استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی

دکتر مهرنوش شمس‌فرد [1]
دانشکدهٔ مهندسی کامپیوتر
دانشگاه صنعتی امیر کبیر

دکتر احمد عبداله زاده بارفروش
دانشکدهٔ مهندسی کامپیوتر
دانشگاه صنعتی امیرکبیر

امروزه هستان‌شناسی‌ها که پایگاه‌های دانش مفهومی هستند، در سیستم‌های اطلاعاتی کاربرد بسیاری دارند. ساخت انواع هستان‌شناسی برای انواع قلمروها و کاربردها، فرآیندی پرهزینه و زمان‌بر است. خودکارسازی این فرآیند، گامی در جهت رفع گلوگاه کسب دانش در سیستم‌های اطلاعاتی و کاهش هزینه ساخت آنهاست. در این مقاله، ابتدا مروری بر روش‌های استخراج دانش مفهومی و ساخت هستان‌شناسی داریم و سپس برای کشف روابط میان مفاهیم، از روی متون زبان طبیعی، به بررسی روش‌های مبتنی بر الگو خواهیم پرداخت. در ادامه، برای استخراج دانش مفهومی از متون زبان فارسی، برخی الگوهای زبانی و معنایی را معرفی می‌کنیم. این الگوها عمومی و مستقل از دامنه و کاربرد می‌باشند و در سطح جمله عمل می‌کنند. آنها جهت استخراج روابط طبقه‌ای و غیر طبقه‌ای و اصول بدیهی، از عبارات و جملات فارسی به کار می‌روند. از جمله روابط استخراج شده از طریق این الگوها، می‌توان به روابط شمول معنایی، جزء ـ کل، ویژگی ـ مقدار، هم‌مرجعی و ... اشاره نمود. در این مقاله، ضمن معرفی الگوهای استخراج روابط مفهومی، در هر مورد مثال‌هایی از روابط قابل استخراج ارائه خواهد شد.

مقدمه

هستان‌شناسی (ontology) مدلی انتزاعی از جهان واقع است که مفاهیم و روابط میان آن را در قلمروی مورد بحث نمایش می‌دهد. هستان‌شناسی‌ها که پایگاه دانش مفهومی (conceptual knowledge) هستند، در محدوده وسیعی از قلمروها کاربرد دارند که برای نمونه می‌توان به شبکه‌های جهان‌گستر معنایی (semantic web)، موتورهای جست‌وجو، تجارب الکترونیکی، پردازش زبان طبیعی، مهندسی دانش، استخراج و بازیابی اطلاعات، سیستم‌های چند عاملی، مدل‌سازی کیفی از سیستم‌های فیزیکی، طراحی پایگاه‌های داده، سیستم‌های اطلاعات جغرافیایی و کتابخانه‌های رقمی اشاره نمود.

در قلمروی کامپیوتر، هستان‌شناسی را می‌توان با یک چهارتایی (C, R, F, A) تعریف کرد (شمس فرد، ۱۳۸۱) که در آن:

- C مجموعه مفاهیم موجود در جهان مدل شده است.
- R مجموعه روابط میان مفاهیم است و خود به دو زیر مجموعه مجزای R_T و R_N افراز می‌شود.

[1] نشانی تماس: تهران، خیابان حافظ، دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر، آزمایشگاه سیستم‌های هوشمند
e-mail: shams@pnu.ac.ir

- R_T مجموعه روابط طبقه‌ای (taxonomic) میان مفاهیم است که سلسله مراتب مشمول را ایجاد می‌کند و دودویی می‌باشد.
- R_N مجموعه روابط غیر طبقه‌ای است که ممکن است n تایی نیز باشد ($n \geq 1$).

• F مجموعه تصریحات هستان‌شناسی در مورد مفاهیم و روابط آنهاست و خود به دو زیر مجموعه F_T و F_N افراز می‌شود:

- F_T مجموعه تصریحات هستان‌شناسی درباره روابط طبقه‌ای مفاهیم است. به عبارت دیگر، سلسله مراتب شمول را نشان می‌دهد.
- F_N مجموعه تصریحات هستان‌شناسی درباره روابط غیر طبقه‌ای مفاهیم است.

• A مجموعه اصول بدیهی (axioms) هستان‌شناسی است که به زبان صوری، مثل منطق بیان می‌شود.

امروزه با توجه به روند رو به رشد استفاده از هستان‌شناسی‌ها در سیستم‌های اطلاعاتی، ساخت هستان‌شناسی، روش‌شناسی ساخت، ابزارهای ساخت و ساخت خودکار و یادگیری هستان‌شناسی‌ها از مباحث مطرح در میان محققان است. هستان‌شناسی‌ها را ممکن است به صورت دستی، با استفاده از ابزارهای مهندسی هستان‌شناسی و یا با روش‌های اکتساب دانش و ساخت (نیمه) خودکار تولید کرد که در مورد هر یک، در زیر نمونه‌هایی ارائه شده است.

• *ساخت دستی:* در این روش، حجم عظیمی از دانش مفهومی به وسیله افراد در ماشین کد می‌شود و پایگاه‌های دانش بزرگ عمومی یا تخصصی ایجاد می‌گردند. Cyc (لنات، ۱۹۹۵) و Mikrokosmos (نیرنبرگ و همکاران، ۱۹۹۵) نمونه‌هایی از هستان‌شناسی‌های ساخته شده با این روش هستند.

• *استفاده از ابزارهای مهندسی هستان‌شناسی:* در سال‌های اخیر ابزارهایی برای پشتیبانی ساخت هستان‌شناسی ساخته شده‌اند. Protégé (اریکسون و همکاران، ۱۹۹۹)، Ontolingua (فارکوهار و همکاران، ۱۹۹۷)، Ontosaurus (اسوارتوت وهمکاران، ۱۹۹۷) و WebOnto (دومینگ، ۱۹۹۸) با فراهم کردن واسط کاربرد مناسب، محیط را برای اکتساب دانش مفهومی از کاربرد آماده و دسته دیگر مانند DODDLE II (یاماگوچی، ۲۰۰۱) و 'SVETLAN (چالندار و گرو، ۲۰۰۰) داده‌ها و ساختارهای لازم برای ساخت هستان‌شناسی را از منابع ورودی استخراج می‌کنند و در اختیار سازنده هستان‌شناسی (انسان یا ماشین) قرار می‌دهند.

• *ساخت (نیمه) خودکار:* برای کاهش مشکلات ساخت هستان‌شناسی‌ها و تولید (نیمه) خودکار آنها دو راهکار پیشنهاد شده است:

الف) یکپارچه‌سازی و استفاده مجدد از هستان‌شناسی‌های موجود (چاپولسکی و همکاران، ۱۹۹۷؛ نوی و موسن، ۲۰۰۰؛ ریوتارو و همکاران، ۲۰۰۱)؛

ب) یادگیری و ساخت خودکار هستان‌شناسی‌ها از روی منابع موجود، مانند سیستم‌های Text-to-onto (میده و استاب، ۲۰۰۱ a)، Syndikate (هان و روماکر، ۲۰۰۱) Asium (فوره و همکاران، ۱۹۹۱) و هستی (شمس‌فرد و عبداله زاده بارفروش، ۲۰۰۲).

ساخت دستی انواع هستان‌شناسی، برای قلمروها و کاربردهای مختلف، پرهزینه، وقت‌گیر و مستعد خطاست و هستان‌شناسی‌هایی که به صورت دستی ساخته می‌شوند، معمولاً، گران، متمایل به نظرات شخصی طراح، در مقابل تغییرات غیر منعطف و دقیقاً خاص منظوری که برای آن تهیه شده‌اند، می‌باشند. لذا خودکارسازی عمل ساخت هستان‌شناسی نه فقط هزینه ساخت را کاهش می‌دهد، بلکه به تولید هستان‌شناسی با انطباق بیشتر با کاربرد آن منجر می‌شود. وجود ابزارهای مهندسی هستان‌شناسی که فقط به صورت واسط کاربرد عمل می‌کنند، نیاز به سازنده انسانی را منتفی نمی‌کنند، بلکه فقط محیط را برای وی مهیا می‌سازند. لذا حرکت به سمت خودکارسازی اکتساب دانش از روی منابع، متون، پایگاه‌های داده و هستان‌شناسی‌های دیگر، مشکلات مهندسی هستان‌شناسی را محدود و هزینه ساخت و

استفاده اشتراکی از هستان‌شناسی‌ها را کاهش می‌دهد.

در این مقاله، ابتدا به روش‌های اکتساب خودکار دانش مفهومی و یادگیری هستان‌شناسی اشاره‌ای خواهیم نمود. در ادامه، برای استخراج دانش مفهومی از متون زبان طبیعی، روش‌های مبتنی بر الگو را مورد بررسی قرار می‌دهیم و ضمن معرفی یک سیستم یادگیر هستان‌شناسی، برای استخراج روابط طبقه‌ای، غیر طبقه‌ای و اصول بدیهی از جملات و عبارات زبان فارسی، الگوهای ارائه شده در این سیستم را معرفی خواهیم کرد.

استخراج دانش مفهومی و یادگیری (نیمه) خودکار هستان‌شناسی

یادگیری هستان‌شناسی به معنی استخراج دانش مفهومی از منابع ورودی و ساخت یک هستان‌شناسی بر اساس آنهاست. یادگیری هستان‌شناسی از روش‌ها و الگوریتم‌های رشته‌های مختلفی چون پردازش زبان طبیعی، مهندسی دانش، یادگیری ماشینی، اکتساب دانش، استخراج اطلاعات، استنتاج خودکار و پردازش نمادین و احتمالاتی بهره می‌گیرد. در دو دهه اخیر، در زمینه یادگیری هستان‌شناسی فعالیت‌هایی شده و روش‌ها، متدولوژی‌ها، ابزارها و سیستم‌های مختلفی نیز ارائه گردیده است. برخی از این سیستم‌ها، سیستم‌های خودمختار یادگیر هستان‌شناسی هستند، در حالی که برخی دیگر ابزارهای پشتیبانی ساخت هستان‌شناسی می‌باشند. سیستم‌های یادگیر هستان‌شناسی را می‌توان بر اساس شش عامل افتراق زیر دسته‌بندی نمود (شمس‌فرد و عبداله زاده بارفروش، ۲۰۰۳):

الف) *عنصر آموختنی:* عنصر آموختنی می‌تواند دانش مفهومی و هستان‌شناسانه به تنهایی و یا ترکیب آن با دانش لغوی باشد. اصلی‌ترین عناصر لغوی که سیستم‌های موجود یاد می‌گیرند، کلمات و اصیل‌ترین عناصر آموختنی هستان‌شناسی، مفاهیم، روابط مفهومی (اعم از روابط طبقه‌ای و غیر طبقه‌ای و اصول بدیهی) هستند. همچنین برخی سیستم‌ها، در مورد نحوه استخراج دانش‌های فوق از ورودی، فرادانشی را می‌آموزند.

