اسامی مکان‌ها قابل استفاده است.

در بعضی مواقع، تنها کافی است که به سادگی تمامی عبارات موجود در یک واژه‌نامه را که در متن ذکر شده‌اند پیدا کرد. برای مثال نشریه نیویورک تایمز[2] برای پیاده سازی صفحه سرفصل‌ها[3]ی خود از چنین روشی استفاده می‌نماید. در این روش یک واژه‌نامه‌ی حاوی اسامی سرفصل‌ها موجود است که به سادگی در هر متنی که یکی از این اسامی یافت شود، این متن به سرفصل مربوطه ارجاع داده می‌شود. این رویکرد در بسیاری از موارد به درستی عمل می‌نماید ولی در دو صورت با ابهام مواجه خواهد شد. یکی زمانی که واژه‌ی موجود در واژه‌نامه به صورت کامل استفاده نشده باشد؛ مثلا اگر در واژه‌نامه کلمه‌ی "وزارت علوم، تحقیقات و فناوری" موجود باشد، این روش قادر به تشخیص متنی که در آن تنها واژه‌ی "وزارت علوم" آورده شده است نخواهد بود. مورد دیگر ابهام در این روش هم زمانی خواهد بود که یک واژه معانی متفاوتی داشته باشد و در واژه نامه به اشتباه به معنی دیگری مرتبط شود. برای مثال ممکن است واژه‌ی "هما" در یک متن به عنوان اسم یک شخص استفاده شده باشد اما به اشتباه به عنوان مخفف "هواپیمایی ملی ایران" تشخیص داده شود (Martin, 2007 & Jurafsky).

نام‌هایی نظیر "50 Cent" و یا نام یک محصول شبیه به "xyx120 dvd player" را به سختی می‌توان توسط روش‌های آماری و یا روش های مبتنی بر قواعد تشخیص داد. معمولا واژه‌نامه می‌تواند در کنار روش‌های تشخیص آماری برای بهبود تشخیص کلماتی که با روش‌های آماری به سختی مشخص می‌شوند، به کار می‌رود.

---

[1] Gazetteer
[2] New York Times
[3] Times Topics

روش دیگر به این صورت می‌باشد که متن اصلی به وسیله‌ی روش‌های آماری پردازش گردد و موجودیت‌های اسمی خروجی آن به صورت یک واژه‌نامه برای پردازش مجدد متن و یافتن موجودیت‌هایی که در مرحله‌ی اول تشخیص داده نشدند استفاده گردد (LingePipe: Named Entity Tutorial, 2003 - 2009).

واژه‌نامه ها می‌توانند به دو صورت دقیق و یا تقریبی جستجو شوند. در روش دقیق عبارت موجود در واژه‌نامه باید عینا در متن آورده شده باشد در صورتی که در روش تقریبی کافی است با درصدی از تقریب دو عبارت با یکدیگر هم‌سان قلمداد شوند (Martin, 2007 & Jurafsky).

روش‌های آماری

این روش‌ها غالباً به عنوان روش‌های اصلی جهت تشخیص اسامی خاص شناخته می‌شوند و سایر روش‌ها به عنوان مکمل در کنار آن مورد استفاده قرار می‌گیرند.

بسیاری از مسائل در پردازش زبان‌های طبیعی قابل بیان به صورت مسائل دسته‌بندی آماری می‌باشند که در آن‌ها هدف تخمین زدن احتمال وقوع حالت a با محتوا b، یا به عبارت دیگر p(a,b) می‌باشد. محتوا در مسائل مربوط به پردازش زبان‌های طبیعی کلمات است که بسته به نوع مسئله ممکن است یک کلمه و یا عبارتی چند کلمه‌ای باشد. (Ratnaparkhi, 1997)

