WI&CRF: روش پیشنهادی برای استخراج اطّلاعات مورد نیاز از متون نظامی

نیما ملایی[1]، احمد عبداله‌زاده[2] و حسین شیرازی[3]
[1 و 3] دانشگاه صنعتی مالک اشتر
[2] دانشگاه صنعتی امیر کبیر، دانشکدهٔ مهندسی فناوری اطّلاعات و کامپیوتر

چکیده

امروزه تکنیک‌های استخراج اطّلاعات مورد نیاز، از متون نظامی، مورد توجه فرماندهان و مدیران نظامی قرار گرفته که به‌دلیل تفاوت‌های ساختاری متون نظامی در مقایسه با متون غیر نظامی، استفاده از روش‌های متداول موجود بر روی متون نظامی، فاقد کارایی لازم است. در این مقاله ضمن مقایسهٔ ساختار متون نظامی با متون غیرنظامی، دسته بندی جدیدی از متون نظامی ارایه‌شده و نتایج استخراج اطّلاعات در هر گروه با استفاده از سامانه‌های موجود، مورد بررسی قرار گرفته است. براساس نتایج حاصل از ارزیابی، روشی نوین بر مبنای ترکیب روش استنتاج پوشش و مدل حوزهٔ تصادفی شرطی ارایه‌شده است. ارزیابی کارایی بر روی بستر تهیه‌شده از گزارش‌های صحنهٔ نبرد، انجام شده است. با به‌کارگیری روش پیشنهادی بر روی بستر ارایه شده، بهبود کارایی در معیارهای فراخوانی و معیار F نمایش داده شده است. همچنین با مقایسهٔ نتایج حاصل از اجرای این سامانه بر روی متون غیر نظامی، امکان به‌کارگیری آن برای استخراج اطّلاعات از متون غیر نظامی نیز بررسی شده است.

واژگان کلیدی: متن نظامی، استخراج اطّلاعات، استنتاج پوشش، مدل حوزهٔ تصادفی شرطی

۱- مقدمه

افزایش چشم‌گیر استفاده از رایانه‌های شخصی و سازمانی باعث انباشت حجم زیادی از داده و اطّلاعات در رایانه‌ها شده و این موضوع سازمان‌ها را با مشکل جدیدی با عنوان استخراج اطّلاعات و کسب دانش، مواجه ساخته است. تاکنون از روش‌های بازیابی اطّلاعات[1] برای حل این مسئله، استفاده شده است. امّا مشکلات موجود در بازیابی اطّلاعات از جمله عدم توجّه به معنا در متن، باعث شده تا شاخهٔ جدیدی از علم رایانه با عنوان استخراج اطّلاعات[2] ایجاد و به‌شدّت مورد توجه قرار گیرد.

از آنجا که به‌طور معمول متون نظامی ساختار خاص دارند و جملات به‌صورت کامل بیان نمی‌شود، استفاده از برچسب‌های[3] نحوی به‌تنهایی کارایی مناسب برای استخراج اطّلاعات از این‌گونه متون را ندارد. پژوهش‌های انجام‌گرفته در زمینهٔ استخراج اطّلاعات مانند *Crystal* (*Ciravegna, 2001*)، *(LP)2* (*Soderland, 1995*)، *iASA* (*Tang, 2005*)، *AutoSlog* (*Riloff, Automatically constructing a dictionary for information extraction tasks, 1993*)، *Whisk* (*Rosenfield, B. et al;, 2009*) که با استفاده از ساختار نحوی متن، به استخراج اطّلاعات می‌پردازند، قادر به استخراج اطّلاعات از این‌گونه متون نیستند. لذا روش پیشنهادی با این فرض که استفاده از حاشیه‌نوشته‌های دیگری مانند اطّلاعات ساختاری و یا نحوی می‌تواند به استخراج بهتر اطّلاعات از این نوع متون کمک کند، پایه‌گذاری شده است.

در این مقاله روشی بر اساس ترکیبی از استنتاج پوشش[4] (*Kushmerick, 1997*) و مدل حوزهٔ تصادفی شرطی[5] (*Lafferty, 2001*) پیشنهاد شده که در آن، حوزهٔ تصادفی شرطی با استفاده از ساختار متن نظامی آموزش دیده و از این طریق کارایی سامانهٔ استخراج اطّلاعات از متن نظامی را بهبود بخشیده است.

در بخش دوم، متون نظامی و ساختار آنها بررسی شده و بخش سوم، مروری بر روش‌ها و

[1] Information Retrieval (IR)
[2] Information Extraction (IE)
[3] Tag
[4] Wrapper Induction
[5] Conditional Random Field (CRF)

الگوریتم‌های استخراج اطّلاعات و همچنین معرّفی تعدادی از سیستم‌های ارایه شده مبتنی بر این روش‌ها می‌باشد. در ادامه تأثیر ویژگی‌های ساختاری متون نظامی بر نتایج استخراج اطّلاعات بررسی و روش پیشنهادی برای بهبود کارایی استخراج اطّلاعات ارایه‌شده است. در پایان روش پیشنهادی با استفاده از داده‌های موجود در پایگاه داده ویکی‌لیکس[1] مورد آزمایش قرار گرفته و نتایج حاصله با نتایج سامانه‌های مشابه مورد مقایسه و ارزیابی قرار گرفته است.

۲- بررسی ساختاری متون نظامی

بررسی ساختاری متن و نشانه‌گذاری آن، ارزش اطّلاعاتی متن را ارتقا می‌بخشد (*Abolhassani M., 2003*) و بدون نشانه‌گذاری، از نقطه‌نظر سامانه‌ای، یک سند فقط یک سلسلهٔ طولانی از کلمات است. بنابراین مجموعه عمل‌گرهایی که می‌تواند بر روی این نوع متن کار کند محدود است. درصورت نشانه‌گذاری متن، سامانه قادر است تا معانی پنهان در متن را به‌کمک حاشیهٔ نوشته‌ها، آشکار کرده و عملیات‌ها را به سطح معنایی و مفهومی نزدیک‌تر کند.

بررسی ساختاری متن در سه سطح انجام می‌پذیرد (*Abolhassani M., 2003*):

الف- بررسی ساختاری سطح کلان[2]: نشانه‌گذاری ساختار کلان سند، براساس نمایش سند یا بر اساس ارتباط بخش‌های سند با یکدیگر (منطق) مرتبط است. به‌منظور استخراج ساختار نمایشی متن، ابزارهایی نیز با عنوان استنتاج پوشش ارایه شده‌اند. مانند استنتاج پوشش[3] (*Kushmerick, 1997*)، استنتاج پوشش افزایشی[4] (*Freitag, 2000*) و استنتاج پوشش سلسله‌مراتبی[5] (*Muslea, 2001*)

ب- بررسی ساختاری سطح خرد[6]: در این حالت، سند از نظر محدودهٔ جملات، بخش‌های هر جمله و نقش‌های نحوی آنها بررسی شده و نشانه‌هایی به گروه کلمات (اسم، فعل و ...) اختصاص می‌یابد. ابزارهایی که به بررسی ساختاری خرد متن می‌پردازند نیز تحت عنوان تجزیه‌کننده‌های نحوی ارایه‌شده‌اند که می‌توان به (*Collins, 1999*) و (*Nivre, 2006*) اشاره کرد.

ج- در سطح نشانه: ساختار متن از نظر وجود نشانه‌ها، به‌منظور تشریح اطّلاعاتی که در محتوای متن به‌صورت صریح وجود ندارد، بررسی می‌شود. (به‌عنوان مثال، در متنی که دارای فرمول‌های ریاضی و شیمی است، به‌منظور تفکیک آنها از نشانه MathML برای فرمول‌های ریاضی و CML برای فرمول‌های شیمی استفاده می‌شود).

به‌منظور برررسی ساختاری متون نظامی می‌بایست آنها را با رویکرد تحلیلی، طبقه‌بندی نیم. دسته‌بندی‌های صورت‌گرفته در منابع موجود مانند *(Hecking, 2003)* غالباً در راستای کاربرد محدود صورت گرفته است. در ادامه به بررسی ساختاری این گروه از متون می‌پردازیم.