ب) *نقطه شروع:* سیستم‌های یادگیر هستان‌شناسی از دانش پیش زمینه خود استفاده می‌کنند و از منابع ورودی، دانش‌های جدید را می‌آموزند. دانش پیش زمینه ممکن است از نوع دانش زبانی (دستور زبان، دانش لغوی، الگوها و ...) و یا از نوع دانش مفهومی (هستان‌شناسی مبنا) باشد. در اکثر سیستم‌های موجود، برای پردازش متن از واژگان معنایی از پیش تعریف شده‌ای مانند DODDLE II , TEXT-TO-ONTO و SYNDIKATE (مانند WordNet (میلر، ۱۹۹۵)) استفاده می‌شود که معمولاً حاوی دانش مفهومی نیز هست. در سیستم‌های مختلف، هستان‌شناسی مبنا (اولیه) ممکن است تنها هسته کوچکی از دانش اولیه (شمس‌فرد، ۱۳۸۱؛ هوانگ، ۱۹۹۹) یا یک هستان‌شناسی تخصصی در یک قلمروی خاص (Syndikate; Text-to-onto) و یا یک هستان‌شناسی کلی بزرگ حاوی دانش عرفی مانند Cyc (لنات، ۱۹۹۵) باشد. منبع ورودی سیستم‌های یادگیر ممکن است داده‌های ساخت یافته مانند ساختار جداول پایگاه داده (کاشیاپ، ۱۹۹۹)، هستان‌شناسی‌های موجود دیگر (ویلیامز و تساتسولیس، ۲۰۰۰) و پایگاه‌های دانش (سویانتو و کومپتون، ۲۰۰۰) یا داده‌های نیمه ساخت یافته (پرنل و همکاران، ۲۰۰۱)، مانند فرهنگ‌های لغات و مستندات XML , HTML و یا داده‌های غیر ساخت یافته مانند متون زبان طبیعی موجود در پیکره‌های زبانی (SVETLAN', SYNDIKATE) و متون موجود در مستندات وب (TEXT-TO-ONTO) باشد.

ج) *پیش پردازش لازم:* متداولترین پیش پردازش استفاده شده در سیستم‌های یادگیر هستان‌شناسی از متن، پیش پردازش زبانی است. این پیش پردازش می‌تواند به صورت درک عمیق متن و یا پردازش سطحی آن باشد. درک عمیق، به کشف روابط خاص و پردازش سطحی به استخراج روابط کلی میان مفاهیم منجر می‌شود. از آنجا که درک عمیق معمولاً مشکلتر است و سرعت ساخت هستان‌شناسی را کاهش می‌دهد، اکثر سیستم‌ها (مانند TEXT-TO-ONTO و ASIUM) برای استخراج ساختارهای مورد نیازشان از متن، از تکنیک‌های پردازش سطحی متن مانند نمونه‌برداری، تعیین اجزای کلام، تحلیل نحوی و ... استفاده می‌کنند. برخی سیستم‌ها (مانند SYNDIKATE) نیز از روش‌های درک عمیق برای استخراج دانش مفهومی بهره

می‌برند. همچنین پیمانه‌هایی وجود دارند که برای استخراج ساختارهای خاص از ورودی، به کار می‌روند. این ساختارها به وسیله سیستم یادگیر برای ساخت هستان‌شناسی استفاده می‌شوند که از آن جمله می‌توان به 'SVETLAN برای استخراج طبقه‌بندی اسامی اشاره نمود.

د) *روش یادگیری:* روش‌های استخراج دانش در دامنه‌ای از روش‌های با دانش ضعیف (مانند تکنیک‌های آماری) تا روش‌های غنی از دانش (مانند استدلال منطقی) گسترده‌اند. این روش‌ها ممکن است به صورت با نظارت و یا بدون نظارت و همچنین به صورت بر خط یا برون خط مورد استفاده قرار گیرند. سیستم‌های بهره‌گیر از روش‌های آماری مانند DODDLE II (واگنر، ۲۰۰۰؛ هایر و همکاران، ۲۰۰۱)، عمدتاً بر اساس فرکانس تکرار و یا فرکانس هم وقوعی و هم مکانی کلمات و عبارات عمل می‌کنند و از تحلیل آماری داده‌های هم وقوع برای یادگیری طبقات و روابط مفهومی از متون استفاده می‌کنند. برخی سیستم‌های دیگر، روش‌های نمادین مانند روش‌های منطقی مبتنی بر الگو و زبان ـ پایه را برای استخراج دانش به کار می‌گیرند. سیستم‌های بهره‌مند از روش‌های منطقی (مانند باورز و همکاران، ۲۰۰۰)، دانش جدید را با استفاده از قیاس و یا استقرا به دست می‌آورند و با گزاره‌ها، منطق درجه اول یا درجات بالاتر نمایش می‌دهند. روش‌های زبان ـ پایه مانند تحلیل نحوی (ASIUM)، تحلیل ساختواژی ـ نحوی (اسدی، ۱۹۹۷)، تجزیه الگوهای لغوی ـ نحوی (فینکلستاین ـ لندو و موریـن، ۱۹۹۹)، پردازش معنایی و درک متن (SYNDIKATE) عموماً وابسته به زبان هستند و برای استخراج دانش از منابع غیر ساخت یافته (زبان طبیعی) به کار می‌روند. در روش‌های مبتنی بر الگو، ورودی (معمولاً متن) به دنبال الگو یا کلمات کلیدی خاص که نشانگر رابطه مفهومی خاصی است، جست‌وجو و اطلاعات مورد نظر از متن استخراج می‌شود. بعضی روش‌های مکاشفه‌ای نیز ممکن است برای تسریع و تسهیل عملکرد هر یک از این روش‌ها و رهیافت‌ها مورد استفاده قرار گیرند. همچنین برخی سیستم‌ها نیز مانندWEB→KB (کراون و همکاران، ۲۰۰۰) و TEXT-TO-ONTO از ترکیب روش‌های فوق برای یادگیری هستان‌شناسی بهره می‌برند.

در ساخت هستان‌شناسی، مرحله اکتساب دانش می‌تواند به صورت دستی، نیمه خودکار و یا خودکار انجام شود. سیستم‌های غیر دستی برای اکتساب دانش مفهومی، از روش‌ها و ابزارهای خودکار (واگنر، ۲۰۰۰) ، نیمه خودکار (TEXT-TO-ONTO) و با همکاری (ASIUM) استفاده می‌کنند. در سیستم‌های نیمه خودکار و با همکاری، معمولاً کاربرد در پیشنهاد هستان‌شناسی اولیه، اعتبارسنجی و تغییر تصمیمات سیستم، انتخاب الگوهای روابط، کنترل سطحی تجرید و مواجهه با نویز، بر چسب‌زنی به مفاهیم جدید و تعیین وزن‌های عناصر هستان‌شناسی دخالت دارد.

هـ) *نتیجه نهایی:* برخی سیستم‌های خودمختار، یادگیر هستان‌شناسی هستند، در حالی که مابقی پیمانه‌هایی هستند که عمل خاصی را انجام می‌دهند و برای ساخت هستان‌شناسی، نتیجه‌ای به صورت مجموعه‌ای از داده‌های میانی تولید می‌کنند. در سیستم‌های پشتیبان (مانند DODDLE II; SVETLAN')، ساختارهای اولیه ساخت هستان‌شناسی اکتساب می‌شوند و هستان‌شناسی نهایی بدون دخالت کاربر یا سیستم‌های دیگر ساخته نخواهد شد. برای سیستم‌های خودمختار که هستان‌شناسی می‌سازند، ویژگی‌های هستان‌شناسی حاصل (مانند سطح پوشش، کاربرد یا منظور، نوع محتویات، درجه جزئیات، ساختار و هم بندی و زبان بازنمایی) از عوامل تمایز سیستم‌های مختلف‌اند.

و) *روش ارزیابی:* برای ارزیابی این سیستم‌ها، دو روش پیشنهاد شده است: ارزیابی روش یادگیری و ارزیابی نتیجه. معمولاً ارزیابی نتیجه از طریق مقایسه دو یا چند هستان‌شناسی مدل‌سازی شده در یک قلمروی خاص با تکنیک‌های ارزیابی ضربدری (مادچ و ستاب، ۲۰۰۱) و یا از طریق کاربردی که مورد استفاده قرار می‌گیرد، انجام می‌شود. ارزیابی مبتنی بر کاربرد بیشتر برای هستان‌شناسی‌های خاص منظوره مناسب می‌باشد. اکثر سیستم‌های یادگیر، عمل ارزیابی را از طریق مقایسه حاصل کار با نتایج مورد نظر فرد خبره و یا با محاسبه معیارهای ارزیابی متداول در استخراج اطلاعات (مثل دقت و توجه) انجام می‌دهند. معیار دقت، نمایشگر نسبت نتایج (مثلاً مفاهیم استخراج شده) صحیح به کل نتایج موجود در مجموعه آزمون، و معیار توجه، نشانگر نسبت نتایج صحیح به کل نتایج استخراج شده است.

در ادامه این بخش مروری خواهیم داشت بر روش‌های مبتنی بر الگو، به عنوان یکی از راه‌های اکتساب دانش مفهومی از متون زبان طبیعی.

