شیوه نامه تهیه پیکره تشخیص گروه‌های اسمی

در این بخش شیوه‌نامه‌ی تهیه‌ی پیکره به صورت کامل ارائه شده است. در تدوین این شیوه نامه، جلسه‌ها خبرگی و متناسب با دسته‌بندی داده‌ی واقعی مورد نیاز بوده است. هدف، تعیین عبارت‌های اسمی در مجموعه‌ی دادگان شامل ۵۰۰ هزار توکن از متون توییتر، اینستاگرام و تلگرام است. در این بخش، فقط نیاز به تشخیص گروه‌‌های اسمی می‌باشد و بنابراین نیاز به تعیین سایر گروه‌ها، مانند گروه‌های قیدی، فعلی و … نمی‌باشد. نیازی به برچسب‌زنی عبارات انگلیسی، عربی، کردی، اردو و ... نمی‌باشد. در شمارش تعداد توکن ها، علایم نگارشی شمارش نمی شوند.

جهت تشخیص محدوده گروه ها روش مختلفی وجود دارد که از متداول ترین آنها می توان به برچسب گذاری با براکت و به فرمت IOB اشاره کرد. در این پژوهش برای تشخیص مرز بین گروهها از فرمت IOB استفاده شده است که از سه نوع برچسب برای کلمات سازنده یک گروه استفاده می کند. برچسب B برای کلمه ابتدای هر گروه، برچسب I برای کلماتی که داخل گروه قرار دارند و برچسبO که معمولا برای نمایه انتهای جمله که اغلب، نقطه یا علامت سوال است استفاده می‌شود.

اجزای گروه‌های اسمی

گروه‌های اسمی در قالب رایج IOB برچسب خواهند خورد. IOB اشاره به یک نوع قالب استاندارد دارد که به شکل زیر است ‎[1]:

۱. B: نشان‌دهنده‌ی واژه‌ی شروع گروه اسمی است.

۲. I: نشان‌دهنده‌ی واژٰه‌های داخل گروه اسمی است.

۳. O: نشان‌دهنده‌ی واژه‌های خارج از گروه اسمی است.

برای نمونه در جمله‌ی «کشور زیبای جذاب ایران، جنگ خونین را دوست ندارد.»، اگر فرض کنیم گروه‌های اسمی به ترتیب عبارت‌انداز «کشور زیبای جذاب ایران» و «جنگ خونین» است، نحو‌ه‌ی برچسب‌زنی به شکل زیر خواهد بود.

مردم در کشور زیبای جذاب ایران جنگ خونین را دوست ندارد

O O O I B I I I B O B

اجزای سخن

اجزای سخن یا ادات کلام، به نقش‌های مختلف واژگان گفته می‌شود که در جمله به واژگان اختصاص می‌یابد. اجزای سخنی که در این مجموعه‌ی دادگان نیاز به مشخص شدن دارند شامل موارد زیر است:

۱. فعل(V)

۲. اسم(N)

۳. ضمیر ((Pro

۴. صفت (Adj)

۵. قید (Adv)

۶. حرف اضافه‌ی پسین (Prep)

۷. حرف اضافه‌ی پیشین (Postp)

۸. حرف ربط (Conj)

۹. منادا (Interj)

۱۰. نشانه‌های نگارشی(Punc)

11. عدد (Num)

۱۲. تاریخ (D)

13.نامشخص(ignore)

نکات تدوین پیکره

لازم به ذکر است که نقش کلمه باید با توجه به جایگاه آن در جمله مشخص گردد و در صورتی که مثلا عدد نقش صفت را گرفت بهتر است نقش صفت برای آن مشخص گردد. در مواردی که تصمیم گیری سخت است، دو برچسب pos با یک جداکننده‌ی کاما از هم جدا شوند. برای مثال:

طلای Ne

24 adjs,nums

عیار adj

برای کاراکترهای چینی و ژاپنی و عباراتی که چند کلمه در یک توکن قرار گرفته است از برچسب ignore استفاده گردد.

چون مجموعه داده شامل متون شبکه‌های اجتماعی است، عباراتی که با هشتگ وجود دارند، دیده می‌شود. این عبارات به صورت خودکار برای آن‌ها برچسبی به جای برچسب ادات سخن تعیین شده است. این برچسب پیش فرض hashtag خواهد بود و در نتیجه نیاز نیست برای آن‌ها برچسب اجزای سخن توسط برچسب گذاران مشخص گردد.

