ساخت هستان‌شناسی دانش عرفی زبان فارسی با رویکردی تلفیقی

مهدی مرادی
کارشناسی ارشد زبان‌شناسی رایانشی؛
دانشگاه صنعتی شریف mehdi.moradi.cl@gmail.com

بهرام وزیرنژاد
دکتری مهندسی پزشکی؛
استادیار؛ دانشگاه صنعتی شریف
پدیدآور رابط bahram@sharif.edu

محمد بحرانی
دکتری هوش مصنوعی؛
استادیار؛ دانشگاه صنعتی شریف bahrani@sharif.edu

دریافت: ۱۳۹۳/۰۳/۲۷ **پذیرش:** ۱۳۹٤/۰۱/۲۳ مقاله برای اصلاح به مدت ۳۹ روز نزد پدیدآوران بوده است.

پژوهشنامهٔ پردازش و مدیریت اطلاعات
فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC و SCOPUS
http://jipm.irandoc.ac.ir
دورهٔ ۳۱ | شمارهٔ ۱ | ص‌ص ۱۰۹- ۱۲٤
پاییز ۱۳۹٤

مقالهٔ پژوهشی

چکیده: تجهیز رایانه‌ها به دانش عرفی بشر همواره یکی از جاه‌طلبانه‌ترین اهداف علم هوش مصنوعی بوده است. میلیون‌ها دلار هزینه و هزاران ساعت زمان صرف شده تا رایانه‌ها بفهمند که «اشیاء بالا نمی‌روند، بلکه می‌افتند» و «دویدن از راه رفتن سریع‌تر است». پایگاه‌های دانش عظیمی ساخته شده، روش‌های خودکار و نیمه‌خودکار متن‌کاوی پیشنهاد شده و از انگیزهٔ همکاری کاربران عادی اینترنت به نفع اکتساب این دانش بهره‌ها برده شده است، ولی رسیدن به روشی خودکار، مؤثر و کم‌خطا همچنان به‌صورت چالشی بزرگ پیش روی جامعهٔ هوش مصنوعی باقی مانده است. هدف این تحقیق ساخت هستان‌شناسی دانش عرفی فارسی به کمک سه روش مبتنی بر الگو، ترجمهٔ ماشینی و استفاده از منابع ساخت‌یافته است. با کمک سه پیکرهٔ مختلف فارسی هفت نوع رابطه و در مجموع هفتاد هزار اظهار (رابطه) استخراج شده و در قالب یک هستان‌شناسی ارائه شد. نتایج بررسی گویشوران فارسی نشان داد که میانگین دقت روابط استخراج‌شدهٔ این هستان‌شناسی برابر با ۷۵ درصد برای روش مبتنی بر الگو، ۷۰ درصد برای ترجمهٔ ماشینی و ۱۰۰ درصد برای اطلاعات استخراج‌شده از جعبهٔ اطلاعات بود.

کلیدواژه‌ها: دانش عرفی؛ هستان‌شناسی؛ استخراج رابطه

۱. مقدمه

با وجود پیشرفت‌های خیره‌کنندهٔ عصر تکنولوژی، هنوز رایانه‌ها به آن درجه از درک و شعور نرسیده‌اند که هفت ژوئن هر سال به احترام پدرشان «آلن تورنیگ»[1] لحظه‌ای سکوت کنند و به سوگ بنشینند. رایانه‌های امروزی قادرند ماهرترین شطرنج‌بازان را شکست دهند، قضایای مهم ریاضی را کشف و اثبات کنند، علائم و مراحل پیشرفت بیماری را تشخیص دهند، ولی هنوز نمی‌دانند که اشیاء بالا نمی‌روند، بلکه می‌افتند، پدران از فرزندانشان بزرگ‌ترند و دادزدن نشان از عصبانیت فرد دارد. عمده‌ترین دلیل برای ناتوان جلوه‌کردن رایانه‌ها، عدم دسترسی به همین حقایق پیش‌پاافتاده و ساده است. همین حقایق ساده وجه تمایز میان انسان و سیستم‌های هوشمند نسل جدید است (Lieberman 2008). هر چند استفاده از روش‌های مبتنی بر کلمات کلیدی و روش‌های آماری در سیستم‌های هوشمند، بازیابی اطلاعات، داده‌کاوی و پردازش زبان طبیعی به موفقیت‌هایی دست یافته است، ولی متخصصان این حوزه معتقدند که این روش‌ها تنها درکی سطحی از متون ارائه می‌دهند و برای پیشرفت در مدیریت اطلاعات و دستیابی به درکی عمیق‌تر از متون، برنامه‌هایمان نیازمند دسترسی به حجم گسترده‌ای از دانش معنایی دنیای پیرامون خود هستند؛ دانشی که دربرگیرندهٔ طیف وسیعی از تجربه و توانایی بشری است، حقایقی در مورد دانش و چگونگی اکتساب آن، حقایقی در مورد رخدادها و پیامدهایشان، اشیاء و ویژگی‌های آنها و نیز عقاید و علایق (McCarthy 1990). میلیون‌ها حقیقت کوچک روزمره و توانایی استفاده از این حقایق در موقع نیاز، گنجینهٔ دانش عرفی[2] (فهم عمومی، فهم عرفی، یا فهم متعارف) بشر را تشکیل می‌دهد. دانش عرفی به مجموعهٔ حقایقی گفته می‌شود که انتظار دانستنشان از هر فرد عادی می‌رود. «سوزن تیز است»، «برای بازکردن در، ابتدا باید دستگیره را چرخاند» و «مردم معمولاً شب‌ها می‌خوابند». محققان هوش مصنوعی از دیرباز به اهمیت بازنمایی محاسباتی این دانش در سیستم‌های هوشمند واقف بوده‌اند (McCarthy 1959) و تلاش‌های گسترده‌ای نیز برای بازنمایی این دانش انجام شده است. مراکز علمی - پژوهشیِ جهان سرمایهٔ مالی- انسانی هنگفتی صرف تجهیز رایانه‌ها به دانش عرفی کرده‌اند (Lenat 1995؛ Sing and et al. 2002) و دههٔ اخیر نیز شاهد گام‌هایی بلند در زمینهٔ استخراج، بازنمایی و استفاده از این دانش بوده است. پایگاه‌های دانش نسبتاً بزرگی ساخته شده و کاربردهای مختلفی برای آنها پیشنهاد شده است. سیستم‌های پرسش و پاسخ (Q/A)[3]، سیستم‌های استنتاج‌گر (RTE)[4]، موتورهای جستجو، سامانه‌های چندعامله[5]، خلاصه‌سازی متن، و تشخیص احساس[6] نمونه‌هایی از کاربردهای این دانش است. با این وجود، تلاش‌های