روش‌های مبتنی بر الگو

روش‌های تطبیق الگو، کلمه کلیدی و یا قالب، به طور وسیعی در قلمروی استخراج اطلاعات کاربرد دارند و به قلمروی یادگیری هستان‌شناسی نیز به ارث رسیده‌اند. در روش‌های مبتنی بر الگو، ورودی (معمولاً متن) به دنبال الگو یا کلمه کلیدی خاص که نشانگر رابطه مفهومی خاصی است جست‌وجو می‌شود. این الگوها انواع مختلفی (اعم از نحوی یا معنایی، و عمومی یا خاص) دارند و برای استخراج عناصر مختلف هستان‌شناسی مثل روابط طبقه‌ای یا غیر طبقه‌ای و یا اصول بدیهی به کار می‌روند. در این بخش، چند نمونهٔ استفاده از الگوها در استخراج دانش را بررسی می‌کنیم.

به عنوان نخستین کار در این زمینه می‌توان به الگوهای هیرست (۱۹۹۲) اشاره نمود. وی برای استخراج روابط شمول معنایی از متن، شش الگوی لغوی ـ نحوی معرفی نموده است. نمونه‌هایی از این الگوها در زیر آمده است:

*Npsuchas {NP,} * (and | or) NP*
*NP {, NP} * {,} (or | and) other NP*
*NP {,} including {NP,} * {or | and} NP*

گامالو و همکاران (۲۰۰۲) نیز برای نگاشت وابستگی‌های نحوی به روابط معنایی مانند شمول، مالکیت، مکان، علیت، عاملیت، موقعیت و ... از الگوهای دستوری استفاده می‌کنند. یک وابستگی با یک رابطه دودویی ($r: w1^{\downarrow}, w2^{\uparrow}$) به شکل نمایش داده می‌شود که در آن r با نشانگرهای دستوری خاص مانند حروف اضافه، رابطه فاعلیت یا مفعولیت نمونه‌دهی می‌شود، فلش‌های «$^{\downarrow}$» و «$^{\uparrow}$» متناظر، نمایانگر موقعیت‌های هسته و مکمل هستند، w_1، کلمه را در موقعیت هسته و w_2، کلمه را در موقعیت مکمل نشان می‌دهد. نشانگرهای دستوری، نشاندهنده روابطه نحوی (فاعل، مفعول، حرف اضافه)، طبقه ساختواژی ـ نحوی دو کلمه مرتبط (فعل و اسم) و حضور یا غیاب تعیین کننده در مکمل می‌باشند. برای مثال یک الگوی دستوری به صورت زیر است:

x=possessed; y=possessor=> [λx$^{\downarrow}$ λy$^{\uparrow}$ (de+; x$^{\downarrow}$, y$^{\uparrow}$)] or [λx$^{\downarrow}$ λy$^{\uparrow}$ (a+; x$^{\downarrow}$, y$^{\uparrow}$)] or [λx$^{\downarrow}$ λy$^{\uparrow}$ (para; x$^{\downarrow}$, y$^{\uparrow}$)]

که در آن به عنوان نمونه نشانگر دستوری *de+* نمایانگر این است که:۱) حرف اضافه *de* موجود است. ۲) تعیین کننده قبل از مکمل حضور دارد. ۳) هسته و مکمل هر دو اسم هستند. با استفاده از این قانون، هسته‌های نشانگرهای *de+* ، *a+* و *para* که با متغیر *x* نشان داده شده‌اند، به نقش معنایی «مملوک» و مکمل‌هایشان که با تغییر *y* نشان داده شده‌اند، به نقش معنایی «مالک» نگاشته می‌شوند.

در این زمینه، کار دیگری (ساندبلاد، ۲۰۰۲) وجود دارد که در آن برای استخراج روابط شمول و جزء ـ کل از پیکره‌های سؤالی برخی الگوهای زبانی مورد استفاده قرار می‌گیرند. این الگوها برای استخراج شمول معنایی عبارت‌اند از:

Who is/was X?
What is the location of X?

و برای استخراج رابطه جزء ـ کل عبارت‌اند از:

What is/was the X of Y?
How many X are in/on Y?

هایر و همکاران (۲۰۰۱) نیز برای استخراج رابطه «نمونه چیزی بودن» و «نام اول کسی بودن» دو الگو به صورت زیر معرفی کرده‌اند:

- الگوی *(last name)* ؟ *(profession)* نشان می‌دهد که طبقه ؟ به احتمال زیاد یک اسم اول (اسم کوچک) است (مانند actress *Julia* Roberts).
- الگوی ؟ *like (class name)* نشان می‌دهد که طبقه ؟ به احتمال زیاد یک نمونه از رده معرفی شده است (مانند metals like nickel, arsenic and lead).

به طور کلی، الگوها بر اساس نمونه، نحوه ساخت و نوع

دانش مورد استخراج با هم متفاوت‌اند. این الگوها ممکن است عمومی و مستقل از قلمرو یا کاربرد خاص باشند، مانند الگوهای مورد استفاده هیرست و ساندبلاد، (۲۰۰۲) و یا تخصصی و در مورد قلمرو یا کاربرد خاص باشند مانند الگوهای پیشنهادی اسدی (۱۹۹۹) برای استخراج دانش از متون مربوط به طراحی شبکه‌های الکتریکی. از سوی دیگر، الگوها ممکن است به صورت دستی تعریف شده باشند (ساندبلاد، ۲۰۰۲؛ گامالو و همکاران، ۲۰۰۲) یا به طور (نیمه) خودکار کشف شوند، مانند سیستم‌های PROMETHEE (ریلوف، ۱۹۹۶)، AutoSlog-TG (فینکستاین ـ لندو و مورین، ۱۹۹۹) و CRYSTAL (سودرلند و همکاران، ۱۹۹۵). همچنین الگوها ممکن است نحوی یا معنایی باشند و جهت استخراج هر یک از روابط طبقه‌ای، غیر طبقه‌ای و اصول بدیهی به کار روند.

در ادامه مقاله، به معرفی یک سیستم یادگیر هستان‌شناسی که بخشی از آن، از روش‌های مبتنی بر الگو جهت استخراج دانش بهره می‌گیرد، می‌پردازیم و الگوهای مورد استفاده آن را بررسی می‌کنیم.

سیستمی برای استخراج دانش از متون فارسی

«هستی» (شمس‌فرد، ۱۳۸۱)، سیستمی برای استخراج دانش مفهومی از متون ساده زبان فارسی و ساخت هستان‌شناسی از روی آنهاست. هستی، از پایه، به ساخت خودکار هستان‌شناسی از یک روش ترکیبی می‌پردازد. منظور از «پایه»، عدم وجود هستان‌شناسی مبنا (اعم از عمومی یا تخصصی) و همچنین نبود واژگان معنایی جهت کمک به فرآیند یادگیری می‌باشد. در ابتدای کار سیستم، واژگان تقریباً تهی و هستان‌شناسی فقط حاوی هسته اولیه یادگیری است که به صورت دستی ساخته خواهد شد. سپس واژگان معنایی (شمس فرد و عبداله‌زاده بارفروش، ۱۳۸۰) و هستان‌شناسی‌های مختلف می‌توانند بر مبنای آن با توجه به ورودی‌هایشان از محیط (که متون زبان فارسی هستند) ساخته شوند. هسته اولیه که تنها بخش هستان‌شناسی است که به وسیله خود سیستم ساخته نمی‌شود، حاوی فرادانش لازم جهت اضافه، حذف و تغییر عناصر هستان‌شناسی می‌باشد. ویژگی این هسته استقلال از زبان، قلمرو و کاربرد است. هستان‌شناسی ساخته شده در این روش، پویا و نسبت به تغییرات محیط (کاربر، کاربرد، قلمرو) انعطاف‌پذیر می‌باشد. در این روش، عناصر لغوی (کلمات) و همه عناصر هستان‌شناسی (اعم از مفاهیم، روابط طبقه‌ای و غیر طبقه‌ای و اصول بدیهی) از متن استخراج و آموخته خواهند شد. هستان‌شناسی حاصل دارای ساختاری مرکب از گراف جهت‌دار (شامل روابط سلسله مراتبی و غیر سلسله مراتبی) و اصول بازنمایی شده با منطق خواهد بود. هستی، از یک رهیافت نمادین ترکیبی، مرکب از روش‌های منطقی، زبان ـ پایه و مبتنی بر الگو و از طریق استفاده‌های متعدد از مکاشفه برای استخراج دانش بهره می‌برد و قادر به ساخت هستان‌شناسی‌های عمومی و تخصصی است.

در این سیستم، ابتدا، درخت تجزیه حاصل از هر جمله (که در یک تجزیه‌گر چارت بالا و پایین ساخته می‌شود)، با استفاده از الگوهای نحوی، به ساختار جمله تبدیل می‌گردد. سپس، ساختارهای جمله ساخته شده با استفاده از الگوهای معنایی به مجموعه‌ای از عناصر هستان‌شناسی که شامل مفاهیم، روابط طبقه‌ای و غیر طبقه‌ای و اصول بدیهی می‌باشند، تبدیل می‌شوند.

ساختارهای جمله حاوی اطلاعاتی در مورد نقش موضوعی (thematic roles) یک جمله می‌باشند. در هستی، جملات فارسی به دو دسته «حالتی» و «فعلی» متناظر با جملات ربطی و غیر ربطی تقسیم شده‌اند. جملات حالتی جملاتی هستند که دارای فعل ربطی (مانند است، بود و شد) می‌باشند و وجود یا ایجاد حالت یا وضعیتی را می‌رسانند. سایر جملات فعلی هستند. ساختار جمله در حقیقت قالب حالتی (case frame) است که برای جملات فعلی شامل نقش‌های کنش، کنشگر یا عامل، کنش‌پذیر یا پذیرا، ذینفع، زمان، مکان، ابزار، توصیف‌گر کنش، مبدأ و مقصد است و برای جملات حالتی حاوی حالت‌پذیر، حالات، زمان و مکان می باشد. ساختارهای جملات فعلی و

حالتی در زیر توضیح داده شده‌اند.