دسته‌بندی متون عامیانه و تهیه پیکره بر اساس شیوه نامه‌ی تهیه شده

پس از جمع آوری داده های مورد نیاز از شبکه های اجتماعی، طبقه بندی داده ها و ایجد پیکره در برنامه‌ی اکسل انجام می‌شود. برای این منظور جلسه‌های با افراد متخصص در این حوزه تشکیل شده و مقالاتی در زمینه تشخیص گروههای اسمی مطالعه شده است. همچنین به کتابهای زبان شناسی مرتبط با اجزای سخن در زبانشناسی از استاد ایران کلباسی مراجعه شده است.

متن‌ها در قالب استاندارد رایج CoNLL در اختیار قرار می‌گیرد و در همین قالب نیز تحویل گرفته می‌شود. قالب CoNLL برای دسته‌بندی همزمانِ اجزای سخن و گروه‌های اسمی به شرح زیر است.

۱. واژه: در ابتدای سطر، خود واژه‌ای که قرار بر برچسب‌زنی آن است قرار می‌گیرد.

۲. دسته‌ی اجزای سخن: در جلوی «واژه»، با فاصله‌ی یک tab، اجزای سخن آن واژه مشخص می‌شود

۳. دسته‌ی گروه اسمی: در جلوی «برچسب اجزای سخن» با فاصله‌ی یک tab، گروه اسمی آن قرار می‌گیرد.

پس از مشخص شدن این سه قسمت، خط بعد شامل واژه‌ی بعدی در جمله است. لازم به ذکر است پس از پایان هر جمله، یک سطر خالی گذاشته می‌شود. و پس از پایان هر توییت/پست/متن نیز دو سطر خالی پشت سرهم قرار خواهد گرفت.

معرفی پیکره

در این بخش، به بررسی پیکره‌ی انتخابی و تعداد آن می‌پردازیم. در این بخش ضمن معرفی پیکره، انواع مختلف برچسب‌های مربوط به واژگان و همین‌طور چانکر مورد بررسی قرار می‌گیرند.

پیکره‌ی انتخابی

این پیکره شامل بیش از ۵۰۰ هزار توکن از بسترهای اینستاگرام، تلگرام و توییتر است. مهم‌ترین ویژگی این پیکره، وجود متن‌های رسمی و غیر رسمی است که باعث می‌شود مدل ابزار چانکر به دست آمده، توانایی تشخیص گروه‌های اسمی را هم در متن‌های رسمی و هم در متن‌های غیر رسمی داشته باشد. این پیکره شامل متن‌های سه بستر یادشده از تاریخ یکم تیرماه ۱۳۹۸ تا پایان اسفند ۱۳۹۹ است.

پیکره‌ی چانکر

متن‌ها برای مدل چانکر با فرمت رایج CONLL برچسب می‌خورند. در واقع ابتدای سطر خود واژه و سپس نقش pos آن و سپس نقش چانکر آن مشخص می‌شود. نقش واژه در گروه اسمی با یکی از سه حرف I، O و B تعیین می‌شود.

* B: شروع گروه اسمی را مشخص می‌کند.
* I: نشان‌دهنده‌ی ادامه‌ی گروه اسمی است.
* O: نشان‌دهنده‌ی آن است که واژه‌ی مورد نظر در گروه اسمی قرار ندارد.

پیکره‌ی برچسب‌زن اجزای سخن

اجزای سخن یا ادات کلام، به نقش‌های مختلف واژگان گفته می‌شود که در جمله به واژگان اختصاص می‌یابد. اجزای سخنی که در این مجموعه‌ی دادگان مشخص می‌شوند شامل موارد زیر است:

۱. فعل که با نماد V مشخص می‌شود.

۲. اسم که با نماد N مشخص می‌شود.

۳. ضمیر که با نماد PRO مشخص می‌شود.

۴. صفت که با نماد ADJ مشخص می‌شود.

۵. قید که با نماد ADV مشخص می‌شود.

۶. حرف اضافه‌ی پسین که با نماد PREP مشخص می‌شود.

۷. حرف اضافه‌ی پیشین که با نماد POSTP مشخص می‌شود.

۸. حرف ربط که با نماد CONJ مشخص می‌شود.

۹. منادا که با نماد INTER مشخص می‌شود.

۱۰. نشانه‌های نگارشی که با نماد PUNC مشخص می‌شود.

۱۱. عدد که با نماد NUM مشخص می‌شود.

۱۲. نامشخص که با نماد IGNORE مشخص می‌شود.