۲-۱- بررسی منابع متنی نظامی

در این بخش به بررسی منابع متنی نظامی با رویکرد چگونگی تحلیل آنها می‌پردازیم. منابع متنی نظامی به‌صورت عمده به دو گروه اخبار نظامی و گزارش‌های صحنهٔ نبرد طبقه‌بندی می‌شوند. اخبار نظامی را در مجلاتی مانند *Jane's Military Journal* (*Jane's Military Journal*) می‌توان مشاهده نمود و گزارش‌های صحنهٔ نبرد در منابعی مانند KFOR (*Hecking M., 2005*) و ویکی‌لیکس (*Afghanestan War Reports*) در دسترس هستند.

به‌منظور آشکارسازی ارزش اطّلاعاتی موجود در متون نظامی و بررسی تفاوت‌های ارزشی موجود در این نوع از متون، در مقایسه با سایر متون غیر نظامی، باید آنها را از نظر ساختاری کلان و خرد مورد بررسی قرار داد. از این‌رو تعدادی از منابع متنی نظامی موجود انتخاب و با منابع مشابه از حوزهٔ اخبار غیرنظامی مقایسه و تفاوت‌های ساختاری آنها مقایسه شده است.

۲-۱-۱- بررسی ساختار کلان متون نظامی

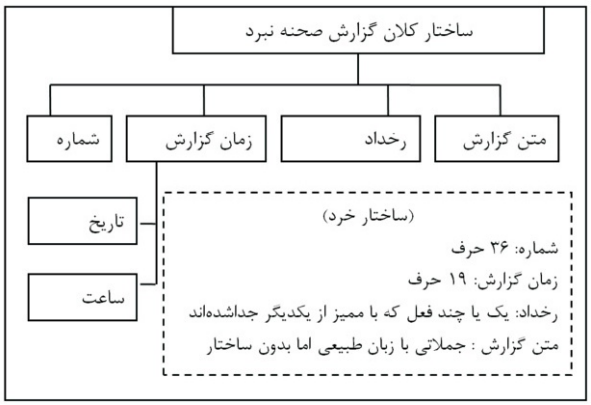
به‌منظور بررسی ساختار کلان گزارش‌های صحنهٔ نبرد، بیش از هفتصد گزارش مربوط به جنگ افغانستان که از سال ۲۰۰۴ جمع‌آوری و در سایت ویکی‌لیکس موجود است، انتخاب و ساختار منطقی آن مطابق (شکل ۱) از آنها استخراج شد.

به‌منظور بررسی ساختار کلان اخبار نظامی، بر روی تعداد حدود ۴۰۰ مقاله و مطلب موجود در ۴۱ شماره مجلهٔ هفتگی نظامی جینز، بررسی ساختاری صورت‌گرفته و ساختار نمایشی و منطقی مندرج در (شکل ۲) برای متون موجود در این منبع اطّلاعاتی استخراج شد.

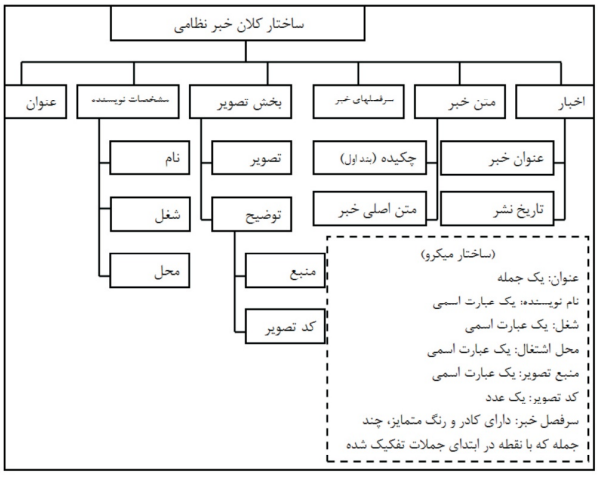
همچنین جهت بررسی ساختاری متون غیر نظامی، مطالب موجود در روزنامهٔ غیر نظامی رویترز (routers) بررسی شد. ساختار موجود در مطالب این نشریه خبری غیر نظامی، همانند اخبار نظامی و مطابق با (شکل ۲) است.

[1] Wikileacks
[2] Macro
[3] Wrapper Induction
[4] Boosted wrapper induction
[5] Hierarchical wrapper induction
[6] Micro

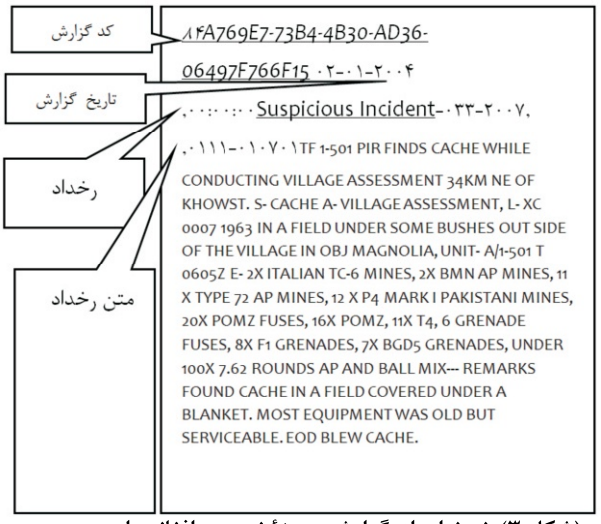
با مقایسهٔ ساختار گزارش‌های نظامی (شکل ۱) و ساختار متون خبری (شکل ۲) می‌توان نتیجه گرفت که ساختار گزارش‌های نظامی در مقایسه با ساختار اخبار نظامی و غیر نظامی از ساخت‌یافتگی کم‌تری دارند. همچنین با توجّه به آنکه (شکل ۱) نمایشی درختی ساختارکلان از (شکل ۳) است، لذا بدون داشتن قالب گزارش نمی‌توان به محتوای گزارش‌های نظامی دسترسی پیدا کرد.



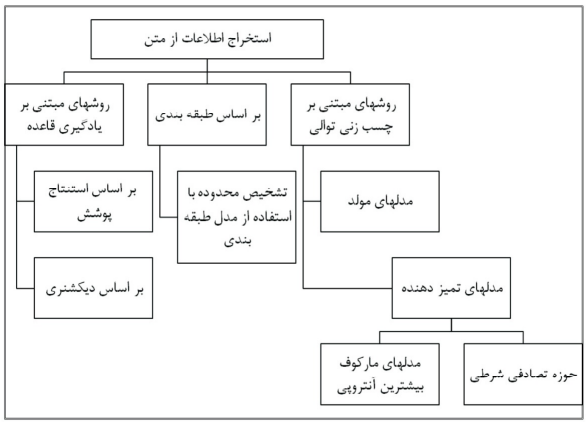
(شکل ۱): نمایش درختی ساختار گزارش نظامی



(شکل ۲): نمایش درختی ساختارمتون خبری نظامی و غیر نظامی



(شکل ۳): نمونه‌ای از گزارش صحنهٔ نبرد در افغانستان موجود در ویکی‌لیکس به همراه ساختار کلان



(شکل ۴): نمایش درختی روش‌های استخراج اطّلاعات

۲-۱-۲- بررسی ساختار خرد [1] در متون نظامی

با بررسی ویژگی‌های مربوط به گزارش‌های نظامی که در توضیحات مندرج در (شکل ۱) نمایش داده شده است و همچنین توضیحات (شکل ۲) که مربوط به ساختار متون خبری است، می‌توان مشاهده کرد که متون خبری (نظامی و غیر نظامی) در مقایسه با گزارش‌های نظامی صحنهٔ نبرد، علاوه‌بر تفاوت در ساختار کلان، در سطح خرد با یکدیگر تفاوت‌هایی دارند. به‌عنوان مثال، عنوان خبر در یک متن، به‌صورت جمله یا عبارتی که تشریح‌کنندهٔ یک خبر باشد، ذکر شده است؛ امّا در گزارش‌های صحنهٔ نبرد با استفاده یک یا چند عبارت فعلی که بیان کننده رخداد یا رخدادهایی باشد، بیان شده است. داین موضوع در (شکل ۲) نمایش داده شده است.