الف) ساختار جملات حالتی

مشخصه اصلی جملات حالتی، داشتن یک مسندالیه یا نهاد و یک گزاره یا مسند است. مسندالیه در این جملات یک گروه اسمی است و مسند می‌تواند هر یک از سه حالت زیر را داشته باشد (مشکوه‌الدینی، ۱۳۷۴) (در هر حالت، زیر مسند خط کشیده شده است):

- گروه اسمی در جایگاه مسند؛ مانند «حافظ، شاعر بزرگی است» و یا «احمد، دوست علی است.»
- گروه صفتی در جایگاه مسند؛ مانند «رنگ لباس علی، سبز روشن بود.» و یا «اتاق مینا، تمیز و مرتب شد.»
- گروه حرف اضافه‌ای در جایگاه مسند؛ مانند «این سینی از مس است.» و یا «گربه مریم زیر میز اتاق بود.»

لذا ساختار جمله حالتی دارای سه بخش اصلی است: یکی برای نهاد، یکی برای گزاره و یکی برای اطلاعات فعل. دو بخش اول، فهرستی از ساختارهای گروه اسمی اند و بخش سوم، زمان و نوع فعل ربطی را مشخص می‌کند. هر ساختار گروه اسمی، اجزای مختلف یک گروه اسمی شامل هسته و وابسته‌ها را به تفکیک در بخش‌های هسته، صفات و اضافات نمایش می‌دهد.

برای یکسان‌سازی نمایش اطلاعات، جهت نمایش یک گروه صفتی، از یک ساختار گروه اسمی با هسته تهی و جهت نمایش یک گروه حرف اضافه‌ای، از یک ساختار گروه اسمی با هسته و صفات تهی استفاده می‌شود. شکل ۱ نمایش صوری ساختار جملات حالتی و شکل ۲، دو مثال از ساختارهای جمله حالتی ساخته شده برای جملات «لباس خواهر مینا، پیراهن زیبایی بود» و «کتاب تمیز مریم روی میز بزرگ است» را نشان می‌دهند.

ب) ساختار جملات فعلی

این ساختار حاوی اطلاعاتی در مورد نقش‌های موضوعی جمله است. نقش‌های مورد توجه در هستی عبارت‌انداز: عامل،

شکل ۱- نمایش صوری ساختار جملات حالتی

ساختار جملات حالتی → < حالت‌پذیر > <حالت> <فعل>
< حالت‌پذیر> → < ساختار گروه اسمی >+
<حالت> → <ساختار گروه اسمی > ا<صفت> ا <ساختار گروه حرف اضافه‌ای>+
<ساختار گروه اسمی> → <هسته > <صفت> * <اضافه> *
<هسته> → <ساختار گروه اسمی>
<اضافه > → <ساختار گروه اسمی>
<ساختار گروه حرف اضافه‌ای > → <حرف اضافه > <ساختار گروه اسمی>

پذیرا، ذینفع، ابزار، مکان، زمان، مبدأ، مقصد، کنش و توصیف‌گر کنش. جملات غیر ربطی، به ساختار جمله‌ای با دوازده بخش تبدیل می‌شوند . ده بخش اول مربوط به ۱۰ نقش موضوعی فوق، بخش یازدهم حاوی فهرست عباراتی که نقش آنها در جمله شناخته نشده و بخش دوازدهم مربوط به مشخصات فعل است. یازده بخش اول هر یک، فهرستی از ساختارهای گروه اسمی اند و بخش دوازدهم، شامل اطلاعاتی در مورد زمان، شخص و ریشه فعل می‌باشد. شکل ۳ نمایش صوری ساختار جملات فعلی را نمایش می‌دهد.

به عنوان مثال، به جمله «پدر امین با میخ و تخته، میز بزرگ را ساخت» توجه کنید. در این جمله، ساختار جمله ساخته شده حاوی بخش‌های عامل، پذیرا، ابزار و کنش خواهد بود که متناظر به عبارات «پدر امین»، «میز بزرگ»، «تخته و میخ» و «ساخت» نسبت داده می‌شوند. شکل ۴ نمایی از ساختار جمله فعلی ساخته شده برای این جمله را نشان می‌دهد.

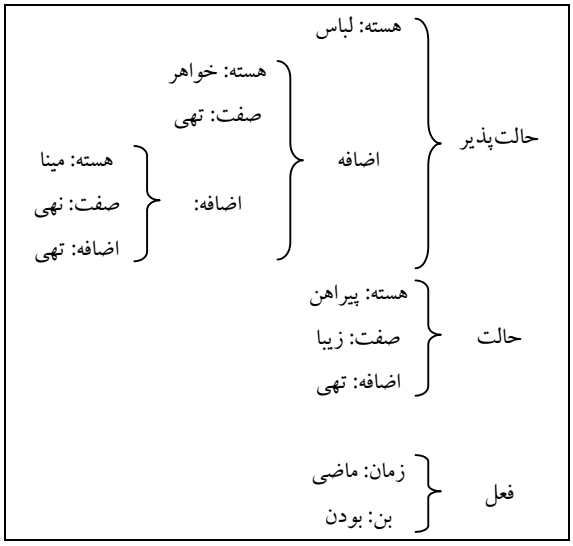
الگوهای استخراج دانش از متون فارسی

هستی هم از الگوهای لغوی ـ نحوی و هم از الگوهای معنایی جهت استخراج روابط شمول معنایی، جزء ـ کل، نقش‌های موضوعی، ویژگی ـ مقدار، مالکیت و ... و همچنین برای ساخت اصول بدیهی استفاده می‌کند. نمونه‌هایی از این الگوها در این بخش معرفی شده‌اند.

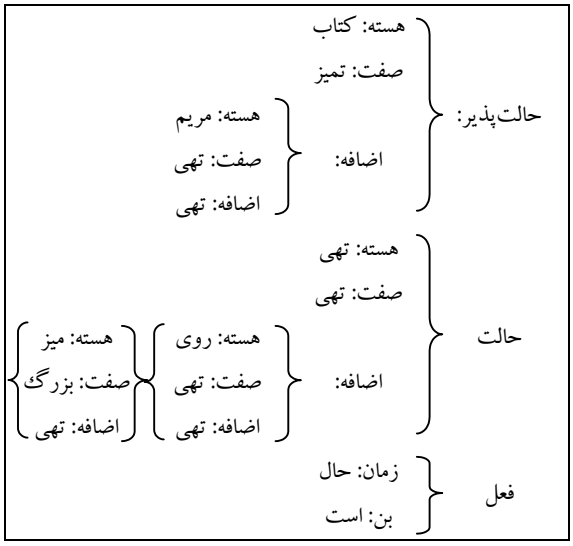
الگوهای نحوی

الگوهای نحوی، توابع متصل به دستور زبان هستند که به سیستم کمک می‌کنند تا ساختار جمله از روی جمله ورودی

شکل ۲- (الف) ساختار جمله حالتی ساخته شده برای جمله «لباس خواهر مینا پیراهن زیبایی بود»
(ب) ساختار جمله حالتی ساخته شده برای جمله «کتاب تمیز مریم روی میز بزرگ است»



(الف)



(ب)

ساخته شود. این الگوها نقش‌های موضوعی اجزای کلام را تعیین می‌کنند و تعیین نقش‌های موضوعی، منجر به استخراج برخی روابط غیر طبقه‌ای جهت ساخت هستان‌شناسی می‌شود.

نقش‌های موضوعی که آنها را نقش‌های معنایی، نقش‌های حالت و یا نقش‌های تتا (theta roles) نیز نامیده‌اند، روابط مفهومی میان مفهوم متناظر با فعل و مفاهیم متناظر با سایر اجزای جمله هستند. محققان مختلف نقش‌های موضوعی متفاوتی برای اجزای یک جمله معرفی کرده‌اند. این تفاوت‌ها ممکن است به دلیل تفاوت در ساختار زبان‌های مختلف، تفاوت در نظریه زبان‌شناسی مورد قبول محقق و یا تفاوت در دیدگاه مفهومی زبان باشد. از این رو، تعداد و نوع نقش‌های لازم برای توصیف روابط معنایی اجزای کلام مورد اتفاق نظر نیست. تعداد نقش‌های پیشنهادی در طیف وسیعی (از تعداد بسیار کم تا صدها نوع) گسترده است. در دو سر این طیف می‌توان از یک سو به نظریه اندرسون (۱۹۷۱) مبنی بر وجود تنها سه نقش مبدأ، موقعیت و هدف و از سوی دیگر، به نظریه پولار و ساگ (۱۹۹۴) مبنی بر وجود نقش‌های موضوعی خاص برای هر فعل (که با نقش‌های فعل دیگر متفاوت‌اند) اشاره نمود. در حالات متعادل میانی، معمولاً تعداد محدودی نقش که اشتراکات بسیاری نیز با هم دارند، پیشنهاد می‌شوند . مثلاً پالمر (۱۹۹۴) نقش‌های اصلی را عامل، پذیرا، ذینفع، ابزار و مکان (یا موقعیت) می‌داند؛ (آرتس، ۱۹۹۷) این نقش‌ها را عامل، پذیرا، موضوع، تجربه‌گر، هدف، ذینفع، مبدأ و موقعیت بر می شمرد و (رادفورد، ۱۹۹۷) آنها را موضوع/ پذیرا، عامل/ علت، تجربه‌گر، دریافت کننده/ مالک و هدف در نظر می‌گیرد.

نقش‌های مورد توجه در هستی، عامل، پذیرا، ذینفع، ابزار، موقعیت (مکان)، زمان، مبدأ، مقصد و توصیف‌گر کنش می‌باشند.

شکل ۳- نمایش صوری ساختار جملات فعلی

ساختار جمله فعلی ← <عامل > <پذیرا > <ذینفع > <ابزار > <مکان > <زمان > <مبدأ> <مقصد> <توصیف‌گر کنش> <کنش>

<عامل> <ساختار گروه اسمی>$^+$ <پذیرا> <ساختار گروه اسمی>$^+$
<مکان> <ساختار گروه اسمی>$^+$ <زمان> <ساختار گروه اسمی>$^+$
<مبدأ> <ساختار گروه اسمی>$^+$ <مقصد> <ساختار گروه اسمی>$^+$
<ابزار> <ساختار گروه اسمی>$^+$ <ذینفع> <ساختار گروه اسمی>$^+$
<توصیف‌گر کنش> <ساختار گروه اسمی>$^+$ <کنش> <فعل>$^+$

این انتخاب محدودیت‌هایی است که فقدان دانش اولیه در مورد دلیل استخراج نقش‌ها، به وجود می‌آورند.

