ارائه یک واژگان برای کلمات فارسی

بهرنگ قاسمی زاده [1] سعید رحیمی [2]
QasemiZadeh@comp.iust.ac.ir

چکیده

در پردازش زبانهای طبیعی، واژگان [3] به منظور ارائه دانش واژگانی کلمات زبان استفاده می شود. اهمیت واژگان در آن است که در پردازش زبان طبیعی مبتنی بر دانش، دانش پایه را جهت پردازش های بعدی فراهم می آورد. در این مقاله خصوصیات و ویژگیهای یک واژگان برای کلمات زبان فارسی تشریح شده است. واژگان معرفی شده، علاوه بر ارائه ویژگیهای صرفی کلمات زبان فارسی، داده های آماری از کاربرد کلمه در جایگاه نحوی خاص را ارائه می نماید. علاوه بر این، از هستان شناسی [4] ها برای ارائه معانی واژه هایی با مقوله اسم و صفت و دسته بندی آنها استفاده شده است.

کلمات کلیدی: واژگان، پردازش رایانه ای زبان، فارسی، هستان شناسی

1 آشنایی

واژگان فهرستی از واژه های یک زبان است که دانشی در رابطه با چگونگی کاربرد واژه با آن همراه شده است. واژگان می تواند عمومی و یا مختص به یک دامنه خاص باشد. یک فرهنگ لغت ساده می تواند مثالی از یک واژگان باشد. یک معیارکلی برای ذخیره دانش در واژگان این است که مشخص شود دانش ذخیره شده به واژه وابسته است یا نه. [1] [2] در این مقاله واژگانی عمومی از واژه های فارسی تشریح شده است. در هر گام از پردازش زبان طبیعی دانشی تولید می شود که مبنا و زیر بنای اطلاعاتی را برای پردازش بعدی فراهم می آورد. واژگان ارائه کننده دانش پایه ای است، که در مراحل پردازش بعدی استفاده می شود. تقریبا برای اکثر زبانهای دنیا، واژگان با هدف استفاده در کاربردهای خاص یا عمومی وجود دارد. شاید شناخته شده ترین واژگان در این میان [3] باشد که در آن هدف بیشتر بیان روابط معنایی میان واژه های ذخیره شده در واژگان است. نوع دانش ذخیره شده در واژگان به کاربردی که برای آن منظور فراهم شده است، وابسته است. به عنوان مثال در سیستم ترجمه ماشینی نیازی به داشتن رشته های آوانگاری کلمات زبان نیست اما در صورتیکه هدف استفاده از واژگان در سیستم تبدیل متن به گفتار باشد، به چنین داده هایی نیاز است.[1]

در این مقاله ساختار یک واژگان تهیه شده برای واژه های فارسی توضیح داده شده است. برای تهیه چنین واژگانی ابتدا دسته بندی جدیدی از واژه های فارسی ارائه شده است. علاوه بر این در این مقاله توضیح داده شده است که چگونه از روشهای خودکار و نیمه خودکار مختلف جهت جمع آوری دانش واژگانی استفاده شده است. واژگان مورد نظر جهت استفاده در کاربردهای متفاوت همانند سیستم های تبدیل متن به گفتار، سیستم های دسته بندی معنایی کلمات، سیستم های پرسش و پاسخ معنایی از متون، مناسب است.

در ادامه مقاله و در بخش دوم، یک دسته بندی جدید از مقوله های واژگانی برای کلمات فارسی ارائه شده است. در بخش سوم، معماری کلی واژگان و ویژگیهای در نظر گرفته شده برای هر یک از واژه ها توضیح داده می شود. چگونگی جمع آوری دانش واژگانی

---

[1] دانشجوی کارشناسی ارشد هوش مصنوعی، دانشگاه علم و صنعت ایران
[2] دانشجوی کارشناسی ارشد ادبیات فارسی، دانشگاه تهران
[3] Lexicon
[4] Ontology

کلمات زبان و مراحل تکمیل واژگان در فصل چهارم توضیح داده شده است. در بخش پنجم، نتیجه گیری و سمت و سوی کارهای آینده تشریح شده است.

