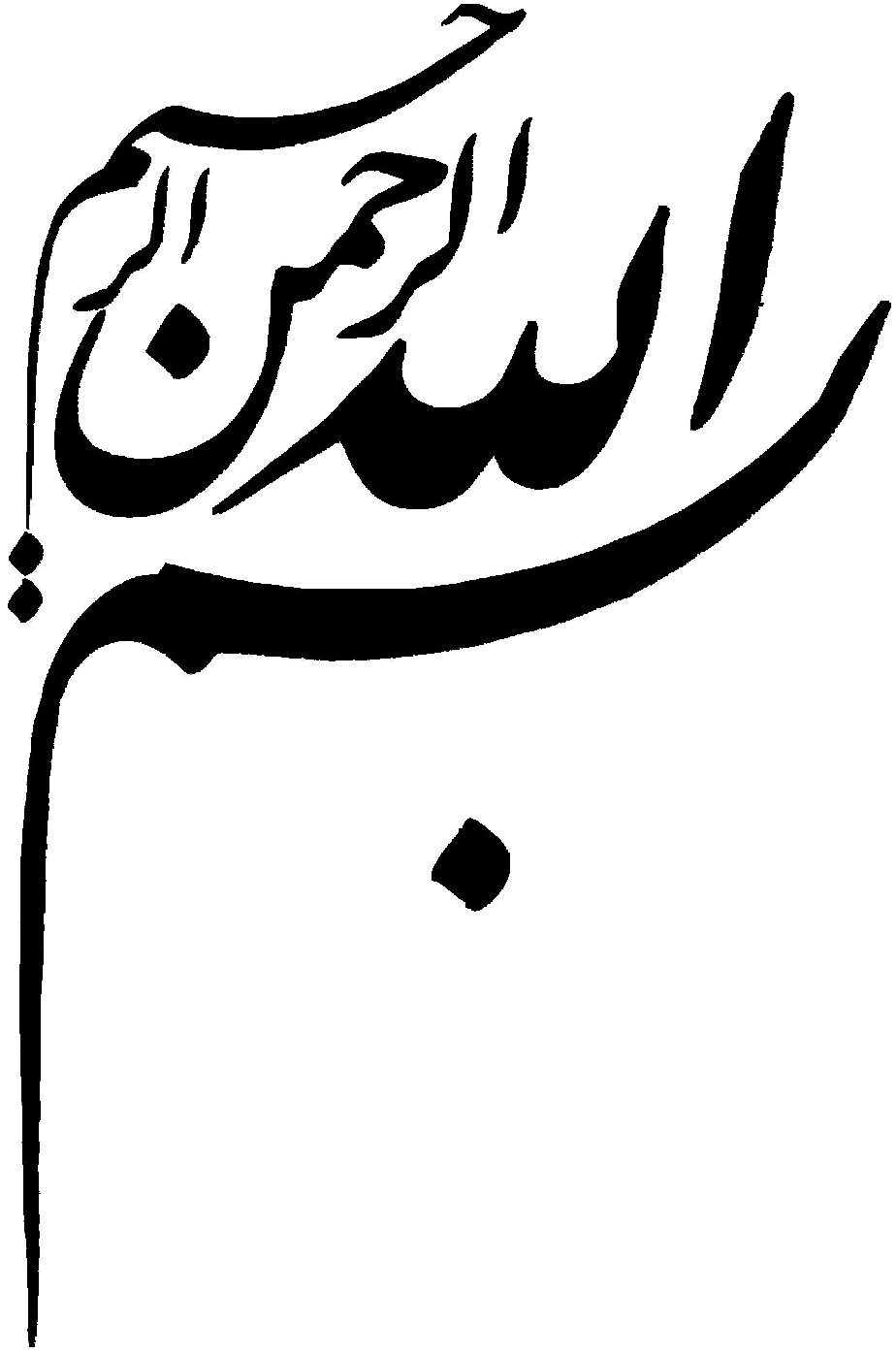
بسم الله الرحمن الرحیم

دانشگاه صنعتی امیرکبیر(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد

در رشته‌ی مهندسی کامپیوتر گرایش هوش ماشین و رباتیک

طراحی و پیاده‌سازی یک سیستم استخراج اطلاعات با استفاده از روش‌های تطبیقی برای متون غیر ساخت‌یافته‌ی زبان فارسی

نگارش:

حمید خدایی

استاد راهنما:

دکتر محمدرضا مطش بروجردی

بهمن ۱۳۸۷

با سپاس فراوان از جناب آقای دکتر

بروجردی که یاری صمیمانه، بی دریغ، و از هر

حیث شان انجام این پروژه را میسر نمود.

