

WI&CRF: روش پیشنهادی برای استخراج

اطلاعات مورد نیاز از متون نظامی

نیما ملایی^۱، احمد عبدالله‌زاده^۲ و حسین شیرازی^۳

^{۱،۲} دانشگاه صنعتی مالک اشتر

^۲ دانشگاه صنعتی امیر کبیر، دانشکده مهندسی فناوری اطلاعات و کامپیوتر

چکیده

امروزه تکنیک‌های استخراج اطلاعات مورد نیاز، از متون نظامی، مورد توجه فرماندهان و مدیران نظامی قرار گرفته که به دلیل 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 مقایسه ساختار متون نظامی با متون غیرنظامی، دسته بندی جدیدی از متون نظامی ارائه شده و نتایج استخراج اطلاعات 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 حوزه تصادفی شرطی ارائه شده است. ارزیابی کارایی بر روی بستر تهیه شده از گزارش‌های صحنه نبرد، انجام شده است. با به کارگیری روش پیشنهادی بر روی بستر ارائه شده، بهبود کارایی در معیارهای فراخوانی و معیار F نمایش داده شده است. همچنین با مقایسه نتایج حاصل از اجرای این سامانه بر روی متون غیر نظامی، امکان به کارگیری آن برای استخراج اطلاعات از متون غیر نظامی نیز بررسی شده است.

واژگان کلیدی: متن نظامی، استخراج اطلاعات، استنتاج پوشش، مدل حوزه تصادفی شرطی

۱- مقدمه

افزایش چشم گیر استفاده از رایانه‌های شخصی و سازمانی باعث انباشت حجم زیادی از داده و اطلاعات در رایانه‌ها شده و این موضوع سازمان‌ها را با مشکل جدیدی با عنوان استخراج اطلاعات و کسب دانش، مواجه ساخته است. تاکنون از روش‌های بازیابی اطلاعات^۱ برای حل این مسئله، استفاده شده است. اما مشکلات موجود در بازیابی اطلاعات از جمله عدم توجه به معنا در متن، باعث شده تا شاخه جدیدی از علم رایانه با عنوان استخراج اطلاعات^۲ ایجاد و به شدت مورد توجه قرار گیرد.

از آنجا که به طور معمول متون نظامی ساختار خاص دارند و جملات به صورت کامل بیان نمی شود، استفاده از برچسب‌های^۳ نحوی به تنهایی کارایی مناسب برای استخراج اطلاعات از این گونه متون را ندارد. پژوهش‌های انجام گرفته در زمینه استخراج اطلاعات مانند (LP)^۲ (Ciravegna, 2001)، Crystal

(Soderland, 1995)، (Tang, 2005) iASA،

(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 اطلاعات می پردازند، قادر به استخراج اطلاعات از این گونه متون نیستند. لذا روش پیشنهادی با این فرض که استفاده از حاشیه نوشته‌های دیگری مانند اطلاعات ساختاری و یا نحوی می تواند به استخراج بهتر اطلاعات از این نوع متون کمک کند، پایه گذاری شده است.

در این مقاله روشی بر اساس ترکیبی از استنتاج پوشش^۴ (Kushmerick, 1997) و مدل حوزه تصادفی شرطی^۵ (Lafferty, 2001) پیشنهاد شده که در آن، حوزه تصادفی شرطی با استفاده از ساختار متن نظامی آموزش دیده و از این طریق کارایی سامانه استخراج اطلاعات از متن نظامی را بهبود بخشیده است.

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

^۴Wrapper Induction

^۵Conditional Random Field (CRF)

^۱Information Retrieval (IR)

^۲Information Extraction (IE)

^۳Tag

الگوریتم‌های استخراج اطلاعات و همچنین معرفی تعدادی از سیستم‌های رایانه شده مبتنی بر این روش‌ها می‌باشد. در ادامه تأثیر ویژگی‌های ساختاری متون نظامی بر نتایج استخراج اطلاعات بررسی و روش پیشنهادی برای بهبود کارایی استخراج اطلاعات ارائه شده است. در پایان روش پیشنهادی با استفاده از داده‌های موجود در پایگاه داده ویکی‌لیکس^۱ 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 اطلاعاتی متن را ارتقا می‌بخشد (Abolhassani M., 2003) و بدون نشانه‌گذاری، از نقطه‌نظر سامانه‌ای، یک سند فقط یک سلسله طولانی از کلمات است. بنابراین مجموعه عمل‌گرهایی که می‌تواند بر روی این نوع متن کار کند محدود است. در صورت نشانه‌گذاری متن، سامانه قادر است تا معانی پنهان در متن را به کمک حاشیه 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: 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 ارائه شده‌اند. مانند استنتاج پوشش^۲ (Kushmerick, 1997)، استنتاج پوشش افزایشی^۳ (Freitag, 2000) و استنتاج پوشش سلسله‌مراتبی^۴ (Muslea, 2001)

ب- بررسی ساختاری سطح خرد: در این حالت، سند از نظر محدوده 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 ارائه شده‌اند که می‌توان به (Collins, 1999) و (Nivre, 2006) اشاره کرد.