در این روش ابتدا سامانه به وسیله‌ی پیکره‌ای از داده‌های آموزشی که به صورت دستی و به وسیله‌ی انسان برچسب‌گذاری شده‌اند آموزش دیده، با یادگیری از طریق این داده‌ها به تشخیص خودکار اسامی خاص در متن می‌پردازد. برای برچسب‌زنی داده‌ی آموزشی از روش برچسب‌زنی شروع-داخل- خارج[1] استفاده می‌شود. در این روش، برای تشخیص موجودیت‌های اسمی، کلمات متن را تک تک برچسب‌زنی می‌نماییم (Erik, Tjong, & Fien, 2003). به این صورت که همزمان با برچسب‌زنی هم متن را بر اساس کلمات قطعه‌بندی نموده و هم با برچسب‌زنی عبارات موجودیت‌اسمی را مشخص می‌نماییم. روش متعارف این عمل برچسب‌زنی شروع-داخل-خارج است که برچسب‌های شروع(B) و داخل (I) هر قسمت از عبارت موجودیت اسمی را نمایش داده، همانگونه که برچسب خارج(O) قطعه‌های متن که خارج از این عبارات می‌باشند را نمایش می‌دهد. در این صورت، تعداد برچسب‌ها برای هر عبارت 2n+1 می‌باشد که در آن n نشان‌دهنده‌ی تعداد عبارات خواهد بود. در

---

[1] BIO tagging

این روش به صورت صریح پایان یک عبارت مشخص نمی‌شود بلکه همواره پایان عبارت با تغییر برچسب از I یا B به B یا O است (Martin, 2007 & Jurafsky).

مثال 2 یک جمله برچسب‌زده شده توسط روش شروع-داخل-خارج را نشان می‌دهد.

American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said.

American	B_{ORG}
Airlines	I_{ORG}
,	*O*
a	*O*
unit	*O*
of	*O*
AMR	B_{ORG}
Corp	I_{ORG}
.	*O*
,	*O*
Immediately	*O*
Matched	*O*
the	*O*
move	*O*
,	*O*
spokesman	*O*
Tim	B_{PERS}
Wagner	I_{PERS}
said	*O*
.	*O*

(مثال 2)

مدل به‌هم‌ریختگی بیشینه

به‌هم ریختگی بیشینه این اصل را بیان می‌دارد که توزیعی از P(a,b) صحیح خواهد بود که بیشترین میزان به‌هم‌ریختگی را با توجه به تموارد مشاهده شده داشته باشد. در واقع هنگام استنتاج بر مبنای اطلاعات نیمه‌کامل باید توزیع احتمالی استفاده گردد که بیشترین میزان به‌هم ریختگی در بین همه‌ی مقادیر معلوم را داشته باشد. این عمل تنها عمل تخصیصی می‌باشد که می‌توان اعمال نمود. برای استفاده از هر توزیع دیگری مجبور خواهیم بود که برخی پیش‌فرض‌های دل‌خواه را که در فرضیات مسئله وجود ندارد در نظر بگیریم (Jayenes, 1957).

به عبارت دیگر، اگر A به عنوان مجموعه حالات ممکن و B به عنوان مجموعه محتوای ممکن باشد، در این صورت p باید دارای به‌هم ریختگی بیشینه باشد.

$$H(p) = -\sum_{x \in \varepsilon} p(x) \log p(x) \qquad \text{(رابطه2)}$$

که در آن $x = (a,b)$، $a \in A$، $b \in B$، و $\varepsilon = A \times B$، و این میزان باید با انجام مشاهدات، یا اطلاعات نیمه‌کامل، ثابت باقی بماند.

یکی از راه‌های نمایش مشاهدات، کدگذاری حقایق مفید به صورت خصوصیات و اعمال محدودیت‌ها به مقادیر امید ریاضی این خصوصیات است. یک خصوصیت، تابعی با مقدار دودوئی بر روی رویدادها تعریف می‌شود.$\left(f_j : \varepsilon \rightarrow \{0,1\}\right)$ اگر تعداد k خصوصیت داشته باشیم، محدودیت‌ها به فرم $E_{\mathrm{p}}\mathrm{f_j} = E_{\tilde{\mathrm{p}}}\mathrm{f_j}$ خواهد بود که در آن $1 \leq j \leq k$ می‌باشد. $E_{\mathrm{p}}\mathrm{f_j}$ مدل امید ریاضی p برای f_j است.