تقدیم به پدر و مادرم

چکیده:

در پردازش زبان طبیعی، استخراج اطلاعات نوعی بازیابی اطلاعات بوده که قادر به استخراج اطلاعات ساختاریافته از متون فاقد ساختار به‌وسیله‌ی الگوهایی از پیش تعریف‌شده است. در این پروژه، یک روش شبه‌بی‌نظارت مبتنی‌بر خوشه-بندی در دو مرحله برای تعیین مناسب‌بودن، و همچنین طبقه‌بندی الگوهای استخراج اطلاعات از متن براساس نوع آنها طراحی، و برای زبان فارسی پیاده‌سازی شده است. مرحله‌ی اول خوشه‌بندی بر اساس فرکانس تکرار الگوها و مرحله‌ی دوم بر طبق محتویات معنایی آنها انجام می‌شود که می‌تواند نوع اطلاعاتی را که هر الگو استخراج می‌کند نیز به صورت خودکار تعیین کند. آزمایش روش برروی یک دامنه‌ی خاص (اخبار زلزله) و یک دامنه‌ی عمومی‌تر (اخبار سوانح) انجام یافته و در حالت دامنه‌ی خاص، چند روش دیگر نیز پیاده‌سازی شده، و نتایج آنها مورد مقایسه قرار گرفته است.

برای پردازش متن، از یک روش تکه‌سازی مبتنی بر ماشین بردار پشتیبان به همراه الگوریتمی برای تعیین محدوده‌ی جملات ساده و شکستن جملات مرکب به چند جمله‌ی ساده استفاده شده است که می‌تواند یک جمله را به گروه‌های دستوری تقسیم، و پس از آن، جملات مرکب را به جمله‌هایی ساده تبدیل نماید. همچنین روش تعیین جزء سخن براساس درخت تصمیم نیز استفاده شده است.

کلمات کلیدی:

استخراج اطلاعات، یادگیری تطبیقی الگو، خوشه‌بندی، پردازش زبان طبیعی، تکه‌سازی متن، تعیین جزء سخن، زبان فارسی، متن بدون ساختار، تحلیل جملات مرکب

