در مجموعه مقالات سومین همایش ملی زبان‌شناسی پیکره‌ای، صفحات ۱۸۱-۲۰۷

معرفی داده استاندارد طلایی در سطح معنا برای هم‌نگاره‌های فارسی

مسعود قیومی
پژوهشگاه علوم انسانی و مطالعات فرهنگی
M.Ghayoomi@ihcs.ac.ir

چکیده

واژه کوچکترین عنصر زبانی است که حاوی اطلاعات وسیعی از ساخت آوایی تا نحو و معنا می‌باشد. واژه‌های یک زبان معمولاً بیش از یک مقوله دستوری و معنا دارند. این ویژگی موجب می‌شود به‌هنگام پردازش زبان با کمک رایانه دچار چالش شویم؛ چراکه نحوه پردازش داده زبانی باید به گونه‌ای باشد که نقش دستوری یا معنای واژه باتوجه‌به بافت محلی در جمله صحیح تشخیص داده شود. این مقاله به معرفی یک داده استاندارد طلایی تهیه‌شده در سطح معنا برای واژه‌های فارسی می‌پردازد که دارای صورت نوشتاری یکسان (هم‌نگاره) بوده و رایانه به‌هنگام تشخیص معنای صحیح واژه با چالش مواجه می‌گردد. این داده برای تعداد ۲۰ واژه هدف فارسی تهیه شده‌است و تعداد ۱۰۰ جمله مستخرج از یک پیکره زبانی برای این واژه‌ها، به‌صورت دستی برچسب‌گذاری معنایی انجام می‌شود. سپس، این داده به ساختار استاندارد تبدیل می‌گردد. از این تعداد داده، نشانه‌گذاری ۴ واژه منتخب به ۵ برابر افزایش می‌یابد تا کارایی روش یادگیری ماشینی بامربی در طبقه‌بندی که در «ابهام‌زدایی معنایی واژه» به‌کار می‌رود را با روش یادگیری ماشینی بی‌مربی در خوشه‌بندی که در «استنتاج استقرایی معنای واژه» به‌کار می‌رود قابل‌مقایسه گردد.

کلیدواژه‌ها: پیکره زبانی، داده استاندارد طلایی، نشانه‌گذاری داده، ابهام معنایی، طبقه‌بندی، خوشه‌بندی

۱. مقدمه

زبان‌شناسی رایانشی یک میان‌رشته‌ای است که به دو حوزه زبان‌شناسی و حوزه فناوری اطلاعات مرتبط است. معمولاً پژوهش‌های انجام‌گرفته در این میان‌رشته از نوع کاربردی است. زبان‌شناسی رایانشی به اکثر حوزه‌های پنج‌گانه زبان‌شناسی، یعنی آواشناسی، صرف، نحو، معناشناسی و تجزیه و تحلیل کلام، ورود پیدا کرده‌است. استفاده از زبان‌شناسی رایانشی در دنیای امروز امری اجتناب‌ناپذیر است. ازآنجاکه پژوهشگران متعددی در حوزه زبان‌شناسی رایانشی مشغول به فعالیت پژوهشی هستند، خروجی آنها که معمولاً یک سامانه پردازش زبان است باید مورد ارزیابی قرار گیرد تا برتری یک پژوهش بر دیگر پژوهش‌ها مشخص گردد. برای رسیدن به این هدف، با استفاده از یک داده ثابت که به «داده استاندارد طلایی»[1] معروف است می‌توان کارایی سامانه پردازش زبان را سنجید و به مقایسه ایده یا الگوریتم معرفی‌شده توسط پژوهشگر پرداخت.

این مقاله به معرفی یک داده استاندارد طلایی می‌پردازد که در آن معنای واژه‌هایی که صورت نوشتاری یکسان دارند در «بافت محلی»[2] خود در جمله تعیین شده‌است. این مقاله در ۷ بخش تهیه شده‌است. در بخش ۲، چارچوب نظری پژوهش معرفی می‌شود. بخش ۳ به چالش پردازش معنایی خط و زبان فارسی می‌پردازد. در بخش ۴، به‌طور فشرده تعدادی از پیکره‌های زبانی تهیه‌شده برای زبان فارسی که به‌عنوان داده استاندارد طلایی شناخته می‌شود معرفی می‌گردد. در بخش ۵، داده‌هایی که برای تهیه داده استاندارد طلایی مورد بحث در این مقاله به‌کار می‌رود معرفی می‌گردد. در ادامه این بخش، جدول معنایی واژه‌های هدف توضیح داده شده و این داده ساختارمند می‌گردد. در بخش ۶، یک تحلیل مقایسه‌ای بین روش یادگیری ماشینی

[1] gold standard data
[2] local context

بامربی[۳] و بی‌مربی[۴] انجام می‌پذیرد. در انتها در بخش ۷، به نتیجه‌گیری می‌پردازیم.

۲. چارچوب نظری تحقیق

دوسوسور (۱۹۱۶) معتقد است زبان دارای صورت[۵] و معنا[۶] است. ممکن است بین صورت (آوایی یا نوشتاری) و معنا رابطه یک‌به‌یک وجود نداشته باشد. ابهام معنایی یکی‌از ویژگی‌های «زبان طبیعی»[۷] در مقایسه با زبان‌های دیگر مانند «زبان برنامه‌نویسی»[۸] یا «زبان صوری»[۹] است. یکی‌از صور ابهام معنایی، واژگانی است. با درنظرداشتن روابط معنایی بین واژه‌ها می‌توان چندمعنایی[۱۰]، هم‌نامی[۱۱] و هم‌نویسگی[۱۲] (لاینز، ۱۹۸۱) را ازجمله عوامل ایجاد ابهام واژگانی که در خط متبلور است معرفی کرد. در چندمعنایی ممکن است یک واژه بیش‌از یک معنای مرتبط به هم داشته باشد، مانند «آب» در مثال‌های «آب‌های آزاد» و «لیوان آب». در هم‌نامی ممکن است دو یا چند واژه از نظر آوایی و نوشتاری مانند یکدیگر باشد ولی از نظر معنایی هیچ ارتباطی با یکدیگر نداشته باشد، مانند «شیر» در مثال‌های «شیر جنگل»، «لیوان شیر» و «شیر آب». در هم‌نویسگی نیز دو یا چند واژه ممکن است مانند یکدیگر

[3] supervised machine learning
[4] unsupervised machine learning
[5] form
[6] meaning
[7] natural language
[8] programming language
[9] formal language
[10] polysemy
[11] homonymy
[12] homography

نوشته شود ولی متفاوت از یکدیگر تلفظ گردد، مانند زنجیره حروف در «مهر» که می‌تواند به‌صورت‌های «مَهر»، «مِهر» و «مُهر» تلفظ می‌گردد.
هریس (۱۹۵۴) معتقد است معنای واژه از بافتی که در آن به‌کار می‌رود به‌دست می‌آید. ازاین‌رو، یکی‌از راه‌های رفع ابهام واژگانی، درنظرداشتن بافت زبانی وقوع واژه هدف است.