ج- در سطح نشانه: ساختار متن از نظر وجود نشانه‌ها، به‌منظور تشریح اطلاعاتی 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 بررسی 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 صحنه نبرد طبقه‌بندی می‌شوند. اخبار نظامی را در مجلاتی مانند *Jane's Military Journal* می‌توان مشاهده نمود و گزارش‌های صحنه نبرد در منابعی مانند KFOR (Hecking M., 2005) و ویکی‌لیکس (*Afghanistan War Reports*) در دسترس هستند.

به‌منظور آشکارسازی ارزش اطلاعاتی 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 حوزه اخبار غیرنظامی مقایسه و تفاوت‌های ساختاری آنها مقایسه شده است.

۲-۱-۱- بررسی ساختار کلان متون نظامی

به‌منظور بررسی ساختار کلان گزارش‌های صحنه 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 مجله هفتگی نظامی جینز، بررسی ساختاری صورت‌گرفته و ساختار نمایشی و منطقی مندرج در (شکل ۲) برای متون موجود در این منبع اطلاعاتی استخراج شد.

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

¹Wikileaks

²Macro

³Wrapper Induction

⁴Boosted wrapper induction

⁵Hierarchical wrapper induction

⁶Micro

(R. et al, 2007) لذا فرایند معنا دار کردن می تواند به صورت تبدیل متن به یک فایل جدول یا به صورت یک متن حاشیه نویسی شده باشد.

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

- **بخش بندی**^۱: با بخش بندی، متن به ساختار تشکیل دهنده خود (کلمه، جمله و پاراگراف) تفکیک می شود.

- پردازش واژگانی و ریخت شناسی^۲: به تشخیص اشکال کلمه صرف شده می پردازد.

- **تحلیل نحوی**^۳: به خروجی مرحله قبل اعمال شده و می تواند به صورت تجزیه کم عمق یا عمیق باشد.

- **تحلیل دامنه**^۴: در نهایت قالب تعیین شده بر اساس خروجی مرحله قبل تکمیل می شود.

همان گونه که در (شکل ۴) نشان داده شده است، برای استخراج اطلاعات از متن، سه رویکرد اصلی وجود دارد. این رویکردها بر یادگیری قاعده^۵، مدل طبقه بندی^۶ و مدل مبتنی بر برچسب زنی توالی^۷ متمرکز هستند. هر یک از این روش ها از نوع یاد گیری ماشین با نظارت و دارای دو مرحله یادگیری و استخراج می باشند.

سامانه های WIEN (Kushmerick, 1997) و BWI (Freitag, 2000) نمونه هایی از سامانه های ایجاد شده به منظور استنتاج پوشش می باشند. همچنین (Dalvi N., et al, 2009) نیز نمونه ای از آخرین تحقیقات صورت گرفته در همین زمینه است.

در برچسب زنی توالی، یک سند به صورت کلمات متوالی دیده شده و برچسب های متوالی برای نشان دادن ویژگی هر کلمه به آن منتسب می شود.

نمونه ای از این نوع برچسب زنی را می توان در پردازش زبان طبیعی مشاهده کرد. در این عمل، هر کلمه با یک حاشیه نوشته که نشان دهنده ادات سخن^۸ متناظرش است برچسب زده می شود.

یکی از روش های متداول برای این نوع برچسب زنی، استفاده از مدل مولد^۹ است. مدل مخفی مارکوف^{۱۰} نمونه ای از یک مدل مولد است. در این مدل یک توزیع احتمالی اتصال $P(x,y)$ که در آن x و y متغیرهای تصادفی که

نشان دهنده "توالی مشاهده"^{۱۱} و "توالی برچسب"^{۱۲} می باشند، تعیین می شود. به منظور تعیین مقدار احتمال اتصال، مدل های مولد می بایست همه توالی های مشاهده ممکن را محاسبه کنند. به عبارت ساده تر، اجزای مشاهده در هر نمونه داده شده در هر زمان می بایست به طور مستقیم وابسته به حالت یا برچسب در همان زمان باشد. این فرضیه برای تعداد کمی از مجموعه های داده ای ساده مناسب است؛ اما در دنیای واقعی، بسیاری از توالی های مشاهده، به صورت عباراتی دارای خصایصی با اثر متقابل و اجزای مشاهده با وابستگی خیلی زیاد نمایش داده می شوند.

مدل مطلوب، مدلی است که استنتاج های ساده ای را پشتیبانی کرده، و داده را بدون آنکه نیاز به ایجاد فرضیات مستقل غیر قابل توجیه باشد، نمایش دهد. یکی از روش های مرتفع کردن این مسئله، استفاده از مدل تصادفی^{۱۳} است. در این مدل احتمال شرطی $p(Y|x)$ بر روی دنباله های برچسب که به توالی مشاهده معین x متعلق است، به جای توزیع اتصال بر روی هر دو دنباله مشاهده و برچسب تعریف می کند. مدل های تصادفی، برای برچسب زنی یک توالی مشاهده جدید x^* ، از طریق انتخاب توالی برچسب y^* که احتمال شرطی $p(y^*|x^*)$ را بیش تر می کند، به کار می رود. در این مدل، خصایص دل خواه از داده مشاهده، بدون نگرانی از ارتباط این خصایص با یکدیگر، می تواند همراه مدل نگهداری شود.