۳- روش‌ها و الگوریتم‌های استخراج اطّلاعات

از دههٔ ۱۹۵۰ میلادی که برای اوّلین‌بار ایدهٔ استخراج اطّلاعات توسط دانشمندی آمریکایی با نام زلاهریس [2] برای تبدیل خلاصهٔ تأییدیه بیماران به ساختار جدولی در بیمارستان مورد استفاده قرار گرفت، تا امروز که کنفراس‌های متعددی مانند کنفرانس استخراج خودکار محتوا [3] برگزار می‌شود، ابعاد مختلفی از استخراج اطّلاعات مورد توجّه دانشمندان قرار گرفته است *(Grishman, 1997)* . امّا آنچه که امروزه همگان بر آن اتفاق نظر دارند، معنادارکردن کلمات، یندها یا ترکیبی از آنها در متن می‌باشد *(Fieldman*

[1] Micro Structure
[2] Zalaharis
[3] Automatic Content Extraction (ACE)

(R. et al, 2007 لذا فرایند معنا دارکردن می‌تواند به‌صورت تبدیل متن به یک فایل جدول یا به صورت یک متن حاشیه نویسی شده باشد.

از این رو برای انجام استخراج اطّلاعات از متن باید حداقل چهار فرایند زیر صورت پذیرد (*Fieldman R. et al, 2007*):

- **بخش‌بندی**[1]: با بخش‌بندی، متن به ساختار تشکیل دهنده خود (کلمه، جمله و پارگراف) تفکیک می‌شود.
- پردازش واژگانی و ریخت شناسی[2]: به تشخیص اشکال کلمه صرف شده می‌پردازد.
- **تحلیل نحوی**[3]: به خروجی مرحلهٔ قبل اعمال شده و می‌تواند به‌صورت تجزیهٔ کم‌عمق یا عمیق باشد.
- **تحلیل دامنه**[4]: درنهایت قالب تعیین‌شده بر اساس خروجی مرحله قبل تکمیل می‌شود.

همان‌گونه که در (شکل ۴) نشان داده شده است، برای استخراج اطّلاعات از متن، سه رویکرد اصلی وجود دارد. این رویکردها بر یادگیری قاعده[5]، مدل طبقه‌بندی[6] و مدل مبتنی بر برچسب زنی توالی[7] متمرکز هستند. هر یک از این این روش‌ها از نوع یاد گیری ماشین با نظارت و دارای دو مرحله یادگیری و استخراج می‌باشند.

سامانه‌های WIEN (*Kushmerick, 1997*) و *BWI* (*Freitag, 2000*) نمونه‌هایی از سامانه‌های ایجاد شده به‌منظور استنتاج پوشش می‌باشند. همچنین (*Dalvi N., et al, 2009*) نیز نمونه‌ای از آخرین تحقیقات صورت گرفته در همین زمینه است.

در برچسب‌زنی توالی، یک سند به‌صورت کلمات متوالی دیده‌شده و برچسب‌های متوالی برای نشان‌دادن ویژگی هر کلمه به آن منتسب می‌شود.

نمونه‌ای از این نوع برچسب‌زنی را می‌توان در پردازش زبان طبیعی مشاهده کرد. در این عمل، هر کلمه با یک حاشیهٔ نوشته که نشان‌دهندهٔ ادات سخن[8] متناظرش است برچسب زده می‌شود.

یکی از روش‌های متداول برای این نوع برچسب‌زنی، استفاده از مدل مولد[9] است. مدل مخفی مارکوف[10] نمونه‌ای از یک مدل مولد است. در این مدل یک توزیع احتمالی اتصال $P(x,y)$ که در آن x و y متغیرهای تصادفی که نشان‌دهندهٔ "توالی مشاهده[11]" و "توالی برچسب[12]" می‌باشند، تعیین می‌شود. به‌منظور تعیین مقدار احتمال اتصال، مدل‌های مولد می‌بایست همهٔ توالی‌های مشاهدهٔ ممکن را محاسبه کنند. به‌عبارت ساده‌تر، اجزای مشاهده در هر نمونهٔ داده‌شده در هر زمان می‌بایست به‌طور مستقیم وابسته به حالت یا برچسب در همان زمان باشد. این فرضیه برای تعداد کمی از مجموعه‌های داده‌ای ساده مناسب است؛ امّا در دنیای واقعی، بسیاری از توالی‌های مشاهده، به‌صورت عباراتی دارای خصایصی با اثر متقابل و اجزای مشاهده با وابستگی خیلی زیاد نمایش داده می‌شوند.

مدل مطلوب، مدلی است که استنتاج‌های ساده‌ای را پشتیبانی کرده، و داده را بدون آنکه نیاز به ایجاد فرضیات مستقل غیر قابل توجیه باشد، نمایش دهد. یکی از روش‌های مرتفع کردن این مسئله، استفاده از مدل تصادفی[13] است. در این مدل احتمال شرطی $p(Y|x)$ برروی دنباله‌های برچسب که به توالی مشاهدهٔ معین x متعلق است، به‌جای توزیع اتصال بروی هر دو دنباله مشاهده و برچسب تعریف می‌کند. مدل‌های تصادفی، برای برچسب‌زنی یک توالی مشاهده جدید x^*، از طریق انتخاب توالی برچسب y^* که احتمال شرطی $p(y^*|x^*)$ را بیش‌تر می‌کند، به‌کار می‌رود. در این مدل، خصایص دل‌خواه از دادهٔ مشاهده، بدون نگرانی از ارتباط این خصایص با یکدیگر، می‌تواند همراه مدل نگهداری شود.

مدل‌های تصادفی شرطی، چارچوبی احتمالی برای برچسب‌زنی بر اساس روش‌های ارایه‌شده در بند بالا هستند. این مدل، شکلی از مدل گرافیکی بدون جهت است که یک توصیف ساده از لگاریتم خطی بر روی توالی‌های برچسب متعلق به یک توالی مشاهده تعریف می‌کند. مزیت اصلی مدل‌های تصادفی شرطی نسبت به مدل مخفی مارکوف، ماهیت شرطی آنهاست که باعث می‌شود از مفروضات استقلال (که برای مدل مخفی مارکوف به منظور تسهیل در انجام استنتاج مورد نیاز است) رها شده و بتوانیم آنها را حذف کنیم. به‌علاوه مدل‌های تصادفی شرطی از مسئله بایاس برچسب[14] نیز دوری می‌کنند (*Lafferty, 2001*). این ضعف در مدل‌های مارکوف بیش‌ترین بی‌نظمی[15] (*McCallum, 2000*) و سایر روش‌های مدل مارکوف که بر روی گراف جهت دار تعریف می‌گردند، مشاهده می‌شود.

اشکال موجود در این نوع از سامانه‌ها، استفاده از برچسب‌هایی از نوع ادات سخن بوده که دارای ماهیت نحوی

[1] Tokenization
[2] Morphological and Lexical Processing
[3] Syntactic Analysis
[4] Domain Analysis
[5] Rule learning based method
[6] Classification model based method
[7] Sequential labeling based method
[8] Part Of Speech (POS)
[9] Generative Model
[10] Hidden Markov Model
[11] Observed Sequence
[12] Label Sequence
[13] Random Model
[14] Label Bias Problem
[15] Maximum Entropy Markov Model

هستند. این برچسب‌ها نسبت به برچسب‌های معنایی[1]، در استخراج اطّلاعات ارزش کمتری دارند. بنابراین استفاده از برچسب‌هایی که دارای ارزش معنایی برای متن مورد بررسی باشد، می‌تواند در بهینه کردن نتایج استخراج اطّلاعات، مفید باشد.

در حوزهٔ تصادفی شرطی (*Lafferty, 2001*)، استخراج اطّلاعات به‌صورت عمل برچسب‌زنی دنباله فرض می‌شود. در برچسب‌زنی دنباله، یک سند به‌صورت بخش‌هایی از جمله به‌صورت متوالی دیده‌شده و برچسب‌ها برای نشان‌دادن ویژگی هر بخش[2] به آن منتسب می‌شود.

در روش حوزهٔ تصادفی شرطی، یک توزیع احتمال از مجموعه‌ای از متغیرهای تصادفی برچسب (Y)، بر روی بخشی از مجموعه نشانه‌ها (X)، تعریف می‌شود. درصورتی‌که اعضای X با x_i و اعضای Y با y_i نشان داده شوند، در این مدل هر برچسب y_i تنها با نشانه x_i و برچسب y_{i-1} مرتبط است.