۳. چالش پردازش معنایی خط و زبان فارسی

صورت آوایی معرفی‌شده توسط دوسوسور (۱۹۱۶) در رایانه ازطریق خط در قالب داده گفتاری و نوشتاری تجلی می‌یابد. در زبان‌شناسی رایانشی این داده (گفتاری و نوشتاری) باید برای کاربردهای متنوع پردازش گردد. پردازش خط در داده نوشتاری فارسی چندان ساده نیست؛ چراکه با چالش‌های متعددی مواجه هستیم. قیومی و همکاران (۲۰۱۰) به‌صورت مبسوط به این مشکلات اشاره کرده‌اند. عمده‌ترین مشکلات، عدم نگارش مصوت‌های کوتاه در نوشتار فارسی و عدم رعایت فاصله‌گذاری صحیح و ایجاد مشکل واژه‌های چندقطعه‌ای (شریفی‌آتشگاه و بی‌جن‌خان، ۲۰۰۹) است. با وجود تلاش‌های انجام‌شده برای تهیه پیکره زبانی فارسی توسط افرادی چون عاصی (۱۹۹۷) و بی‌جن‌خان (۱۳۸۳)، هیچ‌یک از پیکره‌ها از نظر دو مشکل مطرح‌شده کاملاً معیار نشده‌است تا یک شرایط ایده‌آل برای استخراج اطلاعات صحیح فراهم آید. بر فرض وجود چنین پیکره‌ای، نمی‌توان با این پیکره یک سامانه پردازش زبان طبیعی را آموزش داد و از این سامانه برای پردازش داده‌های در شرایط واقعی با کارایی[۱۳] بالا استفاده نمود. با قبول اشکالات در خط تلاش شده‌است برای کاربردی‌شدن سامانه‌های پردازش خط و زبان فارسی، تا حدی از شرایط

[13] performance

ایده‌آل آزمایشگاهی فاصله گرفته و مواجهه با واقعیات مشکلات خط فارسی اجتناب‌ناپذیر است. در این رویکرد ممکن است تناظر یک‌به‌یک بین نظریه و کار عملی برقرار نگردد. نمونه عملی این شیوه، برچسب‌گذاری مقوله دستوری پیکره بی‌جن‌خان (بی‌جن‌خان، ۱۳۸۳) است که در آن تلاش شده‌است نحوه برچسب‌گذاری واژه‌ها با ویژگی‌ها و اشکالات خط فارسی منطبق شود. برای مثال، زنجیره حروف «وبا»، در مثال (۱) دارای برچسب N,COM,SING (اسم عام و مفرد) و در مثال (۲) دارای برچسب CONJ,PREP (حرف ربط، حرف اضافه) است.

(۱) او دچار بیماری وبا است.

(۲) او خرید کرد وبا دوستش به رستوران رفت.

پژوهش حاضر بر برچسب‌گذاری معنایی واژه در داده مستخرج از پیکره زبان فارسی متمرکز است. با ملاک‌قراردادن شکل نوشتاری موجود واژه‌های دارای صورت نوشتاری یکسان در پیکره فارسی و صرف نظر از علل ایجاد ابهام، کار برچسب‌گذاری معنایی واژه‌های هدف انجام می‌پذیرد. در این مقاله، مجموعه واژه‌های چندمعنا، هم‌نام و هم‌نویسه که صورت نوشتاری یکسان دارند و به‌عنوان علل ابهام واژگانی در خط نمود پیدا می‌کنند «هم‌نگاره» می‌نامیم. برای قابل‌استفاده‌شدن این داده برای کاربردهای واقعی، باید با درنظرگرفتن اشکالات در خط و شکل نگارش‌شده موجود در پیکره، کار نشانه‌گذاری انجام پذیرد.

۴. داده‌های استاندارد طلایی تهیه‌شده برای زبان فارسی

زبان فارسی همانند زبان‌های دیگر دارای داده‌های استاندارد طلایی نوشتاری و گفتاری است، با این وجود زبان فارسی همچنان جزء زبان‌های با منابع

محدود تلقی می‌شود. در این بخش، نمونه‌هایی از داده‌های استاندارد طلایی تهیه‌شده برای زبان فارسی به‌طور مختصر معرفی می‌گردد.

در حوزه آوایی، بی‌جن‌خان و همکاران (۱۹۹۴) مجموعه صوتی از عبارات و جملات به‌نام FarsDat تهیه کرده‌اند که از مناطق مختلف کشور ایران توسط گویش‌وران فارسی‌زبان جمع‌آوری شده‌است. مجیری و مینایی بیدگلی (۱۳۸۷) مجموعه داده عروض اشعار فارسی را برای ارزیابی سامانه‌های تشخیص خودکار وزن عروضی اشعار فارسی تهیه کرده‌اند.

برای داده متنی، آل‌احمد و دیگران (۲۰۰۹) داده پیکره همشهری را براساس ویژگی‌های TREC[14] استانداردسازی کرده‌اند تا برای بازیابی اطلاعات مورد استفاده قرار گیرد. بی‌جن‌خان و دیگران (۲۰۱۱) یک پیکره متنی را با هدف مدل‌سازی آماری زبان تهیه کرده‌اند. قسمتی از این داده که حاوی برچسب مقوله دستور است به پیکره بی‌جن‌خان (بی‌جن‌خان، ۱۳۸۳) معروف است.