مدل های تصادفی شرطی، چارچوبی احتمالی برای برچسب زنی بر اساس روش های ارایه شده در بند بالا هستند. این مدل، شکلی از مدل گرافیکی بدون جهت است که یک توصیف ساده از لگاریتم خطی بر روی توالی های برچسب متعلق به یک توالی مشاهده تعریف می کند. مزیت اصلی مدل های تصادفی شرطی نسبت به مدل مخفی مارکوف، ماهیت شرطی آنهاست که باعث می شود 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. به علاوه مدل های تصادفی شرطی از مسئله بایاس برچسب^{۱۴} نیز دوری می کنند (Lafferty, 2001). این ضعف در مدل های مارکوف بیش ترین بی نظمی^{۱۵} (McCallum, 2000) و سایر روش های مدل مارکوف که بر روی گراف جهت دار تعریف می گردند، مشاهده می شود.

اشکال موجود در این نوع از سامانه ها، استفاده از برچسب هایی از نوع ادات سخن بوده که دارای ماهیت نحوی

¹¹ Observed Sequence

¹² Label Sequence

¹³ Random Model

¹⁴ Label Bias Problem

¹⁵ Maximum Entropy Markov Model

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

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

که در آن $Z(x)$ یک تابع نرمال سازی استاندارد بوده که مقادیر احتمالی را بین ۰ و ۱ تضمین می کند.

از سامانه های ارایه شده با استفاده از این مدل می توان به سامانه CARE اشاره کرد که در (Rosenfield, B. et al.; 2009) تشریح گردیده است همچنین در (Ekbali, Haque, 2008) & Bandyopadhyay, جزئیات تحقیقات صورت گرفته در حوزه استخراج اطلاعات در زبان بنگالی ارایه شده است. نتایج هر یک از این سامانه ها نشان دهنده بهبود کارایی سامانه های استخراج اطلاعات نسبت به سامانه های مشابه است. همچنین به منظور توسعه سامانه های مبتنی بر این مدل ابزارهایی مانند (crf++) (CRF++ و McCallum A. MALLET) ارایه شده است.

تحقیقات انجام شده در حوزه نظامی مانند (Hecking, 2003) و (Hecking M., 2005) اغلب بر روی روش های مبتنی بر قاعده متمرکز است. بررسی های صورت گرفته نشان می دهد که تحقیقاتی در حوزه نظامی با استفاده از روش مدل حوزه تصادفی شرطی تاکنون منتشر نگردیده است.

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

به منظور تأثیر تفاوت های ذکر شده در بخش قبل آزمایش هایی را با استفاده از سامانه های استخراج اطلاعات متداول مانند ANNIE (Cunningham, D. et al) که توسط دانشگاه شفیلد تهیه شده و LingPipe (Alias-I corp) که توسط آژانس پروژه های تحقیقاتی پیشرفته دفاعی آمریکا پشتیبانی می شود، بر روی متون نظامی که در بالا بررسی شد، انجام گردید.

مجموعه متون مورد بررسی، شامل تعداد سی گزارش از گزارش های صحنه نبرد ویکی لیکس است.

برای بررسی کارایی سامانه از معیارهای فراخوانی^۳، دقت^۴ و معیار F^۵ که غالباً در استخراج اطلاعات، استفاده می شوند (Fieldman R. et al, 2007)، (Makhoul R. et al, 2007) بهره می گیریم.

تعداد مواردی که به درستی استخراج شده به صورت درصدی از تعداد کل موارد استخراج شده، نشان دهنده دقت می باشد. معیار فراخوانی، تعداد مواردی را که به درستی

هستند. این برچسب ها نسبت به برچسب های معنایی^۱، در استخراج اطلاعات ارزش کمتری دارند. بنابراین استفاده از برچسب هایی که دارای ارزش معنایی برای متن مورد بررسی باشد، می تواند در بهینه کردن نتایج استخراج اطلاعات، مفید باشد.

در حوزه تصادفی شرطی (Lafferty, 2001)، استخراج اطلاعات به صورت عمل برچسب زنی دنباله فرض می شود. در برچسب زنی دنباله، یک سند به صورت بخش هایی از جمله به صورت متوالی دیده شده و برچسب ها برای نشان دادن ویژگی هر بخش^۲ به آن منتسب می شود.