فهرست مطالب

۱. پیشگفتار ۱

۱-۱. استخراج اطلاعات ۱

۱-۲. تفاوت استخراج و بازیابی اطلاعات ۳

۱-۳. تاریخچه‌ی استخراج اطلاعات ۳

۱-۴. کاربردهای استخراج اطلاعات ۶

۱-۵. مسائل مطرح در سیستم‌های استخراج اطلاعات ۷

۱-۵-۱. مساله‌ی قابلیت انتقال ۷

۱-۵-۲. فرضیات هر روش ۷

۱-۵-۳. ورودی و خروجی در سیستم‌های استخراج اطلاعات ۸

۱-۵-۴. کشف الگوها ۹

۱-۶. معماری سیستم‌های استخراج اطلاعات ۹

۱-۷. روش‌های طراحی ۹

۱-۸. الگوریتم‌های استخراج اطلاعات ۱۰

۱-۸-۱. یادگیری قوانین ۱۲

۱-۸-۱-۱. روش‌های بانظارت ۱۲

۱-۸-۱-۱-۱. یادگیری گزاره‌ای ۱۲

۱-۸-۱-۱-۲. یادگیری رابطه‌ای ۱۳

۱-۸-۱-۲. روشهای شبه بی‌نظارت ۱۴

۱-۸-۲. یادگیری با استفاده از مدل‌های آماری ۱۷

۱-۹. سیستم‌های موجود ۱۷

۱-۹-۱. سیستم‌های مبتنی بر مهندسی دانش برای کشف الگوها ۱۸

۱-۹-۱-۱. FASTUS ۱۸

۱-۹-۱-۲. GE NLTOOLSET ۱۸

1-9-1-3. PLUM ۱۹

۱-۹-۱-۴. Proteus ۱۹

۱-۹-۲. سیستم‌های مبتنی‌بر یادگیری بانظارت برای کشف الگوها ۱۹

۱-۹-۲-۱. AutoSlog ۱۹

۱-۹-۲-۲. PALKA ۱۹

۱-۹-۲-۳. CRYSTAL ۲۰

۱-۹-۲-۴. LIEP ۲۰

۱-۹-۲-۵. WHISK ۲۰

۱-۹-۲-۶. RAPIER ۲۱

۱-۹-۲-۷. GATE ۲۱

۱-۹-۳. سیستم‌های مبتنی‌بر کشف شبه بی‌نظارت الگوها ۲۱

۱-۹-۳-۱. AutoSlog-TS ۲۱

۱-۹-۳-۲. Mutual Bootstrapping ۲۲

۱-۹-۳-۳. EXDISCO ۲۳

۱-۹-۳-۴. Snowball ۲۳

۱-۹-۳-۵. QDIE ۲۴
۱-۱۰. نحوه‌ی ارزیابی سیستم‌های استخراج اطلاعات ۲۴
۱-۱۱. پردازش زبان طبیعی ۲۶
۱-۱۱-۱. مقدمه ۲۶
۱-۱۱-۲. اجزای سیستم پردازش زبان طبیعی ۲۷
۱-۱۱-۲-۱. پیش‌پردازش ۲۷
۱-۱۱-۲-۲. تعیین جزء سخن ۲۷
۱-۱۱-۲-۲-۱. روش مدل مخفی مارکوف و الگوریتم ویتربی ۳۰
۱-۱۱-۲-۲-۲. روش درخت تصمیم ۳۱
۱-۱۱-۲-۲-۳. تعیین جزء سخن برای کلمات ناشناخته ۳۲
۱-۱۱-۲-۳. تعیین گروه‌های دستوری ۳۳
۱-۱۱-۲-۳-۱. ماشین بردار پشتیبان ۳۴
۱-۱۱-۲-۳-۲. تعیین ویژگی‌ها ۳۵
۱-۱۱-۲-۴. تحلیل نحوی و تفسیر معنایی ۳۷
۱-۱۱-۲-۵. تحلیل مباحثه‌ای ۳۸
۱-۱۲. خلاصه و جمع‌بندی ۳۸
۲. پیش‌پردازش ۴۰
۲-۱. پیش‌پردازش متن ۴۰
۲-۲. جداسازی جمله‌ها و واژه‌ها ۴۲
۲-۳. تعیین جزء سخن ۴۴
۲-۴. تعیین گروه‌های دستوری ۴۷