1. Alan Turing
2. commonsense knowledge
3. question answering
4. recognizing textual entailment
5. multi-agent system
6. emotion detection (sentiment analysis)

صورت‌گرفته برای استخراج این دانش برای زبان فارسی و نیز تنوع روش‌های مورد استفاده برای این زبان، محدود و انگشت‌شمار بوده است (شمس‌فرد، و عبداله‌زادهٔ بارفروش ۱۳۸۱).

۲. مروری بر منابع

۲-۱. دانش عرفی

فیلسوفان، اولین کسانی بودند که دانش عرفی، ذهن پرسشگرشان را به اندیشه واداشت. از نظر ارسطو دانش عرفی[1] از قوای باطنی است که صوَر حاصل از حواس ظاهری در آن جمع و به‌وسیلهٔ آن ادراک می‌شوند (Lories 1991). رواقیون این دانش را «درک عمومی»[2] می‌نامیدند، یعنی عقایدی که همگان بدون نیاز به اثبات و استدلال بدان اعتقاد دارند (Lories 1998). بنا به عقیدهٔ دکارت دانش عرفی جزو معدود چیزهایی است که عادلانه بین همگان توزیع شده و حریص‌ترین افراد نیز نیازی به اکتساب بیشتر آن نمی‌بینند (Descartes 2008). «جان‌مک‌کارتی» اولین کسی بود که به ضرورت این دانش در سیستم‌های هوشمند پی برد. به نظر وی «برنامه‌ای دارای دانش عرفی است» که بتواند به‌صورت خودکار، برای خود در مورد طیف گسترده‌ای از پیامدهای بی‌درنگ هر چیزی که به آن ارائه می‌شود و نیز هر آنچه از قبل می‌داند، استنتاج کند (McCarthy 1959).

۲-۲. استخراج و بازنمایی دانش عرفی

۲-۲- ۱. سایک

«سایک»[3] جزو اولین پروژه‌ها در زمینهٔ بازنمایی دانش عرفی بود. مهندسان دانش استخدام‌شده، به‌صورت دستی حقایق اولیهٔ مرتبط با فهم عمومی را در قالب زبان «سایک ال» وارد این آنتولوژی می‌کردند. «سایک ال» زبان اختصاصی «سایک» بود که علاوه بر دقت، فاقد ابهامات زبان طبیعی نیز بود. با کمک این حقایق اولیه، حقایق بیشتری (برای مثال، از طریق استنتاج) نیز به پایگاه دانش این پروژه افزوده گشت (Lenat 1995). با این وجود، واردکردن میلیون‌ها حقیقت عرفی، خارج از توان چند ده مهندس دانش بود و این امر پیشرفت پروژه را کند و پرهزینه می‌کرد.

۲-۲-۲. پروژهٔ اوپن مایند کامن‌سنس

پروژهٔ اوپن مایند کامن‌سنس[4] خلاقانه‌ترین روش برای جمع‌آوری این دانش را به‌کار گرفت. این پروژه که در مدیا لب[5] دانشگاه ام‌آی‌تی کلید خورد، برخلاف پروژه «سایک»، متکی بر

1. Sensus communis
2. Ennoai koinai
3. Cyc
4. Open mind commonsense project
5. Media lab

کاربران عادی اینترنت بود. پانزده هزار داوطلب در مدت زمانی کم و بدون هیچ چشم‌داشتی بیش از یک میلیون حقیقت عرفی را به پایگاه این آنتولوژی افزودند (Push Singh et al. 2002).