تشخیص نقش کلمات و نوع رابطه آنها با فعل، علاوه بر ساختار نحوی جمله، به ساختار معنایی آن بستگی دارد و نیازمند داشتن دانش معنایی (هم در مورد کلمه مورد نظر و هم درباره فعل جمله) است. برای مثال، در جمله «علی سیب می‌خورد»، *علی* *عامل* یا *کنش‌کار* است؛ در حالی که در جمله «علی سیب دوست دارد»، علی نقش *تجربه‌گر* را دارد و یا در جمله «او کتاب را دید» کتاب *موضوع* محسوب می‌شود؛ در حالی که در جمله «او کتاب را پاره کرد»، کتاب *پذیرا* یا *کنش‌پذیر* است.

در مورد تعیین نقش موضوعی کلمات در جمله، نظرات مختلفی ارائه و الگوریتم‌هایی پیشنهاد شده است (بیکر، ۱۹۹۷). تعیین نقش‌های موضوعی جملات در اعمالی چون حاشیه‌نویسی پیکره‌های زبانی، جهت استخراج اطلاعات مختلف از آنها نیز کاربرد دارد. در سال‌های اخیر، برای خودکارسازی این عمل و انجام آن به وسیله سیستم‌های پردازش متون، روش‌هایی معرفی و آزموده شده‌اند که معمولاً بر اساس ساختار نحوی و ویژگی‌های معنایی کلمات در جمله عمل می‌کنند. به عنوان نمونه‌ای از این تحقیقات می‌توان به اسمیت (۲۰۰۰) اشاره کرد که از روی طبقات

شکل ۴- ساختار جمله فعلی ساخته شده
برای جمله «پدر امین با میخ و تخته میز بزرگ را ساخت»

نحوی کلی در جملات زبان آلمانی، از توابع نحوی برای تعیین نقش‌های عامل، دریافت‌کننده و پذیرا استفاده می‌کند.

در هستی، با توجه به محدودیت دانش اولیه درباره ویژگی‌های کلمات، از رویه ساده و همسانی برای تشخیص نقش‌ها استفاده می‌شود. در این رویه، عامل، متناظر با فاعل جملات معلوم و پذیرا و متناظر با مفعول جملات دارای فعل متعدی در نظر گرفته شده‌اند. در زیر، الگوی نحوی تعیین عامل (فاعل) و پذیرا (مفعول مستقیم) در جملات دارای فعل متعدی آمده است:

<جمله > ← <گروه اسمی > «عامل» <گروه فعلی>
<گروه فعلی> ←.. I<گروه اسمی > «پذیرا» <فعل متعدی> «کنش»

برای تعیین سایر نقش‌ها، بر اساس نوع حرف اضافه‌ای که پیش از آنها ظاهر می‌شود و ویژگی‌های زیر، افعال (در صورت وجود) طبقه‌بندی می‌شوند. به این منظور، نقش‌های محتمل و همچنین نقش پیش فرض (محتمل‌ترین نقش) برای متمم، بعد از هر حرف اضافه در پایگاه دانش نوشته شده است. هر فعل می‌تواند این مقادیر پیش فرض را در مدخل خود در واژگان بازنویسی کند. مثلاً، گرچه نقش پیش فرض برای متمم پس از حرف «به» نقش مقصد است، ولی برای فعل «دادن» می‌توان آن را به نقش ذینفع بازنویسی نمود. به این ترتیب در جمله «مریم به مدرسه رفت»، مدرسه مقصد است؛ در حالی که در جمله «مریم سیب را به علی داد»، علی ذینفع می‌باشد. در صورتی که این اطلاعات زیر طبقه‌بندی، در واژگان موجود نباشند و یا بیش از یک نقش را برای یک متمم معرفی کنند، کلیه نقش‌های محتمل با تعیین حالت پیش فرض برای پردازش‌های بعدی، در نظر گرفته خواهند شد.

الگوهای معنایی

از الگوهای معنایی برای تبدیل ساختارهای جمله به عناصر هستان‌شناسی استفاده می‌شود. آنها روابط معنایی و مفهومی میان مفاهیم مختلف موجود در جمله را تعیین می‌کنند و برای استخراج روابط طبقه‌ای و غیرطبقه‌ای میان مفاهیم استفاده می‌شوند. این

جدول ۱- الگوهای معنایی برای جملات ربطی

قاعده دستوری: جمله ← نهاد مسند فعل ربط،	**نهاد ← گروه اسمی،**		**گروه اسمی ← هسته + صفات + اضافات**	
	شرایط		**نتیجه**	**مثال**
	شرط ۱	**شرط ۲**	**شرط ۳**	
مسند ← گروه صفتی		هسته نهاد زیر طبقه property است	صفت مسند تحت هسته نهاد است	هسته نهاد مشخصه‌ای برای هسته اضافه آن است و مسند مقدار این مشخصه است.	رنگ ماشین علی سبز است.
			در غیر اینصورت	هسته نهاد مشخصه‌ای برای هسته اضافه آن است و مسند صفتی برای این مشخصه است.	رنگ ماشین علی زیبا است.
	در غیر اینصورت			مسند مقدار مشخصه‌ای (نامعلوم) برای هسته نهاد است.	ماشین علی سبز است.
مسند ← گروه اسمی		هسته نهاد تحت مفهوم Property است.	مسند تحت هسته نهاد است	نمونه‌های ساخته شده برای هسته نهاد و هسته مسند هم مرجع هستند.	نام این درخت نارون است.
	هسته نهاد مترادف مسند است.				لباس او لباس زیبایی بود.
	رابطه isa بین نهاد و مسند برقرار است.				علی مرد است. آن سلاح یک کارد بود.
	در غیر اینصورت				علی برادر احمد است.
مسند ← گروه حرف اضافه‌ای	حرف اضافه نشانگر مکان/ موقعیت است (مانند در، روی، زیر)			مسند مکان نهاد است.	کتاب مریم روی میز است.
	حرف اضافه نشانگر جنس/ گروه است (مانند از)			مسند جنس نهاد است.	این انگشتر از طلا است. علی از انقلابیون است.
	حرف اضافه نشانگر زمان است (مانند قبل، بعد، هنگام)			مسند در رابطه زمانی حرف اضافه با نهاد است.	صبح قبل ظهر است.

الگوها را می‌توان حوزه عملکرد آنها و یا بر اساس نوع عنصری که استخراج می‌کنند، طبقه‌بندی نمود. هر دو نوع این طبقه‌بندی‌ها در زیر آمده است.

طبقه‌بندی بر اساس حوزه عملکرد

در این طبقه‌بندی، الگوهای معنایی در چهار دسته اصلی دسته‌بندی می‌شوند:

الف) الگوهای جملات ربطی، جهت استخراج روابط شمول، جزء ـ کل، ویژگی داشتن و ویژگی بودن؛

ب) الگوهای جملات غیر ربطی، جهت استخراج نقش‌های موضوعی و ارتباطات میان فعل و سایر اجزای جمله؛

ج) الگوهای جملات شرطی و سوردار جهت استخراج اصول بدیهی متن؛

د) الگوهای عبارات اسمی، جهت استخراج روابط مفهومی مفاهیم متناظر با صفت و موصوف (مانند ویژگی داشتن) و یا میان مضاف و مضاف‌الیه (مانند مالکیت).

در زیر در مورد هر یک از دسته‌ها توضیحاتی داده شده است.

الف) الگوهای جملات ربطی

این الگوها که برای استخراج روابط مفهومی از جملات ربطی به کار می‌روند، خود، بر اساس نوع مسند به سه گروه تقسیم می‌شوند:

جدول ۲- مثال‌هایی از ترکیبات اضافی و روابط مضاف و مضاف‌الیه

نوع ترکیب	توضیح	مثال	نوع رابطه
اضافه ملکی	مضاف مالک مضاف‌الیه است. در این ترکیب مضاف معمولاً انسان است.	خانهٔ علی، کتاب من، دست مریم	مالکیت ـ HAS جزء ـ کل ـ HAS –PART
اضافهٔ تخصیصی	مضاف اختصاص به مضاف‌الیه دارد. مضاف معمولاً غیر انسان است.	در باغ، فرش اتاق، شاخهٔ درخت	
اضافه بیانی جنسی	مضاف جنس مضاف‌الیه را بیان می‌کند.	درخت میوه، انگشتر طلا	جنسیت، نوع
اضافه توضیحی یا بیانی نوعی	مضاف نوع مضاف‌الیه را بیان یا توضیحی بر آن ارائه می‌کند.	درخت خرما، کتاب گلستان، شهر تهران	نام ـ NAME-OF، نوع

- گروه اول: الگوهای مربوط به جملات ربطی که در آنها مسند یک گروه صفتی است.
- گروه دوم: الگوهای مربوط به جملات ربطی که در آنها مسند یک گروه اسمی است.
- گروه سوم: الگوهای مربوط به جملات ربطی که در آنها مسند یک گروه حرف اضافه‌ای است.

در جدول (۱) سه گروه فوق همراه با انواع حالات در نظر گرفته شده در الگوها و مثال‌های مربوط به جملاتی که می‌پوشانند، آمده است.

همان‌طور که در جدول دیده می‌شود، در شرایطی که مسند گروه صفتی باشد، صفات موجود در آن به عنوان ویژگی‌های هسته نهاد در نظر گرفته می‌شوند. در شرایطی که مسند گروه حرف اضافه‌ای باشد، نوع حرف اضافه آن رابطه میان هسته نهاد و هسته مسند را تعیین می‌کند. هنگامی که مسند گروه اسمی باشد، هسته نهاد و هسته مسند هم مرجع معرفی می‌شوند و سپس با اعمال قواعد دیگری روی این رابطه هم مرجعی، ممکن است روابط جدیدی میان مفاهیم موجود در نهاد و مسند به دست آیند. به علاوه، در شرایطی که مسند گروه اسمی است و نهاد نه ویژگی است و نه مترادف مسند، همان طور که در جدول آمده، میان نهاد و مسند رابطه *isa* برقرار است. معمولاً تشخیص جهت این رابطه چندان ساده نیست. گاهی از روی نکره یا معرفه بودن نهاد یا مسند و یا عام یا خاص بودن آنها می‌توان به جهت رابطه پی برد مثلاً در جمله «علی انسان است» اگر بدانیم علی اسم خاص و انسان اسم عام است، می‌توان نتیجه گرفت که «علی» در سلسله مراتب تحت انسان واقع می‌شود. اما گاهی این قوانین مکاشفه‌ای کاربرد ندارند. مثلاً، به دو جمله زیر توجه کنید:

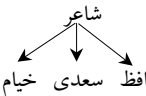
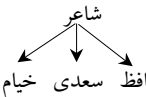
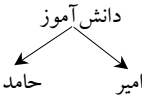
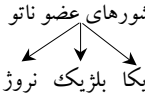
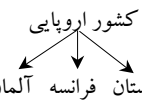
- این اسب یک حیوان است.
- این حیوان یک اسب است.