۲-۵. تحلیل نحوی ۵۰

۲-۵-۱. استفاده از تحلیلگرهای نحوی ۵۱

۲-۵-۲. استفاده از تکه‌سازی در سطح دوم ۵۲

۲-۵-۳. الگوریتم تحلیل نحوی ۵۳

۲-۵-۴. تحلیل جملات محذوف ۵۵

۲-۶. تفسیر معنایی ۵۶

۲-۷. خلاصه و جمع‌بندی ۵۸

۳. یادگیری الگوها ۶۰

۳-۱. استخراج الگوها ۶۰

۳-۲. فیلتر کردن الگوهای مناسب ۶۲

۳-۳. آزمایش و ارزیابی ۶۵

۳-۴. تغییر دامنه ۷۰

۳-۵. خلاصه و جمع‌بندی ۷۱

۳-۶. نتیجه‌گیری ۷۲

۳-۷. کارهای آینده ۷۲

۴. واژه‌نامه ۷۴

۵. منابع و مراجع ۷۶

۶. پیوست ۸۱

۶-۱. دستور زبان فارسی ۸۱

۶-۱-۱ انواع نقش‌های دستوری ۸۲

۶-۱-۱-۱. فاعل ۸۲

۶-۱-۱-۲. مسندالیه ۸۲

۶-۱-۱-۳. مفعول ۸۲

۶-۱-۱-۴. مسند ۸۳

۶-۱-۱-۵. متمم ۸۳

۶-۱-۱-۶. قید ۸۴

۶-۱-۱-۷. فعل ۸۴

۶-۱-۱-۸. حروف عطف و ربط ۸۷

۶-۱-۱-۹. منادا ۸۸

۶-۱-۱-۱۰. حروف شرط ۸۸

۶-۱-۱-۱۱. صفت ۸۸

۶-۱-۱-۱۲. مضافالیه ۸۹

۶-۱-۲. انواع جمله ۸۹

۶-۱-۲-۱. جمله‌ی اسنادی و فعلی ۸۹

۶-۱-۲-۲. جمله‌ی ساده و جمله‌ی مرکب ۹۰

۶-۱-۲-۲-۱. جمله‌ی ساده ۹۰

۶-۱-۲-۲-۲. جمله‌ی مرکّب ۹۰

۶-۱-۲-۳. جمله‌ی کامل و محذوف ۹۲

۶-۲. گرامر زبان فارسی ۹۴

فهرست جداول

جدول ۱-۱: نمونه‌ای از اطلاعات استخراج‌شده ۲

جدول ۱-۲: دسته‌بندی روش‌های مبتنی‌بر یادگیری ماشین ۱۱

جدول ۱-۳: نمونه‌ای از اکتشاف‌های AUTOSLOG ۱۹

جدول ۱-۴: نتایج گزارش‌شده‌ی تعیین جزء سخن با روش درخت تصمیم برای زبان انگلیسی ۳۲

جدول ۱-۵: مثالی از IOB و IOE و نظایر آن‌ها ۳۶

جدول ۲-۱: نگاشت تبدیل کاراکترهای یونیکد به اسکی ۴۱

جدول ۲-۲: برچسب‌های پیکره بیجن‌خان ۴۴

جدول ۲-۳: نتایج آموزش تعیین‌کننده‌ی جزء سخن با روش درخت تصمیم برای زبان فارسی ۴۶

جدول ۲-۴: نتایج گزارش‌شده‌ی تعیین جزء سخن برای چند روش ۴۷

جدول ۲-۵: صحت الگوریتم تشخیص گروه‌های دستوری ۵۰

جدول ۲-۶: نتیجه‌ی تکه‌سازی در سطح دوم ۵۳

جدول ۲-۷: موجودیت‌های نامی تعریف‌شده ۵۶

جدول ۳-۱: درصد نمونه‌ها در هر خوشه در مرحله‌ی اول ۶۷

جدول ۳-۲:نتیجه‌ی تخصیص خوشه به شیارها ۶۸

جدول ۳-۳:نتیجه‌ی ارزیابی نهایی سیستم استخراج اطلاعات ۶۸

جدول ۳-۴: نتیجه‌ی ارزیابی نهایی روش AUTOSLOG-TS ۶۹

جدول ۳-۵: نتیجه‌ی ارزیابی نهایی روش BASILI ۶۹

جدول ۳-۶: نتیجه‌ی ارزیابی نهایی روش خوشه‌بندی با دامنه‌ی گسترده ۷۰

فهرست شکل‌ها

شکل ۱-۱: معماری متداول یک سیستم استخراج اطلاعات ([TURMO 2006]) ۹

شکل ۱-۲: یک گره مفهوم که توسط AUTOSLOG استقرا می‌شود. ([RILOFF 1996]) ۱۳

شکل ۱-۳: فلوچارت AUTOSLOG-TS ۲۲

شکل ۱-۴: اجزای اصلی سیستم SNOWBALL ۲۳

شکل ۱-۵: مثال واژه‌نامه برای واژگان پسوندها ([SCHMID 1994]) ۳۳

شکل ۲-۱: معماری سیستم استخراج اطلاعات ۴۰

شکل ۲-۲: درخت لغات چند قسمتی ۴۲

شکل ۲-۳: درخت نحوی جمله‌ی «زلزله‌ای که ...» ۵۵

شکل ۳-۱: نتیجه‌ی خوشه‌بندی مرحله‌ی اول ۶۷

۱. پیش‌گفتار

۱-۱. استخراج اطلاعات

داده‌های فاقد ساختار، مانند متن ساده، معمولاً بخش اعظم دانش ذخیره‌شده‌ی یک سازمان را تشکیل می‌دهد. اما دسترسی، تحلیل، جستجو، و یا استفاده از آن‌ها توسط افراد، بسیار پرهزینه و زمان‌بر می‌باشد. بنابراین، چنانچه ابزارهایی موجود باشد که بتوان به‌وسیله‌ی آن‌ها، اطلاعات موجود در متن‌ها را به صورت یک ساختار ساده و قابل فهم توسط ماشین استخراج کرد، می‌توان از آن اطلاعات به نحو مطلوب استفاده نمود. روشی که برای این منظور به کار گرفته می‌شود، *متن‌کاوی*[۱] نام دارد.

متن‌کاوی که زیر شاخه‌ای از *داده‌کاوی*[۲]ست، دانشی‌ست درباره‌ی نحوه‌ی جستجوی الگوها در متنِ به زبان طبیعی که به صورت «*کشف اطلاعات به‌وسیله‌ی کامپیوتر و با استفاده از استخراج اطلاعات از منابع نوشتاری مختلف*[Hearst 2004]» تعریف شده است.

استخراج اطلاعات[۳] یکی از برجسته‌ترین تکنیک‌هایی‌ست که در متن‌کاوی مورد استفاده قرار می‌گیرد. به طور خاص، در این فن‌آوری می‌توان با استفاده از ترکیب روش‌ها و تکنیک‌های پردازش زبان طبیعی[۴] کارایی بالایی را برای کاوش متن از دامنه‌های[۵] گوناگون به‌دست آورد. بنابراین، استخراج اطلاعات، یک فن‌آوری منشعب از *پردازش زبان طبیعی*‌ست که به تحلیل متن فاقد ساختار و به زبان طبیعی برای مشخص‌نمودن اطلاعات یا وقایعی می‌پردازد که به صورت صریح یا ضمنی در آن وجود دارد[Cowie 1996].

یکی از دلایل تمایل پژوهشگران به این دانش، نقش آن در فراهم نمودن امکان مقایسه و ارزیابی تکنیک‌های مختلف پردازش زبان طبیعی می‌باشد[Cowie 1996]. علاوه‌بر آن، این فناوری،

---

[1] Text Mining (TM)
[2] Data Mining
[3] Information Extraction (IE)
[4] Natural Language Processing (NLP), Natural Language Understanding (NLU)
[5] Domain

می‌تواند کاربردهای گوناگونی برای صنایع مختلف به خصوص فعالان بورس، بانک‌ها، ناشران، و حتی دولتمردان داشته باشد. به طور کل، هر شخص یا سازمانی که نیاز به مطالعه‌ی روزانه‌ی اخبار نوشتاری برای استخراج اطلاعات از آن داشته باشد، می‌تواند از این‌گونه سیستم‌ها استفاده نماید.