۲-۲- ۳. بازی‌های هدف‌دار[1]

در پروژهٔ «اوپن مایند کامن‌سنس»، مستقیماً از کاربران خواسته می‌شد که دانش عرفی را وارد پایگاه دانش کنند. ولی، اخیراً رویکرد دیگری توجه پژوهشگران را به خود جلب کرده است: استفاده از بازی‌های هدف‌دار که در آن امتیاز هر بازیکن به‌کمک دانشی که ارائه می‌کند، سنجیده می‌شود. پیشگامان این ایدهٔ خلاقانه (رایانش انسان‌محور[2]) Luis von Ahn et al. (2006) بودند که بازی ورباسیتی[3] را برای جمع‌آوری دانش عرفی از کاربران طراحی کردند. «کامن کانسِنسوس»[4] نام بازی دیگری بود که هدف اصلی طراحانش جمع‌آوری اهداف انسان به‌عنوان بخشی از دانش عرفی بود (Lieberman et al. 2007).

۲-۲-٤. روش مبتنی بر الگو

در این روش، که برای اولین بار توسط «هرست» معرفی شد، به استفاده از کلمات خاص و نیز به کمک مقولهٔ نحوی کلمات، دانش مفهومی از متن غیرساخت‌یافته استخراج می‌شود (Hearst 1992). برای افزایش دقت در این متد، از دو روش بهره می‌برند. روش اول، استفاده از چندین الگوی مختلف برای هر رابطه (جهت پوشش حداکثری جملات هدف)، و روش دوم، انجام پس‌پردازش بر روی جملات استخراج شده است. «شمس‌فرد و عبداله‌زادهٔ بارفروش» با معرفی الگوهای زبانی و معنایی که مستقل از دامنه و کاربرد هستند، توانستند روابط طبقه‌ای و غیرطبقه‌ای و اصول بدیهی را از جملات فارسی استخراج کنند (شمس‌فرد و بارفروش ۱۳۸۱). «مرادی» و همکاران توانستند با استفاده از روش مبتنی بر الگو از ۱۵۰ هزار مقالهٔ فارسی ویکی‌پدیا ۳۸۹۲ جملهٔ عمومی استخراج کنند (۱۳۹۱).

۳. مواد و روش‌ها

۳-۱. پیکره‌ها

۳-۱-۱. پیکرهٔ همشهری

این پیکره یکی از معتبرترین پیکره‌ها در زبان فارسی بوده و اساسش متون خبری روزنامهٔ همشهری است. این پیکره در فرمت xml است (آل احمد و دیگران ۲۰۰۹).

1. game with a purpose (GWAP) 2. human computation 3. Verbasity
4. common consensus

جدول ۱. پیکرهٔ همشهری

پیکرهٔ همشهری	
معیار	نسخه ۱
حجم (یونی‌کد[1])	۷۰۰ مگابایت
تعداد اسناد	۱۶۰۰۰۰
تعداد کلمه	۶۳۵۱۳۸۲۷

۳-۱-۲. پیکرهٔ بی‌جن‌خان[2]

حجم داده‌های این پیکره یک‌صد میلیون کلمه و دارای داده‌های نوشتاری (حدود ۴۱۷ مگابایت و شامل نود میلیون کلمه) و داده‌های گفتاری (در حدود ده میلیون کلمه) است. مجموعهٔ نود میلیون کلمه‌ای از پیکرهٔ یاد شده در این تحقیق مورد استفاده قرار گرفت. با توجه به حجم این پیکره از متون نگارشی فارسی و نیز معیاربودن آن از نظر تنوع متنی، این پیکره یکی از جامع‌ترین پیکره‌های فارسی است (امیری و دیگران ۱۳۸۵).

جدول ۲. پیکرهٔ بی‌جن‌خان

پیکرهٔ بی‌جن‌خان	
معیار	نسخه ۱
حجم	۴۷۶
تعداد اسناد	۳۵۰۵۹
نوع سند	متن ساده
تعداد کلمه	۹۹۷۶۱۱۱۴

۳-۱-۳. ویکی‌پدیا

ویکی‌پدیا دانشنامه‌ای اینترنتی، چندزبانه و با محتویات آزاد است که با همکاری افراد داوطلب نوشته می‌شود و مقالات آن می‌تواند توسط هر کسی که به اینترنت دسترسی دارد، ویرایش گردد. ویکی‌پدیای فارسی با داشتن حدود ۲۳۰ هزار مقاله در حوزه‌های مختلف پیکرهٔ متنی مناسبی جهت مطالعات زبانی می‌باشد. ویکی‌پدیا به کاربران این امکان را می‌دهد که نسخهٔ کاملی از تمامی مقالات را در قالب یک فایل با فرمت Xml دانلود کنند. برای انجام این تحقیق نسخه 2012/06/03 این دانشنامه بارگذاری شده و مورد استفاده قرار گرفت.

