عنوان زیرپروژه:

مطالعه و بررسی ابزارهای آماده برای تحلیل پیکره متنی زبان فارسی

فهرست مطالب

عنوان شماره صفحه

1. سازمان‌ها و مراکز تحقیقاتی مهم

1-1. مقدمه

در این فصل به سازمان‌ها و مراکز تحقیقاتی مهم که در حوزه پردازش زبان طبیعی، زبانشناسی محاسباتی و مهندسی زبان فعال می‌باشند، می‌پردازیم. سعی می‌شود که فعالیت‌هایی که این سازمان و مراکز انجام می‌دهند، ماموریت‌هایی که برای هر کدام تعریف شده است و خدماتی که سازمان و مرکز ارائه می‌دهد، پوشش داده شود. همچنین در مواردی که سازمان یا مرکز مورد بحث دارای ساختار گسترده و زیربخش‌هایی باشد، سعی خواهد شد که ساختار آن سازمان یا مرکز و زیربخش‌های هر کدام نیز مورد بررسی قرار گیرد. هدف از این بررسی به چندین دلیل باز می‌گردد. اولا در این‌جا سازمان‌ها یا مراکزی تحت بررسی قرار می‌گیرند که حوزه زبان‌هایی را که تحت پوشش دارند به زبان فارسی بی‌ارتباط نباشد. یعنی به عبارت دیگر حوزه تحت پوشش‌شان زبان‌های هند و اروپایی باشد که البته زبان فارسی نیز در این دسته تقسیم‌بندی می‌شود. ثانیا به دلیل مشابهت این زبان‌ها و زبان فارسی، می‌توان از استانداردها و مدل‌های تولید شده آن‌ها در زبان فارسی استفاده کرد.

همچنین ساختار تشکیلاتی و اهداف و ماموریت‌های هر سازمان می‌تواند الگوی مناسبی برای ایجاد سازمان‌های مشابه برای زبان فارسی باشد، چه این‌که نتایج حاصل از عملکرد این سازمان‌هاست که منجر به پیشرفت و توسعه هم‌جانبه زبان‌های تحت پوشش این سازمان‌ها در سطح گسترده شده است. لذا

انتظار می‌رود ایجاد سازمان‌های مشابه برای زبان فارسی منجر به پیشرفت و توسعه کاربری خط و زبان فارسی در محیط رایانه‌ای گردد و چالش‌های حضور زبان فارسی در رقابت‌پذیری با دیگر زبان‌های مطرح همچون انگلیسی مرتفع گردد.

2-1. انجمن منابع زبانی اروپایی (ELRA)[1]

1-2-1. فعالیت، ماموریت و خدمات

ماموریت ELRA ارتقای منابع زبانی برای بخش تکنولوژی زبان طبیعی[2] و همچنین ارزیابی تکنولوژی‌های مهندسی زبان می‌باشد. ELRA در راستای این دو ماموریت خدمات زیر را ارائه می‌دهد:

- تشخیص منابع زبانی
- ارتقای سطح تولید منابع زبانی
- تولید منابع زبانی
- اعتبارسنجی منابع زبانی
- ارزیابی سیستم‌ها، محصولات و ابزارهای مربوط به منابع زبانی
- توزیع منابع زبانی

[1] European Language Resources Association
[2] Human Language Technology sector

- استاندارد سازی

ارتقای سطح تولید منابع زبانی شامل پشتیبانی از زیرساخت‌ها برای برگزاری مسابقات و پشتیبانی از توسعه حوزه‌های علمی ترکیب و ارزیابی منابع زبانی از طریق کنفرانس‌های LREC (کنفرانس منابع زبانی و ارزیابی)[1] می‌باشد.

از خدمات فوق، ارزیابی و توزیع، توسط ELDA (آژانس ارزیابی و توزیع منابع زبانی)[2] انجام می‌شود.

ELRA همچنین با هدایت و مطالعات در حوزه تکنولوژی زبان طبیعی (HLT)[3] و انتشار خبرنامه[4] در توسعه تکنولوژی‌های زبان طبیعی و ارتقاء آنها در سطح قاره اروپا و در سطح بین‌المللی مشارکت می‌کند.

2-2-1. توزیع منابع زبانی

ELRA در توزیع منابع زبانی تولید شده توسط افراد شرکت‌ها و سازمان‌های نیز فعالیت دارد. برای اینکار ELRA با ارائه پیشنهاد در مورد توزیع محصولات زبانی تولید امکان توزیع محصول را در سطح گسترده‌تر ممکن می‌سازد. به عنوان مثال دادگان گفتاری زبان فارسی (Farsdat) که توسط پژوهشکده هوشمند علائم تهیه شده است توسط ELRA توزیع می‌گردد.

[1] Language Resources and Evaluation Conference
[2] Evaluations and Language resources Distribution Agency
[3] Human Langue Technology
[4] Newsletter

3-1. ELDA (آژانس ارزیابی و توزیع منابع زبانی)

1-3-1. فعالیت، ماموریت و خدمات ELDA

ELDA (آژانس ارزیابی و توزیع منابع زبانی) همزمان با ELRA در فوریه 1995 تشکیل شد. ELDA به عنوان یک شرکت مدیریت همه فعالیت‌های تجاری ELRA را بـه عهـده داشـته و بـه صـورت بـازوی اجرایی ELRA فعالیت می‌کند. ELDA مسئول توسعه و پیاده‌سازی ماموریت‌ها و فعالیـت‌هـای تعریـف شده برای ELRA می‌باشد.

فعالیت‌های ELDA در سه مورد زیر خلاصه می‌گردد:

1- توزیع منابع زبانی: در حال حاضر حدود 850 منبع زبانی نوشتاری و گفتاری توسـط ELDA توزیع مـی‌گـردد. تشـخیص و جمـع‌آوری منـابع زبـانی موجـود بخشـی از فعالیت‌های این مجموعه است.

2- تولید و یا مشارکت در تولید منابع زبانی: ELDA در تولید منابع زبانی و یا مشارکت در تولید منابع زبانی نوشتاری و گفتاری به منظور اهداف توسعه‌ای و یـا تحقیقـاتی فعالیت می‌کند.

3- ارزیابی تکنولوژی‌های زبان طبیعی: ELDA در ارزیـابی و آزمـایش تکنولـوژی‌هـای زبانی با برگزاری مسابقات مشارکت می‌کند. از نمونه فعالیـت‌هـای ELDA در ایـن

زمینه همکاری با CLEF[1] می‌باشد.

ELDA که به منظور ارزیابی و توزیع منابع زبانی در سطح گسترده ایجاد شده است، به صورت مشارکتی محصولات زبانی تولید شده توسط افراد شرکت‌ها و سازمان‌هارا به دو قیمت تجاری و آکادمیک توزیع می‌کند. ELDA منابع زبانی را در چهار طبقه توزیع می‌کند:

1- منابع گفتاری[2]

a) رکوردهای تلفنی[3]: در این بخش بانک‌های صوتی تجاری گفتارهای ضبط شده از روی تلفن (ثابت یا همراه) و یا با استفاده از میکروفون قرار دارد. (حمل دادگان speech Dot)

b) رکوردهای میکروفونی/ رومیزی[4]: در این بخش بانک صوتی ضبط شده با کمک میکروفون قرار دارد. (مثل دادگان BABEL)

c) منابع رادیویی[5]: در این بانک‌های صوتی حاوی گفتار ضبط گویندگان در رادیو، تلویزیون و اینترنت قرار می‌گیرد (مثل بانک پروژه پیکره خبری رادیویی ایتالیایی Italian Broad Gust news corpus)

[1] Cross Language Evaluation Forum، CLEF هر ساله تشکیل می‌شود و در قالب مسابقات در ارتقای تحقیقات در زمینه‌های کاربردی مربوط به زبان تلاش می‌نماید.

[2] Speech Resources

[3] Telephone Recordings

[4] Desktop/Microphone Recording

[5] Broadcast Resources

d) منابع مرتبط با گفتار[1]: در این بخش واژگان‌های آوایی و واجی قرار می‌گیرد. (مثل دادگان‌های MHATLEXC, BDLEX PHONOLEX)

2- منابع زبانی نوشتاری[2]:

a) پیکره‌ها: این بخش شامل پیکره‌ها تک زبانه، چندزبانه (موازی یا غیرموازی)، به صورت برچسب خورده یا برچسب نخورده می‌باشد. (نمونه‌هایی مثل MULTEXT، پیکره‌های موازی و چندزبانه (MLCC)، پیکره علمی فرانسوی، پیکره‌های روزنامه‌ای عربی و...)

b) واژگان‌های تک زبانه: این بخش اختصاص به واژگان‌های تک زبانه دارد که شامل انواع مختلف فرهنگ لغت‌نامه‌ها (دیکشنری‌ها) می‌باشند (مثل دیکشنری افعال فرانسوی، دیکشنری کلمات ژاپنی، چند واژگان PAROLE در بسیاری از زبان‌ها و...)

c) واژگان‌های چند زبانه: در این قسمت واژگان‌ها و دیکشنری‌های دو زبانه و چند زبانه قرار می‌گیرد (مانند Euro wordnet)

3- منابع زبانی اصطلاحات فنی3: دادگان‌های اصطلاحات فنی تک زبانه، دو زبانه و چند زبانه در این طبقه قرار می‌گیرند. این دادگان‌ها تعداد زیادی از دامنه‌های تخصصی را مانند مهندسی خودرو، بیمه، زبان‌شناسی، امور مالی و.... در زبان‌های مختلفی را پوشش می‌دهند.

[1] Speech Related Resources

[2] Written Language Resources

[3] Terminological Language Resources

4- منابع زبانی Multimodal/Multimedia: منابعی که در این بخش قرار دارد با استفاده از modality مختلف تولید شده‌اند. (مانند دادگان تولید شده در پروژه M2VTS)

4-1. EAGLES (گروه مشاوره خبره در استانداردهای مهندسی زبان)[1]

1-4-1. معرفی EAGLES

ایده استفاده مجدد[2] از محصولات زبانی مورد توجه بسیاری از محققان، مهندسان و طراحان تکنولوژی‌های زبان می‌باشد. استفاده مجدد از محصولات زبانی، نقشی اساسی در توانمند سازی توسعه محصولات تکنولوژی‌های زبانی برای پاسخ به نیازهای کاربران دارد.

استفاده مجدد در حوزه تکنولوژی‌های زبان همانند دیگر حوزه‌های تکنولوژی بر وجود استانداردها، راهنماها عملیات مشترک و زیرساخت‌های سازگار و مناسب متکی می‌باشد.

با استانداردهایی که به صورت گسترده پذیرفته شده و مورد استفاده قرار می‌گیرند تغییرات در مؤلفه‌های تکنولوژی زبان ساده و ممکن می‌گردد. ابزارها می‌توانند به گونه‌ای ساخته شوند که ورودی‌هایی را پذیرفته و خروجی‌هایی را تولید کنند که شکلی استاندارد دارند، منابع می‌توانند بر اساس

[1] Expert Advisory Group on Language Engineering Standards
[2] Reusability

یک استاندارد طراحی و تولید شدند و یا در قالب شکلی استاندارد در آیند و محصولات از یک نوع، اگر در استانداردی مناسب باشند قابل مقایسه خواهند بود.

راهنماهای EAGLES نتیجه تلاشی است که توسط جمع بزرگی از مهندسان زبان برای پیشنهاد استانداردها، راهنماها و توصیه‌هایی است برای چند حوزه اساسی زیر:

- واژگان‌های محاسباتی
- پیکره‌های متنی
- فرمالیسم‌های زبانشناسی محاسباتی
- منابع زبانی گفتاری
- ارزش‌گذاری و ارزیابی

بنابراین EAGLES به جنبه‌های زیر از مهندسی زبان می‌پردازند:

- منابع (واژگان، پیکره‌ها)
- روش‌هایی برای توصیف و مدیریت دانش و داده (فرازبان‌ها[1]، فرمالیسم‌ها[2])
- روش‌هایی برای ارزش گذاری و ارزیابی منابع، ابزارها، محصولات
- سیستم‌های قابل حمل و Generic

[1] Metalanguage
[2] Formalism

1-4-2. فعالیت، ماموریت و خدمات

EAGLES یک سازمان اروپایی است پیش زمینه آن به یک برنامه مهندسی و تحقیقاتی زبان‌شناسی (همانند ELRA و ELDA) باز میگردد. هدف EAGLES تسریع در فراهم‌سازی استانداردها برای موارد زیر است:

- منابع زبانی بسیار بزرگ[1] (از قبیل پیکره‌های متنی، واژگان‌های محاسباتی، پیکره‌های گفتاری)
- روش‌های مدیریت منابع زبانی فوق، از طریق فرمالیسم‌های زبانشناسی محاسباتی، زبان‌های نشانه‌گذاری و ابزارهای نرم‌افزاری مختلف
- روش‌های ارزش‌گذاری و ارزیابی منابع زبانی و ابزارها و محصولات

بسیاری از شرکت‌ها، مراکز تحقیقاتی، دانشگاه‌های و اشخاص مشهوری در اتحادیه اروپا در تولید راهنماهای EAGLES[2] مشارکت داشته‌اند. راهنماهای EAGLES شامل توصیه‌ها و پیشنهاداتی برای de fecto و برای فعالیت در حوزه‌های مهندسی زبان (نامبرده شده در بالا) می‌باشند.