یکی دیگر از تعاریفی که می‌توان برای استخراج اطلاعات در نظر گرفت این است که در این دانش سعی بر آن است که بتوان اطلاعات موجود در یک متنِ به زبان طبیعی را به شکل ساختاری سازمان‌یافته و جدول‌گونه[1] ارائه داد[Grishman 1997]. منظور از ساختار جدولی این است که پیش از اعمال یک الگوریتم استخراج اطلاعات، جدولی با دو ستون موجود باشد که در ستون اول، نوع اطلاعاتی که انتظار می‌رود الگوریتم آنها را بیابد قرار دارد و ستون دوم، پس از اجرای الگوریتم، با اطلاعات استخراج‌شده پر گردد. به هر ردیف، که شامل یک زوجِ نوع اطلاعات و اطلاعات استخراج‌شده است یک *شیار*[2] یا شکاف گفته می‌شود. برای نمونه، اگر خبری مانند خبر زیر موجود باشد، سیستم استخراج اطلاعات باید بتواند اطلاعات را به صورت جدول ۱-۱ استخراج نماید:

خبر: انفجار مرکز نیویورک را به لرزه درآورد[3]

انفجار شدیدی در ساعات پر رفت‌وآمد بعد از ظهر در مرکز نیویورک در اثر ترکیدن یک لوله تاسیسات در دل زمین باعث مرگ یک نفر و زخمی شدن دست کم ۱۸ نفر شده است. پلیس تعداد زیادی از ساختمان های اطراف و ایستگاه مرکزی شهر را که انفجار در نزدیکی آن روی داد تخلیه کرد.

انفجار در اثر ترکیدن یک لوله بخار در زیر خیابان لکزینگتون در ناحیه منهتن روی داد. اداره پلیس نیویورک گفت حادثه که باعث ایجاد حفره بزرگی در خیابان شد، تروریستی نبوده است.

انفجار بزرگ روز چهارشنبه که درست قبل از ساعت شش بعد از ظهر روی داد باعث ایجاد وحشت و هرج و مرج در خیابان های بخش شرقی منهتن شد و درحالی که بخار با غلظت زیاد از دل زمین بیرون می زد مردم از صحنه می گریختند. ماموران آتش نشانی و سازمان های اضطراری به نقطه حادثه شتافته و بخش هایی از خیابان، در فاصله ایستگاه "گراند سنترال" و ساختمان کرایسلر را بستند. هزاران عابر از ایستگاه تخلیه شدند.

جدول ۱-۱: نمونه‌ای از اطلاعات استخراج‌شده

نوع حادثه	انفجار
علت	ترکیدن یک لوله تاسیسات در دل زمین
تلفات	مرگ یک نفر و زخمی شدن دست کم ۱۸ نفر
مکان	خیابان لکزینگتون، منهتن، نیویورک
زمان	چهارشنبه-قبل از ساعت ۶ بعد از ظهر

[1] Tabular
[2] Slot

[3] منبع خبر BBC است.

به طوری که در مثال فوق مشاهده می‌شود، به‌ازای یک خبر، ۵ شیار اطلاعاتی نوع، زمان، مکان، تلفات، و علت حادثه استخراج گردیده است.

۱–۲. تفاوت استخراج و بازیابی اطلاعات

به طور سنتی، اطلاعات موجود در متون توسط افراد خبره‌ی مربوط به دامنه‌ی متن و به صورت دستی[1] استخراج می‌شده‌است که بدیهی‌ست این عمل بسیار پرهزینه چه از لحاظ زمانی و چه از لحاظ مسائل مربوط به نیروی انسانی می‌باشد. به همین دلیل، در دهه‌های اخیر، تلاش‌هایی برای ایجاد سیستم‌هایی موسوم به *هوش متنی*[2] انجام گرفته که هدف آن‌ها تغییر متن بوده چنان که بتوان به اطلاعات موجود در آن، دست یافت[JACOBS 1992, Turmo 2006]. متن‌ها معمولاً هنگامی که برای استفاده‌ی ماشینی تولید می‌شوند(مانند زبان‌های برنامه‌نویسی)، به صورت کاملاً ساختاریافته هستند ولی هنگامی که برای استفاده توسط انسان تهیه می‌شوند، به شکل زبان طبیعی بوده که فاقد ساختاری صریح می‌باشد. که این امر موجب می‌شود استخراج اطلاعات از آن‌ها نیاز به دانشی وسیع از زبان و دامنه‌ی متن داشته باشد.

هوش متنی به دو زیرشاخه‌ی اصلی بازیابی اطلاعات و استخراج اطلاعات، تقسیم شده است. در بازیابی اطلاعات، هدف این است که از مجموعه‌ای از مستندات، آنهایی که مرتبط با یک جُستار[3] ورودی‌اند، انتخاب شوند. در حالی که در استخراج اطلاعات، باید از داخل یک متن، اطلاعات مهم آن استخراج شده و به کاربر ارائه شود.

در سیستم‌های بازیابی اطلاعات، هر متن، تنها به صورت مجموعه‌ای از کلید-واژه‌ها پنداشته می‌شود. در حالی که در استخراج اطلاعات، به ساختار زبانی جملات توجه شده و اطلاعات بر اساس الگوهایی مطابق با ساختارهای دستوری زبان طبیعی متن استخراج می‌شوند.