(شکل ۵ الف) نمایی از مدل تصادفی شرطی را برای رشته ' MP unit finds rockets west of KAF ' که از داده‌های ویکی‌لیکس انتخاب‌شده نشان می‌دهد. در این شکل متغیرهای مشاهده با دایره‌های تیره نشان داده شده و متغیرهای نامشخص (بخش‌های جمله) که باید برچسب‌ها به آنها انتصاب داده شوند، به‌صورت دوایر توخالی نشان داده شده است. همچنین لبه‌ها، نشان‌دهندهٔ روابط احتمالی هستند.

در مثال بالا، فضای حالت برچسب‌ها، مشابه با مجموعه زیر است:

Y = { org.name,mil.org. name, verb, noun, state, prep, country }

دو تابع خصیصه نیز به‌صورت زیر تعریف می‌شود:

$f1(y_i, y_{i-1}, x_i)$ = [x_i appears in a country list] .[y_i = country]

$f2(y_i, y_{i-1}, x_i)$ = [x_i is an string] . [y_i = mil.org. name].[y_{i-1} = org.name]

یک بخش‌بندی، مانند $y=\{y_1,...,y_T\}$، می‌تواند یک روش ممکن برای برچسب‌زنی هر بخش جمله در x با یکی از برچسب‌های Y باشد. (شکل ۵ ب) نمونه‌ای از دو بخش بندی از x و احتمالات آنهاست.

اگر $\{f_k(y_i, y_{i-1}, x_i)\}_{k=1}^{K}$ مجموعه‌ای از توابع خصیصه مقادیر حقیقی و $\Delta = \{\lambda_k\} \in R^k$ برداری از پارامترهای با مقادیر حقیقی باشد، آنگاه یک مدل حوزهٔ تصادفی شرطی، توزیعی احتمالی از بخش‌های y با یک توالی بخش جمله خاص x تعریف می‌نماید:

$$p(y|x) = \frac{1}{Z(x)} exp\{\sum_{i=1}^{T}\sum_{k=1}^{k}\{f_k(y_i, y_{i-1}, x_i)\}\} \qquad (۱)$$

که در آن Z(x) یک تابع نرمال‌سازی استاندارد بوده که مقادیر احتمالی را بین ۰ و ۱ تضمین می‌کند.

از سامانه‌های ارایه‌شده با استفاده از این مدل می‌توان به سامانهٔ *CARE* اشاره کرد که در (*Rosenfield, B. et al;, 2009*) تشریح گردیده است همچنین در (*Ekbal, Haque, & Bandyopadhyay, 2008*) جزییات تحقیقات صورت گرفته در حوزهٔ استخراج اطّلاعات در زبان بنگالی ارایه شده است. نتایج هر یک از این سامانه‌ها نشان‌دهندهٔ بهبود کارایی سامانه‌های استخراج اطّلاعات نسبت به سامانه‌های مشابه است. همچنین به‌منظور توسعه سامانه‌های مبتنی‌بر این مدل ابزارهایی مانند *CRF++ (crf++)* و *MALLET (McCallum A.)* ارایه‌شده است.

تحقیقات انجام شده در حوزهٔ نظامی مانند (*Hecking, 2003*) و (*Hecking M., 2005*) اغلب بر روی روش‌های مبتنی‌بر قاعدهٔ متمرکز است. بررسی‌های صورت‌گرفته نشان می‌دهد که تحقیقاتی در حوزهٔ نظامی با استفاده از روش مدل حوزهٔ تصادفی شرطی تاکنون منتشر نگردیده است.

۳-۱-۳- ارزیابی تأثیر ساختار بر نتایج استخراج

به‌منظور تأثیر تفاوت‌های ذکر شده در بخش قبل آزمایش‌هایی را با استفاده از سامانه‌های استخراج اطّلاعات متداول مانند ANNIE (*Cunningham, D. et al*) که توسط دانشگاه شفیلد تهیه شده و *LingPipe* (*Alias-I corp*) که توسط آژانس پروژه‌های تحقیقاتی پیشرفته دفاعی آمریکا پشتیبانی می‌شود، بر روی متون نظامی که در بالا بررسی شد، انجام گردید.

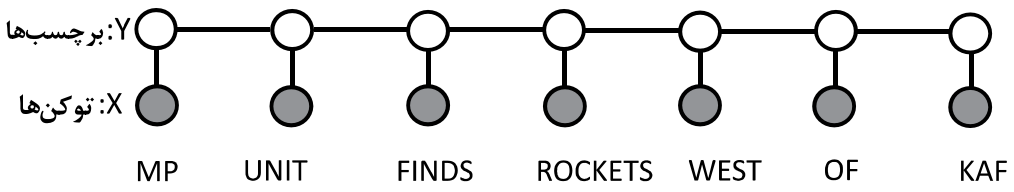
مجموعه متون مورد بررسی، شامل تعداد سی گزارش از گزارش‌های صحنهٔ نبرد ویکی‌لیکس است.

برای بررسی کارایی سامانه از معیارهای فراخوانی[3]، دقّت[4] و معیار F[5] که غالباً در استخراج اطّلاعات، استفاده می‌شوند (*Fieldman R. et al, 2007*)، (*Makhoul R. et al, 2007*) بهره می‌گیریم.

تعداد مواردی که به‌درستی استخراج‌شده به‌صورت درصدی از تعداد کل موارد استخراج‌شده، نشان‌دهندهٔ دقّت می‌باشد. معیار فراخوانی، تعداد مواردی را که به‌درستی

[1] Semantic Tags
[2] Token

[3] Recall
[4] Precision
[5] F-Measure



الف)

ب)

X:	MP	UNIT	FINDS	ROCKETS	WEST	OF	KAF
Y1:	org.name	mill.org.name	v.	noun	state	prep	country
Y2:	org.name	mill.org.name	v.	noun	state	prep	state

(شکل ۵): نمایی ازمدل برچسب‌زنی دنباله با استفاده از مدل حوزهٔ تصادفی شرطی بر روی یکی از جملات ویکی‌لیکس

استخراج شده به‌صورت درصدی از تعداد کل موارد صحیح اندازه می‌گیرد. به عبارت دیگر، تعیین می‌کند که چه تعداد از مواردی که باید به‌صورت صحیح استخراج شوند، بدون در نظر گرفتن اینکه چه تعداد مورد اضافی پیش‌بینی شده، به‌طور واقعی استخراج شده‌اند. نرخ بالاتر معیار فراخوانی، سامانهٔ بهتری از نظر آنکه موردهای صحیح کم‌تری حذف شده است.

معیار F با ترکیب وزن‌دار دو معیار دقت و فراخوانی تعریف شده است. معیار F در ارتباط با معیارهای دقت و فراخوانی به‌صوت میانگین وزن‌دار، از هر دو معیار، به‌کار می‌رود.

$$\text{F - Measure} = \frac{(\beta^2+1)P*R}{(\beta^2 R)+P} \qquad (۲)$$

مقدار β نشان‌دهندهٔ وزن دقّت (P) در مقابل فراخوانی (R) است. اگر مقدار β برابر با ۱ باشد، در این صورت هر دو دارای وزن‌های یکسانی خواهند بود.

درصورتی‌که β برابر ۱ فرض نماییم، داریم:

$$\text{F - Measure} = \frac{(\beta^2+1)P*R}{(\beta^2 R)+P} = \frac{2*P*R}{R+P} \qquad (۳)$$

نتایج حاصل از این ارزیابی در نمودار ۱، نشان‌داده‌شده است. این نمودار بیان می‌کند که نتایج حاصل از اجرای سامانهٔ ANNIE بر روی گزارش‌های نظامی پایین‌تر از نتایج حاصل از اجرای همین سامانه بر روی مجموعه متون خبری *(Chiticariu L. et al, 2010)* (بالا) و همچنین پایین‌تر از نتایج ایده‌آل پیش‌بینی‌شده توسط موسسهٔ تهیه‌کنندهٔ سامانه LingPipe (پایین) است. بنابراین می‌توان نتیجه گرفت که با استفاده از سامانه‌های متداول موجود که بر روی متون غیر نظامی به‌کار می‌رود، نمی‌توان به کارایی لازم در استخراج اطّلاعات از گزارش‌های نظامی دست یافت.