$$E_{\mathrm{p}}\mathrm{f_j} = \sum_{\mathrm{x}\epsilon\varepsilon} \mathrm{p(x)f_j(x)}$$

(رابطه 2)

و با اعمال آن به امید ریاضی مشاهدات، برای $E_{\tilde{\mathrm{p}}}\mathrm{f_j}$ داریم:

$$E_{\tilde{\mathrm{p}}}\mathrm{f_j} = \sum_{\mathrm{x}\epsilon\varepsilon} \tilde{\mathrm{p}}\mathrm{(x)f_j(x)}$$

(رابطه 3)

که در آن $\tilde{\mathrm{p}}$ احتمال x در مشاهداتی است که روی داده‌ی تمرینی صورت می‌گیرد. بنابراین یک مدل p با مشاهدات انجام شده پایدار خواهد بود اگر و تنها اگر تمامی k محدودیت اعمال شده را محقق سازد. حال طبق اصل به‌هم ریختگی بیشینه از p^* به صورت موجود در رابطه 4 استفاده می‌شود.

$$P = \{p|E_p f_j = E_{\tilde{p}} f_j, j = \{1 \dots k\}\}$$

$$p^* = argmax_{p \in P} H(p)$$

(رابطه 4)

با توجه به این که این اصل به‌هم ریختگی را روی مجموعه مدل‌های پایدار P بیشینه می‌کند، p^* باید فرمی معادل با حالت رابطه 5 داشته باشد.

$$p^*(x) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(x)} \ , 0 < \alpha_j < \infty$$

(رابطه 5)

که در آن π ثابت نرمال‌سازی است و α_j ها پارامترهای مدل می‌باشند. هر پارامتر α_j مربوط به دقیقا یک خصوصیت f_j است و به عنوان وزن آن خصوصیت تلقی می‌گردد. (Ratnaparkhi, 1997)

روش استفاده از مدل به‌هم‌ریختگی بیشینه برای یافتن موجودیت‌های اسمی به این صورت شرح داده می‌شود. فرض کنید یک رشته‌ی ورودی طبیعی به صورت $w_1^N = w_1 \dots w_n \dots w_N$ موجود می‌باشد که از داخل آن رشته‌ی برچسب‌های موجودیت‌اسمی $c_1^N = c_1 \dots c_n \dots c_N$ را که بیشترین احتمال را بین تمامی رشته برچسب‌ها دارند را انتخاب می‌نماییم (Bender, Och, & Ney, 2003).

$$\hat{c}_1^N = \underset{c_1^N}{argmax} \{Pr(c_1^N | w_1^N)\}$$

(رابطه 6)

در این مدل برای هر کلمه‌ی از رشته‌ی ورودی احتمال ثانویه به طور مستقیم اعمال شده و برچسب موجودیت اسمی نظیر آن نیز مدنظر قرار می‌گیرد. فرض می‌کنیم که تصمیم‌ها تنها به پنجره‌ای محدود به $w_{n-2}^{n+2} = w_{n-2} \dots w_{n+2}$ از کلمه‌ی فعلی w_n و بر روی دو برچسب قبلی آن خواهد بود. بنابراین، مدل مرتبه دوم به شکل رابطه 7 بدست می‌آید:

$$Pr(c_1^N|w_1^N) = \prod_{n=1}^{N} p(c_n|c_{n-2}^{n-1}, w_1^N) \xrightarrow{model} \prod_{n=1}^{N} p(c_n|c_{n-2}^{n-1}, w_{n-2}^{n+2})$$

(رابطه 7)