در سیستم‌های بازیابی اطلاعات، برخلاف استخراج اطلاعات، از دانش پردازش زبان طبیعی، تنها برای مسائلی نظیر تعیین محدوده‌ی واژه‌ها، تعیین ریشه‌ی کلمه‌ها، تشکیل گروه‌های اسمی، و نظایر آن‌ها استفاده می‌شود. ولی در استخراج اطلاعات، سطوح بالاتر این دانش مانند تحلیل‌گر[4]های نحوی، تفسیرکننده‌های معنایی نیز به کار گرفته می‌شوند.

۱–۳. تاریخچه‌ی استخراج اطلاعات

توسعه‌ی دانش استخراج اطلاعات وابسته به *کنفرانس درک متن*[5] (MUC) بوده که از سال ۱۹۸۷ تا ۱۹۹۸ برگزار می‌شده‌است. هدف اصلی کلی کنفرانس مذکور، برگزاری رقابتی بر اساس

[1] Manually
[2] Text-based Intelligence (TBI)
[3] Query
[4] Parser
[5] Message Understanding Conference(MUC): http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

سیستم‌های استخراج اطلاعات از متون با دامنه‌ی محدود و قالب آزاد بوده است که توسط تیم‌های مختلف طراحی شده بودند. معمولاً برای هر کنفرانس، یک دامنه انتخاب می‌شد. خلاصه‌ای از فعالیت-های انجام‌شده در دوره‌های مختلف، بدین شرح است:

- **MUC-1 (۱۹۸۷)**: اولین MUC به صورت تحقیقاتی برگزار شد. به این صورت که نه الگوریتمی ارائه گردید و نه معیاری برای ارزیابی؛ و هر گروه، به طراحی قالب خود برای ارائه-ی اطلاعات بسنده نمود. دامنه‌ی متون، عملیات تاکتیکی دریایی انتخاب شده بود.
- **MUC-2 (۱۹۸۹)**: در این کنفرانس، دامنه‌ی کنفرانس قبلی انتخاب شد. الگوریتمی نیز برای پر کردن الگوها ارائه گردید. ۱۰ شیار نوع رویداد، عامل، زمان و مکان، اثر، و .. برای پرکردن به شرکت‌کنندگان معرفی شد. برای ارزیابی و شیوه‌ی آن، معیار واحدی ارائه نگردید.
- **MUC-3 (۱۹۹۱)**: دامنه‌ی مستندات به وقایع تروریستی در آمریکای لاتین با ۱۸ شیار نوع حادثه، تاریخ، زمان، عامل، هدف، وسیله، و ... تغییر کرد. تعداد ۱۳۰۰ متن آموزشی و ۳۰۰ متن آزمایشی در اختیار شرکت‌کنندگان قرار گرفته بود. ۴ میزان[1] برای اندازه‌گیری معرفی شده بود: تعداد شکاف‌هایی که درست استخراج شده‌اند (COR)، تعداد شکاف‌هایی که خالی مانده‌اند(INC)، تعداد شکاف‌هایی که اشتباه پر شده‌اند(SPUR)، تعداد اطلاعات موجود در متن که استخراج نشده‌اند(MISS)، و تعداد اطلاعاتی که ناقص استخراج‌شده‌اند(PAR). همچنین دو معیار بازخوانی(R)[2] و دقت(P)[3] برای ارزیابی به صورت زیر تعریف شده بودند:

$$R = \frac{COR + 0.5 \times PAR}{COR + PAR + INC + MISS} \qquad (۱)$$

$$P = \frac{COR + 0.5 \times PAR}{COR + PAR + INC + SPUR} \qquad (۲)$$

- **MUC-4 (۱۹۹۲)**: این کنفرانس مشابه MUC-3 بود با این تفاوت که تعداد شکاف‌ها به ۲۴ افزایش یافت. برای ارزیابی نیز معیار F بر اساس دقت و بازخوانی به صورت زیر تعریف گردید:

$$F = \frac{\left(\beta^2 + 1\right) \times P \times R}{\beta^2 \times P + R} \quad 0 < \beta \le 1 \qquad (۳)$$

- **MUC-5 (۱۹۹۳)**: در کنفرانس‌های قبلی، هدف تنها استخراج اطلاعات از متون به زبان انگلیسی بود. در MUC-5، برای زبان ژاپنی نیز سیستمی ارائه شد. همچنین دو دامنه‌ی مختلف اخبار مالی در زمینه‌ی سرمایه‌گذاری مشترک و پیشرفت‌های محصولات میکرو-الکترونیک، پیشنهاد شد. برای ارزیابی، علاوه‌بر روش‌های کنفرانس‌های قبلی، معیارهایی نیز براساس خطا معرفی شدند.