لذا می‌بایست روشی متفاوت به‌منظور انجام عمل استخراج اطّلاعات بر روی این دسته از متون به‌کار برد.

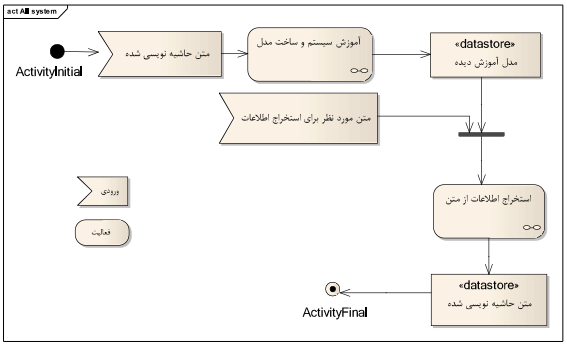
(نمودار ۱): نتایج حاصل از اجرای سامانه‌های استخراج اطّلاعات بر روی گزارش‌های نظامی (بالا: ANNIE و پایین: LingPipe)

۴- روش پیشنهادی برای استخراج اطّلاعات از متون نظامی

در بخش قبل مشاهده کردیم که کارایی سامانه‌های استخراج اطّلاعات موجود بر روی گزارش‌های نظامی محدود بوده و باید روش‌های دیگری برای این نوع از اسناد مورد استفاده قرار گیرد. لذا در اینجا روشی ترکیبی از روش‌های استنتاج پوشش و مدل حوزهٔ تصادفی شرطی، پیشنهاد و مورد ارزیابی قرار گرفته است.

۴-۱- سامانهٔ پیشنهادی برای استخراج اطّلاعات از متون نظامی

در این مقاله برای استخراج اطّلاعات، روشی ترکیبی پیشنهاد شده است که در آن از دو روش استنتاج پوشش و روش مدل حوزهٔ تصادفی شرطی استفاده شده است. علّت استفاده از این دو روش که ایدهٔ اصلی استخراج اطّلاعات از متون نظامی نیز است، اضافه کردن اطّلاعات ساختار کلان متون نظامی به ورودی‌های مدل حوزهٔ تصادفی شرطی است. سامانهٔ پیشنهادی دو بخش آموزش و استخراج دارد. در (شکل ۶)، ساختار سامانهٔ پیشنهادی ارایه شده است.



(شکل ۶): ساختار سامانهٔ پیشنهادی

۴-۱-۱- مرحلهٔ آموزش

همان‌گونه که در نمودار فعالیت[1] (شکل ۷) مشاهده می‌شود، آموزش مدل، از چهار جزء اصلی به‌شرح زیر تشکیل شده است؛ که هر یک به تفضیل بیان می‌شود:

الف- بخش اول، دریافت متن ورودی به سامانه با ساختار متنی (TXT) بوده که هر فایل با استفاده از ابزار استنتاج پوشش، به فایل‌هایی با ساختار XML که ساختار کلان هر فایل در آن مشخّص شده تبدیل می‌شود.

ب- پس از استخراج ساختار کلان و حاشیه‌نویسی در متن، یک نسخه از فایل حاشیه نویسی شده، برای حاشیه‌نویسی خرد متن و تجزیهٔ نحوی در اختیار بخش دوم قرار گرفته تا بر روی جملات هر قسمت از ساختار کلان، تجزیهٔ نحوی انجام و حاشیه‌نویسی نحوی صورت پذیرد.

ج- یک نسخهٔ دیگر از فایل XML، به‌منظور مشخّص کردن اطّلاعات مورد نیاز، در اختیار کاربر قرار گرفته تا با استفاده از ابزارهای حاشیه‌نویسی که در اختیار دارد، اطّلاعات مورد نیاز را که قصد استخراج آن را دارد در جهت آموزش سامانه، ارایه کند.

د- پس از آنکه ساختارهای کلان و خرد متن معین شد و حاشیه‌نویسی‌های لازم بر روی متون به‌منظور تعیین نیاز اطّلاعاتی کاربر، انجام شد، خروجی هر یک از بخش‌ها در اختیار تابعی قرار می‌گیرد تا ورودی آموزشی برای مدل حوزهٔ تصادفی شرطی ایجاد نماید. پس از آنکه فایلی جهت آموزش سامانه، آماده شد، فایل آماده‌شده جهت ساخت مدل، در اختیار ابزار CRF++ (crf++) قرار می‌گیرد. این ابزار به‌منظور توسعهٔ سامانه‌های مبتنی بر حوزهٔ تصادفی شرطی پیاده‌سازی شده است.

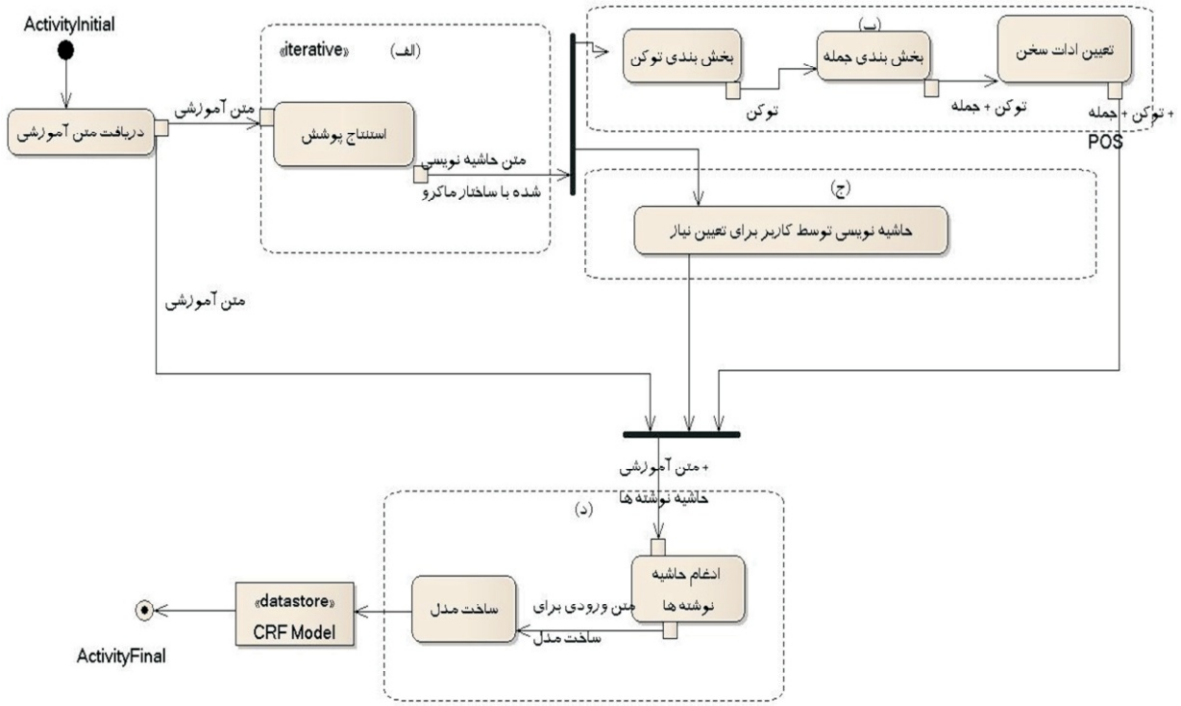
۴-۱-۲- مشخّص کردن اطّلاعات مورد نیاز به‌عنوان ورودی سامانه

دریافت نیار کاربر یکی از با اهمیت‌ترین ورودی‌های یک سامانه استخراج اطّلاعات مورد نیاز از متن است؛ زیرا نیازهای مختلف، توسط افراد متفاوت از سیستم درخواست می‌گردد. بنابراین استراتژی دریافت نیاز از کاربر در حل مسئله استخراج اطّلاعات اهمیت بالایی دارد.

استفاده از زبان طبیعی برای استخراج اطّلاعات به دلیل احتمال عدم انطباق دستور زبان استفاده شده و مجموعه متون مورد آزمایش، مناسب نیست. استفاده از زبان‌های استاندارد، مانند SQL که در پایگاه‌های دادهٔ رابطه‌ای معمول است در برخی از سامانه‌های استخراج اطّلاعات نوین مورد توجه محققان قرار گرفته است. نمونهٔ آن را می‌توان در سامانهٔ ارایه‌شده توسط محقّقان دانشگاه برکلی *(Daisy Z. et al, 2010)* می‌توان مشاهده کرد. استفاده از این روش به‌دلیل محدودیت‌هایی که در عمل‌گرهای استاندارد وجود دارد، کارایی پایینی دارد.

