ارائه‌ی یک روش به منظور تکمیل پایگاه‌های دانش با استفاده از سیستم‌های پرسش‌وپاسخ

عاطفه صافدل[۱]، سید مصطفی فخراحمد[۲]، محمدهادی صدرالدینی[۳]

[۱] دانشجوی کارشناسی ارشد، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز ، شیراز
a.safdel@cse.shirazu.ac.ir

[۲] استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز ، شیراز
fakhrahmad@shirazu.ac.ir

[۳] استاد، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز ، شیراز
sadredin@shirazu.ac.ir

چکیده

مدت‌هاست که پایگاه‌های دانش به‌عنوان منابع مهمی برای دسترسی به اطلاعات ساختاریافته مورد جستجو قرار می‌گیرند. ازجمله کاربردهای آن‌ها می‌توان به سیستم‌های پرسش‌وپاسخ اشاره کرد؛ آن‌هایی که در جستجوی پاسخ پرسش‌ها تنها در قالب قوانین ساختاریافته‌ی سه‌تایی هستند. با وجود کاربردهای زیادی که پایگاه‌های دانش دارند، حتی بزرگ‌ترین نمونه‌های موجود از آن‌ها به‌صورت وسیعی ناقص هستند و احتمال اینکه به آن‌ها مراجعه شود و پاسخ لازم یافت نشود زیاد است. به‌عنوان مثال اطلاعات مربوط به همسر و فرزندان اشخاصی که در Freebase ذخیره شده‌اند، به‌ترتیب برای ۹۲ درصد و ۹۴ درصد افراد وجود ندارد. در این پژوهش سعی شده است راهکاری برای اضافه کردن سه‌تایی‌های جدید به پایگاه‌های دانش ارائه شود که استخراج مقادیر مجهول را بر مبنای سیستم‌های پرسش‌وپاسخ انجام دهد و نگاشت پاسخ‌های استخراج شده به موجودیت‌ها را بر مبنای ماهیت رابطه‌ی مورد بررسی و سایر روابط موجود در پایگاه دانش به انجام برساند. برای ارزیابی روش پیشنهادی، نمونه‌گیری از بین صدهزار شخص مهم و برجسته‌ی جهان انجام گرفته و نتیجه‌ی ارزیابی، بهبود رتبه‌بندی پاسخ‌های صحیح درقالب معیارهای MRR و MAP می‌باشد.

کلمات کلیدی

پایگاه‌دانش، استخراج متن، نگاشت موجودیت، تکمیل اطلاعات، Freebase

۱- مقدمه

از گذشته‌ها مسئله‌ی ذخیره‌سازی مطرح است و بشر همواره به‌دنبال نگه‌داری اطلاعات بوده‌است. ابتدا از فایل‌ها استفاده میشد که قابلیت مدیریت، پردازش و جستجو در آن‌ها بسیارکم بود. پس‌ازآن، مسائل موجود در دنیای واقعی به صورت داده در پایگاه‌های داده[۱] ذخیره می‌شدند. هرچند می‌توان برای داده‌ها در پایگاه‌های داده، ویژگی‌هایی تعریف کرد و از این طریق امکان نزدیکی اطلاعات ذخیره‌شده را به واقعیت بیشتر کرد؛ اما باز هم شرایطی پیش می‌آید

که در آن شرایط، این شیوه‌ی ذخیره‌سازی کافی نیست. زمانی که ارتباط موجود بین داده‌ها ارزشمند باشد، باید به طریقی بین داده‌ها لینک برقرار کرد و لینک‌ها را نیز ذخیره نمود. پایگاه‌های دانش می‌توانند پاسخگوی این مشکل باشند. بنابراین نباید پایگاه‌های داده را با پایگاه‌های دانش اشتباه گرفت.