در تهیه‌ی پیکره نکات زیادی در نظر گرفته شد مانند:

* نقش کلمه باید با توجه به جایگاه آن در جمله مشخص می‌گردد و در صورتی که مثلا عدد نقش صفت را گرفت نقش صفت (ADJ) برای آن مشخص می‌شود و نه نقش عدد (NUM).
* برای کاراکترهای چینی و ژاپنی از برچسب IGNORE استفاده می‌گردد.
* در صورتی که کاربری که متن‌ها را نوشته است به اشتباه دو واژه را سرهم نوشته بود، با نشانه‌ی + نقش دو واژه مشخص می‌شود. برای نمونه: اومرد PRO+V
* ضمیرهای متصل که به پایان اسم‌ها/صفت‌ها می‌چسبند، به عنوان pro در نظر گرفته نمی‌شوند. برای نمونه:

من Pro

گلشان N

را POSTP  
قطع VS

کردم V

* اگر ضمیر متصل، در خط بعد نوشته شده بود، باز هم به عنوان بخشی از اسم در نظر گرفته می‌شود و نه ضمیر! برای نمونه:

من Pro

گل NS

شان N

را POSTP  
قطع VS

کردم V

* در صورت وجود کسره‌ی اضافه از یک نشانه‌ی E به همراه برچسب انتخاب شده استفاده می‌گردد. به عنوان مثال:

کتاب NE B

خوب ADJ I

* در صورتی که برچسب‌گذار با کلمه یا عبارتی مواجه شود که یک کلمه بوده و به اشتباه (خطای قطعه‌بند) در دو سطر مجزا واقع شده است از کاراکتر S کنار برچسب کلمه‌ی اول استفاده می‌کند تا نشان دهد این کلمه و کلمه‌ی بعد باهم تشکیل یک کلمه‌ی واحد را می‌دهند. به عنوان مثال:

رئیس Ns B

جمهورN I

* اگر کلمه‌ای بیش از دو بخش داشت و در چند سطر واقع شده باشد:

سی Ns B

و Ns I

سه Ns I

پل N I

* ممکن است ضمیر نیز کسره‌ی اضافه بگیرد برای مثال:

من Proe

دوست Adjs

داشتنیAdj

معرفی کلاس‌های مختلف واژه

با بررسی پیکره‌ی برچسب‌خورده، می‌توان دریافت که هر واژه می‌تواند نقش‌های مشخص شده در ‏جدول 2-1 را در جایگاه pos به خود بگیرد.

نقش‌های مختلف واژه‌ها در برچسب‌زنی اجزای سخن

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PUNC | HASH | DETS | ADV | DETE |
| N | ADVS | PREPS | DET+N+N | NE+N |
| ADJ | ADV | CONJS | ADJ+PRO | ADJ+N |
| NE | ADJE | PREP+NE | PRO+CONJ | PRO+N |
| POSTP | AD | ADVE | ADJ+POSTP | N+ADJ |
| PREP | INTER | PROS | N+VS | E |
| V | NES | DET+N | N+PREP+N | CONJ+NE |
| NS | NUM | CONJE | ADJ+CONJ | PREP+ADV |
| CONJ | PPRE | NSE | CONJ+VS | PREP+pro |
| DET | PRO+V | N+V | ADJ+ADJS | N |
| PRO | PREP+N | CONJ+N | N+NS | N+N |
| VS | NE+ADJ | NUMS | DVS | N+PRO |
| ADJS | PREPE | PREP+V | ADV+V | NE+ADJE |
| NE+NS | ADJS+NE | PREP+DET | ADJ+V | CONJ+N+PRO |
| ADJ+VS | N+NE | N+PREP | N+PRO+CONJ | POSTP+V |
| N+PO | N+CONJ | CONJ+PREP | ADJS+NS | ADV+N |
| PREP+ADJ | DET+ADJ | PEREP | N+POSTP | ADJS+N |
| PROE | CONJ+ADJ | NE+NE+PRO | V+PRO | NE+N+PRO |

همان‌طور که مشاهده می‌شود هر واژه می‌تواند ۸۹ کلاس مختلف را به خود بگیرد. که این تعداد بسیار زیاد است و باید مدل به قدری پیچیده باشد تا بتواند این حجم از تنوع و پیچیدگی را مدیریت کند.

نقش‌های مختلفی که هر واژه در قالب IOB می‌تواند بگیرد نیز در ‏جدول 2-2 دیده می‌شود.

نقش‌های مختلف واژه‌ها در چانکر

|  |
| --- |
| B |
| O |
| I |
| O+B |
| B+O |
| O+O |
| O+V |
| B+I |
| I+I |

همان‌طور که مشاهده می‌شود هر واژه می‌تواند ۹ کلاس مختلف به خود بگیرد که این نیز دلیل دیگری بر پیچیدگی طراحی چانکر است.

برای دریافت اطلاعات بیشتر از پروژه با ایمیل [zekavat@itrc.ac.ir](mailto:zekavat@itrc.ac.ir) تماس بگیرید.