در حوزه نحو، اولین دادگان درختی سازه‌ای به‌نام PerTreeBank توسط قیومی (۲۰۱۲) تهیه شده‌است. جملات این داده، براساس دستور ساخت سازه‌ای هسته‌بنیان (پولارد[15] و ساگ[16]، ۱۹۹۴) تحلیل شده‌است. در این پیکره، علاوه‌بر تجزیه درختی جملات، بن‌واژه‌ها، مقولات دستوری واژه‌ها و برچسب موجودیت‌های نامدار نیز مشخص شده‌است. قیومی و کوهن[17] (۲۰۱۴) این دادگان درختی سازه‌ای را به دادگان درختی وابستگی به‌نام DepPerTreeBank تبدیل کرده‌اند. سایر دادگان درختی وابستگی توسط سراجی و همکاران (۲۰۱۲) و رسولی و همکاران (۲۰۱۳) تهیه شده‌است. حسین‌نژاد و همکاران (۱۳۹۴)مقداری داده‌ای را به‌صورت سطحی و براساس

[14] http://trec.nist.gov/
[15] Pollard, C.
[16] Sag, I. A.
[17] Kuhn, J.

استاندارد IOB نشانه‌گذاری کرده‌اند. همچنین، در نوع دیگری از تحلیل نحوی، موسوی و همکاران (۲۰۰۹) ارجاع اسمی ضمایر را در حجمی از متون مشخص کرده‌اند.

در حوزه معنا، شمس‌فرد و همکاران (۲۰۱۰) شبکه واژگانی فارس‌نت را به‌روش نیمه‌خودکار تهیه کرده و نوع رابطه واژه‌ها با یکدیگر مانند ترادف، تضاد، شمول معنایی و غیره را مشخص کرده‌اند. خداپرستی (۱۳۷۶) فهرستی از واژه‌های متضاد و مترادف فارسی را تهیه کرده و عبدوس و همکاران (۱۳۹۴) پیکره‌ای به‌نام «پیکره واحدهای اسمی آرمان» برای تشخیص موجودیت‌های نامدار[18] تهیه کرده‌اند.

همان‌گونه که مشاهده می‌گردد، علیرغم تلاش‌های انجام‌شده برای نشانه‌گذاری داده زبان فارسی برای موضوعات مختلف، داده استاندارد طلایی در حوزه معناشناسی و ابهام واژگانی تهیه نشده‌است. پژوهش حاضر به کاهش این خلاء در منابع نشانه‌گذاری‌شده زبان فارسی می‌پردازد.

۵. اقدامات انجام‌شده برای برچسب‌گذاری معنایی واژه‌های هدف

۱.۵. داده ورودی

در انجام این پژوهش از پایگاه داده زبان فارسی (عاصی، ۱۹۹۷) استفاده می‌شود. این پیکره عاری از برچسب معنایی واژه است و به‌دلیل عدم معیارسازی کل داده، در این پژوهش، سخت‌ترین شرایط پردازش داده فارسی مورد توجه قرار گرفته‌است تا بتوان با الگوریتم‌های پردازش زبانی به نتایج قابل قبولی از واقعیت پردازش خط فارسی دست یافت. برای هدف مورد نظر، ۲۰ واژه به‌عنوان واژه‌های هدف از فارس‌نت (شمس‌فرد و همکاران، ۲۰۱۰) استخراج شد و به‌صورت تصادفی تعداد ۱۰۰ جمله از پایگاه داده زبان فارسی

[18] named entity recognition

انتخاب شد. سپس، نشانه‌گذاری معنایی این واژه‌های هدف توسط یک انسان خبره براساس جدول معنایی برای هر واژه انجام پذیرفت. لازم به ذکر است در تهیه این جداول معنایی، صرف نظر از علل ایجاد ابهام واژگانی، فقط صورت نوشتاری واژه هدف در پیکره ملاک بوده‌است.

با نگاهی به معانی ارائه‌شده در فارس‌نت برای این ۲۰ واژه هدف مشاهده شد که بعضی‌از معانی واژه‌ها تکراری است و یا بهتر است بعضی از معانی واژه‌ها کلی‌تر عنوان شود، مانند «شیر خشک» به‌عنوان یکی‌از معانی «شیر». باتوجه‌به استخراج جملات شاهد از پایگاه داده زبان فارسی برای واژه‌های هدف هم‌نگاره، نیاز بود معانی که در فارس‌نت موجود نبود به مجموعه معانی اضافه گردد، مانند واژه «نبرد» که در فارس‌نت ۳ معنا تعریف شده بود ولی به‌هنگام نشانه‌گذاری داده، تعداد معانی به ۸ معنا افزایش یافت. تغییرات در تعداد معانی واژه‌های هدف در جدول ۱ گزارش شده‌است.

جدول ۱: تعداد معانی استخراج‌شده از فارس‌نِت برای واژه‌های هدف

واژه هدف	فارس‌نت	پژوهش حاضر	واژه هدف	فارس‌نت	پژوهش حاضر
برداشت	۴	۵	سیر	۷	۶
پروانه	۴	۴	شیر	۱۳	۷
تار	۹	۸	کرم	۱۲	۹
تن	۸	۶	کره	۷	۵
تند	۷	۵	گرد	۶	۶
تیز	۵	۵	گل	۱۱	۶
روان	۵	۵	ملک	۶	۶
ریش	۲	۶	مهر	۱۰	۸
سبک	۸	۳	نبرد	۳	۸
سر	۲۴	۱۲	نشسته	۳	۳

۲.۵. جدول معانی

معانی مختلف صورت نوشتاری واژه‌های هدف به ذکر مثال در جداول ۲ تا ۲۱ ارائه شده‌است.

جدول ۲: معانی مختلف صورت نوشتاری «برداشت»

ردیف	معنا	مثال
۱	استنباط	برداشت از سخن
۲	جمع‌آوری	برداشت محصول
۳	برداشتن قسمتی از سرمایه یا سود	برداشت از حساب
۴	مفهوم فعل سبک داشتن	او کتاب را برداشت
۵	سایر موارد	برداشت فیلم

جدول ۳: معانی مختلف صورت نوشتاری «پروانه»

ردیف	معنا	مثال
۱	جواز	پروانه کسب
۲	نوعی حشره	بال‌های پروانه
۳	اسم خاص	پروانه خانم
۴	سایر موارد	شنای پروانه

جدول ۴: معانی مختلف صورت نوشتاری «تار»

ردیف	معنا	مثال
۱	رشته‌های باریک	تارعنکبوت، تار مو، تار ابریشم
۲	رشته‌هایی از سیم برای ساخت ساز زهی	سیم تار
۳	ساز زهی مضرابی	سه تار
۴	پارچه‌ای عمودی، فرش و غیره که پود از لابه‌لای آن رد می‌شود	تار و پود پارچه
۵	ناواضح، تیره	تیره و تار
۶	اصلی، پایه، بنیان، ریشه، عمق	این ویژگی اخلاقی در تار و پود او ریشه دوانده است.