یک پایگاه‌دانش با هدف نگه‌داری موجودیت[2]ها و مشخص بودن رابطه‌ی بین موجودیت‌ها تشکیل می‌شود. پایگاه‌دانش می‌تواند به این منظور، از یک یا چند پایگاه‌داده استفاده کند. اطلاعات ذخیره‌شده در هر پایگاه‌دانش اغلب دارای اشتراکاتی هستند و مربوط به موضوع خاصی می‌شوند.

با رویکردی دیگر، مسئله‌ی بازیابی اطلاعات ذخیره‌شده مطرح است. اگر داده‌ها بدون هیچ‌گونه ساختارمندی و به‌صورت متنی ذخیره شوند، فرایند بازیابی آن‌ها مشکل خواهدبود و اگر به جستجوی کلمه‌ای در آن متن پرداخته‌شود، حداقل نیازمند تجزیه‌ی لغوی[3] متن می‌باشد. بنابراین با گذر زمان به سوی داده‌های ساختارمند و نیمه‌ساختارمند روی‌آورده شد. در این نوع از پایگاه‌دانش، داده با فرمت خاصی ذخیره می‌شود. به‌عنوان مثال فرمت RDF[4] به طور متداول مورد استفاده قرار می‌گیرد. این فرمت نمایش داده دارای سه قسمت اطلاعاتی است که عنصر اول و سوم موجودیت‌هایی هستند که از طریق رابطه‌ای که عنصر دوم آن را نمایش می‌دهد، بهم مربوط می‌شوند. با ساختارمند شدن پایگاه‌های دانش، شیوه‌ی جستجو در آن‌ها و بازیابی اطلاعات از آن‌ها هم ساده‌تر می‌شود. این فرمت داده به درک بهتر اطلاعات کمک می‌کند و می‌توان پرس‌وجو[5]های دقیقی روی آن‌ها اعمال کرد.

اطلاعاتی که در پایگاه‌های دانش ساختارمند ذخیره می‌شوند، ممکن است به‌صورت اتوماتیک از متون ساده استخراج شده و یا اینکه به‌صورت دستی و توسط افراد وارد شده باشند. مسلماً دقت پایگاه‌های دانش با استخراج خودکار اطلاعات کم‌تر از حالت دیگر است؛ اما هر دو روش هنوز متداول هستند.

ازجمله نمونه‌هایی که از پایگاه‌های دانش موجود می‌توان نام برد، شامل Freebase [1]، Wikidata [2]، Yago [3]، Probase [4] و SigmaKB [5] می‌شود. برخی از آن‌ها مانند Freebase بسیار بزرگ هستند و شامل موجودیت‌ها و روابط بسیار زیادی می‌شوند. هرچند که به دلیل افزایش روزافزون اطلاعات متنی، حتی پایگاه‌های دانش بزرگ نیز، از ابعاد مهمی ناقص هستند.

موتورهای جستجو[6] قابلیت‌های معنایی خوبی در اختیار می‌گذارند. اگر از جوامع به‌اشتراک‌گذاری دانش[7] مثل ویکی‌پدیا استفاده شود، فرایند «استخراج موجودیت‌ها و روابط» از منابع وب نیمه‌ساختارمند، به‌خوبی منابع زبان طبیعی می‌تواند به‌طور خودکار صورت گیرد. کارهایی نیز از این نوع انجام شده‌است؛ مانند DBpedia [6]، EntityCube [7]، KnowItAll [8]. هدف همه این است که یک پایگاه‌دانش جامع از قوانین[8]، درباره‌ی موجودیت‌های نامدار، دسته‌های معنایی آن‌ها، ارتباط‌های متقابل آن‌ها وجود داشته باشد.