فعالیت EAGLES توسط پنج کارگروه انجام می‌گیرد:

- پیکره‌های متنی
- واژگان‌های محاسباتی

[1] Very Large Scale Language Resources
[2] EAGLES Guidelines

- فرمالیسم‌های دستور زبان[1]
- ارزیابی
- زبان گفتاری[2]

این کارگروه‌ها در وهله به روش شناسی‌های رایج برای پنج حوزه تحت پوشش EAGLES می پردازند که پس از آن استانداردهای de fecto[3] حاصل می‌گردد.

3-4-1. متدولوژی EAGLES

ایده اساسی فعالیت‌های EAGLES عمل کردن به عنوان یک کاتالیزور به منظور جمع آوری نتایج پروژه‌های بزرگ اروپایی است. عملیات مشترک و استانداردهای در حال طراحی در جاهایی که مناسب باشند خصوصا در حوزه واژگان، رمزگذاری متن و گفتار، به عنوان ورودی فعالیت‌های EAGLES در نظر گرفته می‌شوند.

تئوری‌های بسیار زیاد و همچنین هر توصیه و پیشنهادی برای هماهنگ‌سازی نیازها و ساختار تئوری‌های مهم و مختلف باید در نظر گرفته می‌شود. EAGLES همچنین نتایج پروژه‌های مهم که در پیشبرد مسائل مختلف نقش داشته‌اند را مورد توجه قرار می‌دهد. تلاش‌های اساسی در EAGLES حول

[1] Grammar Formalism

[2] Spoken Language

[3] یک استاندارد de facto یک عادت، رسم، قرارداد، محصول یا سیستمی است که از طریق پذیرش عمومی یا نیروهای تجاری جایگاه مستحکمی به دست می‌آورد. de facto یک عبارت لاتینی به معنای "برگرفته شده از واقعیت" یا " حاصل از تجربه و عمل" می‌باشد.

محورهای زیر در کارگروه‌های مختلف دنبال می‌گردد:

- تشخیص حوزه‌هایی به اندازه کافی توسعه یافتن برای استانداردهای کوتاه مدت در برابر حوزه‌هایی که هنوز به تحقیقات بنیادی و توسعه نیازمندند.
- ارزش‌گذاری و کشف حوزه‌هایی مورد توافق همگان در منابع زبانی فرمالیسم‌ها و سیستم‌های گفتاری موجود
- پیشنهاد مشخصات مشترک برای پدیده‌های بنیادی، توصیه عملیات مناسب و موثر و متدولوژی‌های استادندارد در مواردی که توافق همگانی می‌تواند به وجود آید.
- ارائه راهنماهایی برای نمایش ویژگی‌های اساسی و برای نمایش منابع و...
- مطالعات امکان‌سنجی برای حوزه‌های کمتر رشد یافته
- پیشنهاد مراحل فرآیندهایی که منجر به ساخت منابع چند زبانه با قابلیت استفاده مجدد می‌شود و شرح ابزارها و متدولوژی‌های ارزیابی آنها.

1-5. ISLE (استانداردهای بین‌المللی برای مهندسی زبان)[1]

ISLE (استانداردهای بین‌المللی برای مهندسی زبان) هم نام یک پروژه و هم نام یک مجموعه از فعالیت‌های هم راستا در حوزه تکنولوژی زبان طبیعی (HLT) می‌باشد.

ISLE تحت حمایت و پشتیبانی EAGLES ، که توسعه و پیشرفت گسترده در ارائه تعدادی

[1] International Standards for language Engineering

استاندارد de fecto داشته است، می‌باشد.

1-5-1. فعالیت‌ها، ماموریت‌ها و خدمات ISLE

هدف از ISLE توسعه استانداردهای تکنولوژی زبان طبیعی در یک چهارچوب بین‌المللی، با مشارکت محققان بین‌المللی آمریکایی و اروپایی می‌باشد.

اهداف ISLE به طور کلی پشتیبانی پروژه‌های ملی پروژه‌های تکنولوژی زبان از طریق توسعه، گسترش و ارتقای استانداردهای de fecto برای تکنولوژی زبان طبیعی و راهنماهایی برای منابع زبانی، ابزارها و محصولات می‌باشد.

ISLE سه حوزه را به عنوان هدف در نظر دارد: واژگان‌های چند زبان، تعامل طبیعی multimodality (NIMM) و ارزیابی سیستم‌های تکنولوژی زبان طبیعی

این حوزه‌های نه فقط به خاطر ارتباط شان با نیازهای فعلی بلکه به خاطر اهمیت بلند مدتشان انتخاب شده‌اند. برای واژگان‌های محاسباتی چند زبانه ISLE فعالیت زیر را انجام می‌دهد:

- گسترش کار EAGLES در واژگان‌های معنایی که مورد نیاز پیوندهای چندزبانی می‌باشد.
- استانداردهای طراحی برای واژگان‌های چندزبانه
- توسعه ابزارهای اولیه برای پیاده‌سازی راهنماها و استانداردهای واژگان
- ساخت واژگانی نمونه EAGLES-Conformant بسیار بار کیفیت وبرچسب‌گذاری پیکره‌های با کیفیت بالابرای اهداف اعتبارسنجی

- توسعه روال‌های ارزیابی استاندارد شده برای واژگان‌ها

برای NIMM که یک حوزه با تحول‌پذیر زیاد و نیازمنداستانداردسازی سریع می‌باشد، ISLE راهنماهایی برای موارد زیر ارائه می‌دهد:

- ساخت منابع داده‌ای برای NIMM
- نشان‌گذاری تفسیر کننده (تفسیری) داده‌های NIMM ، شامل محاورات گفتاری در زمینه NIMM
- نشان‌گذاری پدیده‌های گفتمان[1]

برای ارزیابی ISLE کارهای زیر را انجام می‌دهد:

- مدل‌های کیفیت برای سیستم‌های ترجمه ماشینی
- حفظ کیفیت راهنماهای قبلی در چهارچوب مبتنی بر ISO (ISO 9126, ISO 14598)

تعامل بسیار خوبی بین گروه‌های فوق در موضوعاتی که مورد توجه و علاقه بیش از یک گروه باشد، وجوددارد. در نتیجه این تعامل توافق گسترده‌ای صورت می‌پذیرد.

[1] Discourse phenomena

1-5-2. متدولوژی فعالیت ISLE

سه گروه ذکر شده و زیرگروه‌های آنها بر طبق متدولوی EAGLES و با همکاری خبره‌هایی از اروپا و آمریکا تحت یک چهارچوب هماهنگ فعالیت می‌کنند و کار را پیش می‌برند.

1-6. LDC (کنسرسیوم داده‌های زبانی)[1]

1-6-1. درباره LDC

کنسرسیوم داده‌های زبانی یک کنسرسیوم از دانشگاه‌ها، شرکت‌ها و آزمایشگاه‌های تحقیقاتی دولتی است. LDC در تولید، جمع‌آوری و توزیع دادگان‌های متنی وگفتاری، واژگان‌ها و دیگر منابع با اهداف توسعه‌ای و تحقیقاتی فعالیت می‌کند. دانشگاه پنسیلوانیا میزبان LDC می‌باشد. LDC در سال 1992 با امتیازی از طرف آژانس پروژه‌های تحقیقاتی پیشرفته (ARPA)[2] تاسیس شد. اگرچه LDC ابتدا در دانشگاه پنسیلوانیا (Pennsylvania) تاسیس شد ولی این کنسرسیوم در حال حاضر شامل پیش از 100 شرکت، دانشگاه و آژانس‌های دولتی میباشد. از زمان تاسیس، داده‌های مختلف از 197 موسسه عضو و 485 موسسه غیرعضو تحویل LDC شده است.

[1] Language Data Consortium

[2] Advanced Research Project Agency

1-6-2. فعالیت، ماموریت و خدمات LDC

مسائل و مشکلات جمع‌آوری، پردازش و برچسب‌گذاری کمیت‌های مورد نیاز داده‌های زبانی مسائل و مشکلات بسیار بزرگی برای شرکت‌ها می‌باشند. به هر ترتیب، دوباره کاری سرمایه‌گذاری، در تحقیقات که هنوز به نتیجه کامل نرسیده‌اند ناکارآمد خواهد بود. علاوه بر آن، فعالیت محققان دانشگاهی و شرکت‌های کوچک که مهمترین عامل نوآوری‌های تکنیکی هستند، ممکن است به دلیل مشکلات ذکر شده کاملاً متوقف شدند.

بنابراین وجود یک کنسرسیوم برای استفاده از خلاقیت‌های فردی و تلاش‌های انجام شده توسط شرکت‌های بزرگ می‌تواند راهی موثر برای حل مسائل فوق باشد. به این ترتیب از دوباره کاری تلاش‌های انجام شده جلوگیری به عمل می‌آید. LDC برای نیل به این هدف با در اختیار نهادن منابع اولیه و ابزارهای موجود فرصتی برای محققان شرکت‌های کوچکتر و آکادمی‌ها فراهم می‌سازد.

یک چنین نهادی همچنین می‌تواند با پیشرفت جدید بین‌المللی در ساختار تحقیقات تکنولوژی‌های زبانی به خوبی سازگار باشد. در سال اخیر فعالیت‌های مشارکتی در حوزه توسعه تکنوولوژی‌های زبانی افزایش یافته است. این مشارکت به اشکال متفاوت در کشورهای مختلف صورت می‌گیرد. در ژاپن آزمایشگاه‌های جدیدی تاسیس شده‌اند که نمونه بارز ان آزمایشگاه‌های ART Interpreting Telephony در کیوتو میباشد. در این آزمایشگاه‌ها محققانی از شرکت‌های مختلف روی پروژه‌ای مشترک کاری می‌کنند. در آلمان پروژه verbmobile یک فعالیت ترجمه گفتار به گفتار مشترک می‌باشد. هدف این پروژه هم فکری مشهورترین دانشگاه‌های آلمان و آزمایشگاه‌های صنعتی در آلمان به منظور ترکیب تکنولوژی‌های زبانی و گفتاری می‌باشد.

در آمریکا برنامه تکنولوژی زبان طبیعی ARPA برای استفاده موثر و سیستماتیک روشی به نام "Commontask" را در پیش گرفته است. در این روش هر پروژه را با مشخص کردن یکTask، تعریف یک معیار ارزیابی کمی و فرمال و فراهم‌سازی یک دادگان مشترک بزرگ با هدف آموزش و آزمایش شروع می‌کنند. سپس هر فرد عضو در پروژه مراحل‌هایی را به صورت مشخص دنبال می‌کند و همه افراد درگیر در پروژه در زمان‌های مشخص برای مقایسه روش‌ها و نتایج (به همراه نتایج ارزیابی) همدیگر را ملاقات می‌کنند.

این روش که از سال 1986 مورد استفاده قرار گرفته منجر به توسعه سریع در حوزه‌های مختلف شده است. برای مثال نرخ خطای تشخیص کلمه در طول 6 سال گذشته در هر دو سال نصب شده است.

به طور مشابه، کارایی درک پیام‌های متنی و سیستم‌های بازیابی اطلاعات در دو معیار ارزیابی صحت[1] و بازخوانی[2] در هر سال بین 20% تا 50% افزایش داشته است. فعالیت‌های مشترک همچنین روشی موثر در ایجاد همکاری و تبادلات سازنده تکنیک‌ها و ایده‌ها بوده است.