۷	پریشان و نیست و نابود	تار و مار
۸	سایر موارد	اسم خاص مانند دریاچه تار

جدول ۵: معانی مختلف صورت نوشتاری «تن»

ردیف	معنا	مثال
۱	کیفیت صحبت کردن فرد	تن صدا
۲	واحد اندازه گیری وزن معادل هزار کیلوگرم	۵ تن محصول
۳	واحد شمارش انسان	۲۰ تن کشته شدند
۴	ساختار فیزیکی و شکل ظاهری یک موجود زنده	تن او آسیب دید.
۵	ماهی تن	تن ماهی
۶	سایر موارد	تن دادن

جدول ۶: معانی مختلف صورت نوشتاری «تند»

ردیف	معنا	مثال
۱	سریع	تند برو و برگرد.
۲	قوی، شدید	آفتاب تند
۳	پرادویه	غذای تند
۴	توأم با خشم. خشمگین	تند خو
۵	زاویه کمتر از ۹۰ درجه	زاویه تند

جدول ۷: معانی مختلف صورت نوشتاری «تیز»

ردیف	معنا	مثال
۱	باهوش	بچه تیز
۲	برنده	چاقوی تیز، نوک نیز
۳	هرچیزی که طعم آن تیز باشد و زبان را بسوزاند	تند و تیز
۴	شتابان	تیز برو و برگرد.
۵	شدید، قوی	گوش تیز

جدول ۸: معانی مختلف صورت نوشتاری «روان»

ردیف	معنا	مثال
۱	جان - روح - روان	روح و روان شاد
۲	روان - سلیس - فصیح	شعر سلیس و روان
۳	جاری - روان - در حال حرکت و جریان	آب روان
۴	آسان کردن کاری، روان کردن روند کار اداری	بررسی پرونده روان در حال انجام است.
۵	سایر موارد	اسم خاص مانند روان اتکینسون

جدول ۹: معانی مختلف صورت نوشتاری «ریش»

ردیف	معنا	مثال
۱	محاسن	ریش سفید پیرمرد
۲	تعهد	ریش گرو گذاشتن
۳	دادن اختیار به کسی	ریش و قیچی دستت است.
۴	پاره پاره شدن	ریش شدن پارچه
۵	جراحت	دلم ریش شد
۶	سایر موارد	اسم خاص یا لقب فرد مانند صادق ریش

جدول ۱۰: معانی مختلف صورت نوشتاری «سبک»

ردیف	معنا	مثال
۱	بی ارزش و ناخوشایند	رفتارسبک شخص
۲	دارای وزن کم	صندلی سبک
۳	روش، چگونگی انجام دادن کاری یا رخ دادن چیزی	سبک زندگی

جدول ۱۱: معانی مختلف صورت نوشتاری «سر»

ردیف	معنا	مثال
۱	عضو بدن	سر او درد می‌کرد.
۲	لیز	زمین سر است.
۳	راز	او سر نگهدار است
۴	اول	سر خط
۵	دهانه	سر بطری؛ سر قابلمه
۶	اندیشه	از سر گذراندن
۷	کرخت	دندانش را سر کرد.
۸	بالای چیزی	سر درخت؛ سر کوه
۹	برتر	او از همه سر است.
۱۰	درباره چیزی	دعوا بر سر قانون است.
۱۱	سمت یا طرف	یک سر فلان افراد هستند، سر دیگر بقیه
۱۲	سایر موارد	به سر بردن

جدول ۱۲: معانی مختلف صورت نوشتاری «سیر»

ردیف	معنا	مثال
۱	دارای حالت سیری، اشباع	گرسنه و سیر
۲	دارای رنگ تند	آبی سیر
۳	گیاهی از خانواده سوسن با بنه‌ای متورم و چند قسمتی	سیر و پیاز
۴	واحد اندازه گیری وزن معادل ۷۵ گرم	یک سیر گوشت
۵	روند، حرکت از جایی به جایی	سیر تغییر، سیر صعودی
۶	بررسی عمیق کردن,در اندیشه حافظ سیر کردن	سیر و سلوک عارف

جدول ۱۳: معانی مختلف صورت نوشتاری «شیر»

ردیف	معنا	مثال
۱	حیوان	شیر جنگل
۲	فلکه	شیر آب
۳	نوشیدنی	شیر گاو
۴	شجاع	دل شیر دارد.
۵	آن روی سکه که دارای تصویر شیر بوده	شیر یا خط
۶	به اشتراک گذاری	شیر کردن
۷	سایر موارد	اسم خاص مانند محمد چرم شیر

جدول ۱۴: معانی مختلف صورت نوشتاری «کرم»

ردیف	معنا	مثال
۱	سخاوت	کرم او بسیار است
۲	نوعی حشره	مثل کرم می‌لولید.
۳	موادی آرایشی	کرم نرم‌کننده
۴	رنگ سفید مایل به زردی	لباس او کرم است.
۵	نوعی خمیر نرم جهت تهیه انواع شیرینی	کرم خامه‌ای
۶	نوعی فلز،نگارش دیگر کلمه کروم	فلز کرم
۷	بدافزاری که بخش‌های رایانه را از کار می‌اندازد	کرم کامپیوتر
۸	اسم خاص	آقای کرم زاده
۹	سایر موارد	کرم کتاب

جدول ۱۵: معانی مختلف صورت نوشتاری «کره»

ردیف	معنا	مثال
۱	مادهٔ سفید یا زردرنگی که از تکان دادن سریع شیر یا دوغ حاصل می‌شود	کره مربا
۲	بچه چهارپایان	کره خر
۳	سیاره	کره زمین

۴	نام کشور	کره شمالی
۵	سایر موارد	اسم خاص مانند کره رود

جدول ۱۶: معانی مختلف صورت نوشتاری «گرد»

ردیف	معنا	مثال
۱	غبار	گرد و خاک
۲	اطراف/پیرامون	گرد میز
۳	هر چیز مدور و دایره شکل	چوب گرد
۴	اشکال دارویی که به شکل پودر عرضه می‌شود	گرد پنی‌سیلین
۵	دلیر ، پهلوان	گرد و ستبر
۶	فعل گشتن	بر گرد