جدول (۱) : درصد عدم وجود اطلاعات درباره‌ی اشخاص ذخیره شده در Freebase در رابطه‌های همسر و فرزندان

رابطه	اطلاعات ناقص (٪)	
	صدهزار نفر برتر	کل اشخاص
همسر	٪۶۸	٪۹۲
فرزندان	٪۸۰	٪۹۴

۲- شرح مسئله

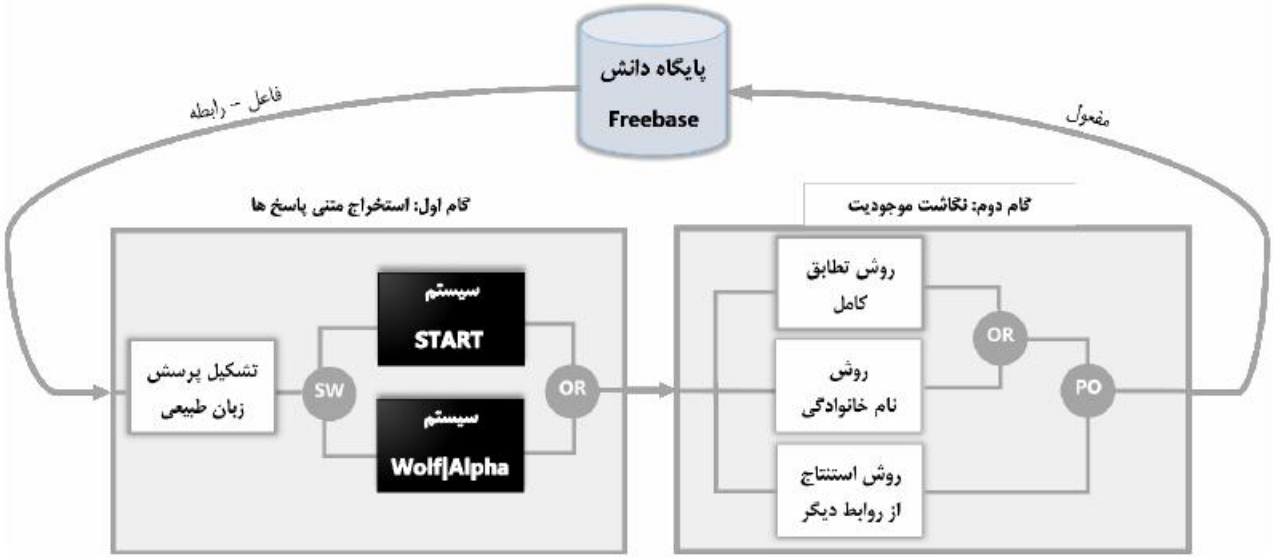
به‌دلیل افزایش مراجعه به پایگاه‌های دانش به عنوان یک منبع اطلاعاتی، اهمیت زیادی دارد که اطلاعات موجود در آن کامل باشد. تکمیل بودن اطلاعات لزوماً به معنی وجود همه‌ی انواع ممکن از موجودیت‌ها و یا وجود همه‌ی نمونه‌های هر نوع موجودیت نیست. بلکه این اهمیت دارد که روابط متداولی که بین انواع موجودیت‌ها اغلب برقرار است، نیز تا آن‌جا که ممکن است کامل باشد. به‌عنوان مثال فرض کنید که از یک سیستم پرسش‌وپاسخ[9] مبتنی بر پایگاه‌دانش پرسیده می‌شود که «نویسنده‌ی مثنوی‌معنوی چه کسی است؟» ازآن‌جایی‌که در این حالت پایگاه‌دانش برای یافتن پاسخ جستجو می‌شود، برای پاسخ‌گویی به این پرسش، باید هر دو موجودیت‌های «مولانا» و «مثنوی معنوی» در پایگاه‌دانش وجود داشته‌باشند و مهم‌تر از همه، لینک مستقیم یا غیرمستقیمی بین آن‌ها تشکیل شده باشد تا «مولانا» به‌عنوان خروجی جستجو در پایگاه‌دانش برگشت داده شود. با اینکه پایگاه‌های دانش بسیار غنی و عظیمی وجود دارند، باز هم کمبود اطلاعاتی شدیدی در آن‌ها وجود دارد که باعث می‌شود در بسیاری از موارد، رجوع به آن‌ها بدون نتیجه باشد. بنابراین همواره سعی بر این است که پایگاه‌های دانش را کامل‌تر کرد و قوانین جدیدی به آن اضافه نمود.