تکنیک‌های حاصل از فعالیت‌های مشترک که با روش‌های مختلف در سه کشور نام برده شده به دلایل زیر کارآمد می‌باشند:

1- توجه به مسائل اساسی

2- توجه به توسعه زیرساخت‌های مشترک لازم

3- رشد و شکوفایی سریع روش‌های موفق

4- فراهم سازی فضایی مناسب برای بحث‌های علمی

[1] Precision

[2] Recall

5- جهت‌دهی راه‌حل‌هابه سوی ایجاد سیستم‌های یکپارچه

6- تشخیص و ارزش‌گذاری شاخص تکنیکی

7- فراهم‌سازی آژانس‌ها و سازمان‌های حامی برای حوزه‌های با پتانسیل نتیجه‌دهی

2. زیرساخت‌ها و چهارچوب‌ها

2-1. مقدمه

در حوزه پردازش زبان طبیعی، زبانشناسی محاسباتی و مهندسی زبان وجود زیرساخت و معماری مناسب می‌تواند منجر به تسهیل استفاده مجدد از محصولات و نتایج فعالیت‌های مختلف گردد. حوزه‌های مذکور با متن سر و کار دارند و چون طبیعت متن به عنوان داده‌ها و اطلاعات با طبیعت دیگر داده‌ها و اطلاعات کاملا متفاوت می‌باشد، استفاده‌پذیری مجدد از نتایج و محصولات حاصل از کاربردها و تحقیقات در این حوزه‌ها در صورت نبود یک زیرساخت و معماری مشترک به سختی صورت می‌گیرد.

در این فصل به چند زیرساخت و معماری مهم که با هدف فوق ایجاد شده‌اند، می‌پردازیم. بعضی از این موارد چون با هدف کلی‌تری ایجاد شده‌اند قابلیت‌های بیشتری ارائه می‌کنند و برخی نیز همچون یک زیرساخت و معماری درون سازمانی هستند که علاقه به استفاده از آن‌ها کمتر خواهد بود. ولی در هر صورت همین زیرساخت‌های و معماری‌های درون سازمانی ایده‌های بسیار مناسبی برای ایجاد موارد مشابه در زبان فارسی ارائه خواهد کرد.

2-2. GATE

GATE زیرساختی است برای توسعه و تهیه مولفه‌های نرم‌افزاری مخصوص پردازش زبان طبیعی. GATE از سه طریق به توسعه دهندگان و دانشمندان این حوزه کمک می‌کند:

1- با مشخص کردن یک معماری[1] یا ساختار سازمانی برای نرم‌افزارهای پردازش زبان

2- با فراهم کردن یک چهارچوب[2] یا کتابخانه کلاس[3] که معماری را پیاده‌سازی می‌کند می‌تواند برای پردازش زبان در کاربردهای گسترده‌ای استفاده شود.

3- با فراهم‌سازی یک محیط توسعه‌ای که بر روی یک چهارچوب متشکل از ابزارهای گرافیکی مناسب با مولفه‌های توسعه‌ای ساخته شده است.

این معماری از توسعه نرم‌افزاری مبتنی بر مولفه[4]، شی‌گرایی[5] و کد سیار[6] استفاده می‌کند.

چهارچوب و محیط توسعه‌ای با جاوا نوشته شده است و به صورت نرم افزار رایگان متن باز تحت لیسانس GNU قابل استفاده می‌باشد. GATE از یونیکد استفاده می‌کند و در زبانهای بسیاری مورد آزمایش قرار گرفته است.

GATE از سال 1995 در دانشگاه sheffield در حال توسعه بوده است و در پروژه‌های توسعه‌ای و تحقیقاتی زیادی استفاده شده است. نسخه یک GATE که در سال 1996 ارائه شد تحت لیسانس چند

---

[1] Architecture
[2] Framework
[3] Class library
[4] Component based
[5] Obejct oriented
[6] Mobile code

صد موسسه بود و در بعد وسیعی از زمینه‌های تحلیل زبانی شامل استخراج اطلاعات در زبان‌های انگلیسی، یونانی، اسپانیایی، سوئدی و آلمانی، ایتالیایی، فرانسوی، بلغاری و روسی و تعدادی از زبان‌های دیگر مورد استفاده قرار گرفت.

GATE می‌تواند به عنوان یک معماری نرم‌افزاری برای مهندسی زبان در نظر گرفته شود. معماری نرم‌افزاری در این‌جا، به معنای زیرساختی نرم‌افزاری است برای توسعه نرم‌افزار و محیط و چهارچوب توسعه‌ای.

برای این‌که مطلب فوق روشنتر شود تعریفی از مهندسی زبان ارائه می‌کنیم. مهندسی زبان (LE)[1] به صورت زیر تعریف می‌شود:

نظام یا فعالیت سیستم‌های نرم‌افزاری مهندی که وظایفی شامل پردازش زبان طبیعی را انجام می‌دهند. فرآیند ساخت و تولید خروجی، هر دو قابل اندازه‌گیری و قابل پیش‌بینی‌اند.

نتایج علمی مربوط در حوزه تحلیل زبان به طور کلی خروجی‌های زبانشناسی محاسباتی (CL)[2]، پردازش زبان طبیعی (NLP)[3] و هوش مصنوعی (AI)[4] هستند. برخلاف این سه نظام گفته شده (AL,NLP, CL)، مهندسی زبان به عنوان یک نظام مهندسی قابلیت پیش‌بینی را هم در فرآیند ساخت نرم‌افزارهای مبتنی بر مهندسی زبان شامل می‌شود و هم کارآیی آن نرم‌افزارها را بعد از کامل شدن دارا می‌باشد.

در فعالیت‌های علمی در حوزه پردازش زبان طبیعی و زبانشناسی محاسباتی نقش GATE پشتیبانی

[1] Language Engineering
[2] Computational Linguistics
[3] Natural Language Processing
[4] Artificial Intelligence

از نتایج تجربی می‌باشد. در این مرحله ویژگی مهم GATE پشتیبانی از حصول خودکار نتـابج و کـاهش سربار تحقیقات می‌باشد.

2-3. ULMA (معماری مدیریت اطلاعات غیر ساخت‌یافته)[1]

2-3-1. خواستگاه (Motivation)

اطلاعات غیرساخت یافته بیشترین منابع اطلاعاتی در حال رشد برای شرکت افراد و دولت‌ها می‌باشد. و بسا نمونه‌ای از این اطلاعات است. حجم اطلاعات در سطح دنیا و اشکال مختلف مـتن، صـوت و ویـدئو موید این نظر است. محتوای با ارزش در این مجموعه‌های بزرگ اطلاعـات غیـر سـاخت یافتـه در حجـم زیادی از نویزها پنهان میباشد. جستجو برای آنچه کار نیاز دارد یا انجام عمـل داده‌کـاوی در ایـن منـابع اطلاعات غیرساخت یافته خود مسأله و چالشی جدید است.

یک کاربرد مدیریت اطلاعات غیرساخت یافته (UIM) می‌تواند به صورت کلی تحـت عنـوان سیسـتم نرم‌افزاری توصیف گردد که حجم بزرگی از اطلاعات غیرساخت یافته (متن، صوت، ویدئو، تصـویر و....) را تحلیل کرده تا دانش مناسب را برای کاربر نهایی کشف، سازماندهی و استخراج کند. یک نمونه از چنـین کاربردی پردازش میلیون‌ها پرونده پزشکی برای کشف تاثیرات داروها می‌باشد. نمونه دیگر به هر ترتیـب

[1] Unstructured Information Management Architecture

داده‌های غیرساخت یافته باید برای تفسیر، تشخیص و یافتن مفاهیم مورد علاقه مانند مداخل اسمی(اشخاص، سازمان‌ها، اماکن، ابزارها، محصولات و...) که به صور مشخص برچسب‌گذاری و نشان‌گذاری نشده‌اند، مورد تحلیل قرار گیرند.

تحلیل‌های پیچیده‌تر تشخیص چیزهایی شبیه، عقایدریال نارضایتی‌ها، تهدیدها و واقعیت‌ها در اطلاعات غیرساخت یافته می‌باشد. همچنین تشخیص ارتباطات بین اشیاء نیز بسیار مهم است. لیست مفاهیم مهم در کاربردها که نیاز به کشف و بازیابی داشته باشند در اطلاعات غیرساخت یافته زیاد، مختلف و اغلب وابسته به حوزه و دامنه می‌باشند. تحلیل‌کننده‌های مختلف ممکن است بخش‌های مختلفی از فرایند تحلیل کلی را انجام دهند.

این تحلیل‌کننده‌ها با هم در ارتباط باشد و بتوانند به سادگی با هم ترکیب شوند تا امکان کاربردهای توسعه یافته UIM (مدیریت اطلاعات غیرساخت یافته) را موجب گردند.

نتایج تحلیل‌ها برای تبدیل به شکل ساخت یافته استفاده می‌شوند به صورتی که پردازش داده‌ها و تکنولوژی‌ها جستجو شبیه موتورهای جستجو، بتوانند موتورهای بانک‌های اطلاعاتی یا موتورهای OLAP (پردازش تحلیلی برخط ویا داده‌کاوی)[1] بتوانند محتوای کشف شده جدید را در پاسخ کارآمد به پرسمان‌ها و نیازهای کاربران ارائه دهند.

در تحلیل محتوای غیرساخته، کاربردهای UIM از تکنولوژی‌های تحلیلی مختلفی استفاده می‌شود که شامل موارد زیر است:

- پردازش زبان طبیعی به صورت آماری یا مبتنی بر قاعده

[1] On-Line Analytical Processing

- بازیابی اطلاعات
- یادگیری ماشین
- هستان شناسی‌ها
- استدلال خودکار[1]
- منابع دانش (مثال: CYC، WorldNet، FrameNet و...)

امکانات تحلیلی خاص با استفاده از این تکنولوژی‌ها مستقلاً با استفاده از تکنیک‌های واسطها و سکوهای[2] مختلف توسعه می‌یابند.

گپ بین جهان غیرساخت‌یافته و جهان ساخت یافته با ترکیب و توسعه این امکانات تحلیلی پر می‌شود. این یکپارچه‌سازی اغلب مسئله‌ای پرهزینه میباشد.

UIMA یک معماری و زیرساخت نرم‌افزاری است که کمک می‌کند این گپ بین جهان غیرساخت یافته و جهان ساخت یافته پر شود. UIMA از ساخت، کشف، ترکیب و توسعه دامنه‌ای گسترده از امکانات تحلیلی و برقراری ارتباط آن‌هابا سرویس‌های اطلاعات ساخت یافته پشتیبانی می‌کند.

UIMA برای تیم‌های توسعه امکانی فراهم می‌کند تا مهارت‌ها را با بخش‌های مختلف یک راه‌حل منطبق سازند و همچنین کمک می‌کند تا یکپارچه‌سازی سریع تکنولوژی‌ها و سکوها به اشکال مختلف و مطلوب صورت گیرد.

[1] Automatic Reasoning
[2] Platform

2-3-2. UIMA چیست؟

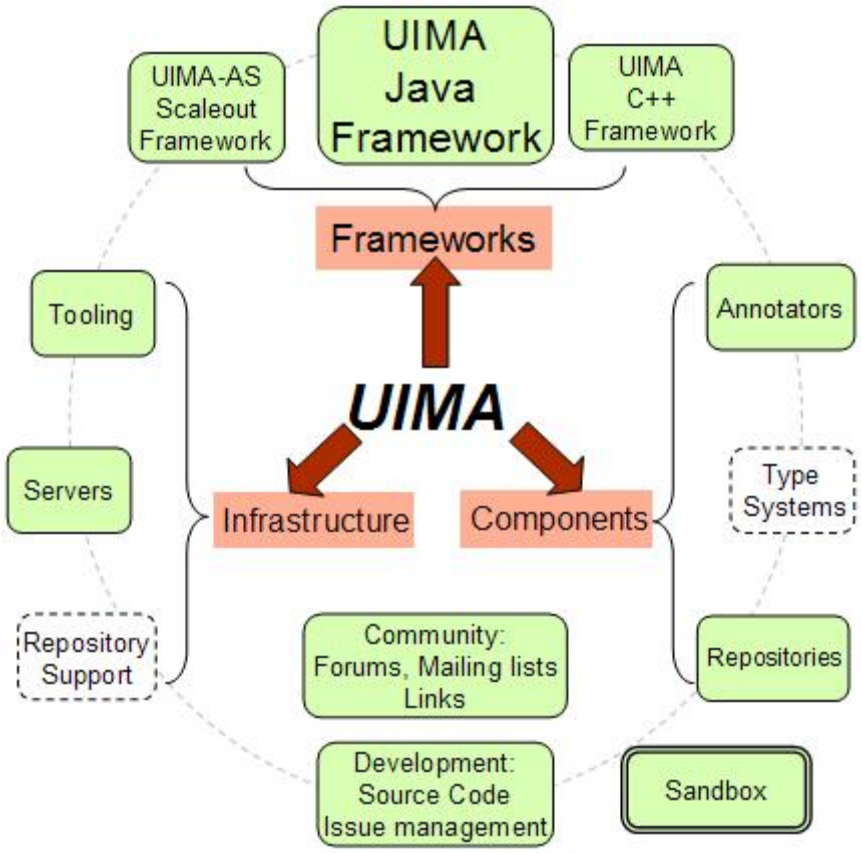
کاربردهای مدیریت اطلاعات غیرساخت یافته سیستم‌های نرم‌افزاری هستند که حجم زیادی از اطلاعات غیرساخت یافته به منظور کشف و استخراج دانش مورد استفاده کاربر تحلیل می‌کند.