جدول ۱۷: معانی مختلف صورت نوشتاری «گل»

ردیف	معنا	مثال
۱	خاک به آب آمیخته	گل کاری
۲	گردن	گل میخ
۳	وارد شدن توپ به دروازهٔ حریف	آنها گل زدند.
۴	گیاه	گل سرخ
۵	مرغوب و شاخص	گل میوه
۶	اسم خاص	ابراهیم گل

جدول ۱۸: معانی مختلف صورت نوشتاری «ملک»

ردیف	معنا	مثال
۱	نامی از نامهای صفات الهی	مَلِک
۲	نامی از نامهای صفات الهی	مَلِک
۳	پادشاه	مَلِک,ملک عبدالله
۴	هرآنچه در تصرف کسی باشد و مالک آن بود	مِلْک
۵	پادشاهی	مُلک,ملک سلیمان
۶	اسم خاص	آقای ملک محمدی

جدول ۱۹: معانی مختلف صورت نوشتاری «مهر»

ردیف	معنا	مثال
۱	پول یا چیزی که هنگام عقد نکاح بر ذمهٔ مرد مقرر می‌شود	مهر دختر
۲	مهربانی	مهر مادرانه
۳	ابزار	مهر دفتر اسناد، مهر نماز
۴	آفتاب	آفتاب مهر
۵	ماهی از سال	ماه مهر
۶	نشان	مهر بدنامی بر او زدند.
۷	خبرگزاری	خبرگزاری مهر
۸	اسم خاص	مهر و ماه

جدول ۲۰: معانی مختلف صورت نوشتاری «نبرد»

ردیف	معنا	مثال
۱	جنگ	نبرد با زندگی؛ نبرد حق علیه باطل
۲	بردن در مسابقه (فعل منفی شده)	او مسابقه را نبرد.
۳	بریدن (فعل منفی شده)	مواظب باش تیغ دستت را نبرد.
۴	بردن در جابجایی (فعل منفی شده)	کتاب را همراه خودش نبرد.
۵	مسابقه	نبرد استقلال و پرسپولیس
۶	فعل سبک	دشمن را از بین نبرد.
۷	اسم خاص	خیابان نبرد
۸	فعل سبک (فعل منفی شده)	او نامی از دانش‌آموزان ممتاز نبرد

جدول ۲۱: معانی مختلف صورت نوشتاری «نشسته»

ردیف	معنا	مثال
۱	شستن (فعل منفی شده)	لباس‌های نشسته را شستم.
۲	نِشستن	روی زمین نشسته است.
۳	فعل سبک	تلاش‌ها به بار نشسته است.

۵ . ۳. استانداردسازی داده

فناوری اطلاعات (لویت[19] و وسلر[20]، ۱۹۵۸) شاخه‌ای از علوم طبیعی است که در آن باید اطلاعات به‌صورت منظم در رایانه ساماندهی شود. داده زبانی به‌عنوان مواد لازم در پژوهش‌های زبان‌شناختی برای کاربرد در حوزه رایانه، باید به‌طور منظم پیکربندی گردد. قیومی (۱۳۹۷) به اهمیت این موضوع اشاره کرده‌است و شیوه استانداردسازی داده برای سطح زبان‌شناسی آوایی، صرفی و نحوی را توضیح داده‌است. ازآنجاکه این مقاله در حوزه معناشناسی متمرکز است نیاز به ارائه استاندارد لازم برای این سطح زبان‌شناسی است.

در سال ۱۹۹۸ کارگاهی به‌نام «کارگاه ارزیابی معنایی» با هدف بررسی معنای واژه با کمک رایانه برگزار شد. از آن تاریخ به بعد، همه‌ساله، این کارگاه با اهداف خاص برگزار می‌گردد. در «کارگاه ارزیابی معنایی سال ۲۰۱۰» که به SemEval2010 معروف است چندین فعالیت در حوزه معناشناسی رایانشی مورد توجه قرار گرفت. فعالیت شماره ۱۴ این کارگاه در قالب یک مسابقه به موضوع ابهام‌زدایی واژه و استنتاج استقرایی معنایی واژه پرداخت. ازاین‌رو، داده آموزش و داده آزمون به زبان انگلیسی برای این مسابقه تهیه و در دسترس شرکت‌کنندگان قرار گرفت. این داده براساس زبان نشانداری XML ساختارمند شده‌است. در تصویر (۱) و (۲) نحوه ساختارمندسازی داده آموزش و داده آزمون برای واژه هدف «access» نمایش شده است.

برای این مسابقه، ابزار محاسبه کارایی سامانه‌های طراحی‌شده توسط شرکت‌کنندگان تهیه شده‌است.[21] این ابزار دو شاخص F (ون‌ریجسنبرگ[22]،

[19] Leavitt, H.
[20] Whisler, T.L.
[21] https://www.cs.york.ac.uk/semeval2010_WSI/datasets.html
[22] Van Rijsbergen, C. J.

<access.n.train><access.n.1>The Commission seeks comment on whether the analytical framework that was used to streamline AT & T 's services should be applied to incumbent LEC access services. In particular , the Commission seeks comment on which of the factors that it used in examining AT & T 's pricing behavior could be used to determine when to remove incumbent LEC access services from price cap regulation. It cites demand elasticity , supply elasticity , market share , and the pricing of services under price cap regulation as relevant factors . </access.n.1></access.n.train>

تصویر ۱: ساختارمندی داده آموزش در SemEval2010

<access.n.test><access.n.1>The port is famous for wind shelter , deep water , non-accumulation of mud and sand , short sea - routes and long , usable coastline . <TargetSentence>In more than four years , 2.2 billion yuan has been invested in the construction of harbors and docks , storage fields , support facilities and infrastructure of the ports and city , creating good conditions for building access to the sea for the Great Southwest . </TargetSentence>Currently , there are 25 berths in the whole city , of which 10 berths are of 10,000 ton level . </access.n.1></access.n.test>

تصویر ۲: ساختارمندی داده آزمون در SemEval2010

۱۹۷۹) و شاخص V (روزنبرگ[۲۳] و هیرشبرگ[۲۴]، ۲۰۰۷) را برای داده‌ها محاسبه می‌کند. شاخص F فقط همگنی[۲۵] و شاخص V علاوه‌بر همگنی، کلیت[۲۶] را نیز در بر می‌گیرد. نحوه محاسبه این دو شاخص خارج از این بحث است ولی فقط به تفاوت این دو شاخص بسنده می‌شود. در شاخص F فقط توزیع دسته‌ها[۲۷] در خوشه‌ها[۲۸] مورد توجه قرار می‌گیرد و به این نکته که در

[23] Rosenberg, A.
[24] Hirschberg, J.
[25] homogeneity
[26] completeness
[27] class
[28] cluster

هر خوشه تعداد دسته‌ها به حداقل رسیده‌است یا خیر توجهی ندارد. درحالی‌که در شاخص V، علاوه‌بر توزیع دسته‌ها در خوشه‌ها، به این نکته که در هر خوشه تعداد دسته‌ها به حداقل رسیده‌است یا خیر توجه دارد و بهترین حالت کلیت زمانی اتفاق می‌افتد که تمام نمونه‌های داخل یک دسته در یک خوشه قرار گیرد.