2 مقوله های واژگانی و معماری واژگان

به‌طور کلي دسته‌بندي هاي انجام شده از مقوله‌هاي زباني در زبان فارسي در روش های سنتي بر پایه دیدگاه‌های قدیمي ساخت زبان استوار است. در یکی از این دسته‌بندي ها، اجزاي کلام فارسي به هشت گروه تقسیم بندي شده است[4]. از آنجا که این روش سنتي بر پایه دیدگاه‌های قدیمي استوار است، چندان به رابطه صرف و نحو تکیه نداشته و بیشتر به تحلیل واژه مستقل از روابط نحوي مي‌پردازد [5][6] دسته‌هاي ارائه شده در روش سنتي عبارتند از[4]: اسم، صفت، فعل، قید، ضمیر، عدد، صوت و حرف.

این دسته‌بندي دو مشکل اساسي دارد. اول این‌که اطلاعات کافي را جهت تحلیل نحوي زبان به کمک روش‌هاي نوین زبان‌شناسي فراهم نمي‌آورد. مشکل دوم این دسته‌بندي، عدم وجود دسته و گروهي خاص براي برخي از اجزاي کلام در زبان فارسي است. این کلمات در زبان‌شناسي سنتي با نام ادات شناخته مي‌شود. علاوه بر این وجود چنین دسته‌بندي سبب فراهم آوردن اطلاعات افزوده‌اي نخواهد بود و کمکي در جهت تسهیل تحلیل رایانه‌اي زبان فراهم نمي‌آورد. [7]

با توجه به توضیحات ارائه شده یک دسته بندی جدید برای واژگان فارسی ارائه شده است. این دسته بندی در [7] شرح داده شده است. بر اساس دسته بندی صورت گرفته، واژه های فارسی در سیزده گروه مختلف جای گرفته اند. این گروه ها در جدول 1 به همراه تعدادکل ویژگیهای در نظر گرفته شده برای آنها لیست شده اند.

واژگان ارائه شده بر اساس دسته بندی اخیر سازماندهی شده است. برای هریک از مقوله های واژگانی ارائه شده، مجموعه ای از ویژگیهای صرفی در نظر گرفته شده است که در ادامه و در بخش بعد توضیح داده شده است. بنا بر دسته بندی صورت گرفته واژه های فارسی به یکی از دسته های ارائه شده نسبت داده می شوند و با توجه به مقوله واژگانی، اطلاعات بیشتر در مورد آنها ذخیره می شود. توجه به این نکته الزامی است که کلمات همنویسه ممکن است چندین بار در واژگان ذخیره شده باشند با توجه به اینکه چنین واژه هایی ویژگیهای صرفی، نحوی و معنایی متفاوت و خاص خود را دارند.

جدول شماره 1. مقوله های واژگانی که واژگان بر اساس آن شکل گرفته است

مقوله واژگانی	تعداد ویژگیها[1]	تعداد مدخل ها
اسم	41	37000
صفت	35	13000
قید	9	2000
فعل (بن ماضی و مضارع)	33	437
عدد	6	88
حرف اضافه	3	219
حرف ربط	4	122
شاخص	2	85
ممیز	2	159
ضمیر	17	139
صوت (شبه جمله)	1	350
حرف تعریف[2]	6	50

[1] تعداد ویژگیها بدون در نظر گرفتن ویژگیهای معنایی است علاوه بر این کلیه ویژگیها به شکل بولین (دودویی) نمایش داده شده اند. بنابراین به عنوان مثال مفرد و جمع بودن، هر یک مستقلا در شمارش ویژگیها لحاظ شده اند.

[2]. معادل Determiner، وابسته پیشن 3 و شامل صفت های پیشین در دسته بندی متعارف.