1. unicode 2. Bijakhan

جدول ۳. پیکرهٔ متنی ویکی‌پدیا

پیکرهٔ ویکی‌پدیا	
معیار	نسخه ۱
حجم	۲۷/۱ گیگا
تعداد اسناد	۱۷۸۰۰۰
نوع سند	XML
تعداد کلمه	۲۹۷۳۴۹۲۳

۳-۲. روش استخراج

در این تحقیق سعی شده است که الگوهای مورد جستجو با توجه به روابط تعریف‌شده در «کانسپت‌نت»[1] انگلیسی استخراج شود (ضمایم جدول ۲). روش استخراج دانش عرفی در این پروژه یک روش ترکیبی مشتمل بر رویکردهای مختلف نظیر استخراج جملات دارای الگوهای خاص از پیکره‌های زبانی، ترجمهٔ دانش عرفی موجود در «کانسپت‌نت» و نیز استفاده از اطلاعات ساخت‌یافتهٔ «ویکی‌پدیا» بوده است. هر یک در ادامه توضیح داده شده است.

۳-۲-۱. استخراج به روش مبتنی بر الگو

در روش‌های مبتنی بر الگو، با کمک الگوها و کلیدواژه‌های خاص روابط مورد نظر استخراج می‌شود رابطه "Is A"، "Used For"، "Made Of"و روابط تضاد و هم‌معنایی از این طریق استخراج شد. از تکنیک «بوت استرپینگ»[2] برای استخراج الگوها استفاده شد.

◇ تکنیک «بوت‌استرپینگ» برای استخراج الگوها:

۱) استخراج الگو:

۱. انتخاب چند نمونه که به‌خوبی رابطهٔ مورد نظر را نشان دهد
۲. جستجوی عبارات فوق در موتور جستجوی گوگل
۳. ذخیرهٔ ۱۰۰ صفحهٔ اول حاوی این جستجوها
۴. تقطیع هر سند به جملات سازندهٔ آن
۵. ذخیرهٔ جملاتی که حاوی کلمات جستجو بودند و نیز نرمال‌سازی نویسه‌ها و فاصله‌ها و نیز حذف برچسب‌های html.

۲) بازنویسی الگوهای استخراج‌شده در قالب عبارات منظم

1. ConceptNet
2. Bootstrapping

۳) استخراج جملات دارای الگوی مورد نظر

۴) انجام پس‌پردازش به‌منظور حذف عبارات نامرتبط

برای مثال: برای رابطه "Is A" پرس‌وجوهای ارسالی به گوگل عبارات زیر بودند:

الف- «شیر، ماست، پنیر، خامه، کره، بستنی، شیرخشک»

ب- «دیابت، قلب، کلیه، کیسهٔ صفرا»

ج- «عدس، لوبیا، نخود»

د- «مس، نقره، پلاتین»

پس از بررسی جملات استخراج‌شده از نتایج گوگل، الگوهای زیر برای استخراج رابطه "Is A" بازنویسی شد.

جدول ٤. الگوهای رابطه "Is A"

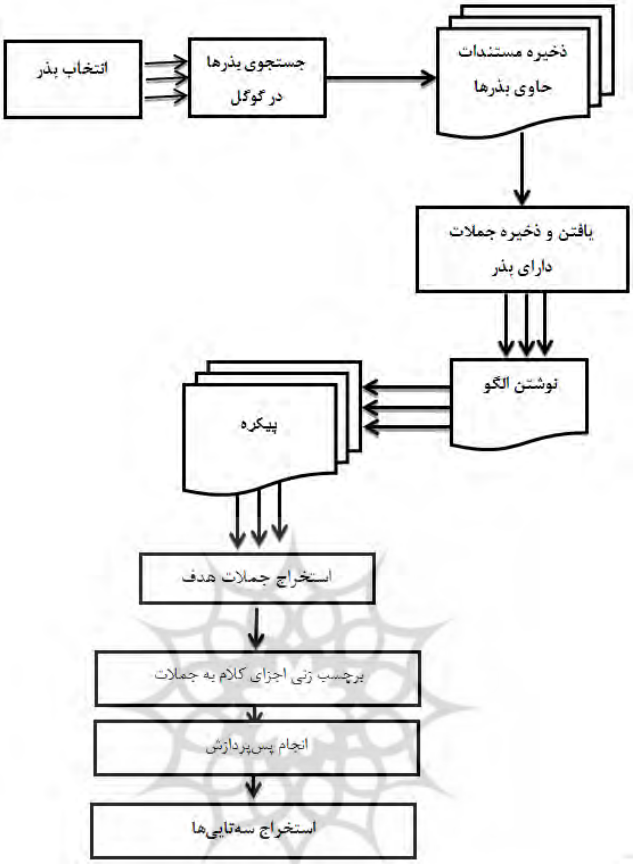
شماره	الگوی مورداستفاده	مثال
۱	Xهایی (نظیر| مانند| همانند| همچون| چون| از قبیل| مثل| ازجمله| از قبیل) y1, y2, y3	گازهایی مانند دی‌اکسید کربن، سولفید، هیدروژن، آمونیاک، مونوکسید
۲	Y1, y2, y3 (ازجمله)، xهای	«آرماگدون»، «گودزیلا» و «پارک ژوراسیک» از جمله فیلم‌های شاخص
۳	Xها عبارت‌اند از y1, y2, y3	غذاهای دیرهضم عبارت‌اند از: حبوبات و غلات مثل جو، گندم، جو دو سر
۴	Xها اعم از y1, y2, y3	دانه‌ها اعم از گندم، جو یا عدس
۵	انواع Xها (y1, y2, y3)	انواع فلزات (تیتانیوم، مولیبدن، طلا، نقره، استیل، آهن، برنز، برنج
۶	Xها (به‌خصوص، به‌ویژه، مخصوصاً) y1, y2, y3	همهٔ کشورهای اروپایی، به‌خصوص انگلستان، فرانسه و آلمان ...