استفاده از متون آموزشی و با استفاده از محیط رابط کاربر برای انجام عمل استخراج اطّلاعات یکی دیگر از روش‌های مورد استفاده برای دریافت نیاز کاربر است. در این روش، به‌منظور مشخّص کردن اطّلاعات مورد نیاز، حاشیه‌نویسی متن توسط کاربر انجام گرفته و سپس متن حاشیه‌نویسی شده به‌منظور آموزش در اختیار سامانه قرار می‌گیرد. حاشیهٔ نوشته‌ها در متن می‌تواند معرف یک یا چند نیاز کاربر باشد. مزیت استفاده از این روش در مقایسه با سایر روش‌های موجود، در آن است که اوّل این‌که دستور زبان درخواست در مقایسه با روش اوّل به‌طور کامل یکنواخت با ساختارهای متن مورد آزمایش است، دوم این‌که از محدودیت‌های موجود در پرس‌وجوهای استاندارد نیز اجتناب کند.

[1] Activity Diagram



(شکل ۷): مراحل فعّالیت مربوط به مرحلهٔ آموزش و ساخت مدل

روش‌های حاشیه‌نویسی متن می‌تواند به‌صورت کلان یا خرد انجام شود. بدین‌منظور ابزار حاشیه‌نویسی متن می‌بایست در اختیار کاربر قرار گرفته تا کاربر قادر به مشخّص کردن (یا به‌عبارت دیگر حاشیه‌نویسی) نیاز یا نیازهای خود در متن باشد. از آنجا که این متون به‌عنوان ورودی سامانه بوده و به‌منظور آموزش مورد استفاده قرار می‌گیرند، پرواضح است که افزایش تعداد متون حاشیه‌نویسی‌شده نسبت مستقیم با دقّت سامانه خواهد داشت.

۴-۱-۳- مرحلهٔ استخراج

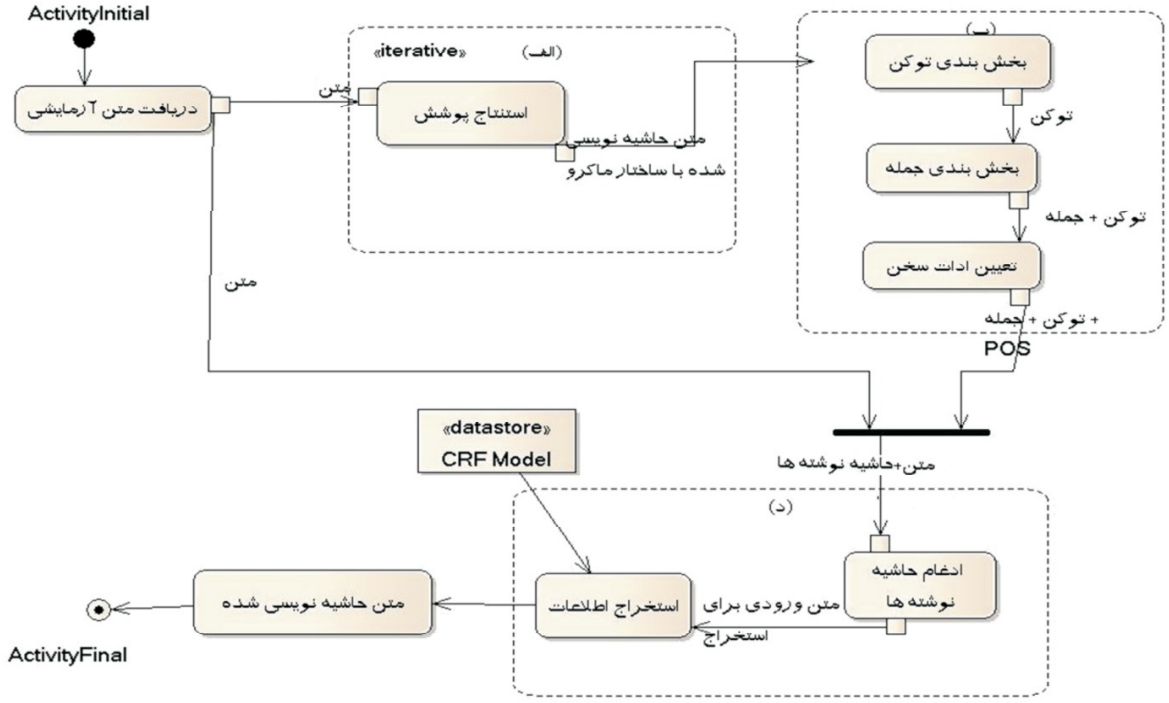
اکنون با ساخت مدل، قادر هستیم تا اطّلاعات مورد نیاز کاربر را از متون جدید استخراج کنیم. همان‌گونه که در (شکل ۸) مشاهده می‌شود، متون مورد نظر همانند آنچه که در فرایند آموزش سیستم توضیح داده شد، به سامانه اعمال و ساختارهای کلان و خرد آنها استخراج شده و برای استخراج اطّلاعات و به‌کارگیری در مدل حوزهٔ تصادفی شرطی آماده می‌شوند. سپس با کمک مدل آموزش‌دیده در مرحلهٔ قبل، فرایند استخراج اطّلاعات با استفاده از ابزار CRF++ صورت می‌پذیرد.

۵- پیاده‌سازی سامانه

همان‌گونه که در ساختار سامانهٔ پیشنهادی بیان شده است، این سامانه متشکل از دو مرحلهٔ اصلی است. یک مرحلهٔ وظیفهٔ آموزش مدل و دیگری وظیفهٔ استخراج اطّلاعات و حاشیه‌نویسی متن را به‌عهده دارد. در هر یک از زیرسلمانه‌ها، فرایندها همان‌گونه که در (شکل ۷) مشاهده می‌شود، به‌ترتیب از استخراج ساختار کلان آغاز و با حاشیه‌نویسی گزارش نظامی پایان می‌یابد. در پیوست ۱ جزییات هر یک از این زیرسامانه‌ها تشریح شده است.

۶- ارزیابی سامانهٔ پیشنهادی

قدم بعد، پس از طراحی و پیاده‌سازی روش پیشنهادی، ارزیابی آن است. به‌منظور بررسی کارایی سامانهٔ معیارهای فراخوانی، دقت و معیار F را که درقبل معرفی و مورد استفاده قرار گرفت، به‌کار می‌بریم.



(شکل ۸): مراحل فعّالیت در مرحلهٔ استخراج

۶-۱- بستر آزمایش

از آنجا که این سامانه به‌منظور استفاده در حوزهٔ نظامی طراحی و پیاده‌سازی شده است، می‌بایست حداقل در یک بستر، دادهٔ نظامی مورد ارزیابی قرار گیرد. بدین منظور منابع مرتبط ازجمله سایت آژانس تحقیقات پیشرفته نظامی ایالات متحده *(DARPA)* و سایت پیمان آتلانتیک شمالی (ناتو) *(north Atlantic Treaty Organization)* مورد بررسی قرار گرفت و درنهایت تنها یک مجموعه داده نظامی از کشور آلمان یافت شد. این مجموعه داده با نام *KFOR*[1] (*Hecking M., 2005*) مربوط به جنگ کوزوو می‌باشد. در پروژهٔ[2] زنون[3] (*ZENON project*) و در مقاله آقای دکتر ماتیاس هکینگ[4] (*Hecking M., 2005*) و (*Hecking, 2003*) بدان اشاره شده است. این مجموعه‌داده شامل اطّلاعات اخبار صحنهٔ نبرد به دو زبان آلمانی و انگلیسی بوده که آقای دکتر هکینگ در گزارش [۳۱] اعلان کرد که این مجموعه‌داده طبقه‌بندی داشته و امکان انتشار آزاد آن وجود ندارد. بنابراین اقدام به تهیه بستر آزمایش از داده‌های صحنه نبرد موجود در ویکی‌لیکس کردیم. این مجموعه داده مربوط به اطّلاعات نبرد نیروهای ناتو در افغانستان از سال ۲۰۰۴ و حاوی بیش از هفتصد هزار گزارش است. اطّلاعات مربوط به این مجموعه‌داده در سایت آزمایشگاه سامانه‌های هوشمند دانشگاه صنعتی امیر کبیر *(mollaei, 2010)* موجود است.