UIMA این امکان را فراهم می‌کند که کاربردها به مولفه‌هایی تقسیم شوند برای مثال تشخیص زبان ⇐ بخش‌بندی زبان ⇐ تشخیص کردان جملات ⇐ تشخیص مداخل اسمی (شخص مکان و...). هر مولفه واسط‌هایی را با استفاده از معماری پیاده‌سازی می‌کند و فرادادهایی[1] به صورت خودتعریف[2] توسط فایل‌های توصیفی XML فراهم می‌آورد. معماری مولفه‌ها و جریان دادهٔ بین آن را مدیریت می‌کند. مولفه‌ها به زبان جاوا یا C++ نوشته شده‌اند و داده‌ها که بین مولفه‌ها جریان می‌یابند برای نگاشت کارآمد بین این زبان‌ها طراحی شده‌اند.

علاوه بر آن، UIMA امکاناتی را فراهم میسازد تا مولفه‌ها به عنوان سرویس‌های تحت شبکه قابل استفاه باشند، مولفه‌ها قابل استفاهد در پردازش داده‌های با حجم خیلی بزرگ، در پردازش‌های Pipeline در یک کلاستر باشند. Apache UIMA یک پیاده‌سازی منبع باز از مشخصات UIMA میباشد (این مشخصات به طور همزمان در یک کمیته تکنیکی در OASIS، یک سازمان استانداردسازی در طراحی می‌باشد) در UIMA به سه مورد: 1- زیرساخت، 2- مولفه‌ها 3- زیرساختارها برمی‌خوریم. (شکل زیر، نقاط خط‌چین محل افزودن موارد جدید در آینده می‌باشد)

[1] Metadata

[2] Self-describing



شکل2-1. UILM، چهارچوب‌ها، مولفه‌ها و زیرساختار

زیرساخت‌ها مولفه‌ها را اجرا می‌کند و به دو زبان Java و C++ موجود می‌باشند. زیرساخت جاوا اجرای مولفه‌های جاوایی و غیرجاوایی (با استفاده از زیرساخت C++) را پشتیبانی می‌کند. زیرساخت C++ در کنار پشتیبانی نشانه‌گذارهای نوشته شده در C/C++، نشانه‌گذاری‌های[1] Perl، Python و TCL را نیز پشتیبانی می‌کند.

زیرساخت UIMA-AS Scoleout یک زیرساخت نصب شده روی زیرساخت جاوا است که از یک

[1] Annotator

امکان Scoleout انعطاف‌پذیر مبتنی بر (java messaging service) و Active Ma پشتیبانی می‌کند.

زیرساخت‌ها از پیکربندی و اجرای Pipeline مولفه‌های نشانه‌گذار پشتیبانی می‌کنند. این مولفه‌ها کار واقعی تحلیل اطلاعات غیرساخت یافته را انجام می‌دهند. کاربران می‌توانند نشان‌گذاری خود را بنویسند یا از نشان‌گذاری‌های از قبل تهیه شده را و پیکره‌بندی و استفاده کنند. بعضی از این نشان‌گذارها در حال حاضر موجودند و برخی دیگر در منابع اینترنتی قابل دسترسی‌اند. زیرساختار اضافه شده از مولفه‌ها شامل یک سرور ساده پشتیبانی می‌کند که می‌تواند تقاضاهای REST را دریافت کند و نتایج نشان‌گذاری را برای استفاده سرویس‌های دیگر وب بازگرداند.

Sandbox محلی است که ایده‌های جدید برای مشارکت و همکاری در پروژه توسعه می‌یابد.

2-4. FEX و SNoW

2-4-1. FEX (یک زبان استخراج ویژگی‌های ارتباطی)[1]

FEX ابزاری است برای استخراج ویژگی‌های از متن، این ویژگی‌ها می‌توانند برای تولید نمونه‌هایی به کار روند که در نرم‌افزارهای یادگیری ماشین مانند Snow (بخش بعد) قابل استفاده باشند.

FEX یک مجموعه از داده‌های متنی (به عبارت یک پیکره) را با استفاده از یک مجموعه از دستورات

[1] A Relational Feature Extraction language

(اسکریپ) پردازش می‌کند. این دستورات تعیین می‌کنند چه ویژگی‌هایی باید تولید شوند. FEX واژگانی می‌سازدکه ویژگی‌های استخراج شده از متن را به یک عدد صحیح (integer) نگاشت می‌کند.

FEX به عنوان خروجی نمونه‌ها (لیست اعداد صحیح منطبق بر ویژگی‌ها در واژگان) را به عنوان خروجی تولید می‌کند. اگر در Script دستور مناسب قرار داده شده باشد، نمونه‌ها (علاوه بر اعدادا) نیز خواهند داشت. این نمونه‌ها طوری تهیه شده‌اند که با Snow سازگار باشند.

FEX می‌تواند متن‌ها را در چندین فرمت مختلف شامل متن ساده، متن برچسب خورده با برچسب‌های اجزا واژگانی کلام و فرمت ستونی تولید کند. این اشکال مختلف خروجی داده‌های متنی را در مراحل مختلف پردازش نشان میدهد. فرمت ستونی امکان نشان‌گذاری با مجموعه‌ای غنی‌تر از نشانه‌هارا نسبت به فرمت متنی ساده و متن برچسب‌خورده با برچسب‌های اجزای واژگانی فراهم می‌سازد.

FEX در سیستم‌های یونیکس و لینوکس قابل استفاده است و برای استفاده از آن در محیط ویندوز به cygwin نیاز می‌باشد.

۲-۴-۲. SNOW[1]

چهارچوب‌های معماری یادگیری Snow یک شبه اسپارس از توابع خطی است که روی یک فضای ویژگی از قبل تعریف شده یا در حال یادگیری و رشد قرار دارد. Snow به خاص مناسب یادگیری در حوزه‌ای است که در آن تعداد ویژگی‌های درگیر در فرآیند تصمیم‌گیری بسیار زیاد می‌باشند ولی احتمالاً

[1] Sparse Network of Winnows

از قبل مشخص نیستند. برخی از مشخصات این چهارچوب یادگیری به این قرار است: واحدهایی که به صورت اسپارس با هم در ارتباط‌اند، اختصاص ویژگی‌ها و پیوندها به روش مبتنی[1] بر داده، وابستگی محاسباتی به تعداد ویژگی‌های فعال به جای وابستگی به همه ویژگی‌ها و بهره‌گیری از یک قاعده به روزرسانی کارآمد ویژگی‌ها.

SNOW به طور موفق در فعالیت‌های مختلف یادگیری در حوزه‌های از قبیل زبان طبیعی، بیوانفورماتیک و پردازش بصری مورد استفاده قرار گرفته است.

چندین قاعده به روزرسانی ممکن است در Snow استفاده شود:

Winnow، Percoptron کلاسیک، نسخه‌های مختلفی از یک winnow تنظیم شده و یا یک Perception تنظیم شده، الگوریتم‌های رگراسیون بر مبنای کاهش گرادیان[2] و الگوریتم Naïve Bayes.

SNOW را می‌توان به عنوان یک رده‌بند[3] که منظوره چندکلاسه در نظر گرفته و در نسخه اخیر آن امکان چندکلاسه بودن به سیاست یادگیری یکی در برابر هم[4] افزوده شده‌است.

علاوه بر برچسب کلاس پیش‌بینی شده، SNOW می‌تواند یک درجه اطمینان پیش‌بینی[5] به هر برچسب نسبت دهد. این مقدار می‌تواند به عنوان تابعی از فاصله بین فعال‌سازی خود هدف و حد آستانه محاسبه گردد.

SNOW همچنین باید به عنوان یک چهارچوب معماری یادگیری در نظر گرفته شود. کاربر یک

[1] Data Driven
[2] Gradient Descent
[3] Classifier
[4] One-vs-all
[5] Prediction confidence

معماری در چهارچوب طراحی می‌کند. یعنی علاوه بر تعریف تعدادی نمایش کلاس برای یادگیری، امکان تعریف پارامترهای بیشتری از معماری (شامل قواعد به روزرسانی و پارامترهای آن‌ها، پارامترهای تنظیم، سیاست‌های یادگیری و...) نیز تعریف شود.

در مستندات SNOW از معماری تعریف شده توسط کاربر و همه داده‌های جمع‌آوری شده در آن به طور کلی تحت عنوان شبکه یاد می‌شود. در شبکه، برچسب کلاس‌ها هدف‌ها (Targets) نامیده می‌شوند. به عنوان توابع خطی اسپارس روی ویژگی‌های ورودی آموزش داده می‌شوند. منطور از اسپارس در متن این است که هر هدف ممکن است به عنوان یک تابع از زیرمجموعه‌ای از همه ویژگی‌ها در فضای ویژگی و به صورت مبتنی بر داده آموزش داده شود. این زیرمجموعه به صورت جزئی توسط مجموعه پارامترها تعریف توسط کاربر کنترل می‌گردد. هنگامی که SNOW را به طور ساده به عنوان یک سیستم رده‌بندی می‌بینیم. ورودی معمول سیستم حجمی از نمونه‌های برچسب خورده شامل ویژگی‌های بولین یا عددی در فرمت خاص تعریف شده برای آن می‌باشد. مهمترین قواعد به روز رسانی تعریف شده برای SNOW به شرح زیر است:

- § اندازه ورودی متغیر
- § روش‌های خالص کردن[1] ویژگی‌ها
- § مکانیسم درجه اطمینان پیش‌بینی
- § تخصیص مبتنی بر داده ویژگی‌ها و پیوندها
- § مکانیسم پشتیبان تصمیم[2]

[1] Pruning

[2] decision support mechanism

§ یکپارچه‌سازی با یک مکانیسم استخراج ارتباطی ویژگی (FEX) که قابلیت ترکیب منابع اطلاعات خارجی (ویژگی‌ها) را به صورت قابل انعطافی فراهم می‌سازد.

SNOW و FEX در گروه محاسبات شناختی دانشگاه الینویز توسعه یافته است. از پروژه‌های مرتبط با SNOW از پروژه سیستم‌های پاسخگوی سوالات و از ابزارهای مرتبط با SNOW از خلاصه یاب آماری SNOW نام برده می‌شود.

2-5. OPEN NLP

2-5-1. شرح OPEN NLP

OPEN NLP سازمانی است برای پروژه‌های متن‌باز مرتبط با پردازش زبان طبیعی. نقش اول OPEN NLP جذب و ایجاد امکان مشارکت محققان و توسعه‌دهندگان در چنین پروژه‌هایی است. پروژه‌های زیادی در OPEN NLP تولید انجام شده است که برخی از آن را نام می‌بریم : Maxnet، کتابخانه OPENNLPCCG، Word Freak، AGTK، Arithmetic Coding، Weka، Ngram statistics و...

همچنین OPENNLP میزبان ابزارهای مختلف جاوایی مخصوص پردازش زبانی طبیعی است که تشخیص جملات، واحدسازی، و coreference برچسب‌گذاری اجزای واژگانی کلام، تقطیع[1]، تجزیه[2]،

[1] Chunking
[2] Parsing

تشخیص مداخل اسمی انجام می‌دهند.

این ابزارها به خودی خود مفید نمی‌باشند اما می‌توانند در ترکیب با دیگر نرم‌افزارهای پردازش متن مفید واقع شوند.

2-5-2. مدل‌ها در OPEN NLP

مدل‌های زیادی برای مولفه‌های مختلف آموزشی داه شده‌اند و مورد نیاز میباشند مگر اینکه کاربر خود بخواهد مدل را با استفاده از داده‌های خود آموزش دهد. این مدل‌ها آماده برای همگان قابل دسترس می‌باشد. مدل‌ها عمدتاً بزرگ می‌باشد خصوصاً مدل تجزیه‌گر[1] که از بقیه بزرگ‌تر می‌باشد و مدل‌های مختلف در چند زبان تهیه شده‌اند که در زیر می‌آید:

- زبان انگلیسی
  - تقطیع‌گر[2]
  - Co reference
  - تشخیص مداخل اسمی
  - تجزیه‌گر
  - تشخیص جمله

[1] Parser
[2] Chunker

- واحدساز[1]

- زبان اسپانیایی
  - برچسب‌گذار اجزای واژگانی کلام
  - تشخیص جملات
  - واحدساز
- زبان آلمانی
  - برچسب‌گذار اجزای واژگانی کلام
  - تشخیص جملات
  - واحدساز
- زبان Thai
  - تشخیص جملات
  - واحدساز
  - برچسب‌گذار اجزای واژگانی کلام

3-5-2. اجرای ابزارها

برای اجرای ابزارها نیاز است که مدل‌های فوق را در دسترس داشته باشیم. ابزارها، مدل‌های فوق را به

[1] Tokenizer

عنوان پارامتر دریافت می‌کند. اکثر ابزارها فقط با یکی از مدل‌های سروکار دارند ولی تجزیه‌گر به بیش از یک مدل نیازمند است. این ابزارها همان تشخیص دهنده جملات، واحدساز، برچسب گذار اجزای واژگانی کلام، تقطیع‌گر، تشخیص مداخل اسمی، تجزیه‌گر و Co reference هستند که مدل‌های آن در بالا ذکـر گردید.