3 ویژگیهای ارائه شده برای واژها در واژگان معرفی شده

به طور کلی ویژگیها و خصوصیاتی که برای واژه های ذخیره شده در واژگان ارائه شده است، در سه دسته جای می گیرند. این سه دسته خصوصیت برای هر مقوله واژگانی متفاوت است. در گروه نخست، ویژگیهای متداول صرفی با توجه به مقوله واژگانی در نظر گرفته شده است. جدول شماره دو تعداد ویژگیهای صرفی برای هر یک از مقوله های واژگانی و مثال هایی از آنها را نمایش می دهد. برخی از این ویژگیها صرفا دانش صرفی را ارائه می نمایند و گروهی دیگر از آنها برای اطمینان از درستی نتایج در فرایند تحلیل صرفی و گروهی دیگر برای کاربردهای واژه سازی استفاده می شوند. به عنوان مثال برای مقوله واژگانی اسم، گروهی از ویژگیها همانند اینکه کلمه عام است یا خاص، مشتق است یا مرکب و یا ساده؛ ارائه دهنده دانش صرفی اند. در عین حال در واژگان به صراحت مشخص شده است که هر یک از اسامی موجود، آیا توانایی جمع بسته شدن با نشانه جمع خاص را دارند یا نه.

جدول 2. ویژگی های صرفی در نظر گرفته شده برای مقوله های واژگانی

مقوله واژگانی	تعداد ویژگیهای صرفی	مثال از ویژگی ها
اسم	30	جامد،مشتق و..
صفت	24	جمع،مفرد،درجه پذیر و...
قید	8	مختص،مشترک و...
فعل (بن ماضی و مضارع)	32	متمم پذیری،معلوم،مجهول و...
عدد	6	شمارشی نوع اول،شمارشی نوع دوم و...
حرف اضافه	2	حرف اضافه گروهی
حرف ربط	4	همپایگی،وابستگی و...
شاخص	-	
ممیز	-	
ضمیر	14	شخص،شمار و...
صوت	-	
حرف تعریف	6	مبهم،اشاره،پرسشی و...

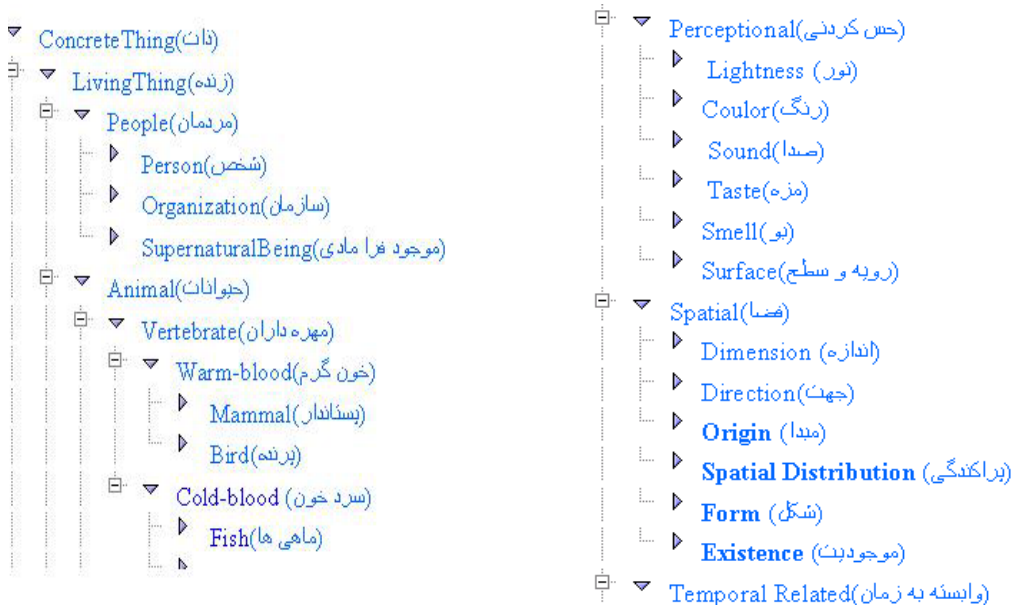
دسته دوم از ویژگیها شامل ویژگیهای آماری جهت استفاده در مراحل مختلف پردازش زبان است. این اطلاعات شامل فرکانس استفاده وابسته[1] از واژه ها در پیکره های فارسی است. این اطلاعات به شکل نیمه خودکار و با نظارت کارشناس، جمع آوری شده است. در ادامه، این ابزار جانبی تشریح شده است. علاوه بر آن، امکان ظهور واژه در جایگاه های نحوی مختلف مشخص شده است که از آن می توان برای تسریع پردازش نحوی استفاده نمود. علاوه بر ویژگیهای آماری عمومی که برای کلیه مقوله های واژگانی وجود دارد، برخی از ویژگیهای آماری خاص یک مقوله واژگانی نیز در نظر گرفته شده است. به عنوان مثال برای واژه هایی با مقوله صفت، امکان استفاد به جای مقوله واژگانی اسم نیز برای واژه مورد نظر در واژگان در نظر گرفته شده است. این دو آیتم با کمک شم زبانی متخصیصین ادبیات و استفاده از پیکره های فارسی ارائه شده است.