پس از استخراج جملات پس پردازش‌هایی بر روی آنها انجام گرفت تا بخش‌های زائد جملات حذف گردد.

الگوهای زیر برای استخراج رابطه "Used For" بازنویسی شدند:

جدول ۵. الگوهای مورد استفاده برای استخراج رابطه "Used For"

شماره	الگوی مورد استفاده	مثال
1	X (برای| به‌منظور| جهت) y به‌کار برده می‌شود	ناقوس‌ها برای فراخوان دفاع شهری و اعلام خطر آتش‌سوزی نیز به‌کار برده می‌شود.
2	X برای Y ضروری (است| می‌باشد).	اتصال کلسیم به میوزین برای انقباض عضلات صاف جدار عروق ضروری است.
3	برای X از Y (کمک| استفاده|)	برای بریدن ورق‌ها از انواع قیچی‌های دستی و یا ماشینی کمک می‌گیرند.
4	از X برای Y (استفاده)	از http برای دسترسی به برگه‌های وب نرمال استفاده می‌شود.
5	هدف از X، Y است	هدف از عکاسی اچ‌دی‌آر، افزایش محدودهٔ دینامیکی است.
6	X برای (درمان) Y تجویز	هالوپریدول برای درمان روان‌پریشی تجویز می‌شود.
7	X وسیله است برای Y	پَرگار وسیله‌ای برای کشیدن دایره است.



شکل ۱. روند کلی استخراج چندتایی‌ها با روش مبتنی بر الگو

۳-۲-۲. ترجمهٔ ماشینی

بخش بزرگی از دانش عرفی در قالب جملات عمومی[1] بیان می‌گردد؛ جملاتی مانند: الف) فلزات جامدند، ب) پستانداران بچه‌زا هستند. به چنین جملاتی که قاعده‌ای عمومی، برای مثال، راجع به فلزات یا پستانداران بیان می‌کنند و معمولاً ساختار زبانی ساده‌ای دارند، جملات عمومی گفته می‌شود (Carlson et al. 1995). لذا، استفاده از مترجم‌های ماشینی در این مورد، علیرغم ضعف آنها در ترجمهٔ عبارات پیچیده، گزینهٔ مناسبی خواهد بود. از این‌رو، با توجه به ۱) ساختار سادهٔ جملات عمومی ۲) هزینه و زمان مورد نیاز برای ساخت یک آنتولوژی نو برای زبان فارسی ۳) دقت پایین روش مبتنی بر الگو در استخراج بعضی از روابط (مانند رابطهٔ توالی

1. Generic sentence

رخدادها، محرک‌ها و علی‌ها (جدول ۴)) و ۴) غنای مفاهیم و روابط موجود در «کانسپت‌نت» انگلیسی، تصمیم گرفته شد که اظهارات این آنتولوژی به‌زبان فارسی ترجمه شود. از بین مترجم‌های ماشینی موجود در بازار، ۵ مترجم گوگل، بابیلون، مترجم بینگ (مایکروسافت)، مترجم پدیده از زبان فارسی پشتیبانی می‌کنند. پس از انتخاب ۲۰ اظهار تصادفی از هر رابطهٔ موجود در «کانسپت‌نت»، از ۱۰ مترجم خواسته شد که نتایج ترجمهٔ این اظهارات توسط مترجم‌های ماشینی را رتبه‌بندی کنند (۰ برای ترجمهٔ نامفهوم، ۱ برای ترجمه‌ای که مفهوم آن را باید حدس زد، ۲ ترجمهٔ قابل قبول، ۳ ترجمهٔ خوب). نتایج رتبه‌بندی نشان از کیفیت بهتر مترجم گوگل داشت (ضمایم نمودار ۱). از این‌رو، در این پروژه از این ترجمه‌کننده برای ترجمهٔ اظهارات «کانسپت‌نت» استفاده شد.

پیش‌پردازش «کانسپت‌نت» و پس‌پردازش اظهارات ترجمه‌شده

بعد از انجام پیش‌پردازش‌های لازم، اظهارات با مترجم گوگل ترجمه شدند. از بین روابط ترجمه‌شده توسط گوگل، فقط روابط Location OF، Part Of، Made Of، Property Of که به‌ترتیب شامل ۱۶۰۰۰، ۱۳۰۰۰، ۱۵۰۰، ۹۱۰۰ اظهار بودند، به‌دلیل کیفیت بالای ترجمه (شکل ۳ ضمایم) برای قرارگرفتن در آنتولوژی فارسی انتخاب شد. جدول شماره ۱، بخش ضمایم، نمونه‌ای از ترجمهٔ اظهاری از هر رابطه توسط گوگل را نشان می‌دهد.