در روش حوزه تصادفی شرطی، یک توزیع احتمال از مجموعه ای از متغیرهای تصادفی برچسب (Y)، بر روی بخشی از مجموعه نشانه ها (X)، تعریف می شود. در صورتی که 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 داده های ویکی لیکس انتخاب شده نشان می دهد. در این شکل متغیرهای مشاهده با دایره های تیره نشان داده شده و متغیرهای نامشخص (بخش های جمله) که باید برچسب ها به آنها انتساب داده شوند، به صورت دایره توخالی نشان داده شده است. همچنین لبه ها، نشان دهنده روابط احتمالی هستند.

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

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

دو تابع خصیصه نیز به صورت زیر تعریف می شود:

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

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

یک بخش بندی، مانند $y = \{y_1, \dots, y_T\}$ ، می تواند یک روش ممکن برای برچسب زنی هر بخش جمله در X با یکی از برچسب های Y باشد. (شکل ۵ ب) نمونه ای از دو بخش بندی از X و احتمالات آنهاست.

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

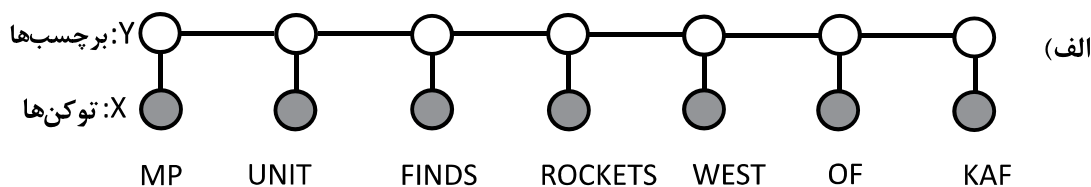
^۱ Semantic Tags

^۲ Token

^۳ Recall

^۴ Precision

^۵ F-Measure



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 از مدل برچسب‌زنی دنباله با استفاده از مدل حوزه تصادفی شرطی

بر روی یکی از جملات ویکی‌لیکس

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



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

۴- روش پیشنهادی برای استخراج اطلاعات از متون نظامی

در بخش قبل مشاهده کردیم که کارایی سامانه‌های استخراج اطلاعات 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 حوزه تصادفی شرطی، پیشنهاد و مورد ارزیابی قرار گرفته است.

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، سامانه 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.

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

مقدار β نشان‌دهنده وزن دقت (P) در مقابل فراخوانی (R) است. اگر مقدار β برابر با ۱ باشد، در این صورت هر دو دارای وزن‌های یکسانی خواهند بود. در صورتی که β برابر ۱ فرض نماییم، داریم:

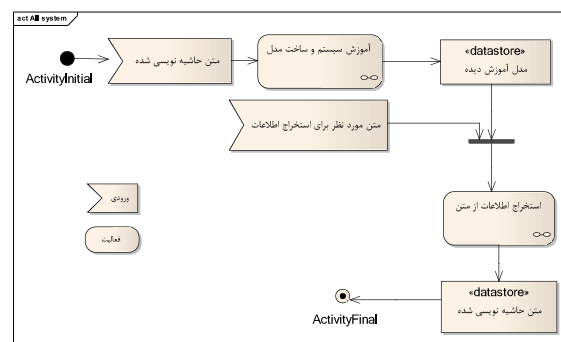
$$F - Measure = \frac{(\beta^2 + 1)P * R}{(\beta^2 R) + P} = \frac{2 * P * R}{R + P} \quad (3)$$

نتایج حاصل از این ارزیابی در نمودار ۱، نشان‌داده شده است. این نمودار بیان می‌کند که نتایج حاصل از اجرای سامانه 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 موسسه تهیه‌کننده 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 اطلاعات از گزارش‌های نظامی دست یافت.

۴-۱- سامانه پیشنهادی برای استخراج

اطلاعات از متون نظامی

در این مقاله برای استخراج اطلاعات، روشی ترکیبی پیشنهاد شده است که در آن از دو روش استنتاج پوشش و روش مدل حوزه تصادفی شرطی استفاده شده است. علت استفاده از این دو روش که ایده اصلی استخراج اطلاعات از متون نظامی نیز است، اضافه کردن اطلاعات ساختار کلان متون نظامی به ورودی‌های مدل حوزه تصادفی شرطی است. سامانه پیشنهادی دو بخش آموزش و استخراج دارد. در (شکل ۶)، ساختار سامانه پیشنهادی ارائه شده است.



(شکل ۶): ساختار سامانه پیشنهادی

۴-۱-۱- مرحله آموزش

همان گونه که در نمودار فعالیت^۱ (شکل ۷) مشاهده می شود، آموزش مدل، از چهار جزء اصلی به شرح زیر تشکیل شده است؛ که هر یک به تفصیل بیان می شود:

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

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

ج- یک نسخه دیگر از فایل XML، به منظور مشخص کردن اطلاعات مورد نیاز، در اختیار کاربر قرار گرفته تا با استفاده از ابزارهای حاشیه نویسی که در اختیار دارد، اطلاعات مورد نیاز را که قصد استخراج آن را دارد در جهت آموزش سامانه، ارائه کند.