دسته سوم از ویژگیهای ارائه شده در واژگان، ویژگیهای معنایی برای مجموعه واژه های واژگان است. برای ارائه ویژگیهای معنایی از مقوله های واژگانی اسم و صفت از هستان شناسی های سطح بالا[2] استفاده شده است که در آن معانی در ساختاری سلسله مراتبی سازماندهی شده اند. شکل بعد قسمتی از این هستان شناسی ها را نمایش می دهد. هستان شناسی های ارائه شده بر اساس نمونه های ارائه شده در [8][9][10] برای واژه های فارسی گزینش و ارائه شده است. قالب ویژگیهای معنایی ارائه شده برای صفت ها از [9] و برای اسامی از [10] است. علاوه بر مقوله های واژگانی اسم و صفت، برای دیگر مقوله های واژگانی نیز، یک مجموعه از ویژگیهای معنایی با توجه به کاربرد آنها در یک ساختار مسطح در نظر گرفته شده است. جدول 3 تعداد ویژگیهای معنایی در نظر گرفته شده برای این مقوله ها و نمونه ای از آنها را ارائه می نماید.

---

[1] Relative Frequency Of Use

[2] Upper Ontology



شکل 1. قسمتی از هستان شناسی های استفاده شده برای ارائه معنای واژه ها. در سمت چپ هستان شناسی مرتبط با مقوله واژگانی اسم و در سمت راست مقوله واژگانی صفت را نشان می دهد.

4 چگونگی جمع آوری واژه ها

جمع آوری دانش در سیستم های مبتنی بر دانش، کاری مشکل، هزینه بر و زمان بر است. این کار در صورت استفاده از ابزارهایی جهت استخراج و جمع آوری خودکار دانش با سهولت، هزینه و زمان کمتری انجام خواهد شد. با توجه به مقدمه گفته شده، برای جمع آوری مدخل های واژگان، سعی بر آن شد تا مراحل انجام کار، حداکثر به شکل خودکار و ماشینی انجام شود. به طور کلی جمع آوری واژه ها، در دو مرحله به شکل دستی و سپس خودکار صورت گرفت. به منظور اطمینان از صحت داده ها نیز هم از ابزارهای خودکار و هم از بازنگری مداخل توسط کارشناسان بهره گرفته شده است.

جدول 3. تعداد ویژگی های معنایی در نظر گرفته شده به همراه مثال از آنها.

مقوله واژگانی	تعداد ویژگیهای معنایی	مثال
قید	29	حالت،تأکیدو...
فعل	-	-
عدد	-	-
حرف اضافه	8	استثناء،مالکیت و..
حرف ربط	3	علی،زمانی و...
شاخص	-	-
ممیز	-	-
ضمیر	3	جاندار،بی جان و..
صوت	15	درد،تنفر،تحسین و..
حرف تعریف	-	-