این دو جمله دارای ساختار یکسان هستند، ولی در اولی، هسته نهاد تحت هسته مسند واقع می‌شود و در دومی، هسته مسند تحت هسته نهاد قرار می‌گیرد. در چنین شرایطی که قوانین مکاشفه‌ای دیگر عمل نکنند، پیش فرض این است که هسته نهاد تحت هسته مسند واقع خواهد شد.

ب) الگوهای جملات غیر ربطی

این الگوها که برای استخراج روابط مفهومی از جملات غیر ربطی به کار می‌روند، ارتباطات میان فعل و سایر اجزای جمله را بر اساس نقش‌های موضوعی استخراج می‌کنند. میزان جزئی بودن روابط استخراج شده به وسیله پارامترهای سیستم قابل تنظیم است. برای این میزان دو حالت متصور است که در یکی فقط روابط فعل با سایر اجزا و در دیگری روابط همه اجزا با هم (فعل و غیر فعل) استخراج می‌شوند که در هر دو حالت، رابطه میان نمونه فعل با نمونه هر جزء، همان نقش موضوعی جزء مورد نظر است و در حالت دوم، رابطه سایر اجزا با هم به فعل بستگی دارد. مثلاً در جمله «علی سیب را خورد» که «علی» کنش کار (عامل)، «سیب» کنش‌پذیر (پذیرا) و «خورد» کنش جمله است، میان نمونه ساخته شده برای «علی» و نمونه

جدول ۳- الگوهای تطبیق یافته هیرست برای فارسی

شماره	الگوی تطبیق یافته	مثال	روابط استخراجی
۱	{چون} <گروه اسمی ش> (مانند ا مثل ا همچون ا چون) { <گروه اسمی >، }* {و ا یا } <گروه اسمی>	شاعرانی مانند حافظ، سعدی و خیام	
۲	{چنین} <گروه اسمی ش> (منجمله ا از جمله) { <گروه اسمی >، }* {و ا یا } <گروه اسمی>	شاعران از جمله حافظ، سعدی و خیام	
۳	<گروه اسمی> { ، <گروه اسمی}* > {،} یا <گروه اسمی ش> دیگر	اردک، غاز یا پرندگان دیگر	
۴	<گروه اسمی> { ، <گروه اسمی}* > {،} و <گروه اسمی ش> دیگر	امیر و حامد و دانش‌آموزان دیگر	
۵	<گروه اسمی ش > {،} شامل { <گروه اسمی >، } * {و} <گروه اسمی>	کشورهای عضو ناتو شامل امریکا، بلژیک و نروژ	
۶	<گروه اسمی ش > {،} شامل (به خصوص ا مخصوصاً) { <گروه اسمی >، }* {و} <گروه اسمی>	همه کشورهای اروپایی، به خصوص انگلستان، فرانسه و آلمان ..	

ساخته شده برای «خورد»، رابطه *کنش‌کاری* و میان فوق رده «علی» (طبقه‌ای که علی متعلق به آن است) و فوق رده «خورد» رابطه قوه *کنش‌کاری* برقرار می‌شود. به همین ترتیب میان نمونه «سیب» و نمونه «خورد» رابطه کنش‌پذیری و میان فوق رده «سیب» و فوق رده «خورد»، رابطه *قوه کنش‌پذیری* ایجاد خواهد شد. همچنین در این مثال «علی» رابطه «خورنده» با «سیب» دارد و «سیب» رابطه «خورده شونده به وسیله» با «علی» خواهد داشت.

ج) الگوهای جملات شرطی و سوردار

این الگوها، به منظور استخراج اصول بدیهی موجود در متن از جملات شرطی یا سوردار به کار می‌روند. در حالتی که جمله به ساختار شرطی تبدیل شده باشد، ابتدا هر دو بخش شرط (مقدم و تالی) به عناصر هستی‌شناسی خاص خود تبدیل می‌شوند. سپس با استفاده از الگوهای جملات شرطی با رابطه استلزام، روابط زمانی (تقدم و تأخر) و یا علی، این دو مجموعه به هم مربوط می‌شوند.

د) الگوهای عبارات اسمی

این الگوها روابط مفهومی موجود در عبارات اسمی را استخراج می‌کنند. هر عبارت اسمی دارای یک هسته و تعدادی وابسته است. هر وابسته که به هسته اضافه می‌شود، ممکن است یک گروه صفتی یا یک گروه اسمی باشد که متناظراً نوع ترکیب را وصفی یا اضافی می‌نامیم. در ترکیبات وصفی، رابطه میان هسته و صفات آن یک رابطه ویژگی داشتن/ ویژگی بودن است که در صورتی که نوع ترکیب (وصفی) درست تشخیص داده شده باشد، استخراج این رابطه بر اساس الگوی وصف به سهولت انجام می‌شود.

در ترکیبات اضافی، تشخیص نوع رابطه مفهومی میان هسته و وابسته به سادگی ترکیبات وصفی نیست. جدول (۲) مثال‌هایی از

وجود روابط مفهومی مختلف میان ساختارهای ترکیبی همسان را نشان می‌دهد.

همان طور که در این مثال‌ها مشهود است، رابطه معنایی میان مضاف و مضاف‌الیه یا هسته و وابسته در یک ترکیب اضافی ممکن است یکی از روابط مالکیت، جزء ـ کل، جنسیت، نام و نوع باشد. برای تشخیص نوع این رابطه مفهومی، به داشتن دانش معنایی و کاربردی در مورد کلمات موجود در ترکیب اضافی و همچنین بافتار متن مورد نظر نیاز داریم. اما از آنجا که در مدل پیشنهادی، فرض بر عدم وجود دانش اولیه و پس زمینه است، لذا برای تشخیص انواع اضافه، از روش‌های مبتنی بر پیکره‌های زبانی استفاده می‌شود.

طبقه‌بندی بر اساس نوع استخراجی

الگوهای معنایی را می‌توان بر اساس آنچه استخراج می‌کنند نیز به صورت زیر دسته‌بندی نمود:

الف) الگوهای استخراج روابط طبقه‌ای

ب) الگوهای استخراج روابط غیر طبقه‌ای

ج) الگوهای استخراج اصول بدیهی

الگوهای استخراج اصول بدیهی، همان الگوهای جملات شرطی و سوردار هستند. در مورد الگوهای استخراج روابط طبقه‌ای و غیر طبقه‌ای در زیر به تفصیل توضیح داده شده است.

الف) الگوهای استخراج روابط طبقه‌ای

روابط طبقه‌ای، روابطی هستند که عناصر مرتبط را در یک طبقه‌بندی یا سلسله مراتب جا می‌دهند؛ مانند رابطه شمول معنایی. معروفترین الگوهای استخراج این روابط، شش الگوی هیرست (هیرست، ۱۹۹۲) هستند که در شکل (۵) معرفی شده‌اند.

در این الگوها گروه‌های اسمی (NP) موجود در هر الگو در رابطه شمول با گروه اسمی که با اندیس H مشخص شده است، قرار می‌گیرند. به عبارت دیگر، در هر الگو NP_H پدر سایر NP ها در سلسله مراتب شمول خواهد بود.

در هستی، این الگوها برای زبان فارسی نیز تطبیق یافته و در جدول ۳ به نمایش در آمده‌اند. در جدول الگوی تطبیق یافته، مثال‌هایی از روابط قابل استخراج مطابق الگو و پاره‌ای توضیحات (در صورت وجود) ارائه شده‌اند. در هر الگو گروه‌های اسمی موجود در رابطه شمول با گروه اسمی که با اندیس «ش» مشخص شده‌اند، واقع می‌شوند.

دراین جدول، الگوی شماره یک، اقتباس لفظی از الگوی اول و دوم هیرست و الگوی شماره دو، اقتباس معنایی از الگوی اول و دوم هیرست است والگوهای شماره سه تا شش متناظرا الگوهای سوم تا ششم هیرست می‌باشند. در این الگوها، فرض بر آن است که کسره اضافه صراحتاً در متن قید شود، در غیر این صورت مثلاً در الگوی ششم عبارت «عطر بخصوص گل‌های بهاری» منجر به استخراج رابطه شمول میان عطر و گل‌های بهاری می‌شود که نادرست است. در ضمن، ذکر این نکته لازم است که این الگوها روی عبارات در یک سازه عمل می‌کنند و در صورتی که مرز عبارت به درستی مشخص نشود، ممکن است استفاده از الگو منجر به خطا شود. مثلاً، در موارد تمثیل مثل جمله «علی چون ابر بهاری می‌گریست»، عبارت «چون ابر بهار» قید است و جزء عبارت نهاد نمی‌باشد و اگر این نکته به وسیله ماشین تشخیص داده نشود، میان علی و ابر بهار رابطه طبقه‌ای ایجاد می‌شود که نادرست خواهد بود.

الگوی دیگر استخراج روابط طبقه‌ای که در هستی معرفی شده، الگوی استثنائات است که شکل کلی آن به صورت زیر است:

{هر ا همه} < گروه اسمی ش > [بجز ا جز] < گروه اسمی >
{(و ا،) < گروه اسمی >}*

مثال: همه پرندگان به جز پنگوئن پرواز می‌کنند.