^۱Activity Diagram

د- پس از آنکه ساختارهای کلان و خرد متن معین شد و حاشیه نویسی های لازم بر روی متون به منظور تعیین نیاز اطلاعاتی کاربر، انجام شد، خروجی هر یک از بخش ها در اختیار تابعی قرار می گیرد تا ورودی آموزشی برای مدل حوزه تصادفی شرطی ایجاد نماید. پس از آنکه فایلی جهت آموزش سامانه، آماده شد، فایل آماده شده جهت ساخت مدل، در اختیار ابزار CRF++ (crf++) قرار می گیرد. این ابزار به منظور توسعه سامانه های مبتنی بر حوزه تصادفی شرطی پیاده سازی شده است.

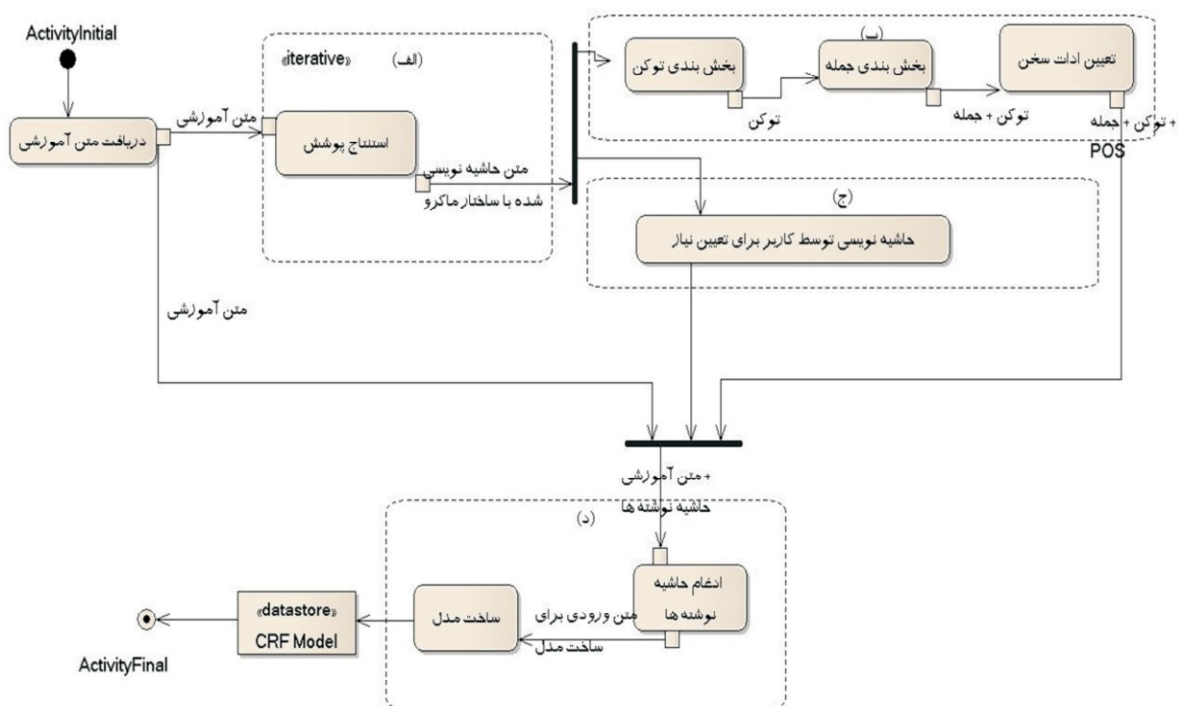
۴-۱-۲- مشخص کردن اطلاعات مورد نیاز به عنوان

ورودی سامانه

دریافت نیاز کاربر یکی از با اهمیت ترین ورودی های یک سامانه استخراج اطلاعات مورد نیاز از متن است؛ زیرا نیازهای مختلف، توسط افراد متفاوت از سیستم درخواست می گردد. بنابراین استراتژی دریافت نیاز از کاربر در حل مسئله استخراج اطلاعات اهمیت بالایی دارد.

استفاده از زبان طبیعی برای استخراج اطلاعات به دلیل احتمال عدم انطباق دستور زبان استفاده شده و مجموعه متون مورد آزمایش، مناسب نیست. استفاده از زبان های استاندارد، مانند SQL که در پایگاه های داده رابطه ای معمول است در برخی از سامانه های استخراج اطلاعات نوین مورد توجه محققان قرار گرفته است. نمونه آن را می توان در سامانه آرایه شده توسط محققان دانشگاه برکلی (Daisy Z. et al, 2010) می توان مشاهده کرد. استفاده از این روش به دلیل محدودیت هایی که در عمل گره های استاندارد وجود دارد، کارایی پایینی دارد.