۳-۲-۳. استفاده از جعبهٔ اطلاعات ویکی‌پدیا

بعضی مقالات ویکی‌پدیا یک یا چند جعبهٔ اطلاعات در خود دارند (شکل ۴). جعبهٔ اطلاعات به الگوهایی گفته می‌شود که معمولاً در بالای مقاله قرار می‌گیرند و اطلاعاتی کلی در مورد موضوع مقاله ارائه می‌دهند. هر سطر از جعبهٔ اطلاعات یک ویژگی و یک مقدار دارد. برای مثال، جدول مربوط به «غلامحسین ابراهیمی دینانی» فیلسوف، دارای ویژگی زادگاه با مقدار اصفهان، ایران، دین با مقدار مسلمان دارد. بعد از استخراج تمامی این جعبه‌ها، ۲۸۰۰ رابطه "Is A" با بازخوانی ۱۰۰ درصد از این جعبه‌ها استخراج شد.

شکل ۲. نمونه‌ای از یک جعبهٔ اطلاعات ویکی‌پدیا

٤. بحث و نتیجه‌گیری

به‌زعم نویسندگان، علیرغم کارهای صورت گرفته برای استخراج بعضی از روابط مفهومی (مانند رابطهٔ Is A) از متن فارسی، تاکنون تحقیق جامعی که هدف اصلی‌اش را استخراج دانش عرفی قرار داده باشد، انجام نشده است. نتایج به‌دست‌آمده در این تحقیق نشان می‌دهد که استفاده از روش ترکیبی (روش‌های متن‌کاوی، ترجمهٔ ماشینی و استفاده از جعبهٔ اطلاعات ویکی‌پدیا) به‌منظور استخراج دانش عرفی، علیرغم کاستی‌های موجود در حوزهٔ رایانش زبان فارسی و نبودِ ابزارهای مورد نیاز حوزهٔ زبان‌شناسی رایانشی فارسی می‌تواند نویدبخش حوزه‌ای پویا در استخراج و بازنمایی دانش عرفی باشد. ما در این تحقیق توانستیم از سه پیکرهٔ مختلف فارسی، هفت نوع رابطه استخراج کرده و الگوهایی برای استخراج روابط مربوط به دانش عرفی فارسی معرفی کنیم که در دیگر حوزه‌های مرتبط با بازیابی اطلاعات و اکتساب دانش مفهومی نیز کاربرد داشته باشد. در مقایسه با کارهای صورت‌گرفته بر روی زبان فارسی (Shamsfard et al. 2010) مزیت این تحقیق تازگی موضوع، استفاده از چند روش مختلف استخراج و نیز تعداد پیکره‌های مورد استفاده

بوده است. علاوه بر این، در این تحقیق تعداد رابطه‌های استخراج‌شده نسبت به تحقیقات پیشین به‌مراتب بیشتر بوده است. در مقایسه با کارهای صورت‌گرفته بر روی زبان انگلیسی دقت استخراج‌ها کمتر بوده است که دلیل این امر هم نبودِ پیکره‌های استاندارد فارسی و عدم توسعهٔ ابزارهای اولیه و استاندارد پردازش متن فارسی است (Herdagdelen 2010; Sangun et la.2007). برای ارزیابی میزان دقت استخراج این نوع روابط، از دو گویشور زبان فارسی خواسته شد که نسبت به دانش عرفی‌بودن یا نبودن ۱۰۰ جمله از هر رابطه که به‌صورت تصادفی انتخاب شده بودند، قضاوت کنند (جدول ۷). نتایج به‌دست‌آمده نشان می‌دهد که در گروه روابط استخراج‌شده به روش مبتنی بر الگو هر دو رابطه دقت خوبی داشته‌اند و در این گروه دقت رابطهٔ Used For نسبت به رابطه Is A (رابطه Is A استخراج‌شده از جعبهٔ اطلاعات ویکی‌پدیا دقت ۱۰۰ درصد داشت) بیشتر است، هر چند جملات این گروه نسبت به رابطهٔ قبلی نیاز به پس‌پردازش‌های بیشتری دارند. در گروه روابط ترجمه‌شده از «کانسپت‌نت» بیشترین دقت‌ها به‌ترتیب مربوط به Property Of، Part Of، Made Of و Location Of بود.

به‌عنوان پیشنهاد برای تحققیات آینده و به‌منظور استخراج روابط بیشتر و نیز افزایش میزان دقت می‌توان از دیگر تکنیک‌های پردازش زبان طبیعی مانند برچسب‌زن گروه، برچسب‌زن معنایی، پیکره‌های مناسب‌تر و بزرگ‌تر، الگوهای غنی‌تر و بیشتر استفاده کرد. علاوه بر این، استفاده از روش‌های خلاقانه مانند رایانش انسان‌محور نیز می‌تواند به‌عنوان روشی مناسب جهت جمع‌آوری دانش عرفی فارسی مورد استفاده قرار گیرد. راه‌اندازی بخش فارسی پروژهٔ «اوپن مایند کامن‌سنس» نیز می‌تواند کمک شایانی به گسترش دانش عرفی محاسباتی فارسی بکند.