رابطه طبقه‌ای قابل استخراج: پرنده
↑ isa
پنگوئن

جدول ۴- مثال‌هایی از اعمال قاعده اول هرم مرجع‌ها

جمله شاهد	روابط استخراجی اولیه	رابطه جدید حاصل از اعمال قاعده
علی برادر مریم است	علی ↔ برادر → مریم Has	علی ← مریم *Is-brother-of*
کتاب فردوسی شاهنامه است.	فردوسی ← کتاب ↔ شاهنامه Has	فردوسی → شاهنامه *Is-book-of*

الگوهای تعیین روابط طبقه‌ای که تاکنون دیدیم، الگوهای در سطح عبارات بودند. برای تشخیص رابطه شمول، الگوهای دیگری نیز وجود دارند که در سطح جمله عمل می‌کنند. برخی الگوهای جملات ربطی که در جدول (۱) آمده‌اند، از این قبیل اند. برای مثلاً، در جملات ربطی با الگو " <گروه اسمی ـ۱ > <گروه اسمی ـ۲ > <فعل ربطی >" گروه اسمی ـ۱ تحت رابطه هم مرجعی با گروه اسمی ـ۲ قرار می‌گیرد. سپس با اعمال قاعده دوم هم مرجع‌ها (رجوع به بخش بعد)، طبق شرایطی، رابطه شمول، میان هسته اصلی این دو گروه اسمی برقرار خواهد شد.

ب) الگوهای استخراج روابط غیر طبقه‌ای

برخی از روابط غیر طبقه‌ای جزو هسته هستان‌شناسی و از پیش تعریف شده می‌باشند. الگوهای استخراج این روابط، مفهوم جدیدی برای این روابط ایجاد نمی‌کنند و فقط به استخراج موارد رخداد آنها و کشف مفاهیم مرتبط به وسیله آنها می‌پردازند. مهمترین روابط از پیش تعریف شده، روابط میان مفهوم (کنش) متناظر با فعل و مفاهیم متناظر با نقش‌های موضوعی موجود در جملات غیر ربطی هستند که به وسیله الگوهای جملات غیر ربطی استخراج می‌شوند. همچنین روابط دیگری چون تعلق، هم مرجعی و ویژگی داشتن نیز از روابط هسته می‌باشند. سایر روابط، از پیش تعریف شده نیستند و الگوهای معنایی نه تنها موارد رخداد آنها، بلکه مفهوم متناظر با آنها را نیز ایجاد می‌کنند. به عنوان مثال‌هایی از این گروه می‌توان به روابط برادر، کارمند، هم مکان و رنگ اشاره نمود. در بخش‌های قبل، برای استخراج روابط نقش موضوعی، تعلق، هم مرجعی و ویژگی داشتن/ بودن الگوهایی را معرفی کردیم. در این بخش به معرفی دو قاعده برای استخراج روابط جدید بر اساس رابطه هم مرجعی می‌پردازیم و مثال‌هایی از هر یک ارائه می‌کنیم. در مثال‌ها علامت --- علامت رابطه هم مرجعی و ↔ علامت رابطه نمونه بودن است.

الف) قاعده اول هم مرجع‌ها:

(=> (and (HAS a b) (EQUAL b c))
(IS-B-OF a c))

بر اساس این قاعده، اگر یکی از طرفین (مثلاً طرف اول) رابطه هم مرجعی، تحت رابطه HAS با مفهومی باشد، آن طرف (طرف اول) به صورت رابطه‌ای میان مفهوم مورد نظر و طرف دیگر ظاهر خواهد شد. در جدول ۴، دو مثال از اعمال این قاعده آورده شده است.

ب) قاعده دوم هم مرجع‌ها:

(=> (and (equal a b) (instance-of a c) (instance-of b c))
(merge a b))

بر اساس این قاعده، اگر پدر دو مفهوم هر مرجع یکی باشد، می‌توان آنها را در هم ادغام کرد و ویژگی‌هایشان را به هم افزود. دو مثال از این وضعیت در جدول ۵ آمده است.

جدول ۵- دو مثال از اعمال قاعده دوم هم مرجع‌ها

جمله شاهد	روابط استخراجی اولیه	شکل روابط پس از اعمال قاعده
لباس مریم لباس زیبایی بود.	مفهوم لباس لباس ۱ ↔ لباس ۲ ↑ Has ↓ Has-prop مریم زیبایی	مفهوم لباس لباس – ۱ Has Has-prop مریم زیبایی
خانه حمید خانه مجید است.	مفهوم خانه خانه – ۱ ↔ خانه – ۲ ↑Has ↑Has حمید مجید	مفهوم خانه خانه – ۱ Has Has حمید مجید

ارزیابی

برای ارزیابی عملکرد سیستم از دو معیار دقت و توجه استفاده نمودیم و برای ساخت ساختار جمله و استخراج مفاهیم و روابط طبقه‌ای و غیر طبقه‌ای، این دو معیار را محاسبه کردیم. معیار دقت، نسبت نتایج صحیح تولید شده به کل نتایج تولید شده و معیار توجه، نسبت نتایج صحیح تولید شده به کل نتایج صحیح موجود را نشان می‌دهند.

شکل ۵- الگوهای هیرست برای استخراج روابط طبقه‌ای

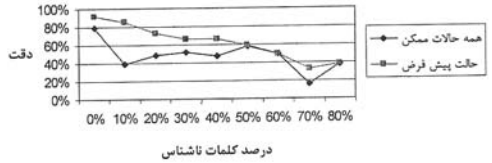
1) NP_H such as {NP,} * {(or | and)} NP
2) Such NP_H as {NP,} * {(or | and)} NP
3) NP {,NP} * {,} or other NP_H
4) NP {,NP} * {,} and other NP_H
5) NP_H {,} including {NP,} * {or | and} NP
6) NP_H {,} especially {NP,} * {or | and} NP

رویه ارزیابی در این سیستم، مقایسه حاصل عملکرد سیستم با حاصل تفکر افراد خبره و غیر خبره است و هستان‌شناسی مرجع، به منظور مقایسه و ارزیابی این افراد تهیه می‌شود. ارزیابی عملکرد بخش مبتنی بر الگو را در هستی، به دو بخش ارزیابی الگوهای نحوی در تولید ساختارهای جمله از روی درخت‌های تجزیه و ارزیابی الگوهای معنایی در استخراج دانش مفهومی از ساختار جملات تقسیم نموده‌ایم. در ضمن، برای ارزیابی هر مرحله، فرض را بر صحت اطلاعات ورودی گذاشتیم. مثلاً، برای ارزیابی مرحله ساخت هستان‌شناسی، فرض شده است که ساختار جمله حاصل از جمله جاری صحیح می‌باشد.

الف) ساخت ساختار جمله

دیدیم که برای ساخت ساختار جمله از روی درخت تجزیه، از الگوهای نحوی استفاده می‌کنیم. این الگوها میزان پیش فرض نقش اجزای کلام را تعیین می‌کنند. به علاوه، مقادیر پیش فرض زیر طبقه‌بندی متمم‌های بعد از هر حرف اضافه نیز در سیستم موجودند. در آزمون هستی بخش، ساختار ساز را در دو حالت استفاده و عدم استفاده از مقادیر پیش فرض فوق آزموده‌ایم. در حالتی که از مقادیر پیش فرض استفاده نشود، سیستم کلیه حالات محتمل را برای نقش کلمه بر اساس جایگاهش در جمله تعیین

شکل ۶- نمایش تغییرات دقت در ساخت ساختار جمله بر اساس تغییرات درصد کلمات ناشناس در جمله



می‌کند. شکل‌های ۶ و ۷ تغییرات متوسط دقت و توجه را بر اساس

جدول ۶- ارزیابی سیستم در مرحله ساخت هستان‌شناسی از روی ساختارهای جمله برای دو نمونه

معیار	متن کتاب فارسی کلاس اول دبستان	متن فنی ساده در قلمروی کامپیوتر
درصد کلمات ناشناس	٪۸۶	٪۸۱
دقت	٪۹۷	٪۷۸
توجه	٪۸۸/۵	٪۷۹/۶

تغییرات درصد کلمات ناشناس جمله نمایش می‌دهند. همان طور که دراین شکل‌ها دیده می‌شود، در استفاده از پیش فرض‌ها با از دست دادن درصدی از معیار توجه، درصد دقت را افزایش می‌دهیم. این کار با توجه به کاهش حجم محاسبات، کارآیی سیستم را افزایش خواهد داد.

بر اساس آزمایش‌های مختلف، در حالتی که از مقادیر پیش فرض استفاده کنیم متوسط معیار دقت برای تشخیص نقش‌های موضوعی و ساخت ساختار جمله ۶۲/۵ درصد و متوسط معیار دقت ۸۵/۱۴ درصد می‌باشند و در حالتی که تمام حالات محتمل را در نظر بگیریم، متوسط دقت ۴۷/۵ درصد و معیار توجه ۱۰۰ درصد خواهد بود. بنابراین، مقایسه نشان می‌دهد که مقادیر پیش فرض انتخاب شده در درصد بالایی از حالات، مقادیر صحیحی هستند.

ب) استخراج دانش مفهومی از ساختارهای جمله

در این مرحله، معیارهای دقت و توجه برای استخراج مفاهیم و روابط طبقه‌ای و غیر طبقه‌ای محاسبه می‌شوند. این مرحله در سطح متن ارزیابی می‌شود. به این منظور دو نمونه متن یکی برگرفته از کتاب فارسی کلاس اول دبستان و دیگری یک متن علمی ساده در قلمروی کامپیوتر را برگزیدیم. جدول (۶) معیارهای دقت و توجه را برای این دو نمونه نشان می‌دهد. در این آزمایش، هسته هستان‌شناسی فقط دارای هفت مفهوم و هفت رابطه اولیه بوده است.

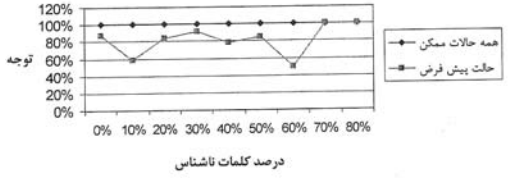
نتیجه‌گیری

در این مقاله، ضمن معرفی سیستمی برای استخراج دانش مفهومی از متون زبان فارسی و ساخت هستان‌شناسی از پایه، به بررسی عملکرد مبتنی بر الگو در این سیستم پرداختیم. در این راستا، الگوهایی برای استخراج دانش از زبان فارسی معرفی و نمونه‌هایی از عملکرد آنها را بررسی نموده‌ایم. در بخش ارزیابی، معیارهای دقت و توجه را برای اعمال الگوهای نحوی و معنایی به طور جداگانه اندازه گرفتیم. نتایج نشان می‌دهند که الگوهای معرفی شده قادر به استخراج درصد بالایی از روابط مورد نظر از متن می‌باشند.