استفاده از متون آموزشی و با استفاده از محیط رابط کاربر برای انجام عمل استخراج اطلاعات یکی دیگر از روش های مورد استفاده برای دریافت نیاز کاربر است. در این روش، به منظور مشخص کردن اطلاعات مورد نیاز، حاشیه نویسی متن توسط کاربر انجام گرفته و سپس متن حاشیه نویسی شده به منظور آموزش در اختیار سامانه قرار می گیرد. حاشیه نوشته ها در متن می تواند معرف یک یا چند نیاز کاربر باشد. مزیت استفاده از این روش در مقایسه با سایر روش های موجود، در آن است که اول این که دستور زبان درخواست در مقایسه با روش اول به طور کامل یکنواخت با ساختارهای متن مورد آزمایش است، دوم این که از محدودیت های موجود در پرس و جوهای استاندارد نیز اجتناب کند.



(شکل ۷): مراحل فعالیت مربوط به مرحله آموزش و ساخت مدل

۵- پیاده سازی سامانه

همان گونه که در ساختار سامانه پیشنهادی بیان شده است، این سامانه متشکل از دو مرحله اصلی است. یک مرحله وظیفه آموزش مدل و دیگری وظیفه استخراج اطلاعات و حاشیه نویسی متن را به عهده دارد. در هر یک از زیرسامانه ها، فرایندها همان گونه که در (شکل ۷) مشاهده می شود، به ترتیب از استخراج ساختار کلان آغاز و با حاشیه نویسی گزارش نظامی پایان می یابد. در پیوست ۱ جزئیات هر یک از این زیرسامانه ها تشریح شده است.

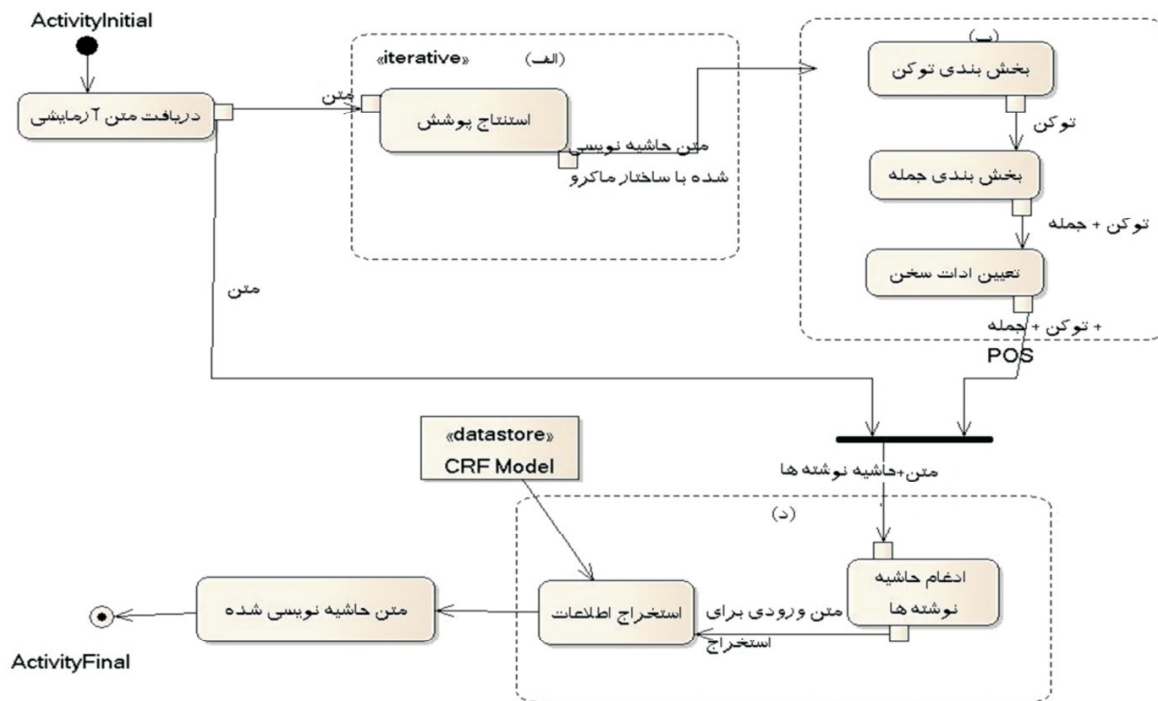
۶- ارزیابی سامانه پیشنهادی

قدم بعد، پس از طراحی و پیاده سازی روش پیشنهادی، ارزیابی آن است. به منظور بررسی کارایی سامانه معیارهای فراخوانی، دقت و معیار F را که در قبل معرفی و مورد استفاد قرار گرفت، به کار می بریم.

روش های حاشیه نویسی متن می تواند به صورت کلان یا خرد انجام شود. بدین منظور ابزار حاشیه نویسی متن می بایست در اختیار کاربر قرار گرفته تا کاربر قادر به مشخص کردن (یا به عبارت دیگر حاشیه نویسی) نیاز یا نیازهای خود در متن باشد. از آنجا که این متون به عنوان ورودی سامانه بوده و به منظور آموزش مورد استفاده قرار می گیرند، پرواضح است که افزایش تعداد متون حاشیه نویسی شده نسبت مستقیم با دقت سامانه خواهد داشت.