3. ابزارهای تحلیل زبان

3-1. مقدمه

در این فصل ابزارهای که در طول دوره گسترش کاربردها و تحقیقات در حوزه زبان مهم بوده و مفید واقع شده‌اند مورد بررسی اجمالی قرار می‌گیرد. امکان دسته بندی مشخص این ابزارها عملا ممکن نمی‌باشد. زیرا ابزارهای مختلف امکانات مختلف کابردی و تحلیل در سطوح مختلف ارائه می‌کنند و برخی نیز این گونه نیست و تنها در یک کاربرد و سطح خاص عمل می‌کنند ولی در همان کاربرد و سطح نیز چندین روش، مدل یا الگوریتم مختلف را به کار می‌گیرند. بنابراین به نظر نمی‌رسد دسته مناسبی بتوان برای این ابزارهای به کار گرفت. به همین دلیل در این‌جا سعی خواهد شد ابزارها بر اساس امکاناتی که ارائه می‌دهند و به نظر می‌رسد قابلیت بیشتری در استفاده در کاربردهای دیگر داشته باشند مورد بررسی قرار گیرند.

اکثر این ابزارها حتی ابزارهای مبتنی بر روش‌ها کاملا آماری در حوزه زبان انگلیسی یا زبان‌های اروپایی می‌باشد و برای استفاده از آن‌ها در زبان‌هایی مثل زبان فارسی باید معمولا مدل‌های جدید برای این زبان‌ها ایجاد کرد تا بتوان از هسته اصلی ابزارهای استفاده کرد. ولی به هر ترتیب روش‌ها و الگوریتم‌ها مورد استفاده و نتایج و محصولات به دست آمده روشن‌گر فعالیت و کارهای جدید در حوزه

زبان فارسی خواهد بود.

2-3. CMU-SLM (مجموعه ابزارهای مدل‌سازی آماری زبان دانشگاه کمبریج)[1]

نسخه یک مجموعه ابزارهای مدل‌سازی آماری زبان دانشگاه کمبریج در سال 1994 در دانشگاه مذکور نوشته شد. مجموعه ابزار CMU-SLM مجموعه‌ای از ابزارهای نرم‌افزاری تحت لینوکس است که به منظور ایجاد مدل‌های زبانی برای محققان طراحی شده است.

برخی از ابزارهای CMU-SLM داده‌هایی متنی به منظور تهیه موارد زیر پردازش می‌کنند:

- لیست فراوانی کلمات و مجموعه کلمات
- دنباله‌های دوتایی و سه تایی کلمات و تعداد تکرار آن‌ها
- دنباله‌های دوتایی و سه تایی کلمات یک مجموعه کلمه خاص و تعداد آنها
- آمارهای مرتبط با دنباله‌های دوتایی و سه‌تایی
- مدل‌های مختلف زبانی دنباله‌های دوتایی و سه‌تایی برگشتی [2]
- برخی دیگر از ابزارها مدل‌های زبانی به دست آمده را برای محاسبه موارد زیر به کار می‌برند:
- سرگشتگی[1]

[1] Cambridge Statistical Language Modeling Toolkit

[2] Backoff bigram and trigram

- نرخ کلمات خارج از مجموعه لغات[۲]
- Bigrams and trigram hit ratios
- توزیع موارد برگشتی
- نشانگذاری داده‌های آزمایش با آمارهای زبانی به دست آمده[۳]

3-3. SRILM (ابزار مدل‌سازی زبان SRI)[۴]

SRILM ابزاری است برای ساخت و به کارگیری مدل‌های آماری زبان که خصوصاً در تشخیص گفتار برچسب‌گذاری و قسمت‌بندی[۵] آماری و ترجمه ماشینی کاربرد دارد. این ابزار در آزمایشگاه تحقیقات و تکنولوژی گفتار SRI از سال 1995 در حال توسعه می‌باشد.

SRILM شامل مولفه‌های زیر می‌باشد:

- مجموعه‌ای از کتابخانه‌های کلاس C++ که مدل‌های زبانی را پیاده‌سازی می‌کنند که از ساختمان داده‌ها و توابع کارآمد و مفیدی پشتیبانی می‌کند.
- مجموعه‌ای از برنامه‌های قابل اجرا که روی این کتابخانه ساخته‌اند و برای انجام فعالیت‌هایی از قبیل آموزش مدل‌های زبانی و آزمایش آن‌ها روی داده، برچسب‌گذاری یا قسمت‌بندی متون و ...

[1] Perplexity
[2] Out of vocabulary rate
[3] Language score
[4] The SRI Language Modeling Tools
[5] Segmentation

به کار می‌روند.

- مجموعه‌ای از اسکریپت‌های متفرقه که انجام فعالیت‌های موردی را ممکن می‌سازند.

SRILM با هدف نیل به یک ابزار مدل‌سازی زبان که موارد زیر را برآورده سازد تهیه شده است:

- پیاده‌سازی دقیق و کارآمد الگوریتم موجود و کارای مدل زبانی برای پشتیبانی از توسعه سیستم‌های رقابتی
- انعطاف‌پذیری گسترش‌پذیری، به طوری که تحقیقات در انواع جدید مدل‌های زبانی را ممکن سازد و در عین حال قابلیت استفاده از مولفه‌های موجود ممکن باشد.
- طراحی نرم‌افزاری مناسب و منطقی[1] که هم یک واسط برنامه‌نویسی کاربردی (API) را فراهم سازد و هم ابزاری راحت متشکل از دستورات ساخت و آزمایش مدل‌های زبانی باشد.

3-4. Ling Pipe

3-4-1. درباره Ling Pipe

Ling Pipe ابزار پردازش زبان طبیعی است که به زبان جاوا نوشته شده است. Ling Pipe واحدسازی،

[1] Rational

تشخیص جمله، تشخیص مداخل اسمی، co reference، رده‌بندی[1]، خوشه‌بندی، برچسب‌گذاری اجزای واژگانی کلام، تقطیع عمومی[2]، انطباق فازی فرهنگ لغت[3] را انجام می‌دهد. همانطور که مشاهده می‌شود Ling Pipe دامنه گسترده‌ای از کاربردها را پشتیبانی می‌کند.

3-4-2. ابزارهای استخراج اطلاعات و داده‌کاوی در Ling Pipe:

ابزارهای استخراج اطلاعات و داده‌کاوی در Ling Pipe به شرح زیر است:

- Track mentions of entity (افراد یا پروتیین‌ها)
- پیوند مداخل ذکر شده با مداخل بانک اطلاعاتی
- کشف ارتباطات بین مداخل و فعالیت‌ها (Actions)
- رده‌بندی بخش‌های متن بر اساس زبان، رمزگذاری حروف، گونه، موضوع یا عقاید و احساسات (Sentiment)
- تصحیح خطاها با توجه به یک مجموعه از متون
- خوشه‌بندی اسناد با موضوعات غیرصریح و کشف روندهای (trends) مهم در طول زمان
- فراهم‌سازی برچسب‌گذاری اجزای واژگانی کلام و تقطیع عبارت

[1] Classification
[2] General chucking
[3] Fuzzy dictionary matching

3-4-3. معماری Ling Pipe

معماری Ling Pipe با اهداف کارایی، قابلیت استفاده با حجم‌های مختلف داده، قابلیت استفاده مجدد و robustness طراحی شده است، ویژگی مهم این معماری به صورت زیر است:

- APIهای جاوایی به صورت کدباز
- مدل‌های چندزبانه[1]، چنددامنه[2]، چندگونه[3]
- N تا از بهترین خروجی‌ها با تخمین آماری درجه اطمینان
- آموزش برخط[4] یعنی به صورت Learn-a-little-tag-a-little
- ورودی / خروجی حساس به رمزگذاری[5] کاراکترها

3-5. TiMBL (Tilburg Memory-Based Learner)

یادگیری مبتی بر حافظه (MBL)[6] یک روش ساده و مستحکم یادگیری ماشین است که در حوزه

[1] Multilingual
[2] Multi-domain
[3] Multi-genre
[4] Online
[5] Encoding sensitive
[6] Memory Based Learner

وسیعی از فعالیت‌های پردازش زبان طبیعی قابل استفاده می‌باشد. یادگیری مبتنی بر حافظه برگرفته شده از روش نزدیکترین K همسایه (K-NN)[1] برای طبقه‌بندی می‌باشند. روش نزدیکترین همسابه به عنوان یکی از الگوریتم‌های قوی و شناخته شده طبقه‌بندی نمونه‌ها برای داده‌های عددی می‌باشند. در فعالیت‌های معمول یادگیری پردازش زبان طبیعی نیز تمرکز بر روی داده‌ها گسسته ، تعداد بسیار زیاد نمونه‌ها و تعداد زیاد خاصیت‌ها با ارتباطات مختلف می‌باشد. سرعت طبقه‌بندی مسئله‌ای اساسی در کاربردهای عملی یادگیری مبتنی بر حافظه می‌باشد. این محدودیت‌ها نیاز به ساختمان داده‌های خاص و بهینه‌سازی سرعت طبقه‌بند K-NN دارد. روش طراحان TiMBL در تولید یک معماری حاصل شده است که سازمان فایل‌های معمولی که در پیاده‌سازی‌های K-NN استفاده می‌شود، در قالب یک ساختار درخت تصمیم فشرده سازند. در حالیکه این درخت تصمیم می‌تواند برای بازیابی دقیقاً نزدیکترین k همسایه استفاده می‌شود (همانطور که در الگوریتم IBT ، TiMBL استفاده می‌شود) ، همچنین می‌تواند به عنوان یک طبقه‌بندی درخت تعمیم پیمایش شود (روشی که به وسیله الگوریتم IGTREE انجام می‌شود)، به این صورت TiMBL یکی از سریع‌ترین پیاده‌سازی‌های K-NN برای مقادیر گسسته می‌باشد.

ویژگی‌ها TiMBL به شرح زیر است:

- پیاده‌سازی سریع و مبتنی بر درخت تصمیم الگوریتم طبقه‌بندی نزدیکترین k همسایه
- پیادسازی الگوریتم‌های IBI، IB2 ، IGTee ، TRIBL و TRIBL2.
- به کارگیری معیارهای شباهت: overlap، MRDM و Jeffrey Divergence ، Dot product ، Cosine

[1] K-Nearest Neighbor

- به کارگیری معیارهای وزن‌دهی ویژگی‌ها : بهره اطلاعاتی[1]، gain ratio، chi squared ، shared variance
- به کارگیری معیارهای وزن‌دهی فاصله: inverse ، inverse linear، exponential decay
- گزینه‌های زیادی برای بررسی مجموعه‌های نزدیکترین همسایه
- قابلیت به کارگیری به صورت server و API بسط داده شده
- قابلیت آزمون سریع leave-one-out و internal cross-validation
- مدیریت وزن‌دهی نمونه‌ها که توسط کاربر تعریف شده است.

3-6. WOPR

WOPR یک پیش‌بینی کننده کلمات و سازنده مدل زبانی مبتنی بر حافظه می‌باشد. WOPR یک طبقه‌بند با روش نزدیکترین k همسایه در TiMBL می‌باشد که امکانات پیش‌بینی کلمه و مدل‌سازی زبان را ارائه می‌کند. با آموزش آن بر روی یک پیکره متنی، WOPR می‌تواند کلمات جا افتاده را پیش‌بینی کند، سرگشتگی‌ها را در سطح کلمه و در سطح متن گزارش دهد و فرضیات[2] تصحیح املا[3] را تولید کند.