در اولین فاز از مراحل جمع آوری مدخل ها، پربسامدترین واژه ها در پیکره های فارسی گزینش شدند. بـرای ایـن منظـور از یک عامل لغزنده[1] بر روی صفحات وب و متون الکترونیکی استفاده شد. برای اسـتخراج واژه هـا، آدرس صـفحات وب و یـا فایـل هـای مربوط به مستندات فارسی ارائه می شود و عامل واژه های فارسی را از آنها استخراج و یک نمایه از واژه ها و فرکانس تکـرار واژه هـا را به عنوان خروجی تولید می نماید. در این مرحله از کار، در حدود 8000 واژه بسیط پربسامد به دست آمده، به واژگان اضافه شده و بـه صورت دستی ویژگیهای مرتبط به آنها نسبت داده شد. برای این منظور کلمات صرف شده و همچنـین مشـتق از لیسـت واژه هـای بـه دست آمده زدوده شدند. علاوه بر این با توجه به محدود بودن تعداد بن های ماضی و مضارع افعـال فارسـی، ایـن واژه هـا نیـز از منـابع [11]، استخراج و به شکل دستی در واژگان ثبت شدند. علاوه بر این واژه ها به شکل دستی آوانگاری شدند.

در مرحله دوم از جمع آوری مداخل، عامل معرفی شده در مرحله قبل، به یک ابزار تحلیل صرفی برای کلمـات فارسـی مجهـز شد. این تحلیل گر در [7] تشریح شده است. در این مرحله نیز پربسامدترین واژه های فارسی که در واژگان وجـود نداشـتند، از منـابع مختلف و پیکره های متفاوت استخراج شدند. علاوه بر اینکه در این مرحله، با توجه به وجود ابزار تحلیلگر صرفی، ریشه یابی از کلمـات صرف شده، به صورت خودکار انجام شد. تعیین ویژگیهای واژه های بدست آمده در این مرحله هم به شکل خودکـار و هـم بـه صـورت دستی صورت گرفت. ویژگیهای واژه های مشتق شده از وند افزائی به شکل خودکار تعیین شدند. برای این منظور کلیـه وندها و شـبه وندهای فارسی جمع آوری شدند[12]و[13]. این وندها با توجه به تاثیر خود، در دسته های جداگانه ای قرار داده شدند و به هر دسـته مجموعه ای از قوانین برای تعیین ویژگیهای صرفی و معنایی واژه های مشتق به دست آمده نسبت داده شد. یک ابزار مبتنی بر قـوانین برای تعیین خودکار ویژگیها با توجه به وند ظاهر شده در ساختمان واژه استفاده شد. علاوه بر این ویژگی مابقی کلمات به شکل دسـتی در واژگان ثبت شد. این کار چندین بار تکرار شد. در پایان فاز دوم، واژگان مشتمل بر 50000 لغت بوده است.

در مرحله سوم، از ابزاری خودکار جهت استخراج اسامی خاص به عنوان مثال نام محل، نام و نـام خـانوادگی و ... اسـتفاده شـد. ابزار استفاده شده در این مرحله هم از قوانین واژه سازی فارسی و هم از قوانین هیوریستیک برای استخراج اسامی خاص استفاده نموده است. بیشتر قوانین هیوریستیک از [12] استخراج شده اند. قوانین واژه سازی عموما مشتمل بر قوانین واژه سازی مبتنی بر وند افزایـی بوده است. در پایان این فاز قریب به 8000 اسم خاص به واژگان افزوده شد.

به منظور اطمینان از صحت داده های وارد شده در واژگان، مداخل توسط کارشناسان زبان فارسـی بـازنگری و در صـورت نیـاز تصحیح شده اند. علاوه بر اینکه در این مرحله نیز از ابزارهایی جهت پیدا نمودن اشتباهات و یا تناقض های ممکن استفاده شـده اسـت. در این مرحله نیز مداخل جدیدی به واژگان افزوده شده است به خصوص کلمات همنویسه بسیاری در ایـن مرحلـه بـه واژگـان افـزوده شده است. تصحیح و افزایش مداخل و بازنگری آن برای بار دوم در حال حاضر در حال انجام است.