---

[1] Measure
[2] Recall (R)
[3] Precision (P)

- **MUC-6 (۱۹۹۵)**: در این کنفرانس، دامنه مالی مورد استفاده قرار گرفت و اهدافی نظیر استقلال از دامنه، قابلیت انتقال[1]، و فهم عمیق‌تر متن عنوان شد. و دو معیار بازخوانی و دقت به صورت زیر تغییر کردند:

$$R = \frac{COR}{COR + INC + MISS} \quad (۴)$$

$$P = \frac{COR}{COR + INC + SPUR} \quad (۵)$$

- **MUC-7 (۱۹۹۸)**: در این کنفرانس، دامنه‌ی سوانح هوایی استفاده شد و زبان چینی نیز مورد بررسی قرار گرفت.

به موازات MUC، اتحادیه‌ی اروپا بر روی برنامه‌ی LRE[2] که در مورد روش‌ها و ابزارهای استخراج و بازیابی اطلاعات است، سرمایه‌گذاری نمود که نتایج آن، پروژه‌های ECRAN[3]، SPARKLE[4]، FACILIE[5]، و AVENTINUS[6] می‌باشد.

پس از کنفرانس‌های MUC، از سال ۱۹۹۹، پژوهش بر روی فناوری استخراج اطلاعات به برنامه‌ی TIDES[7] با سرمایه‌گذاری DARPA منتقل شد. تعدادی از پروژه‌های نهایی شامل RIPTIDES[8]، PROTEUS[9]، CREST[10]، Coreference.com[11]، و UMass[12] می‌باشد.

رقابت اولیه‌ای TIDES، ACE[13] نام داشت که از سال ۱۹۹۹ با هدف استخراج خودکار محتوا از متون به زبان طبیعی، شروع به کار کرد. منابع استفاده‌شده شامل مطالبی بود که از منابعی نظیر مجلات الکترونیکی، اخبار همگانی، و روزنامه‌ها(از طریق OCR) گردآوری شد بود. فعالیت‌های انجام‌شده از طریق اضافه‌کردن موجودیت‌های سلسله‌مراتبی پیچیده‌تر از MUC شد.

در فاز دوم ACE (۲۰۰۱ تا ۲۰۰۲)، ارزیابی بر میزان تشخیص روابط و توصیفات اضافه شد. رقابت ۲۰۰۳، برای اولین بار شامل زبان‌های غیر انگلیسی (چینی و عربی) نیز شد. در سال ۲۰۰۴، برنامه به توضیح روابط عربی و نرمال‌سازی زمان و تاریخ گسترش یافت. همچنین یک الگوریتم برای تشخیص رویدادها ارائه شد. مفهوم رویداد در ACE ساده‌تر از MUC است و شامل موجودیت‌ها،

---

[1] Portability
[2] Linguistic Research and Engineering (http://www.echi.lu/language/en/lehome.html)
[3] http://www.dcs.shef.ac.uk/intranet/research/networks/Ecran/
[4] http://www.informatics.susx.ac.uk/research/nlp/sparkle/sparkle.html
[5] http://tcc.itc.it/research/textec/projects/facile/facile.html
[6] http://www.dcs.shef.ac.uk/nlp/funded//aventinus.html
[7] Tran-lingual Information Detection, Extraction, and Summarization: http://www.darpa.mil/ipto/programs/tides/
[8] http://www.cs.cornell.edu/home/ cardie/tides
[9] http://nlp.cs.nyu.edu/
[10] http://crl.nmsu.edu/Research/Projects/Crest
[11] http://www.coreference.com/lingpipe/
[12] http://ciir.cs.umass.edu/research/tides.html
[13] Automatic Content Extraction: http://www.nist.gov/speech/tests/ace/

مقادیر، و زمان می‌باشد. در سال ۲۰۰۵، منابع جدیدی به هر سه زبان اضافه شد(مکالمات تلفنی، گروه‌های خبری، وبلاگ‌ها و ...).

علاوه‌بر کنفرانس‌های فوق، تلاش‌های متفرقه‌ی دیگری هم در زمینه‌ی دامنه‌های دیگر انجام شده است. به عنوان مثال، [Soderland 1995] سیستمی را برای استخراج اطلاعات پزشکی از گزارش‌های ترخیص بیمارستان‌ها طراحی نمود و [Holowczak 1997] از اطلاعات استخراج‌شده‌ی متون حقوقی، برای دسته‌بندی آنها استفاده کرد. همچنین [Glasgow 1998] برای دامنه‌ی بیمه‌ی عمر سیستمی را ارائه نمود.

برخی از تلاش‌ها نیز برروی کاربردهای استخراج اطلاعات برروی متون نیمه‌ساخت‌یافته انجام گرفت. مانند دامنه‌ی آگهی اجاره‌ی آپارتمان توسط [Soderland 1999]، دامنه‌ی biomedical توسط [Craven 1999].

به موازات TIDES، اتحادیه‌ی اروپا بر روی Pascal Network of Excellence[1] سرمایه-گذاری نمود. این سازمان همراه با پروژه‌ی Dot.Kom European[2]، پروژه‌ی Challenge on Evaluation of Machine Learning for Information Extraction from Documents را اجرا می‌کند. اولین رقابت آن در ژوئن ۲۰۰۴ شروع و ارزیابی‌های رسمی در نوامبر ۲۰۰۴ انجام گرفت. تفاوت عمده‌ی آن با رقابت‌های MUC در مبتنی بر یادگیری ماشین بودن آن است.