ساختار داده ورودی این ابزار ارزیابی داده، حاوی سه ستون است: ستون اول حاوی واژه هدف و مقوله دستوری آن واژه؛ ستون دوم حاوی واژه هدف، مقوله دستوری آن واژه و اندیس جمله در داده آزمون؛ و ستون سوم حاوی واژه هدف، مقوله دستوری آن و اندیس معنایی واژه هدف در جمله مورد نظر. در این پژوهش تلاش شده‌است داده تهیه‌شده برای واژه‌های دارای ابهام فارسی با کمی تغییر براساس استاندارد SemEval2010 بدون احتساب مقوله دستوری واژه ساختارمند گردد. در تصویر (۳) و (۴) ساختارمندی داده واژه هدف «برداشت» در داده آموزش و داده آزمون نمایش داده شده‌است. برای نادیده‌گرفتن مقوله دستوری تعریف‌نشده واژه unkn به‌عنوان برچسب مقوله دستوری برای همه واژه‌ها به‌کار رفته‌است تا امکان استفاده از ابزار ارزیابی در SemEval2010 میسر گردد.

۶. تحلیل مقایسه‌ای روش‌های یادگیری ماشینی

در بخش پیشین گفته شد که برای ۲۰ واژه هدف فارسی با صورت نگارشی یکسان که دارای ابهام معنایی هستند، ۱۰۰ جمله نشانه‌گذاری و براساس SemEval2010 ساختارمند شده‌است. برای انجام یک بررسی مقایسه‌ای، نشانه‌گذاری معنایی جملات ۴ واژه «برداشت»، «کرم»، «ملک» و «مهر»، بدون هم‌پوشانی، از ۱۰۰ جمله به ۶۰۰ جمله افزایش پیدا کرد. هدف از افزایش تعداد جملات، استفاده از ۵۰۰ جمله نشانه‌گذاری‌شده به‌عنوان داده آموزش است.

<train.برداشت.1>
کشور قطر از سال 1991 برداشت گاز از این میدان را آغاز کرده و این در حالی است که ایران ، برداشت از این منبع خدادادی را از دو سال پیش آغاز کرده‌است .
</train.برداشت.1>

تصویر ۳: ساختارمندی داده آموزش فارسی در SemEval2010

<test.برداشت.1>
به گزارش خبرنگار همشهری به نقل از یکی‌از نمایندگان در ابتدای این جلسه آقای نیازی گفت : برخوردهای اولیه ما با متهمان به دلیل <targetWord senseID="1">برداشت</targetWord> ما از حزب اللهی بودن آنها دقیق و بازجویی‌ها نیز نسبت به دیگران متفاوت بوده‌است .
</test.برداشت.1>

تصویر ۴: ساختارمندی داده آزمون فارسی در SemEval2010

در این بررسی، دو موضوع ابهام‌زدایی معنایی واژه و استنتاج استقرایی معنای واژه با یکدیگر مقایسه می‌شود. کار ابهام‌زدایی معنایی واژه این است که ازمیان مجموعه مشخص معانی، مدل بتواند معنای صحیح واژه در آن بافت محلی در جمله را مشخص نماید. روشی که برای اجرای ابهام‌زدایی معنایی واژه به‌صورت خودکار به‌کار می‌رود استفاده از یک طبقه‌بند[۲۹] است که جزء روش‌های یادگیری ماشینی بامربی محسوب می‌شود. در این روش، طبقه‌بند یک مدل از داده نشانه‌گذاری‌شده می‌سازد و از آن مدل برای نشانه‌گذاری داده‌های دیگر دیده‌نشده استفاده می‌نماید. از میان انواع الگوریتم‌های طبقه‌بندی، از الگوریتم ماشین بردار پشتیبان[۳۰] (کورتس[۳۱] و واپ‌نیک[۳۲]، ۱۹۹۵) استفاده می‌نماییم. دلیل استفاده از این الگوریتم در مقایسه با سایر

[29] classifier
[30] Support Vector Machine (SVM)
[31] Cortes, C.
[32] Vapnik, V.N.

الگوریتم‌های طبقه‌بندی این است که تعمیم زیادی را برای داده آموزش قائل می‌شود (کودو[۳۳] و ماتسوموتو[۳۴]، ۲۰۰۱).

پس‌از آموزش طبقه‌بند با ۵۰۰ جمله برچسب‌دار، با نادیده‌گرفتن برچسب معنایی ۱۰۰ جمله داده آزمون، مدل به‌دست‌آمده باید کار نشانه‌گذاری داده آزمون را انجام دهد. از مقایسه داده آزمون برچسب‌گذاری‌شده توسط ماشین و داده آزمون حاوی برچسب استاندارد طلایی می‌توان کارایی مدل را براساس شاخص‌های F و V معرفی‌شده در SemEval2010 مورد ارزیابی قرار داد.