به این ترتیب، نه فقط با فراهم آمدن امکان ساخت خودکار هستان‌شناسی از متون زبان طبیعی، مشکلات ساخت دستی این پایگاه‌های دانش بر طرف می‌شود، بلکه با توجه به معرفی روش ساخت از پایه (بدون نیاز به پایگاه دانش مبنا) درگیر گلوگاه اکتساب دانش اولیه نیز نخواهیم شد.

الگوهای معرفی شده نه فقط برای ساخت هستان‌شناسی، بلکه در کلیه کاربردهای استخراج اطلاعات و اکتساب دانش مفهومی و یا استخراج ساختارهای اطلاعاتی خاص از متون زبان فارسی کاربرد دارند. برای تکمیل و ادامه این پژوهش انجام آزمون‌های گسترده‌تر روی پیکره‌های زبانی، یافتن الگوهای بیشتر برای پردازش جملات پیچیده‌تر و کار روی استخراج خودکار الگوهای جدید پیشنهاد می‌گردد.

شکل ۷- نمایش تغییرات توجه در ساخت ساختار جمله بر اساس تغییرات درصد کلمات ناشناس در جمله



منابـــع

شمس فرد، م.(۱۳۸۱). *طراحی مدل یادگیر هستان شناسی: نمونه سازی در یک محیط درک متن فارسی*، رسالهٔ دکترای مهندسی کامپیوتر - هوش مصنوعی، دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر.

شمس فرد، م.، و عبدالله زادهٔ بارفروش، ا.(۱۳۸۰). واژگان محاسباتی: ساختار مرکزی در سیستم های پردازش زبان طبیعی، *مجلهٔ امیرکبیر*، (۴۸)۱۲، تهران: انتشارات دانشگاه صنعتی امیرکبیر.

مشکوه الدینی، م.(۱۳۷۴).*دستور زبان فارسی بر پایهٔ نظریهٔ گشتاری*، مشهد: انتشارات دانشگاه فردوسی.

Aarts, B. (1997). *English syntax and argumentation*, London: Macmillan Press.

Agirre, E., Ansal, O., Hovy, E., & Martinez, D. (2000). Enriching very large ontologies using the WWW, *Proceedings of the Ontology Learning Workshop,* ECAI, Berlin, Germany.

Anderson, J.M. (1971). The grammar of case: *Towards a localistic theory,* Cambridge: Cambridge University Press.

Assadi, h. (1997). Knowledge acquisition from texts: Using an automatic clustering method based on noun-modifier relatioship, *Proceedings of 35th Annual Meeting of the Association for Computational Linguistic*, Madrid, Spain.

Assadi, H. (1999). Construction of a regional ontology from text and its use within a documentary system, *Proceedings of the International Confetence on Formal Ontologies and Information System, (*FOIS, 98), Torento, Italy.

Baker, M. (1997). Thematic roles and syntactic structure, in L. Haegemann. (ed), *Elements of grammar:* Handbook in generative syntax, Dordresht; Kluwer,73-137.

Bowers, A.F., Giraud-Carrier, C., & Loyd, J.W. (2000). Classification of individuals with complex structure. *Proceedings of the 17th International Conference on Machine Learning:* M. Kaufmann,81-8.

Chalendar, G., & Grau, B. (2000). SVETLAN: System to classify nouns in context *Proceedings of the First Workshop on Ontology Learning (OL 2000), In Conjunction with the 14th European Conference on Artificial Inteligence, (ECAI 2000),*Berlin, Germany.

Chapulsky, H., Hovy, E., & Russ, T.(1997). Progress on an automatic ontology alignment methodology.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., & Nigam, K.(2000). Learning to extract symbolic knowledge from the Word Wide Web, *Artificial Intelligence, 118,*69-113.

Domingue, J., & Tadzebao, (1998).WebOnto: Discussing, browsing, and editing ontologies on the Web, *Proceedings of the Eleventh Workshop on Knowledge Acquisition, Modeling and Management,* KAW,98,Banff,Canada.

Eriksson, H., Fergerson, R.W., Shahar, Y., & Musen, M.A.(1999). *Automatic generation of ontology editors for knowledge-based system workshop*. Banff, Alberta, Canada.

Farquhar, A., Fikes, R., & Rice, J.(1997). The ontolingua server: Tool for collaborative ontology construction, *IJHCS,46*(6),707-28.

Faure, D., Nedellec, C., & Rouveirol, C.(1998). Acquisition of semantic knowledge using machine learning methods: The system ASIUM, *Technical Report Number ICS-TR-88-16,* Universite Paris-Sud.

Finkelstein-Landau, M., & Morin, E. (1999). Extracting semantic relationships between terms: Supervised vs. Unsupervised methods, actes, *International Workshop on Ontological Engineering on the Global Information Infrastructure,*71-80,Dagstuhl-Castle, Germany.

Gamallo, P., Gonzalez, M., Agustini, A., Lopes, G.P., & de Lima, V.(2002). Mapping syntactic dependencies onto semantic relations. *Workshop OLT,2002,*Lyon.France.

Hahn, U., & Romacker, M.(2000). Content management in the SynDIKATE system: How technical documents are automaticaly transformed to text knowledge bases. *Knowledge Engineering,35*(2), 137-59.

Hearst, M.(1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the Fourteenth*

International Conference on Computational Linguistics, Nantes, France.

Heyer, G., Lauter, M., Quasthoff, U., Wittig, T., & Wolff, C.(2001). Learning relations using collocations, *Proc. Of IJCAI,2001, Workshop on Ontology Learning,* Seattle, Washington, USA.

Hwang, C.H.(1999). Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. *Proceedings of the 6th International Workshop on Knowledge Representation Meets Databases(KRDB,99),* Linkoping, Sweden.

Kashyap, V.(1999). Design and creation of ontologies for environmental information retrieval. *Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management (KAW, 99),*Banff, Alberta, Canada.

Lenat,D.B.(1995). CYC: A large-scale investment in knowledge infrastructure, *Communications of the ACM,38*(11),33-8.

Maedche, A., & Staab, S.(2001). Ontology learning for the semantic Web,*IEEE Intelligent System,16*(2),72-9.

Maedche, A., & Staab, S.(2001b). Comparing ontologies-similarity measure and a comparison study, *Internal Report,408,*Institute AIFB, University of Karlsruhe.

Miller, G.A. (1995). A lexical database for english, *Communications of the ACM,38*(11), 39-41.

Nirenburg, S., Raskin, V., & Onyshkevych, B.(1995). *Apologiae Ontologiae,* Memoranda in computer and cognitive science ,MCCS,95-281.

Noy, N.F., & Musen, M.A.(2000). PROMPT: Algorithm and tool for automated ontology merging and alignment, *Seventeenth National Conference on Artificial Intelligence(AAAI-2000),* Austin,TX.

Palmer, F.R.(1994). *Grammatical roles and relations,* Cambridge, Cambridge University Press.

Pernelle, N., Rouset, M.C., & Ventos, V.(2001). Automatic construction and refinement of a class hierarchy over semi-structured data, *IJCAI,2001 Workshop on Ontology Learning(OL,2001),* Seattle, USA.

Pollard, C., & Sag, I.A.(1994). *Head-driven phrase structure grammar,* Chicago: University of Chicago Press.

Radford, A.(1997).*Syntactic theory and the structure of english: A minimalist approach.*(Cambridge Text books in Linguistics.) Cambridge: Cambridge University Press.

Riloff, E.(1996). Automatically generating extraction patterns from untagged text. *Proceedings of the Thirteenth Conference on Artificial Intelligence,*1044-49,AAAI/MIT Press.

Ryutaro, I., Hideaki, T., & Shinichi, H.(2001). Rule induction for concept hierarchy alignment, *Proceedings of the 2nd Workshop on Ontology Learning at the 17th Int. Joint Conf. On AI(IJCAI).*

Shamsfard, M., & Barforoush, A.A.(2002).An introduction to hasti: An ontology learning system, *6th IASTED International Conference on Artificial Intelligence and Soft Computing, (ASC,2002),* Banff, Canada.

Shamsfard, M., & Barforoush, A.A.(2003). The state of the art in ontology learning: A framework for comparison, *Knowledge Engineering Review,* Accepted *for publication.*

Smith, G.(2000). Encoding thematic roles via syntactic functions in a german treebank, *Workshop on Syntactic Annotation of Electronic Corpora,* Tubingen, Germany.

Soderland, S., Fisher, D., Aseltine, J., & Lehnert, W.(1995). Issues in inductive learning of domain-specific text extraction rules, *Proceedings of the Workshop on New Approaches to Learning for Natural Language Processing at the Fourteenth International Joint Conference on Artificial Intelligence.*

Sundblad, H.(2002). Automatic acquisition of hyponyms and meronyms from question corpora. *In Proceeding and Machine Learning for Ontology Engineering at ECAI,2002,*Lyon,France.

Suryanto, H., & Compton, P.(2000). Learning classification taxonomies from a classification knowledge based system, *Proceedings of the Workshop on Ontology Learning,14th European Conference on Artificial Intelligence, ECAI,2000,* Berlin, Germany.

Swartout, B., Patil, R., Knight, K., & Russ, T.(1997). Toward distributed use of large-scale ontologies. *Spring Symposium on Ontological Engineering,* Stanford, California.

Williams, A.B., & Tsatsoulis, C.(2000). An instance-based approach for identifying candidate ontology relations within a multi-agent system, *In Proceedings of the First Workshop on Ontology Learning (OL-2000), In Conjunction with the 14th European*

Conference on Artificial Intelligence (ECAI, 2000), Berlin, Germany.

Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. *In Proceedings of the ECAI-2000 Workshop on ontology learning,* Berlin, Germany.

Yamaguchi, T.(2001). Acquiring coceptual relations from domain-specific texts, *Proceedings of IJCAI Workshop on ontology learning,* Seattle, Washington, USA.