بستر مناسبی که بتواند مستقیما احتمال ثانویه‌ی $p\left(c_n|c_{n-2}^{n-1}, w_{n-2}^{n+2}\right)$ مدل به‌هم‌ریختگی بیشینه می‌باشد. در این بستر، یک مجموعه از M تابع خصوصیت موجود است که $h_m(c_{n-2}^{n-1}, c_n, w_{n-2}^{n+2})$ و m = 1,...,M می‌باشد. برای هر تابع خصوصیت[1] h_m، یک پارامتر مدل λ_m وجود دارد. در این صورت احتمال ثانویه به صورت رابطه 8 مدل خواهد شد:

$$\Pr\left(\lambda_1^M, c_n|c_{n-2}^{n-1}, w_{n-2}^{n+2}\right) = \frac{exp\left[\sum_{m=1}^{M} \lambda_m h_m(c_{n-2}^{n-1}, c_n, w_{n-2}^{n+2})\right]}{\sum_{c'}\left[\sum_{m=1}^{M} \lambda_m h_m(c_{n-2}^{n-1}, c', w_{n-2}^{n+2})\right]}$$

(رابطه 8)

تابع‌های خصوصیت

در این قسمت به معرفی برخی از توابع خصوصیت مورد استفاده برای پیاده‌سازی مدل‌های به‌هم ریختگی بیشینه پرداخته می‌شود.

خصوصیات لغوی: کلمات w_{n-2}^{n+2} با یک واژه‌نامه مقایسه می‌شود. کلماتی که کم‌تر از دو بار در داده‌ی تمرینی آمده باشند در دسته‌ی کلمات ناشناخته قرار می‌گیرند. در واقع این خصوصیت زمانی فعال است که کلمه‌ی w_{n+d} با کلمه‌ی w در واژه‌نامه برابر باشد و محاسبه برچسب فعلی موجودیت اسمی برابر c باشد.

$$h_{w,d,c}(c_{n-2}^{n-1}, c_n, w_{n-2}^{n+2}) = \delta(w_{n+d}, w) \cdot \delta(c_n, n), d \in \{-2, \dots, 2\}$$

(رابطه 9)

[1] Feature function

خصوصیت‌های ظاهری کلمه: ویژگی‌های ظاهری کلمات در این توابع خصوصیت پوشانده می‌شوند. این ویژگی‌ها عبارت‌اند از:

- بزرگی حروف: این خصوصیت زمانی فعال می‌شود که W_n دارای یک حرف داخلی بزرگ باشد و یا کاملا با حروف بزرگ نوشته شده باشد.
- اعداد و رقم‌ها: رشته‌های کد اسکی[1] و اعداد این خصوصیت را فعال می‌سازند.
- پیشوند و پسوندها: اگر یک پیشوند یا پسوند از W_n برابر با پیشوند یا پسوند داده شده باشد، این خصوصیت فعال می‌شود.

خصوصیت‌های انتقال: خصوصیت‌های انتقال وابستگی را به دو برچسب پیشین مدل می‌کند:

$$h_{\acute{c},\,d,c}(c_{n-2}^{n-1}, c_n, w_{n-2}^{n+2}) = \delta(c_{n-d}, \acute{c}).\,\delta(c_n, n)\,,\ d \in \{1,2\}.$$

(رابطه 10)

خصوصیت‌های پیشینی: عبارات پیشینی یک موجودیت اسمی منفرد توسط این خصوصیت‌ها شامل می‌شوند. این خصوصیت‌ها تنها برای برچسبی که در حال مشاهده شدن در حال حاضر است فعال می‌گردد.

$$h_c(c_{n-2}^{n-1}, c_n, w_{n-2}^{n+2}) = \delta(c_n, n)$$

(رابطه 11)

[1] ASCII

مدل پنهان مارکوف

مدل‌های پنهان مارکوف[1] یکی از قوی‌ترین ابزارها برای پردازش سیگنال‌ها می‌باشند. انواع مختلف مدل‌های پنهان مارکوف علیرغم محدودیت‌هایی که دارند، هنوز پراستفاده‌ترین تکنیک در سیستم‌های مدرن بازشناسی گفتار و تشخیص متون هستند. مدل پنهان مارکوف، کل الگوی ورودی را به عنوان یک بردار ویژگی تکی مدل نمی‌کند، بلکه رابطه بین بخش‌های متوالی یک الگو را استخراج می‌کند، زیرا هر بخش نسبت به کل ورودی کوچک‌تر و بنابراین مدل‌سازی آن ساده‌تر است (حاجی، 1384).