همچنین این سامانه در یک بستر داده دیگر به‌منظور بررسی کارایی سامانه با متون خبری، مورد آزمایش قرار می‌گیرد. این بستر مربوط به سمینارهای *CMU* (*Data set for CMU seminars*) بوده و این مجموعه‌داده، شامل اخبار سمینار از دانشگاه کارنیجه ملون[5] واقع در ایالت پنسیلوانیاست. موجودیت‌های مشخّص‌شده شامل شروع و پایان سمینار، محل برگزاری و سخنران آن است. در این مقاله، نسخه‌ای از مجموعهٔ داده را که برچسب‌های ادات سخن آن مشخّص‌شده است، استفاده کرده‌ایم (*Brill E., 1994*).

۶-۲- انجام آزمایش

به‌منظور انجام آزمایش و محاسبهٔ هر یک از معیارهای ارزیابی، نیاز اطّلاعاتی برای استخراج در هر دو مجموعه داده تعریف می‌کنیم. بنابراین از مجموعه متون نظامی، نام

[1] Kosovo Force

[2] این پروژه در کشور آلمان با موضوع استخراج اطلاعات از گزارشات زبان طبیعی به زبان انگلیسی فعال می‌باشد.

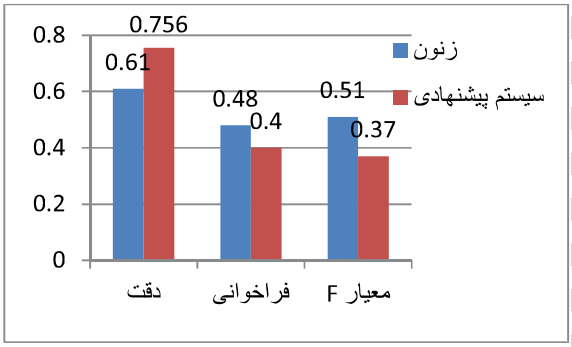
[3] ZENON Project

[4] Dr. Matthias Hecking

5 Carnegie Mellon

سازمان‌ها، تاریخ، رخداد، تجهیزات نظامی را به‌عنوان نیاز در نظر گرفته و با نتایج چند سامانهٔ مشابه و همچنین نتایج آقای دکتر هکینگ در (*Hecking, 2003*) که بر روی داده‌های KFOR که در دوازدهمین کنفرانس بین‌المللی تحقیقات و فناوری فرماندهی و کنترل[1]، ارایه شد، مقایسه می‌کنیم. لذا تعداد سی گزارش از مجموعه گزارش‌های نظامی مربوط به جنگ عراق و نیز تعداد سی خبر غیر نظامی برای داده‌های آموزشی و آزمایش انتخاب کردیم. از مجموعه دادهٔ غیر نظامی، اطّلاعات مربوط به شروع و پایان سمینار و محل برگزاری، به‌عنوان نیاز درنظر می‌گیریم. در هر دو مجموعه ۸۰٪ از داده‌ها را به‌عنوان دادهٔ آزمایشی استفاده می‌کنیم.

با استفاده از هر یک از مجموعه‌های داده، سامانه آموزش دیده و سپس عملیات استخراج اطّلاعات بر روی آنها صورت گرفته و معیارهای دقّت و فراخوانی و F محاسبه شدند.



(نمودار ۲): مقایسهٔ نتایج آزمایش‌های ارزیابی استخراج اطّلاعات سامانهٔ پیشنهادی و سامانهٔ پروژه زنون

مقایسهٔ نتایج حاصل از اجرای سامانهٔ پیشنهادی بر روی مجموعه داده نظامی ویکی‌لیکس با نتایج پروژه زنون در (جدول ۱) و (نمودار ۲) نشان داده شده است.

(جدول۱): مقایسهٔ نتایج سامانهٔ پیشنهادی با نتایج سامانهٔ زنون

مورد استخراج		سازمان	تاریخ	رخداد	وسایل	میانگین
دقت	WI&CRF	۰/۸۳	۱	۰/۹۲	۰/۸	۰/۶۱
	Zenon	۰/۴۳	۰/۹۲	–	۰/۹۲	۰/۷۵۶
فراخوانی	WI&CRF	۰/۵۵	۰/۸	۰/۴۳	۰/۱	۰/۴۸
	Zenon	۰/۴۳	۰/۳۷	–	۰/۹۲	۰/۴
معیار F	WI&CRF	۰/۶۶	۰/۸۸	۰/۵۸	۰/۱۷	۰/۵۱۳
	Zenon	۰/۵۶	۰/۵۳	–	۰/۹۲	۰/۳۷

[1] International Command and Control Research and Technology Symposium (ICCRTS)

(جدول ۲): مقایسهٔ نتایج حاصل از اجرای سامانهٔ پیشنهادی بر روی متون خبری در مقایسه با دو سامانهٔ استخراج اطّلاعات

معیار ارزیابی	مورد استخراج / نام سیستم	نام سازمان	تاریخ	رخداد	وسایل نظامی	میانگین
دقت	ANNIE	۰/۸۳	۱	۰	۰/۶۲	۰/۶۱
	LingPipe	۰/۲	۱	۰	۰/۴	۰/۴
	WI&CRF	۰/۸۳	۱	۰/۹۲	۰/۸	۰/۸۸
فراخوانی	ANNIE	۰/۵۵	۱	۰	۰/۴۸	۰/۵
	LingPipe	۰/۲۵	۰.۸	۰	۰/۰۹	۰/۲۸۵
	WI&CRF	۰/۵۵	۰/۸	۰/۴۳	۰/۱	۰/۴۷
معیار F	ANNIE	۰/۶۶	۱	۰	۰/۵۴	۰/۵۵
	LingPipe	۰/۲۲	۰/۸	۰	۰/۱۴	۰/۲۹
	WI&CRF	۰/۶۶	۰/۸۸	۰/۵۸	۰/۱۷	۰/۵۷

با مقایسهٔ نتایج سامانهٔ پیشنهادی با نتایج پروژه زنون (جدول ۱) می‌توان به این نتیجه رسید که سامانهٔ پیشنهادی در مقایسه با سامانهٔ مشابه، دقّت پایین‌تر، امّا فراخوانی بالاتر دارد. بهبود معیار *F* بیان‌کنندهٔ بهبود کارایی کلی استخراج توسط سامانهٔ پیشنهادی است. همچنین هر یک از متون انتخاب‌شده با استفاده از دو روش مبتنی بر قاعده، مدل مخفی مارکوف مورد ارزیابی قرار گرفته و ابزار *ANNIE* برای روش مبتنی بر قاعده و ابزار *LingPipe* برای روش مدل مخفی مارکوف استفاده شده است.

مقایسهٔ نتایج حاصل از انجام آزمایش توسط سامانهٔ پیشنهادی بر روی داده‌های غیر نظامی با نتایج حاصل از اجرای سامانه‌های متداول نشان می‌دهد که این سامانه قابلیت اجرا بر روی مجموعه‌های غیر نظامی را نیز دارد. (جدول ۲)

۷- نتیجه‌گیری

در این مقاله روش‌های استخراج اطّلاعات از متون نظامی مورد بررسی قرار گرفت. در این راستا متون نظامی به دو دسته متون خبری و گزارش‌های صحنهٔ نبرد تقسیم شده و با بررسی بر روی ساختارهای متون نظامی و مقایسهٔ آنها با متون خبری غیر نظامی مشخّص شد که ساختار متون خبری نظامی با ساختار متون غیر نظامی مشابه و با ساختار گزارش‌های صحنهٔ نبرد تفاوت دارد. با ارزیابی صورت‌گرفته با سامانه‌های متداول استخراج اطّلاعات، تأثیر این تفاوت بر نتایج استخراج بر روی گزارش‌های نظامی، نشان داده شد.