موضوع استنتاج استقرایی معنای واژه به تشخیص معنای واژه می‌پردازد؛ ازاین‌رو، هدف استنتاج استقرایی معنای واژه با ابهام‌زدایی معنایی واژه کاملاً متفاوت است. روشی که برای اجرای استنتاج استقرایی معنای واژه به‌صورت خودکار به‌کار می‌رود استفاده از یک الگوریتم خوشه‌بندی[۳۵] است که جزء روش‌های یادگیری ماشینی بی‌مربی محسوب می‌شود. بنابراین در این روش به هیچ داده آموزشی نیاز نیست و الگوریتم خوشه‌بندی باید باتوجه‌به کاربرد واژه در بافت محلی واژه در تعدا زیادی جملات، تفاوت یا تشابهات معنایی واژه هدف را مشخص نماید و واژه را در خوشه‌های مختلف قرار دهد. تعداد این خوشه‌ها مبیّن تعداد معانی واژه‌ها است. الگوریتم خوشه‌بندی که برای اهداف این بخش از پژوهش مورد استفاده قرار گرفته‌است الگوریتم کی‌مینز (مک‌کویین[۳۶]، ۱۹۶۷) است. این الگوریتم، پارامتری است که به مشخص‌شدن تعداد خوشه‌ها به‌هنگام اجرای الگوریتم نیاز دارد. این پارامتر براساس تعداد معانی موجود در داده آزمون مشخص می‌شود. داده خامی که به این الگوریتم داده می‌شود ترکیبی از جملات خام شامل واژه هدف و جملات آزمون بدون برچسب معنایی آنها است و الگوریتم خوشه‌بندی باید معنای واژه هدف را

[33] Kudo, T.
[34] Matsumoto, Y.
[35] clustering algorithm
[36] MacQueen, J.

مشخص نماید. در مرحله بعد باید جملات مربوط به داده آزمون به‌همراه اندیس خوشه مشخص‌شده توسط ماشین، از سایر جملات، در خروجی به‌دست‌آمده جدا گردد تا بتواند با داده استاندارد طلایی مقایسه گردد.

در چارچوب فرایند معرفی‌شده و مقایسه خروجی دو مدل حاصل از دسته‌بندی و خوشه‌بندی، جزئیات نتایج به‌دست‌آمده از عملکرد این دو روش یادگیری ماشینی برای واژه‌های هدف به‌همراه میانگین عملکرد مدل‌ها در جدول ۲۲ براساس دو شاخص F و V گزارش شده‌است.

جدول ۲۲: عملکرد الگوریتم‌های خوشه‌بندی و طبقه‌بندی

	خوشه‌بندی		**طبقه‌بندی**	
	شاخص F	**شاخص V**	**شاخص F**	**شاخص V**
برداشت	۴۶/۷۳	۳۴/۹۰	۲۹/۴۵	۷/۶۰
کرم	۵۲/۲۳	۵۸/۴۰	۲۴/۶۱	۱۳/۶۰
ملک	۶۰/۵۴	۶۱/۳۰	۳۰/۳۵	۱۲/۶۰
مهر	۴۲/۹۲	۴۸/۹۰	۲۷/۱۶	۱۳/۴۰
میانگین نتایج	۵۰/۶۱	۵۰/۸۸	۲۷/۸۹	۱۱/۸۰

براساس نتایج گزارش‌شده، از میان دو شاخص ارزیابی معرفی‌شده، شاخص V از شاخص F قابل‌اعتمادتر است و همان‌طور که دیده می‌شود شاخص V سه واژه «کرم»، «ملک» و «مهر» که تعداد معانی بیشتری نسبت‌به واژه «برداشت» دارند، امتیاز بالاتر نسبت‌به معیار F به‌دست آورده‌است. همان‌گونه که از میانگین نتایج به‌دست‌آمده از ارزیابی کارایی خوشه‌بندی و طبقه‌بندی برای چهار واژه هدف براساس دو شاخص F و V مشاهده می‌شود، کارایی خوشه‌بندی در یادگیری بی‌مربی بالاتر از طبقه‌بندی در یادگیری بامربی است.

در پژوهش حاضر برای هر چهار واژه هدف، میزان داده آموزش در الگوریتم یادگیری بامربی برابر ۵۰۰ جمله بوده‌است. درحالی‌که میزان جملات به‌کارگرفته‌شده در الگوریتم یادگیری بی‌مربی، مطابق جدول ۲۳ برای واژه‌های مختلف متفاوت است.

جدول ۲۳: تعداد جملات خام در الگوریتم‌های خوشه‌بندی و دسته‌بندی

واژه	خوشه‌بندی	دسته‌بندی
برداشت	۱۲۲۷۷	۵۰۰
کرم	۱۹۳۸	۵۰۰
ملک	۶۱۰۰	۵۰۰
مهر	۱۴۰۱۱	۵۰۰

حجم داده در روش یادگیری بی‌مربی نشان می‌دهد که در این رویکرد به‌دلیل عدم نیاز به داده برچسب‌خورده، تنها داده خام مورد نیاز است که تهیه آن به‌آسانی امکان‌پذیر است و می‌تواند منجر به نتایج قابل‌قبول در استنتاج استقرایی معنای واژه شود. بدیهی است با گسترش داده برچسب‌دهی‌شده می‌توان افزایش دقت سیستم را انتظار داشت.

۷. خلاصه و نتیجه‌گیری

در این مقاله به معرفی یک داده استاندارد طلایی که برای دو کار مهم در حوزه درک زبان طبیعی بسیار پرکاربرد است پرداختیم. این داده با حجم ۲۰۰۰ جمله برای ۲۰ واژه هدف فارسی با صورت نوشتاری یکسان و معانی متفاوت به‌صورت دستی تهیه شد. همچنین برای چهار واژه دیگر، ۲۰۰۰ جمله دیگر به حجم داده نشانه‌گذاری اولیه اضافه شد که در مجموع به یک دادگان ۴۰۰۰ جمله‌ای منجر شد. سپس، این داده براساس استاندارد شناخته‌شده SemEval2010 ساختارمند گردید. داده تهیه‌شده در دو

بستر یادگیری بامربی و بی‌مربی که در دو فعالیت اصلی مربوط به درک زبان طبیعی با عناوین ابهام‌زدایی معنای واژه و استنتاج استقرایی معنای واژه مطرح می‌شود، مورد استفاده و مقایسه قرار گرفت.

براساس نتایج به‌دست‌آمده، کارایی خوشه‌بندی که از یادگیری بی‌مربی برای استنتاج استقرایی معنای واژه استفاده می‌کند بالاتر از طبقه‌بندی است که از یادگیری بامربی برای ابهام‌زدایی معنایی واژه استفاده می‌کند. این تفاوت بیانگر دو نکته است:

نکته اول: می‌دانیم معمولاً روش یادگیری بامربی کارایی بالاتری از یادگیری بی‌مربی دارد و دلیل آن دسترسی به داده استاندارد طلایی برای ساخت مدل اولیه است؛ ولی تهیه این داده سخت، طاقت‌فرسا و زمان‌بر است و به همین دلیل فراهم نمودن حجم بالایی از داده بسیار دشوار می‌باشد.