[1] Information Gain
[2] Hypothesis
[3] Spelling

ویژگی‌های WOPR به صورت زیر است:

- تولید مدل‌های زبانی
- انجام آزمایش مدل‌های زبانی روی متون جدید، گزارش دادن سرگشتگی ، توزیع‌های پیش‌بینی، اینتروپی‌های در سطح کلمه و سرگشتگی‌ها
- تولید خروجی (به صورت اختیاری) فایل‌های مدل زبانی با فرمت ARPA
- فیلتر کردن (به صورت اختیاری) خروجی برای نامزدهای تصحیح املا

3-7. پروژه XTAG

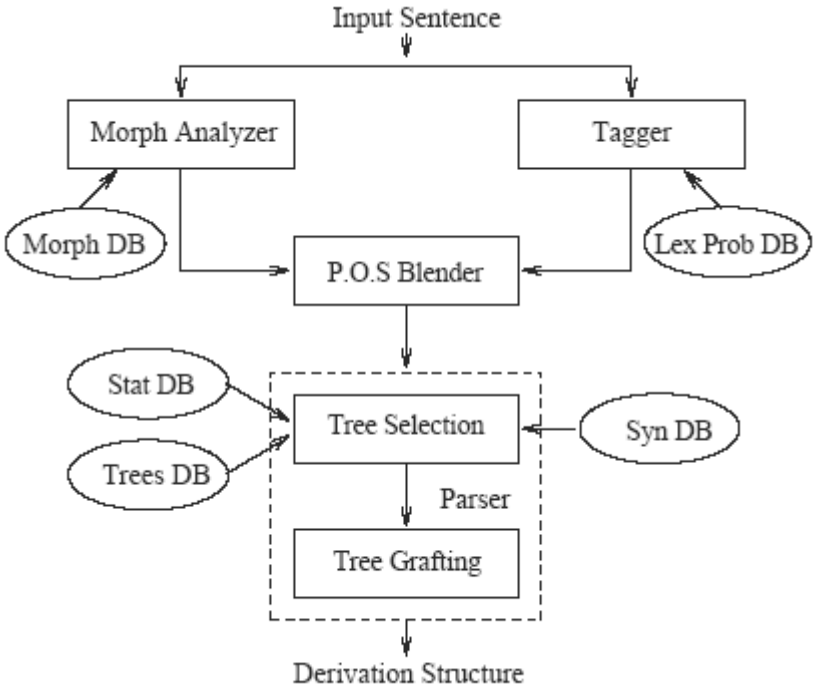
XTAG پروژه‌ای برای توسعه یک گرامر است که زبان انگلیسی را به صورت گسترده پوشش دهد. در این پروژه از یک فرمالیسم گرامر اتصال درختی لغت-ساخت (LTAG)[1] استفاه می‌شود. از XTAG همچنین می‌توان به عنوان یک سیستم توسعه گرامرهای اتصال درختی (TAG) استفاده کرد. XTAG شامل یک تجزیه‌گر، یک واسط توسعه و ایجاد گرامر و یک تحلیل‌گر ساختواژی می‌باشد.

تاکنون از نتایج این پروژه در ترجمه ماشینی استفاده شده است. به عنوان مثال در ترجمه ماشینی انگلیسی و کره‌ای از نتایج و محصولات XTAG استفاده کرده‌اند.

آقای دکتر فیلی نیز از نتایج پروژه XTAG برای یک مترجم ماشینی مبتنی بر ترجمه ماشینی انتقالی استفاده کرده است.

[1] Lexicalized Tree Adjoining Tagger

گرامر پروژه XTAG دارای 1227 درخت اولیه می‌باشد. شکل زیر جریان کلی سیستم را هنگامی که یک جمله تجزیه می‌شود نشان می‌دهد:



شکل3-1. شمایی از سیستم XTAG

در زیر اجزای سیستم را شرح می‌دهیم و مشخصات کلی آن‌ها را بیان می‌کنیم:

تحلیل‌گر ساختواژی و بانک پشتیبان آن (Morphological Analyzer and Morphology Database) : شامل 3170.000 نمونه تعریفی است که از بیش از 90000 ریشه مشتق شده‌اند. در بانک پشتیبان تحلیل‌گر ساختواژی مداخل بر اساس اشکال تجزیه شده شاخص‌گذاری شده‌اند و بنابراین قابلیت برگردادن ریشه برچسب اجزای واژگانی و اطلاعات تصریفی را ممکن می‌سازد.

برچسب‌گذار اجزای واژگانی کلام و بانک احتمالات واژگانی (POS Tagger and Lexical Probabilities Database): برچسب‌گذار، یک برچسب‌گذار بر اساس مدل Trigram (مدل مارکوف مرتبه دو) می‌باشد که قابلیت تولید بهترین n دنبالهٔ برچسب‌های اجزای واژگانی را دارد.

بانک نحوی (Syntactic Database) : بیش از 30.000 مدخل دارد. هر مدخل شامل شکل غیرتصریفی کلمه، برچسب اجزای واژگانی آن، لیست درخت‌ها و خانواده‌های درخت‌های مرتبط با آن کلمه و لیستی از ویژگی‌هایی که مشخصات واژگانی را نمایان می‌سازند.

بانک درخت (Tree Database): شامل 10045 درخت است که در 53 خانواده درخت و 221 درخت مجرد تقسیم شده‌اند. خانواده‌های درخت قالب‌های طبقه‌بندی‌های جزئی‌تر را نشان می‌دهند. درخت‌ها در یک خانواده به هم مرتبط بوده و با یک روش مشخص قابل تبدیل به یکدیگرند.

واسط X (X-Interface): این واسط امکان گرافیکی مناسبی برای ساخت و تصحیح فایل‌های درخت می‌باشد. امکان تنظیم پارامتری تجزیه توسط کاربر در این واسط موجود می‌باشد. پارامترهای کلام، امکانات ذخیره و بازیابی درخت‌های تجزیه شده و اولیه، امکان ترتیب دستی درخت‌ها از طریق الحاق یا جایگزینی با هدف توسعه گرامر، قابلیت انتخاب دستی برچسب‌های اجزای واژگانی کلام قبل از تجزیه را فراهم می‌سازند.

3-8. Multext: ابزارها و پیکره‌های متنی چندزبانه

Multext مجموعه از پروژه‌هارا در برمی‌گیرد که هدف آن ایجاد استانداردها و مشخصات[1] برای رمزگذاری و پردازش پیکره‌های زبانشناسی، و ایجاد ابزارها، پیکره‌ها و منابع زبانشناسی که این استانداردها را در برمی‌گیرد. Multext به ایجاد ابزارها، پیکره و منابع زبانشناسی برای زبان‌های مختلفی شامل Bambara، بلغاری، Catalan ،Czech، Dutch، انگلیسی، Estonian، فرانسوی، آلمانی، Hungarian، ایتالیایی، Kikongo، Occitan، Romanian، Slovenian، اسپانیایی، Swedish و Swahili پرداخته است. نتایج Multext به صورت رایگان برای همگان قابل دسترسی و استفاده می‌باشد.

3-8-1. پیشنهاد استانداردها

یک جنبه مهم در فعالیت‌های Multext مشارکت در استانداردسازی داده‌ها، ابزارها و منابع زبانشناسی به منظور به حداکثر رساندن قابلیت استفاده مجدد از آن داده‌ها و ابزارها و منابع در تحقیقات و کاربردهای مهندسی زبان بوده است. در برخی موارد Multext با موسسات و سازمان‌های استانداردسازی بزرگی همچون EAGLES و TEI[2] همکاری داشته است.

[1] Specifications

[2] Text Encoding Initiative

3-8-2. ابزارهای Multext

Multext یک مجموعه از ابزارها را برای دسترسی و مدیریت پیکره‌ها شامل پیکره‌های در قالب SGML و همچنین ابزارهایی برای انجام برخی از فعالیت‌های نشان‌گذاری پیکره‌ها شامل تشخیص کران جملات و واحدها، برچسب‌گذاری نحوی– ساختواژی[1]، هم‌ترازی متون موازی[2] و نشان گذاری پروزودی[3] در اختیار قرار می‌دهد.

همه ابزارهای Multext در نهایت از مشخصات نرم‌افزاری و معماری داده‌ای ایجاد شده در پروژه تبعیت می‌کنند. اگر چه ابزارها در مراحل مختلفی از ایجاد و توسعه قرار داشته باشند. در هر مرحله با مشخصات Multext تطابق دارند.

3-9. FreeLing

FreeLing یک بسته نرم‌افزاری متن‌باز از تحلیل‌گرهای زبان می‌باشد. FreeLing به صورت یک کتابخانه خارجی طراحی شده است تا در کاربردهایی که به این تحلیل‌گرها و سرویس‌ها نیاز دارند مورد استفاده قرار گیرد. با این وجود این با یک برنامه ساده که به عنوان یک واسط برای کتابخانه عمل کند کاربر را قادر می‌سازد تا فایل‌های متنی را مورد تحت قرار دهد.

[1] Morphosyntacitc Tagger
[2] Parallel text alignment
[3] Prosody markup

سرویس‌ها و تحلیل‌گرهای اصلی ارائه شده توسط کتابخانه FreeLing به این شرح می‌باشد:

- واحدساز متن
- تشخیص‌گران جملات
- تحلیل ساختواژی
- تحلیل رفتار پسوندها، واحدسازی مجدد ضمائر متصل و واژه بست‌ها
- تشخیث واحد چند کلمه‌ای
- شکستن واحدهای ادغام شده
- پیش‌بینی احتمالی مقولات کلمات ناشناخته
- تشخیص مداخل اسمی
- تشخیص تاریخ‌ها، اعداد، نسبت‌های کسری، واحدهای پولی، اندازه‌های فیزیکی (سرعت، وزن، درجه دما، چگالی و ...)
- برچسب‌گذاری اجزای واژگانی کلام
- Chart-based shallow parsing
- طبقه‌بندی مداخل اسمی
- نشان‌گذاری معنای کلمات بر اساس WorldNet
- تجزیه وابستگی متنی بر قاعده

اکثر سرویس‌های فوق برای زبان‌هایی که در حال حاضر FreeLing پشتیبانی می‌کند (اسپانیایی، انگلیسی، ایتالیایی، و Galiciab) فراهم می‌باشد.

3-10. NLTK: Natural Language Toolkit

NLTK یک بسته از ماژول‌های متن باز، دستورالعمل‌های آموزشی و مجموعه مسائل است که برای مطالعه در حوزه زبانشناسی محاسباتی استفاده می‌شود. NLTK پردازش آماری و سمبلیک زبان طبیعی را پوشش می‌دهد، و به عنوان یک واسط پیکره‌های نشانه‌گذاری شده می‌باشد.

ابزارهای NLTK به عنوان یک مجموعه از ماژول‌های مستقل است که هر کدام یک ساختار داده‌ای خاص را تعریف می‌کنند یا برای انجام یک فعالیت خاص ایجاد شده‌اند.

یک مجموعه از ماژول‌های اصلی وجود دارد که انواع داده‌ای و سیستم‌های پردازشی پایه‌ای را تعریف می‌کنند که در همه ابزارها مورد استفاده قرار می‌گیرند. ماژول Token کلاس‌های پایه را برای تک تک اجزای پردازش متن فراهم می‌سازد (مانند کلاس کلمات و جملات). ماژول Tree ساختمان داده‌هایی را برای نمایش ساختارهای درختی بر روی متن تعرف می‌کند. (مانند درخت‌های نحوی و درخت‌های ساختواژی). ماژول Probability کلاس‌هایی را پیاده‌سازی می‌کند که توزیع فراوانی‌ها و توزیع احتمالات (شامل تعداد تکنیک‌های زیادی برای هموارسازی آماری) را پیاده‌سازی می‌کند.

بقیه ماژول‌ها ساختمان داده‌ها و واسط‌هایی برای انجام فعالیت‌های خاص پردازش زبان طبیعی می‌باشند که به مرور در NLTK توسعه یافته و می‌یابند. لیست آنها به صورت زیر است:

- ماژول تجزیه (تجزیه‌گر)
- ماژول برچسب‌گذاری
- اتوماتای حالت محدود (Finite site Automaton)
- Type checking (که برای خطایابی کدهای برنامه مورد استفاده قرار می‌گیرد)

- ماژول‌های Visualization(که برای نمایش و مدیریت ساختمان داده‌ها و برای نمایش خروجی‌های فعالیت‌ها مورد استفاده قرار می‌گیرند.)
- رده‌بندی متون

3-11. Emdros

3-11-1. مشخصات کلی Emdros

Emdros یک موتور دادگان متنی[1] متن باز است که برای ذخیره و بازیابی متون تحلیل شده یا نشان‌گذاری شده استفاده می‌شود. Emdros به یک زبان پرسمان قدرتمند برای درخواست داده‌های مورد نیاز کاربر مجهز می‌باشد. Emdros کاربرد گسترده در حوزهٔ متون تحلیل شده و نشان‌گذاری شده دارد. این کاربردها شامل زبانشناسی، نشر، پردازش متن و هر حوزه‌ای که با متون نشان‌گذاری سروکار دارد، می‌باشد.