۴-۱-۳- مرحله استخراج

اکنون با ساخت مدل، قادر هستیم تا اطلاعات مورد نیاز کاربر را از متون جدید استخراج کنیم. همان گونه که در (شکل ۸) مشاهده می شود، 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 اطلاعات و به کارگیری در مدل حوزه تصادفی شرطی آماده می شوند. سپس با کمک مدل آموزش دیده در مرحله قبل، فرایند استخراج اطلاعات با استفاده از ابزار CRF++ صورت می پذیرد.



(شکل ۸): مراحل فعالیت در مرحله استخراج

۶-۱- بستر آزمایش

از آنجا که این سامانه به منظور استفاده در حوزه نظامی طراحی و پیاده سازی شده است، می بایست حداقل در یک بستر، داده نظامی مورد ارزیابی قرار گیرد. بدین منظور منابع مرتبط از جمله سایت آژانس تحقیقات پیشرفته نظامی ایالات متحده (DARPA) و سایت پیمان آتلانتیک شمالی (ناتو) (*north Atlantic Treaty Organization*) مورد بررسی قرار گرفت و در نهایت تنها یک مجموعه داده نظامی از کشور آلمان یافت شد. این مجموعه داده با نام KFOR^۱ (Hecking M., ۲۰۰۵) مربوط به جنگ کوزوو می باشد. در پروژه^۲ زنون (ZENON project) و در مقاله آقای دکتر ماتیاس هکینگ^۴ (Hecking M., ۲۰۰۵) و (Hecking, ۲۰۰۳) بدان اشاره شده است. این مجموعه داده شامل اطلاعات اخبار صحنه نبرد به دو زبان آلمانی و انگلیسی بوده که آقای دکتر هکینگ در گزارش [۳۱] اعلان کرد که این مجموعه داده طبقه بندی داشته و امکان انتشار آزاد آن وجود ندارد. بنابراین اقدام به تهیه بستر آزمایش از داده های صحنه

نبرد موجود در ویکی لیکس کردیم. این مجموعه داده مربوط به اطلاعات نبرد نیروهای ناتو در افغانستان از سال ۲۰۰۴ و حاوی بیش از هفتصد هزار گزارش است. اطلاعات مربوط به این مجموعه داده در سایت آزمایشگاه سامانه های هوشمند دانشگاه صنعتی امیر کبیر (mollaei, ۲۰۱۰) موجود است.

همچنین این سامانه در یک بستر داده دیگر به منظور بررسی کارایی سامانه با متون خبری، مورد آزمایش قرار می گیرد. این بستر مربوط به سمینارهای CMU (*Data set for CMU seminars*) بوده و این مجموعه داده، شامل اخبار سمینار از دانشگاه کارنیجیه ملون^۵ واقع در ایالت پنسیلوانیاست. موجودیت های مشخص شده شامل شروع و پایان سمینار، محل برگزاری و سخنران آن است. در این مقاله، نسخه ای از مجموعه داده را که برچسب های ادات سخن آن مشخص شده است، استفاده کرده ایم (Brill E., ۱۹۹۴).

۶-۲- انجام آزمایش

به منظور انجام آزمایش و محاسبه هر یک از معیارهای ارزیابی، نیاز اطلاعاتی برای استخراج در هر دو مجموعه داده تعریف می کنیم. بنابراین از مجموعه متون نظامی، نام

5 Carnegie Mellon

^۱ Kosovo Force

^۲ این پروژه در کشور آلمان با موضوع استخراج اطلاعات از گزارشات زبان طبیعی به زبان انگلیسی فعال می باشد.

^۳ ZENON Project

^۴ Dr. Matthias Hecking

(جدول ۲): مقایسه نتایج حاصل از اجرای سامانه پیشنهادی بر روی متون خبری در مقایسه با دو سامانه استخراج اطلاعات

معیار ارزیابی	مورد استخراج	نام سازمان	تاریخ	رخداد	وسایل	میانگین
دقت	ANNNIE	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	ANNNIE	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	ANNNIE	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 مقایسه نتایج سامانه پیشنهادی با نتایج پروژه زنون (جدول ۱) می توان به این نتیجه رسید که سامانه پیشنهادی در مقایسه با سامانه مشابه، دقت پایین تر، اما فراخوانی بالاتر دارد. بهبود معیار F بیان کننده بهبود کارایی کلی استخراج توسط سامانه پیشنهادی است. همچنین هر یک از متون انتخاب شده با استفاده از دو روش مبتنی بر قاعده، مدل مخفی مارکوف مورد ارزیابی قرار گرفته و ابزار ANNNIE برای روش مبتنی بر قاعده و ابزار LingPipe برای روش مدل مخفی مارکوف استفاده شده است.