5 خلاصه و نتیجه گیری

در سیستم های پردازش زبان طبیعی مبتنی بر روش های دانش مدار، واژگان دانش اولیه را جهـت پـردازش زبـان فـراهم مـی آورد. در این سیستم ها، واژگان جهت ارائه دانش مرتبط با واژه های زبان با توجه به کاربرد خاص مورد نظر استفاده مـی شـود. در ایـن مقاله، واژگان جدیدی برای کلمات زبان فارسی ارائه شده است که از آن می توان برای کاربردهای متفاوت همانند سیستم های تبـدیل متن به گفتار و بالعکس، سیستم های پرسش و پاسخ مبتنی بر زبان طبیعی و سیسـتم هـای دسـته بنـدی خودکـار متـون الکتریکـی استفاده نمود. در واژگان ارائه شده، واژه های زبان فارسی در 13 گروه مختلف بر اساس دسته بندی جدید از مقوله های واژگانی، جای گرفته اند. برای هر دسته از واژه ها، ویژگی های صرفی، معنایی و آماری مختلفی در نظر گرفته شده است. استفاده از واژگان ارائه شـده در کاربردهای مختلف، به سبب فراهم آوردن دانش در حوزه ای وسیع، سبب افزایش دقت و سرعت در دیگر مراحل پردازش رایانـه ای زبان فارسی خواهد بود.

---

[1] Crawler Agent

6 منابع

[1] Hirst, Graeme. ``**Ontology and the lexicon.**'' In: Staab, Steffen and Studer, Rudi (editors), Handbook on Ontologies, Berlin: Springer, 2004, 209--229.

[2] Briscoe, Ted; de Paiva, Valeria; and Copestake, Ann (editors) (1993): **Inheritance, Defaults,and the Lexicon.** Cambridge University Press.

[3] Fellbaum, Christiane (1998): **WordNet: An electronic lexical database.** The MIT Press,Cambridge, Mass.

[4] انوري، احمدي گيوي، **دستورزبان فارسي 2 (** ويرايش دوم)، انتشارات فاطمي، تهران، 1382.

[5] قریب،عبدالعظیم و بهار، ملک الشعرا؛ فروزانفر، بدیع الزمان؛ همایی، جلال؛ رشیدیاسمی، غلامرضا. **دستور زبان فارسی (معروف به پنج استاد)**، جلد1و2، تهران، انتشارات بی تا.

[6] خیام پور، عبدالرسول. **دستور زبان فارسی**، تبریز، 1334.

[7] قاسمی زاده، بهرنگ، رحیمی، سعید، نم نبات، مجید، سالاریان، مرتضی، کوچاری، عباس، **روشی نوین برای صرف واژه های فارسی**، مجموعه مقالات یازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران، تهران، 1384.

[8] Lenci, Alessandro (2001): **Building an ontology for the lexicon: Semantic types and word meaning.** In Jensen, Per Anker and Skadhauge, Peter (eds.), *Ontology-based Interpretation of Noun Phrases: Proceedings of the First International OntoQuery Workshop*, University of Southern Denmark, 103–120.

[9] Hamp, B. & H. Feldweg (1997): **GermaNet - a Lexical-Semantic Net for German"**. In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications". Madrid, 1997.

[10] Thatsanee Charoenporn, Canasai Kruengkrai, Virach Sornlertlamvanich and Hitoshi Isahara. **Acquiring Semantic Information in the TCL's Computational Lexicon,** Proceedings of the Fourth Workshop on Asia Language Resources, IJCNLP-04, Sanya City, Hainan Island, China, pp. 47-53, 25 March 2004.

[11] طباطبایی، علاءالدین. **فعل بسیط فارسی و واژه سازی**. مرکز نشر دانشگاهی. تهران.1376.

[12]کلباسی، ایران. **ساخت اشتقاقی واژه در فارسی امروز**. پژوهشگاه علوم انسانی و مطالعات فرهنگی. تهران.1380.

[13]کشانی، خسرو. **اشتقاق پسوندی در زبان فارسی امروز** .مرکز نشر دانشگاهی. 1371.

[14] Rezaie, S. 2001. **Tokenizing An Arabic Script Language**, Arabic Language Processing: Status And Prospects, Acl/Eacl.