در ادامه سامانه‌ای برای استخراج اطّلاعات از گزارش‌های نظامی مبتنی بر دو روش استنتاج پوشش و مدل حوزهٔ تصادفی شرطی پیشنهاد شد؛ سپس با معرّفی معیارهای ارزیابی دقّت، فراخوانی و معیار F، آزمایش‌هایی برای ارزیابی سامانهٔ پیشنهادی انجام دادیم. با استفاده از

بستر داده‌ای که به‌منظور ارزیابی این سامانه از سایت ویکی‌لیکس تهیه شد، سامانهٔ پیشنهادی، مورد آزمایش قرار گرفت. نتایج این آزمایش با نتایج سامانهٔ زنون مقایسه شد. ارزیابی حاصله بیان‌کنندهٔ بهبود کارایی در معیارهای فراخوانی و معیار F بوده و تنها در معیار دقّت از نتایج سامانه‌های مشابه کم‌تر است.

۸- تقدیر و تشکر

از آقای مهندس رسول‌زادگان که در تهیه و تدوین این مقاله همکاری کرده و از صرف وقّت و ارایهٔ راهنمایی‌های لازم دریغ نورزیدند، کمال تشکّر و قدردانی را دارم.

۹- منابع

Abolhassani M. 2003. Information Extraction and Automatic Markup for XML documents.

Afghanestan War Reports. Retrieved 11 20, 2010, from WikiLeaks: http://www.wikileaks.com

Alias-I corp.. LingPipe. Retrieved 10 15, 2010, from Natural language processing software for text analytics, text data mining and search: LingPipe: http://Alias-I.com/lingpipe

Brill E. 1994. Some advances in rule-based part of speech tagging. AAAI.

Chiticariu L. et al. 2010. SystemT: An Algebraic Approach to Declarative Information Extraction. (pp. 128-137). ACL 2010.

Ciravegna, F. 2001.(LP)2 An adaptive algorithm for information extraction from Web-related texts. IJCAI-2001 Workshop on Adaptive Text Extraction and Mining held in conjunction with 17th International Joint Conference on Artificial Intelligence (IJCAI). Seattle, Washington.

Collins M. 1999. Head-driven Statistical Models for Natural Language Parsing. University of Pennsylvania, Ph.D. thesis. Philadelphia: University of Pennsylvania.

Collins, M. 1999. Head Driven Statistical Model for Natural Language Parsing. Philadelphia: Phd Thesis, University of Pennsylvania.

crf++. Retrieved from Source forge: http://crfpp.SourceForge.org

Cunningham, D. et al., The GATE User Guide. Retrieved 10 20, 2010, from GATE: http://gate.ac.uk

Daisy Z. et al. 2010. Querying Probablistic Information Extraction. VLDB Endowment, 3, pp. 1057-1068.

Dalvi N., et al. 2009. Robust Web Extraction: An Approach based on Probulistic tree-edit model. SIGMOD, (pp. 335-348).

DARPA. USA Defense advanced research Project Agancy. Retrieved 09 13, 2010, from http://www.darpa.mil

Data set for CMU seminars. Retrieved 11 22, 2010, from http://www.cs.umass.edu/_mccallum/data/sa-tagged.tar.gz

Ekbal, A., Haque, R., & Bandyopadhyay, S. (2008). Named Entity Recognition in bengali: A Conditional Random Field Approach. Proceedings of the 3rd International Joint Conference on Natural Language Processing, (pp. 589-594).

Fieldman R. et al., 2007. Text Mining Hand Book Advanced Approaches in Analizing Unstructured Data. Cambridge university press.

Freitag, D. e., 2000. Boosted wrapper induction. 17th National Conference on Artificial Intelligence, (pp. 577-583).

Grishman, R. 1997. Information Extraction:Techniques. Proceeding of the Information Extraction International Summer School , (pp. 10-27).

Hecking M. 2005. KFOR-Korpus – Annotierungsvorschrift. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN).

Hecking M. 2007. KFOR-Korpus. 12th ICCRTS “Adapting C2 to the 21st Century”.

Hecking, M. 2003. Information Extraction from Battlefield Reports. 8th International Command and Control Research and Technology Symposium (pp. 17-19). Washington, DC: National Defense University .

Jane's Military Journal. Retrieved 11 14, 2010, from Jane's Military Journal: http://www.JeansJournal.com

Kushmerick, N. e., 1997. Wrapper induction for information extraction. International Joint Conference on Artificial Intelligence (IJCAI’97), (pp. 729-737).

Lafferty, J. e. 2001. onditional random fields: Probabilistic models for segmenting and labeling sequence data. (pp. 282-289). 8th International Conference on Machine Learning (ICML’01).

Makhoul R. et al., 2007. Performance measures for information extraction . Proceedings of the DARPA Broadcast News Workshop.

McCallum, A. e. 2000. Maximum entropy Markov models for information extraction and segmentation. 17th International Conference on Machine Learning (ICML'00), (pp. 591-598).

McCallum, A. Mallet:A machine Learning for language toolkit. Retrieved 06 2011, from Mallet: http://mallet.cs.umass.edu

mollaei, N. 2010. Amir Kabir University Intelligent System Lab. Retrieved from http://Ciet.aut.ac.ir/islab/

Muslea, S. I. 2001. Hierarchical wrapper induction for semistructured information sources. Autonomous Agents and Multi-Agent Systems , 4, pp. 93–114.

Nivre, J. e., 2006. Maltparser:AData-Driven Parser-Generator for Dependency Parsing. Language Resources and Evaluation Conference (LREC-06), (pp. 2216–2219). Genoa, Italy.

north Atlantic Treaty Organization. Science for peace and security. Retrieved from nato: http://www.nato.int/science

Riloff, E. 1993. Automatically constructing a dictionary for information extraction tasks. Eleventh National Conference on Artificial Intelligence, (pp. 811-816).

Rosenfield, B. et al; 2009. Conditional Random Field (CRF) Based Relation Extraction System,. United State Patent Application Publication .

routers. routers Daily Newspaper. (routers) Retrieved 11 22, 2010, from routers Daily Newspaper: http://www.Reuters.com

Soderland, S. e. 1995. CRYSTAL: Inducing a conceptual dictionary. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95), (pp. 1314-1319).

Tang, J. e. 2005. iASA: Learning to annotate the semantic Web. Journal on Data Semantic , 4, 110-145.

ZENON project. Retrieved 2010, from http://www.fkie.fraunhofer.de/

احمد عبداله‌زاده بارفروش همکاری خود را با دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر در سال ۱۳۷۰ و پس از اخذ مدرک دکتری مهندسی کامپیوتر از دانشگاه بریستول انگلستان، آغاز کرده است و هم‌اکنون استاد دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر است. ایشان از سال ۱۳۷۹ الی ۱۳۸۱ به‌عنوان استاد مدعو در دانشگاه‌های Maryland College Park آمریکا و ORASY (Paris 11) فرانسه و در سال ۱۳۸۸ تا ۱۳۹۰ به عنوان استاد مهمان در دانشگاه‌های Trento ایتالیا و Loughborough انگلستان مشغول به کار بوده است. دکتر عبداله‌زاده کتاب‌های "مقدمه‌ای بر هوش مصنوعی توزیع شده" و "کلیات متدولوژی تأمین کیفیت" را نیز تألیف کرده‌اند. زمینه‌های تحقیقاتی مورد علاقهٔ ایشان عبارتند از: تکنیک‌های هوش مصنوعی، هوش مصنوعی توزیع‌شده، مذاکره خودکار، سامانه‌های خبره، پردازش زبان طبیعی، سامانه‌های تصمیم‌یار، هوش تجاری، پایگاه داده تحلیلی، داده‌کاوی، و مهندسی نرم افزار.

نشانی رایانامهٔ ایشان عبارت است از:

ahmad@aut.ac.ir

نیما ملایی تحصیلات خود را در مقطع کارشناسی مهندسی نرم‌افزار رشتهٔ مهندسی نرم‌افزار در سال ۱۳۷۹ و کارشناسی ارشد فناوری اطّلاعات در گرایش مهندسی سامانه‌های اطّلاعاتی را در سال ۱۳۸۹ به پایان رساند. زمینه‌های تحقیقاتی مورد علاقهٔ ایشان عبارت است از: داده‌کاوی، استخراج اطّلاعات، مهندسی نرم افزار.

نشانی رایانامهٔ ایشان عبارت است از:

nima.mollaei@yahoo.com