امکاناتی که Emdros ارائه می‌دهد به شرح زیر است:

- تحلیل‌های زبانشناسی که هدف اصلی Emdros می‌باشد و شامل همه سطوح تحلی از قبیل ساختواژ، نحو، تحلیل گفتمان و حتی آواشناسی میباشد.

[1] Text database engine

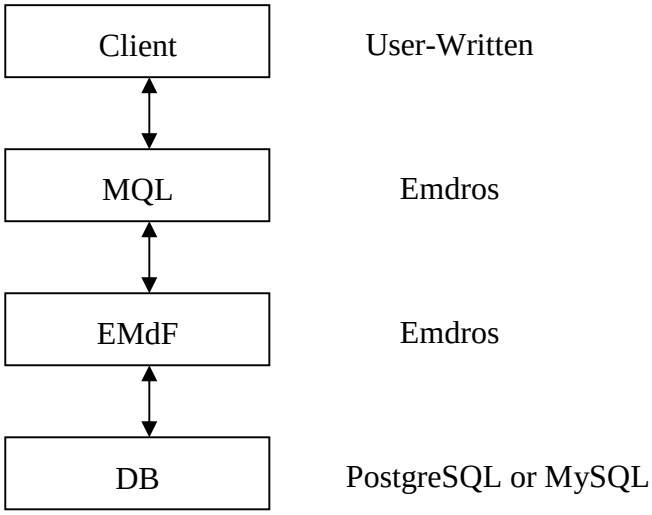
- نشر نیز حوزه‌ای است که می‌توان در آن از Emdros استفاه کرد. Emdros عملیات خرد کردن یک متن به صفحات، فصول، پاراگراف‌ها و ... را پشتیبانی می‌کند.
- در پردازش متن اگر مسئله شامل نشان‌گذاری متن باشد می‌تواند از Emdros استفاده کرد.

Emdros یک مدل مفهومی از متن ایجاد می‌کند پس از ساختن آن بسیار مفید خواهد بود. Emdros همچنین می‌تواند هم در زبانشناسی پیکره‌ای[1] (حجم زیاد متن) و هم در زبانشناسی حوزه‌ای[2] (حجم کم متن) مفید واقع شود.

همچنین Emdros شامل یک مدل خاص از متن به نام مدل EMdF است. مزیت اولیه مدل داده‌ای XML آن است که انواع اشیاء (از قبیل صفحات و فصول) نیاز نداشته باشند. فقط به صورت سلسله مراتبی ساختاربندی شده باشند و امکان روی هم افتادگی نیز داشته باشند. به علاوه اشیاء (از قبیل عبارات) نیاز به پیوستگی و پشت سر هم بودن نداشته باشند و امکان ایجاد فاصله نیز بین آن‌ها باشد. Emdros می‌تواند نتایج را با کمک EMdF در قالب XML تولید کند.

معماری Emdros به صورت زیر می‌باشد:

[1] Corpus linguistics
[2] Field-linguistics



شکل3-2. معماری Emdros

در سطح بالا کاربر قرار دارد که می‌تواند از مزایای سرویس‌های Emdros استفاده کند و نیازهای خود را در حوزه موردنظر تامین سازد.

بعد از آن دو لایه دیگر Emdros قرار دارد: لایه MQL و لایه EMdf. لایه MQL یک واسط با زبان MQL را فراهم می‌سازد. لایه MQL به صورت خودکار از لایه EMdf که پرسمان‌های MQL را به پرسمان‌های SQL (در روی دادگان) تبدیل میکند بهره می‌جوید.

3-11-2. MQL[1]

زبان پرسمان Emdros ، MQL نامیده می‌شود. این زبان یک زبان پرسمان قدرتمند با توانایی‌های ساخت، به‌روزرسانی، حذف می‌باشد. خصوصاً امکان برآورده ساختن تقاضاها را از طریق پرسمان به طرز شایانی انجام می‌دهد.

MQL یک نسخه تعمیم یافته از زبان پرسمان QL است که نتیجه تلاش Crist-jan Doeden در پایان‌نامه دکتری وی می‌باشد[2]. QL یک زبان پرسمان کاملا قوی است که با مدل Mdf سازگار می‌باشد.

پیاده‌سازی عملی QL به چندین دلیل بسیار مشکل می‌باشد. بنابراین MQL به وجود آمد. MQL در واقع مخفف MiniQL (QLکوچک) می‌باشد زیرا MQL نسخه ساده شده QL است.

MQL با دارا داشتن اکثر امکانات QL ایده‌های جدید دیگر ایجاد شد. MQL یک زبان با امکان دسترسی بسیار بالا است که ساخت، به روزرسانی، حذف و پرسمان همه حوزه‌های داده‌ای در مدل EMdF را ممکن می‌سازد.

[1] Emdros Query-Language

[2] Doedens, Crist-Jan [Christianus Franciscus Joannes]. (1994) Text Databases. One Database Model and Several Retrieval Languages. Language and Computers, Number 14. Editions Rodopi Amsterdam. Amsterdam and Atlanta, GA. ISBN: 90-5183-729-1.

EMdF .3-11-3

EMdF یک مدل دادگان متنی است. به عبارت روشنتر، یک مدل ریاضی از متن است که مفاهیم درگیر با نحوه تعامل EMdF با متن را شرح و توضیح می‌دهد.

چهار مفهوم اساسی در مدل EMdF وجود دارد. برای یک مرتبه که این چهار مفهوم درک و تبیین شوند بقیه موارد بسیار ساده درک و همچنین خواهند شد. این چهار مفهوم در زیر بیان می‌گردد:

1- Monads : یک monads به طور ساده یک عدد صحیح یا به عبارت دیگر یک عدد طبیعی است (1، 2، 3، 4، ...)

2- Objects (شی): یک شی یک مجموعه از monad هاست. این مجموعه می‌تواند شامل هر monad شود. یعنی مجموعه {1و2و3} مثل مجموعه {1و2و5و6و7} معتبر میباشد. این شی‌ها هستند که وظیفه نشان دادن متن به علاوه اطلاعات درباره آن متن را به عهده دارند، این کار به وسیله Object type (نوع شی) و ویژگی‌هایشان شرح داده می‌شود.

3- Object type (نوع شی): شی‌ها در انواعی دسته‌بندی می‌شوند. یک نوع می‌تواند به عنوان مثال "phrase"، "clause"، "word" ، "chapter" ، "book" و... باشد. یک شی همیشه از یک نوع خاص است.

4- Features (ویژگی): در نهایت انواع اشیا ویژگی‌هایی دارند. یک ویژگی یک خاصیت یا مقداری است که با یک شی همراه می‌شود. مقدار یک ویژگیِ شئی است که داده تحلیلی را در دادگان ذخیره می‌کند. برای مثال نوع شی ، "phrase" می‌تواند یک

ویژگی"phrase-value" داشته باشد که بیان کند آیا عبارت یک عبارت اسمی است یا فعلی یا حرف اضافه‌ای.

همچنین مقدار یک ویژگیِ شئی است که داده‌های متنی را در دادگان ذخیره می‌کند. برای مثال نوع شی، "word" می‌تواند یک ویژگی به نام "surface" داشته باشد که بیان می‌کند کدام کلمه متنی برای آن شی word خاص می‌باشد.

3-12. IMS Corpus Workbench (CWB)

به منظور پشتیبانی از فعالیت‌های در حوزه فرهنگ‌نگاری و اصطلاح‌نامه نگاری IMS یک Workbench برای بازیابی از منابع متنی بزرگ یا همان پیکره‌ها تهیه نموده است.

3-12-1. کاربردهای CWB

IMS Corpus Workbench برای موارد زیر استفاده می‌شود:

- زبانشناسی مبتنی بر داده: استخراج دانش زبانشناسی از منابع متنی یا Cross-checking فرضیات زبانشناسی در متون بسیار بزرگ
- فرهنگ نگاری: شواهد مبتنی بر پیکره برای توصیف‌های واژگانی
- اصطلاح نامه: استخراج ترم‌ها و یادگیری نیمه خودکار از منابع دارای اصطلاحات فنی

3-12-2. ویژگی‌ها CWB

CWB یک زبان پرسمان دارد که امکانات زیر را ممکن می‌سازد:

- تعداد غیرمحدود خواص در هر موقعیت از پیکره
- عبارات منظم روی مقادیر خاصیت‌ها در موقعیت‌های خاصی از پیکره
- عبارات منظم روی جملات پیکره
- پشتیبانی (جزئی) از نشانه‌گذاری‌های ساختاری (مثل SGML)
- استفاده از یک پرسمان برای همه موارد یک لیست
- خاصیت‌های مجازی، به عبارت دیگر دسترسی زبان اجرا به منابع خارجی (مثل گنجواژه)
- پرسمان‌ها روی متون ترجمه شده موازی

امکانات نمایش نتایج در CWB شامل موارد زیر است:

- نمایش کلمات کلیلدی در بافت متن که توسط کاربر قابل تعریف است.
- خطوط کلمات کلیدی در بافت متن می‌تواند به روش‌های مختلف مرتب شوند.
- مقادیر فراوانی‌ها برای مثال برای ترکیبات کلمات
- تطبیق همزمان چند زبانه از پیکره‌های هم تزار شده
- تولید خروجی به صورت html و latex
- تاریخچه پرسمان‌های انجام شده

برای مدیریت و آماده‌سازی پیکره‌ها از امکانات زیر می‌توان بهره برد:

- ایجاد و ثبت پیکره‌ها
- رمزگذاری پیکره‌هابه عبارت دیگر شاخص‌گذاری و فشرده‌سازی
- امکان افزودن انواع نشان‌گذاری‌های پیکره (مناسبت‌ها) برای مثال مقادیر اجزای واژگانی کلام برای یک پیکره زمانی که یک برچسب‌گذار اجزای واژگانی کلام موجود است.

امکانات بازیابی

- زبان پرسمان بوسیله پردازش‌گر پرسمان پیکره (CQP) تفسیر می‌شود. CQP نیاز دارد پیکره‌ها ایجاد و ثبت شده و در یک روش خاص رمزگذاری شده باشد.

13-3. CRFClassfier: تشخیص دهنده مداخل اسمی دانشگاه آکسفورد

CRFClassfier یک پیاده‌سازی جاوایی برای تشخیص مداخل اسمی است. تشخیص دهنده مداخل اسمی دنباله‌ای از کلمات در یک متن را که نام اشیاء می‌باشند (از قبیل: نام شخص، نام شرکت، نام پروتئین‌ها و ژن‌ها) را برچسب مناسب منتسب می‌کند. نرم‌افزار CRFClassfier یک پیاده‌سازی کلی و

عمومی از مدل‌های دنباله CRF[1] زنجیره کلی فراهم می‌کند که این مدل‌هابا یک استخراج کننده ویژگی برای تشخیص دهنده مداخل اسمی ترکیب شده‌اند.

در حال حاضر نرم‌افزار تشخیص دهنده مداخل اسمی براس سه کلاس (شخص، سازمان، مکان) برای انگلیسی را شامل می‌شود و دو مدل آموزش دیده روی داده‌های آموزشی برای زبان انگلیسی در CoNLL-2003[2] را در بردارد. مدل‌های تشخیص دهنده سه کلاس (شخص، سازمان و مکان) هم با ویژگی‌های توزیعی شباهت و هم بدون آن موجود باشند. ویژگی‌های توزیعی شباهت کارایی را افزایش می‌دهند اما مصرف حافظه را بسیار بالا می‌برند.

از ویژگی‌های CRFClassfier این است که مدل‌های جدیدی را می‌توان برای کلاس‌های مختلف ایجاد کرد. همچنین از این تشخیص‌دهنده مداخل اسمی به صورت مولفه‌ای در کاربردی‌های دیگر پردازش زبان طبیعی مثل تجزبه‌گرها استفاده نمود.

14-3. **Yamcha**

Yamcha یک تقطیع‌گر متن است که به صورت متن باز، با قابلیت استفاده مجدد است که در بسیاری از کارهای پردازش زبان طبیعی از قبیل برچسب‌گذاری اجزای واژگانی کلام، تشخیص مداخل

[1] Conditional Random Field

[2] Conference on Computational Natural Language Learning

اسمی، تقطیع عبارت اسمی ساده و تقطیع متون استفاده می‌شود. Yamcha از الگوریتم یادگیری SVM[1] استفاده می‌کند.