۱-۴. کاربردهای استخراج اطلاعات

از آنجایی که استخراج اطلاعات قادر است یک متن را به ساختار جدولی تبدیل کند، می-توان از نتیجه‌ی آن، در هرگونه عملیات پردازشی استفاده نمود. زیرا جدول، ساختاری است که به سادگی قابل پردازش می‌باشد. به عنوان مثال، می‌توان اطلاعات بازار بورس را در یک دوره‌ی مشخص از روی اخبار موجود استخراج کرد و پردازش‌های لازم را بر روی آن انجام داد.

پایگاه‌های داده‌ی رابطه‌ای براساس ساختاری جدولی تعریف می‌شوند. از این رو به‌وسیله‌ی یک سیستم استخراج اطلاعات، تبدیل متون به یک پایگاه داده‌ی حاوی اطلاعات مورد نیاز امکان‌پذیر است.

در کاربرد خلاصه‌سازی، باید بخش‌هایی از متن که حاوی اطلاعات مهم است مشخص شود. از این رو می‌توان سیستم استخراج اطلاعاتی را به‌کار گرفت که جملات دارای اطلاعات را برچسب-گذاری کند. زیرا معمولاً جملاتی مهم هستند که دارای اطلاعات باشند.

در زمینه‌ی بازیابی اطلاعات نیز استخراج اطلاعات مورد استفاده قرار گرفته است. زیرا با استفاده از این دانش می‌توان به بخش‌هایی از متن بچسب‌ها معنایی تخصیص داد که باعث می‌شود

[1] http://www.pascal-network.org/
[2] http://www.dot-kom.org

استفاده از جستارهای غنی‌تر میسر گردد. چرا که امکان ارائه‌ی آن را به زبان طبیعی میسر می‌سازد. به عنوان مثال با دادن جستار «چه تعداد زلزله‌ی بالای ۶ ریشتر در دو سال گذشته در اندونزی رخ داده است» می‌توان در مستندات مربوط به زلزله که قبلاً اطلاعات آنها شامل «شدت، زمان، مکان، و ...» استخراج شده است و پس از فهم کلمه‌ی «چه تعداد»، اطلاعات مورد نیاز را با یک جستجوی بسیار ساده به کاربر ارائه داد.[1]

در سیستم‌های پاسخ به پرسش که وظیفه دارند با خواندن یک متن به پرسش‌های کاربر در مورد آن پاسخ گویند نیز می‌توان از استخراج اطلاعات استفاده کرد. زیرا اغلب پرسش‌های کاربر در مورد اطلاعات موجود در متن است و اگر این اطلاعات از قبل استخراج شده باشند، پاسخ به پرسش-های کاربر به‌سادگی میسر خواهد بود.

در سیستم‌های اخذ دانش[2] از متن، می‌توان از استخراج اطلاعات به‌منظور فراگیری دانشِ از نوع واقعیت‌ها[3] استفاده نمود. زیرا هر شیار استخراج‌شده مانند <Slot,Value> می‌تواند به صورت منطقی Slot(Value) در نظر گرفته شود.

۱–۵. مسائل مطرح در سیستم‌های استخراج اطلاعات

۱–۵–۱. مساله‌ی قابلیت انتقال[4]

یکی از مسائلی که در کاربردی‌شدن استخراج اطلاعات نقش فراوانی دارد، مساله‌ی قابلیت انتقال یا هزینه‌ی تطبیق‌دادن یک سیستم با دامنه‌های جدید می‌باشد. از آنجایی که سیستم استخراج اطلاعات شدیداً وابسته به دامنه می‌باشد، لذا برای انتقال به یک دامنه‌ی جدید می‌بایست سیستم جدیدی ایجاد نمود که این عمل، دارای هزینه است و یکی از عوامل موثر بر این هزینه، میزان تغییراتی‌ست که باید در سیستم داده شود. عامل دیگر، مقدار اطلاعاتی‌ست که باید مجدداً از کاربر دریافت گردد. هر چه این هزینه‌ها کمتر باشد، سیستم از این لحاظ مطلوب‌تر خواهد بود.

۱–۵–۲. فرضیات هر روش

برای انتخاب یا طراحی یک الگوریتم استخراج اطلاعات و این که بهتر است از روش‌های دستی استفاده شود و یا روش‌های مبتنی‌بر یادگیری ماشین، فرضیات زیر باید مد نظر قرار گیرند:

۱. **موجود بودن داده‌های آموزشی:** اگر داده‌های آموزشی در دسترس یا دست‌یابی به آن کم‌هزینه باشد، روش‌های یادگیری می‌توانند مورد استفاده قرار گیرند.

---

[1] ر.ک. [Moens 2006]

[2] Knowledge Acquisition
[3] Fact
[4] Portability