جدول ٦. دقت و تعداد هر کدام از روابط

شماره	رابطه		دقت	حجم روابط استخراج‌شده
۱	ترادف		۷۵ درصد	۱۵۴۰
۲	Is A	مبتنی بر الگو	۷۵ درصد	۹۲۵۵
		جعبهٔ اطلاعات	۱۰۰ درصد	۱۲۴۴
۳	Used For		۹۰ درصد	۶۴۰۰
۴	Part Of		۷۱ درصد	۱۲۹۳۴
۵	Made Of		۶۸ درصد	۱۵۱۹
۶	Location Of		۶۳ درصد	۲۸۸۰۵
۷	Property Of		۷۷ درصد	۹۱۳۵

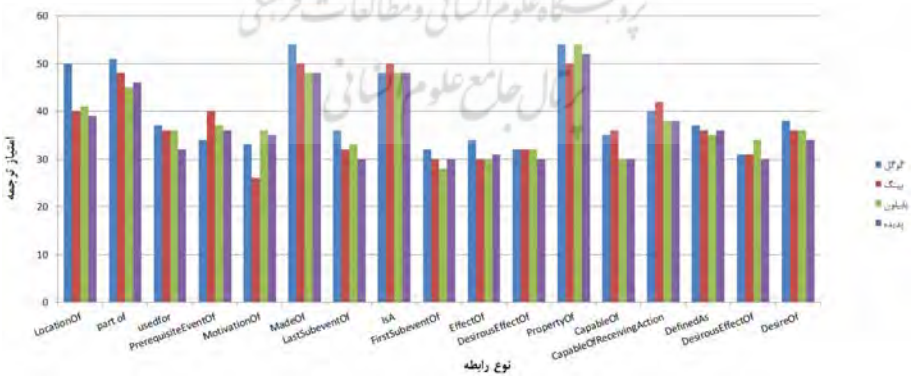
۵. ضمایم

جدول ۱. نمونه‌ای از ترجمه‌های گوگل برای هر رابطه

ش	ترجمهٔ گوگل	متن اصلی	نوع رابطه
۱	«فنلاند» و «اتحادیهٔ اروپا»	“Finland” “European union”	Part Of
۲	«تلفن» «دستگاه الکتریکی»	“phone” “electrical device”	Is A
۳	«خانه» «آجر و سنگ»	“House” “brick and stone”	Made Of
۴	«مسجد»، «مسلمان»	"mosque" “muslim”	Property Of
۵	«عقاب» و «نماد ایالات متحده آمریکا»	“Eagle” “emblem of usa”	Defined As
۶	«کنسرت راک» و «جذب جمعیت»	“rock concert” ”attract crowd”	Capable Of
۷	«فرد خانواده خوراک» «فرد طبخ مواد غذایی»	“feed person 's family” “cook person 's food”	Subevent Of
۸	«شستن خاک از بین فرد پا» و «فرد» کفش »	'' wash dirt from between person 's toe'' "take off person 's shoe	Prerequisite Event Of
۹	«رقابت در مقابل فرد» «سعی کنید به شل»	'' compete against person" ''try to loose''	LastSubevent Of
۱۰	«برو به خواب» «نزدیکان» چشم»	“go to sleep” “close person 's eye”	First Subevent Of
۱۱	«هوا» در آسمان»	“Air” “in sky”	Location Of
۱۲	«فرد» و «کافی خواب»	“Person” “adequate sleep”	Desire Of
۱۳	«معاینه» «دانستن» دولت»	“have examination” “know person 's state”	Motivation Of
۱۴	«نوشیدن» و «خماری»	“Drinking” “hangover”	Effect Of
۱۵	«غار و غور از معده» «را کاهش گرسنگی فرد را خود»	“growl from stomach” “diminish person 's own hunger”	Desirous Effect Of
۱۶	«پوست حیوانات» «چرم»	“animal skin” “make leather”	Used For
۱۷	«برف» «سعی در حذف کنید»	“Snow” ”try to remove”	Capable Of Receiving Action

جدول ۲. روابط تعریف‌شده در «کانسپت‌نت»

طبقه		نوع رابطه	مثال
چیزها	۱	Part Of	«بمبئی»، «هند»
	۲	Is A	«تابستان»، «زمان»
	۳	Made Of	«آجار»، «فلز»
	۴	Property Of	«امینم»، «همجنس‌گرا»
	۵	Defined As	«خدا»، «خالق»
عامل‌ها	۶	Capable Of	«پدر، مادر»، «فرزند»
رخدادها	۷	Subevent Of	«نوشتن»، «شکستن نوک مداد»
	۸	Prerequisite Event Of	«طلاق گرفتن»، «رفتن به دادگاه»
	۹	Last Subevent Of	«ترک اتاق»، «خاموش‌کردن چراغ»
	۱۰	First Subevent Of	«نوشتن مقاله»، «پژوهش»
فضایی	۱۱	Location Of	«شراب»، «کلیسا»
محرک‌ها	۱۲	Desire Of	«سگ»، «توجه زیاد»
	۱۳	Motivation Of	«تف‌کردن»، «نفرت از چیزی»
علّی	۱۴	Effect Of	«گیتارزدن»«آرتروز»
	۱۵	Desirous Effect Of	«حسادت» و «رقابت»
کارکردها	۱۶	Used For	«خوردن سبزی»، «حفظ سلامت»
	۱۷	Capable Of Receiving Action	«صلح»، «مذاکره»