ویژگی‌های Yamcha :

- § کارایی نسبت بالا بر مبنای الگوریتم SVM
- § قابلیت آموزش / تست با هر داده
- § استفاده از PKE/PKI که عملیات رده‌بندی (تقطیع) را سریعتر انجام می‌دهد.
- § قابلیت بازتعریف مجموعه ویژگی‌ها (اندازه پنجره)، مسیر تجزیه (پیش رو / عقب رو) و الگوریتم‌های مسائل چند کلاسه (دوتایی / یکی در برابر بقیه[2])
- § سرعت نسبتاً خوب
- § قابلیت استفاده برای تقطیع اجمالی[3]
- § کتابخانه C/C++

3-15. CRF++

CRF++ یک نرم‌افزار متن‌باز ساده و با قابلیت استفاده مجدد است که CRF ها (فیلدهای شرطی

[1] Support Vector Machine
[2] One vs. rest
[3] Partial Chunking

تصادفی) را برای قسمت‌بندی و یا برچسب‌گذاری داده‌های ترتیبی پیاده‌سازی می‌کند.

CRF++ با هدف عمومیت طراحی شده است و در فعالیت‌های مختلف پردازش زبان طبیعی از قبیل تشخیص مداخل اسمی، استخراج اطلاعات (بازیابی اطلاعات) و تقطیع متن قابل استفاده می‌باشد.

از ویژگی‌های CRF++ به موارد زیر می‌توان اشاره کرد:

- قابلیت بازتعریف مجموعه ویژگی‌ها
- نوشته شده در C++
- آموزش سریع بر مبنای الگوریتم LBFGS (یک الگوریتم کواسی–نیوتن[1] برای مسائل بهینه‌سازی عددی در اندازه بزرگ)
- مصرف حافظه کمتر هم در آموزش و هم در تست
- رمزگذاری و رمزگشایی در زمان کوتاه
- قابلیت اعمال بهترین n خروجی
- قابلیت ارائه خروجی احتمالات Marginal برای همه کاندیداها
- متن باز بودن نرم‌افزار

3-16. fnTBL

fnTBL یک ابزاری یادگیری ماشین متن باز، قابل حمل و با قابلیت استفاده مجدد است که هدف

[1] quasi-newton

آن انجام فعالیت‌های مربوط به پردازش زبان طبیعی می‌باشد. این فعالیت برچسب‌گذاری اجزای واژگانی کلام، تقطیع گروه‌های اسمی پایه، تقطیع متن، تشخیص توان کلمات، ابهام زدایی از معنای کلمات (WSD) می‌باشد. fnTBL می‌تواند فعالیت‌های طبقه‌بندی با مقادیر گسسته و با اندازه نسبتاً خوب را انجام دهد. نمونه‌ها می‌تواند به صورت برداری در اجزای گسسته مشخص شوند.

fnTBL یک تکنیک یادگیری ماشین به نام یادگیری مبتنی بر تبدیل[1] را استفاده می‌کند. این روش ابتدا در سال 1992 توسط Eric Brill پیشنهاد شد و ایده اصلی آن تبدیل مناسب داده‌هاست به طوری که خطاها تصحیح شود و بیشترین موفقیت با توجه به پارامتر نرخ خطا حاصل گردد. قوانین تبدیل معمولاً کم ولی بسیار کارآمد می‌باشند.

fnTBL در گروه پردازش زبان طبیعی دانشگاه Johns Hopkins ساخته شده است.

در حال حاضر برای موارد زیر مجموعه قواعد موجود می‌باشند:

- برچسب‌گذاری اجزای واژگانی کلام برای زبان انگلیسی
- برچسب‌گذاری اجزای واژگانی کلام برای زبان سوئدی
- تقطیع گروه‌های اسمی پایه در زبان انگلیسی
- تقطع متون انگلیسی

[1] Transformation Based Learning

3-17. تقطیع‌گر گروه‌های اسمی Greenwood

این ابزار یک پیاده‌سازی جاوایی برای تقطیع گروه‌های اسمی است که نسخه اولیه آن به زبان C++ توسط Ramshaw و Maraus پیاده‌سازی شده است. این تقطیع‌گر سعی می‌کند که گروه‌های اسمی را با قراردادن براکت در متن مشخص کند.

امکان استفاده از این تقطیع‌گر در چهارچوب GATE نیز وجود دارد.

3-18. Rainbow و کتابخانه Bow

رده‌بندی Naive Bayes یکی از الگوریتم‌های مشهور و موفق در بین الگوریتم‌های یادگیری ماشین برای رده‌بندی متون می‌باشند. Bow (یا Libbow) یک کتابخانه از کدهای C است که برای تحلیل آماری متون، مدلسازی زبان و برنامه‌های بازیابی اطلاعات مفید می‌باشد.

امکاناتی که این کتابخانه فراهم می‌کند به شرح زیر است:

- واحدسازی فایلی متنی بر اساس چندین روش مختلف
- دنباله‌های n تایی در بین واحدها
- نگاشت رشته‌های کاراکتری به اعداد و برعکس
- ساخت یک ماتریس اسپارس از آمارهای اسناد و واحدها
- خالص‌سازی مجموعه لغات با توجه به فراوانی کلمات و بهره اطلاعاتی

- ساخت و مدیریت بردارهای کلمات
- تنظیم وزن‌های بردار کلمات بر طبق روش Naïve Bayes ، TFIDF و چندین روش دیگر
- هموارسازی احتمال رخداد کلمات بر اساس روش‌های Laplace (Dirichlet uniform)، M-estimate، Witten-Bell و Good-Turning
- امتیاز دهی به پرسمان‌ها برای بازیابی و رده‌بندی
- انجام تقسیم داده به داده‌های آموزش و تست و انجام خودکار تست رده‌بندی

پس از شرح کتابخانه Bow به RainBow می‌پردازیم. RainBow یک برنامه رده‌بندی متون می‌باشد که بر اساس کتابخانه Bow تهیه شده است. استفاه از RainBow در دو مرحله می‌باشد:

1- خواندن اسناد و تهیه مدل که شامل آمارهای مورد نیاز می‌باشد.

2- استفاده از مدل برای انجام رده‌بندی و تشخیص.

امکان استفاده از واحدساز، رده‌بند اسناد متنی در RainBow فراهم می‌باشد.

19-3. ابزار Wordsmith اکسفورد

ابزار Wordsmith یک مجموعه برنامه است که نحوه رفتار کلمات در متن را نشان دهد. با استفاه از آن کاربر می‌تواند چگونگی استفاده کلمات را در متن دریابد.

Wordsmith سه بخش عمده دارد: wordlist ، Concord ، keyword

Wordlist: این برنامه لیست کلمات را از یک یا چند فایل متنی (با یونیکد ASCII یا ANSI) استخراج می‌کند. لیست کلمات به صورت خودکار با ترکیب الفبایی و با ترتیب بر اساس فراوانی تولید می‌شود. به صورت اختیاری نیز می‌توان یک لیست شاخص کلمه[1] تولید کرد.

بنابراین با استفاده از این برنامه می‌توان:

- به مطالعه مجموعه کلمات پرداخت.
- خوشه‌های کلمات رایج را تشخیص داد.
- فرکانس کلمات را در فایل‌ها و گونه‌های مختلف متن مقایسه کرد.
- یک Concordance از یک یا چند کلمه در لیست کلمات ایجاد کرد.

با استفاه از wordlist می‌توان دو لیست را مقایسه کرد یا یک تحلیل سازگاری با هدف مقایسه سبک[2] انجام داد. این لیست‌های کلمات همچنین می‌توانند به عنوان ورودی برنامه keyword مورد استفاده قرار گیرند.

Concord: این برنامه برای ایجاد یک Concordance از فایل‌های متنی مورد استفاده قرار می‌گیرد. برای استفاده از آن باید یک کلمه جستجو مشخص شود و Concord آن کلمه را در همه فایل‌های متنی جستجو خواهد کرد. سپس concord یک Concordance را نمایش می‌دهد که امکان دسترسی به

[1] Word index

[2] Stylistic

اطلاعات درباره با هم آیی‌های کلمه جستجو، نمایش محل توزیع کلکه جستجو در فایل و تحلیل‌های خوشه‌ای نمایش دهنده خوشه های تکرار شده کلمات (عبارات) را ممکن می‌سازد.

هدف از ایجاد یک Concordance به دست آوردن مثال‌های زیاد از یک کلمه یا عبارت در بافت‌های مختلف است. با دیدن مثال‌ها می‌توان ایده بهتری نسبت به نحوه استفاده از کلمه به دست آورد. این مورد در موارد مختلفی می‌تواند کاربرد داشته باشد. به عنوان مثال معناهای مختلف یک کلمه در بافت‌های مختلف ظهور می‌یابد.

Keywords: این برنامه برای مشخص کردن کلمات کلیدی متن می‌باشد. کلمات کلیدی در این برنامه کلماتی در نظر گرفته شده‌اند که فرکانس آنها در مقایسه با یک نرم مقدار بالایی داشته باشد. کلمات کلیدی یک روش مفیدی برای مشخص کردن یک متن یا یک گونه می‌باشد. این مورد نیز کاربردهای مختلف مثلا در بازیابی اطلاعات دارد.

4. نتیجه‌گیری

پیش از آن‌که مستقیما به مطالعه و بررسی ابزارهای آماده برای تحلیل پیکره‌های متنی بپردازیم نکات بارز دیگری به نظر می‌رسد که خود تولید گسترده این ابزارهای آماده را منجر شده و مهمتر از آن باعث پیشرفت و توسعه گسترده محصولات زبانی در دیگر زبان‌ها مانند زبان‌های اروپایی شده است. بنابراین در گزارش حاضر نیز ما ابتدا به آن موارد پرداختیم و پس از آن وارد مبحث مطالعه و بررسی ابزارهای آماده برای تحلیل پیکره‌های متنی شدیم.

اولین مسائلی که باید مورد توجه قرار گیرد این است که پیکره‌ها باید با چه روشی تولید شوند و چه استانداردهایی را دارا باشند. این استانداردها توسط چه سازمان، نهاد و یا تشکیلاتی تهیه شود. متولی ایجاد پیکره و نظارت بر آن چه سازمان، نهاد و یا تشکیلاتی باشد. بنابراین در فصل اول این گزارش سازمان‌ها و نهادهای بین‌المللی را مورد بحث قرار دادیم که به این فعالیت‌ها می‌پردازند. سعی شد که سازمان‌ها و نهادهایی که از شهرت بسزایی برخوردار باشند و همچنین فعالیت آن‌ها در حوزه زبان‌های هند و اروپایی (که زبان فارسی نیز در این طبقه‌بندی قرار دارد) نام برده شود. ساختار تشکیلاتی این سازمان‌ها و نهادها می‌تواند الگوی مناسبی برای ایجاد سازمان‌ها و نهادها کشور باشد. علاوه بر آن به دلیل مشابهت زبان فارسی و زبان‌های مورد مطالعه آن سازمان‌ها و نهادها می‌توان از استانداردهای تولید شده توسط آن‌ها استفاده کرد.

موردی دیگری که توجه به آن از اهمیت زیادی برخوردار است ایجاد بسترهای مناسب نرم‌افزاری

برای تحلیل پیکره‌ها و تولید ابزارهای زبانی است. اگر بستر مناسبی موجود باشد افراد مختلف با توجه به نیازهای مختلف می‌توانند به تولید ابزار مورد نیاز خود بپردازند و اشتراک‌گذاری این ابزارها در کابردهای مختلف تسهیل می‌گردد. در فصل دوم به زیرساخت‌ها و چارچوب‌های پرداختیم که این امکان را برای کاربران مختلف ممکن می‌سازند. این زیرساخت‌ها و چارچوب‌ها نیز می‌تواند مورد بررسی قرار بگیرد تا میزان انطباق آن با زبان فارسی مشخص شده و در صورت نیاز به ایجاد نمونه‌های مشابه اقدام شود.

در فصل سوم ابزارهای آماده برای تحلیل پیکره‌های متنی تحت بررسی قرار گرفت. اگرچه برخی از ابزارهای بررسی شده مستقیما قابل استفاده برای تحلیل پیکره‌ها متنی زبان فارسی نیستند ولی ایده‌های ارائه می‌کنند که می‌تواند به ایجاد ابزارهای مناسبتر در زبان فارسی منجر شود. عملا امکان دسته‌بندی دقیق این ابزارها ممکن نبود. زیرا برخی از ابزارها امکانات تحلیلی در سطوح مختلف ارائه می‌کنند و برخی از ابزارها تنها تحلیل در یک سطح را ممکن می‌سازند و یک هدف خاص را دنبال می‌کنند. همچنین در سطوح مختلف نیز روش‌های مختلفی وجود دارد که یک با برخی از آن‌ها در ابزارها تعبیه شده است. بنابراین سعی شد ترتیب بررسی ابزارها بر اساس تلفیقی از سطوح تحلیل و گستردگی امکانات ارائه شده صورت پذیرد.