مقایسه نتایج حاصل از انجام آزمایش توسط سامانه پیشنهادی بر روی داده های غیر نظامی با نتایج حاصل از اجرای سامانه های متداول نشان می دهد که این سامانه قابلیت اجرا بر روی مجموعه های غیر نظامی را نیز دارد. (جدول ۲)

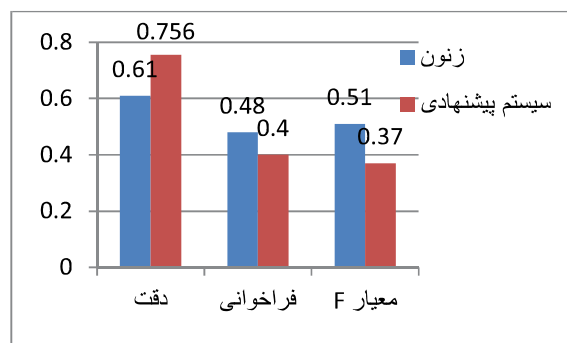
۷- نتیجه گیری

در این مقاله روش های استخراج اطلاعات از متون نظامی مورد بررسی قرار گرفت. در این راستا متون نظامی به دو دسته متون خبری و گزارش های صحنه نبرد تقسیم شده و با بررسی بر روی ساختارهای متون نظامی و مقایسه آنها با متون خبری غیر نظامی مشخص شد که ساختار متون خبری نظامی با ساختار متون غیر نظامی مشابه و با ساختار گزارش های صحنه نبرد تفاوت دارد. با ارزیابی صورت گرفته با سامانه های متداول استخراج اطلاعات، تأثیر این تفاوت بر نتایج استخراج بر روی گزارش های نظامی، نشان داده شد.

در ادامه سامانه های برای استخراج اطلاعات از گزارش های نظامی مبتنی بر دو روش استنتاج پوشش و مدل حوزه تصادفی شرطی پیشنهاد شد؛ سپس با معرفی معیارهای ارزیابی دقت، فراخوانی و معیار F ، آزمایش هایی برای ارزیابی سامانه پیشنهادی انجام دادیم. با استفاده از

سازمان ها، تاریخ، رخداد، تجهیزات نظامی را به عنوان نیاز در نظر گرفته و با نتایج چند سامانه مشابه و همچنین نتایج آقای دکتر هکینگ در (Hecking, 2003) که بر روی داده های KFOR که در دوازدهمین کنفرانس بین المللی تحقیقات و فناوری فرماندهی و کنترل^۱، ارایه شد، مقایسه می کنیم. لذا تعداد سی گزارش از مجموعه گزارش های نظامی مربوط به جنگ عراق و نیز تعداد سی خبر غیر نظامی برای داده های آموزشی و آزمایش انتخاب کردیم. از مجموعه داده غیر نظامی، اطلاعات مربوط به شروع و پایان سمینار و محل برگزاری، به عنوان نیاز در نظر می گیریم. در هر دو مجموعه ۸۰٪ از داده ها را به عنوان داده آزمایشی استفاده می کنیم.

با استفاده از هر یک از مجموعه های داده، سامانه آموزش دیده و سپس عملیات استخراج اطلاعات بر روی آنها صورت گرفته و معیارهای دقت و فراخوانی و F محاسبه شدند.



(نمودار ۲): مقایسه نتایج آزمایش های ارزیابی استخراج اطلاعات سامانه پیشنهادی و سامانه پروژه زنون

مقایسه نتایج حاصل از اجرای سامانه پیشنهادی بر روی مجموعه داده نظامی و یکی لیکس با نتایج پروژه زنون در (جدول ۱) و (نمودار ۲) نشان داده شده است.

(جدول ۱): مقایسه نتایج سامانه پیشنهادی با نتایج سامانه زنون

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

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

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

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

DARPA. USA Defense advanced research Project Agency. 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

Ekbali, 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 Analyzing 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. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (pp. 282-289). 8th International Conference on Machine Learning (ICML'01).

بستر داده‌ای که به‌منظور ارزیابی این سامانه از سایت ویکی‌لیکس تهیه شد، سامانه پیشنهادی، مورد آزمایش قرار گرفت. نتایج این آزمایش با نتایج سامانه زنون مقایسه شد. ارزیابی حاصله بیان‌کننده بهبود کارایی در معیارهای فراخوانی و معیار F بوده و تنها در معیار دقت 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 وقت و ارائه راهنمایی‌های لازم دریغ نورزیدند، کمال تشکر و قدردانی را دارم.

۹- منابع

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

Afghanistan 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>



احمد عبدالله زاده بارفروش همکاری
خود را با دانشکده 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 دانشکده
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 به‌عنوان استاد مهمان در
دانشگاه‌های Trento ایتالیا و Loughborough انگلستان
مشغول به کار بوده است. دکتر عبدالله زاده کتاب‌های
"مقدمه‌ای بر هوش مصنوعی توزیع شده" و "کلیات
متدولوژی تأمین کیفیت" را نیز تألیف کرده‌اند. زمینه‌های
تحقیقاتی مورد علاقه 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 رایانامه ایشان عبارت است از:

ahmad@aut.ac.ir



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

نشانی رایانامه ایشان عبارت است از:

nima.mollaei@yahoo.com

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: A Data-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/>