نکته دوم: برای به‌دست‌آوردن کارایی بالا در روش یادگیری بامربی، به حجم قابل قبول نیاز است تا به ساخت مدل مناسب منجر شود؛ در صورت وجود داده آموزش کم، علیرغم تلاش برای تهیه داده اولیه، الزاماً به نتیجه مطلوب منتهی نمی‌شود.

منابع

بی‌جن‌خان، م. (۱۳۸۳) "نقش پیکره زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای". *مجله زبانشناسی*، ۱۹ (۲): ۴۸ـ۶۷.

حسین‌نژاد، ش.، ط. امامی آزادی، و ی. شکفته (۱۳۹۴) "تهیهٔ پیکرهٔ گروه‌های نحوی استاندارد برای تجزیهٔ سطحی جملات فارسی". در *مجموعه مقالات اولین همایش ملی زبان‌شناسی پیکره‌ای*.

خداپرستی، ف. (۱۳۷۶) *فرهنگ جامع واژگان مترادف و متضاد زبان فارسی*. شیراز: دانشنامهٔ فارس.

عبدوس، م.، ب. مینایی بیدگلی، و ح.ر. قدمنان (۱۳۹۴) "تولید پیکرهٔ واحدهای اسمی فارسی". در *مجموعه مقالات اولین همایش ملی زبان‌شناسی پیکره‌ای*.

قیومی، م. (۱۳۹۷) "پیکرهٔ زبانی و ضرورت ساختارمندسازی داده". در *مجموعه چکیده‌های دهمین همایش زبان‌شناسی ایران*.

مجیری، م. م.، و ب. مینایی بیدگلی (۱۳۸۷) "تشخیص وزن عروضی اشعار فارسی: کاربرد جدیدی از متن‌کاوی". ارائه‌شده در *دومین کنفرانس داده‌کاوی ایران*.

AleAhmad, A., H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian (2009) “Hamshahri: A standard Persian text collection”. *Knowledge-Based Systems*, 22(5): 382–387.

Assi, S.M. (1997) “Farsi linguistic database (FLDB)”. *International Journal of Lexicography*, 10(3):5.

Bijankhan, M., J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi. “Lessons from building a Persian written corpus: Peykare”. *Language Resources and Evaluation*, 45(2):143–164, 2011.

Bijankhan, M., J. Sheikhzadegan, M.R. Roohani, Y. Samareh, C. Lucas, and M. Tebyani (1994) “FARSDAT-The speech database of Farsi spoken language”. In *Proceedins of the Australian Conference on Speech Science and Technology*, 826-831.

Cortes, C., and V. N. Vapnik (1995) "Support-vector networks". *Machine Learning*. 20 (3): 273–297.

de Saussure, F. (1916) *Cours de linguistique générale*, Lausanne, Paris: Payot.

Ghayoomi, M. (2012) “Bootstrapping the development of an HPSG-based treebank for Persian”. *Linguistic Issues in Language Technology*, 7(1).

Ghayoomi, M., and J. Kuhn (2014) “Converting an HPSG-based treebank into its parallel dependency-based treebank”. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 802–809.

Ghayoomi, M., S. Momtazi, and M. Bijankhan (2010) "A study of corpus development for Persian", *International Journal on Asian Language Processing*, 20(1): 17-33.

Harris, Z.S. (1954) "Distributional Structure". *Word*, 23, 146-162.

Kudo, T., and Y. Matsumoto (2001) “Chunking with support vector machines”. In *Proceeding of the 2nd Meeting of the NAACL on Language technologies*, pp: 1-8.

Leavitt, H. J., and T. L. Whisler (1958) “Management in the 1980’s,” *Harvard Business Review*.

Lyons, J. (1981) *Language, Meaning and Context*. London: Fontana.

MacQueen, J. (1967) “Some methods for classification and analysis of multivariate observations”. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, pp: 281–297.

Moosavi, N. S. and Gh. Ghassem-Sani (2009) “A ranking approach to Persian pronoun resolution”. *Advances in Computational Linguistics. Research in Computing Science*, 41, 169-180.

Pollard, C. and I. A. Sag (1994) *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Rasooli, M.S., M. Kouhestani, and A. Moloodi (2013) "Development of a Persian syntactic dependency treebank". In *Proceedings of the HLT Conference of the NAACL*, pp: 306-314.

Rosenberg, A., and J. Hirschberg. (2007). “V-measure: A conditional entropy-based external cluster evaluation measure”. In *Proceedings of the 2007 Joint Conference on EMNLP and CoNLL*, pp: 410-420.

Seraji, M., B. Megyesi, and J. Nivre (2012) “Bootstrapping a Persian dependency Treebank”. *Linguistic Issues in Language Technology*, 7.

Shamsfard, M., A. Hesabi, H. Fadaie, A. Famian, S. Bagherbeigi, N. Mansoory, E. Fekri, M. Monshizadeh, and S.M. Assi (2010) “Semi-automatic development of FarsNet: The Persian WordNet”. In *Proceedings of the 5th Global WordNet Conference*, Mumbai, India.

SharifiAtashgah, M. and M. Bijankhan (2009) "Corpus-based analysis for multi-token units in Persian", In *Proceedings of the 3rd Workshop on Computational Approaches to Arabic Script-based Languages [at] MT*.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

در مجموعه مقالات سومین همایش ملی زبان‌شناسی پیکره‌ای، صفحات ۱۸۱-۲۰۷

Introducing a Gold Standard Data at the Semantic Level for Persian Words with Similar Written Forms

Abstract

A word is the smallest unit of a language that contains wide information, from phonological construction to syntax and semantics. Usually the words of a language have more than one sense. This property causes to face a challenge when we process the language automatically; because the methodology used for processing of the language should be in a form that the syntactic or semantic role of the word to be recognized through the local context of the word in a sentence. This paper introduces a gold standard data developed in the semantic level for Persian words which have more than one written form and a computer faces a challenge to determine the correct word's sense. The data is developed for 20 Persian target words and 100 sentences for each target word extracted from a corpus is annotated manually. Then, the data is standardized. In this data set, the annotation of 4 selected words is increased to 5 times to evaluate the performance of the supervised machine learning method in the classification task used in word sense disambiguation and compare it with the unsupervised machine learning method in the clustering task used in word sense induction.

Keywords: linguistic corpus, gold standard data, data annotation, sense ambiguity, classification, clustering