یک مدل پنهان مارکوف را در واقع می‌توان یک ماشین حالت محدود[2] احتمالاتی به حساب آورد که هر حالت با یک تابع تصادفی مرتبط است. فرض می‌شود که در یک دوره‌ی گسسته‌ی از زمان t، مدل در یک حالت است و با یک تابع تصادفی از آن حالت یک خروجی تولید می‌کند. بر مبنای احتمال انتقال حالت جاری، مدل پنهان مارکوف در زمان t+1 تغییر حالت می‌دهد. دنباله‌ی حالت‌هایی که مدل از آن می‌گذرد معمولا پنهان است، و تنها یک تابع احتمالاتی از آن آشکار است، که مشاهدات تولید شده به وسیله تابع تصادفی مربوط به حالت‌ها است؛ به همین دلیل در نام‌گذاری این مدل‌ها از صفت پنهان استفاده می‌شود. در واقع می‌توان یک مدل پنهان مارکوف را یک فرآیند تصادفی به طور ناتمام مشاهده شده در نظر آورد. یک مدل پنهان مارکوف با عناصر زیر توصیف می‌شود:

1- N: تعداد حالت‌های مدل

2- $S = \{s_1, s_2, ..., s_N\}$: مجموعه‌ی حالت‌ها

3- $\Pi = \{\pi_i = P(s_i \text{ at } t = 1)\}$: احتمالات حالت اولیه

4- $A = \{aij = P(sj \text{ at } t+1 \mid si \text{ at } t)\}$: احتمالات تغییر حالت

5- M: تعداد علائم قابل مشاهده (تولید شده)

6- $V = \{v1, v2, ..., vM\}$: مجموعه علائم قابل مشاهده

7- $B = \{bi(vk) = P(vk \text{ at } t \mid si \text{ at } t\}$: احتمالات تولید علائم قابل مشاهده

8- O_t: علامت مشاهده شده در زمان t

---

[1] Hidden Markov Models (HMM)
[2] Finite State Machine

9- T: طول دنباله مشاهدات

10- :$\lambda = (A, B, \Pi)$ نماد خلاصه‌ی برای مدل پنهان مارکوف

واضح است که بر احتمالات Π، A و B سه قید وجود دارد: $\sum_{i=1}^{N} \pi_i = 1$، $\forall i$ ، $\sum_{i=1}^{N} a_{ij} = 1$ و $\sum_{k=1}^{M} b_i(v_k) = 1$ ، $\forall i$.

منابع و مراجع

[1] Erik, F., Tjong, K. S., & Fien, D. M. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *CoNLL-2003* (p. 6). CNTS - Language Technology Group University of Antwerp.

[2] Jurafsky, D., & Martin, J. H. (2007). *Speech and Language Processing: An introduction to natural language processing,.* Draft.

[3] Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity namesin text. *Bioinformatics*,

[4] McCallum, A., Freitag, D., and Pereira, F. C. N. (2000). Maximum entropy Markov models for information extraction and segmentation.

[5] *LingePipe: Named Entity Tutorial*. (2003 - 2009). Retrieved June 28, 2010, from allias-i.com: http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html

[6] Ratnaparkhi, A. (1997). *A Simple Introduction to Maximum Entropy Models for Natural Language Processing.* Pennsylvania: Institute for Research in Cognitive Science.

[7] Jayenes, E. T. (1957). *Information Theory and Statistical Mechanics.*

[8] Bender, O., Och, F. J., & Ney, H. (2003). Maximum Entropy Models for Named Entity Recognition. p. 4.

[9] حاجی, م. (1384). *مدل سازی آماری زبان فارسی*. شیراز: دانشکده مهندسی دانشگاه شیراز.