شکل ۱. نمودار مربوط به کیفیت ترجمه به تفکیک موتور جستجو و رابطه (حداکثر امتیاز ۶۰)

فهرست منابع

امیری، هادی، حسین حجت، و فرهاد ارومچیان. ۱۳۸۵. بررسی پیکره‌ای مناسب برای برچسب‌زنی کلمات در زبان فارسی. *ارائه‌شده در دوازدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران*، تهران.

شمس‌فرد، مهرنوش، و احمد عبداله‌زادهٔ بارفروش. ۱۳۸۱. استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی. *مجلهٔ تازه‌های علوم‌شناختی* ۱ (۱۳): ۴۸.

مرادی، مهدی، بهرام وزیرنژاد، پروانه خسروی‌زاده، و هادی عبدی قویدل. ۱۳۹۱. استخراج جملات عمومی از ویکی‌پدیای فارسی: تلاشی در جهت ساخت آنتولوژی دانش عرفی. *چهارمین کنفرانس فناوری اطلاعات و دانش*. بابل: انجمن فناوری اطلاعات و ارتباطات ایران و دانشگاه صنعتی نوشیروانی بابل.

AleAhmad, Abolfazl, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2009. Hamshahri: A Standard Persian Text Collection. *Knowledge-Based Systems* 22 (5): 382–387.

Carlson, G. N. and F. J. Pelletier, eds. 1995. The Generic Book. Chicago: The University of Chicago Press.

Descartes, R. 2008. Discourse on the Method for Reasoning Well and for Seeking Truth in the Sciences (The Gutenberg Project).

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on computational linguistics, Nantes, France.

Herdagdelen, A. 2010. Collecting common sense from text and people. Ph.D. Dissertation, University of Torento.

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

Lieberman, H. 2008. Usable AI requires commonsense knowledge. In Workshop on Usable Artificial Intelligence, ACM Conference on. Florence, Italy.

_____, D. Smith, and A. Teeters. 2007. Common consensus: A web-based game for collecting commonsense goals. In Proceedings of the workshop on common sense and intelligent user interfaces held in conjunction with the 2007 international conference on intelligent user interfaces, Honolulu, USA.

Lories, Danielle. 1991. *Des sensibles communs dans le 'De anima' d'Aristote. Revue Philosophique. De Louvain* (3) 89: 401-420.

_____. 1998. *Le sens commun et le jugement du phronimos: Aristote et les stoiciens. Louvain-la-Neuve*, Editions Peeters.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. *Verbosity: a game for collecting common-sense facts.* In Proceedings of ACM CHI 2006 Conference on Human Factors in Computing System, Quebec, Canada (pp. 75-78).

McCarthy, J. 1959. Programs with common sense. In Proceedings of Teddington Conference on the Mechanization of Thought Processes, London, England (pp. 75–91).

Shamsfard, M., A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, 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.

Sangun, P., Juyoung, K., Wooju, K.: A Framework for Ontology Based Rule Acquisi-tion from Web Documents in Web Reasoning and Rule Systems (2007).

Singh, P., T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. In Proceedings of First International Conference on ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, Irvine, California.

Suh, S., H. Halpin, and E. Klein. 2006. Extracting common sense knowledge from Wikipedia. In Proceedings of the ISWC-06 Workshop on Web Content Mining with Human Language Technologies, Athens, USA.

مهدی مرادی

متولد سال ۱۳۶۴، دارای مدرک تحصیلی کارشناسی ارشد زبان‌شناسی رایانشی از دانشگاه صنعتی شریف است. وی هم‌اکنون مدرس رایانه در گروه علوم تربیتی دانشگاه آزاد اسلامی زنجان است.

هستان‌شناسی، استخراج رابطه و الگو، شناسایی خودکار نویسنده از جمله علایق پژوهشی ایشان هستند.

بهرام وزیرنژاد

متولد سال ۱۳۵۷، دارای مدرک دکتری مهندسی پزشکی از دانشگاه صنعتی امیرکبیر است. وی هم‌اکنون عضو هیئت علمی دانشگاه صنعتی شریف و مدیر آزمایشگاه پردازش گفتار و زبان است.

پردازش زبان طبیعی و گفتار از جمله علایق پژوهشی ایشان هستند.

محمد بحرانی

متولد سال ۱۳۵۶، دارای مدرک دکتری در رشتهٔ هوش مصنوعی از دانشگاه صنعتی شریف است. وی هم‌اکنون عضو هیئت علمی گروه زبان‌شناسی رایانشی در مرکز زبان‌ها و زبان‌شناسی دانشگاه صنعتی شریف است.

زبان‌شناسی رایانشی، پردازش زبان طبیعی، پردازش و بازشناسی گفتار و مدل‌سازی زبانی از جمله علایق پژوهشی ایشان هستند.