یکی از انواع اطلاعاتی که مردم علاقه‌مند به جستجوی آن‌ها هستند و پایگاه‌های دانش باید در این زمینه نیز غنی باشند، قوانینی است که درباره‌ی «اشخاص» است. این اشخاص می‌توانند آن قدر معروف باشند که به دلیل شهرت و علاقه‌ی مردم به بیشتر دانستن درباره‌ی آن‌ها، مرتب مورد پرسش قرار بگیرند. به‌عنوان مثال راجع‌به محل‌تولد، همسر، فرزندان، شغل و حرفه، ملیت و تحصیلات آن‌ها به شکل متناوبی پرسش‌هایی مطرح شود. بنابراین پایگاه‌دانش هم باید آن اشخاص را به عنوان موجودیت داشته باشد و هم اینکه رابطه‌های ذکرشده، بین آن‌ها و موجودیت‌های دیگر وجود داشته باشد تا پرسش‌ها به پاسخ برسد و درصد پاسخگویی سیستم بالا برود.

اگر مسئله‌ی اضافه کردن اشخاص جدید به پایگاه‌دانش دغدغه‌ی اصلی نباشد و فقط بخواهیم اشخاص موجود در آن را درباره‌ی این روابط کامل کنیم، با آماری از نقص اطلاعات درباره‌ی رابطه‌های «همسر» و «فرزندان» در Freebase روبه‌رو هستیم که در جدول (۱) نشان داده شده است. اطلاعات این جدول از [9] استخراج شده‌است.

آماری که جدول (۱) نشان می‌دهد، درباره‌ی شدت نقص اطلاعات برای همه‌ی ۳ میلیون شخصی است که به صورت موجودیت در Freebase وجود دارند. به صورتی که حتی ۹۴ درصد از آن‌ها در رابطه‌ی «فرزندان» شرکت نداشته‌اند. شاید این مسئله مطرح شود که وجود چنین قوانینی در پایگاه‌دانش برای همه‌ی این ۳ میلیون شخص اهمیتی ندارد و اگر تنها درمورد افرادی که زیاد مورد جستجو قرار می‌گیرند، غنی باشد، کفایت می‌کند. همان‌طور که از اطلاعات موجود در ستون دیگر جدول مشخص است، این مشکل درباره‌ی آن‌ها کم‌تر است ولی همچنان به شکل وسیعی وجود دارد. آمار موجود در این ستون برای ۱۰۰ هزار فردی است که بیشترین جستجو درباره‌ی آن‌ها در سطح وب صورت گرفته است. شدت ناکامل‌بودن Freebase در حوزه‌ی رابطه‌های اساسی «اشخاص» باعث می‌شود که رفع این نقص اهمیت پیدا کند.

آقای West و همکاران [9] راهکاری برای تکمیل اطلاعات موجود در Freebase ارائه دادند که سیستم جدید طراحی‌شده را با آن مقایسه می‌کنیم.



شکل (۱) : شمای کلی سیستم طراحی شده به‌منظور تکمیل اطلاعات ناقص.

SW: ورودی را به یکی از خروجی‌ها می‌فرستد.
OR: همه‌ی ورودی‌ها را به خروجی می‌فرستد.
PO: فقط ورودی مربوط به "استنتاج از روابط دیگر" به خروجی می‌رود. درصورت خالی بودن آن ورودی، ورودی دیگر به خروجی می‌رود.

۳- روش پیشنهادی

مسئله به این صورت است که یک فاعل و یک رابطه با شناسه‌ی مشخص داریم و می‌خواهیم مفعول آن را بیابیم. به صورتی که سیستم، شناسه‌ی یک موجودیت در پایگاه‌دانش را به عنوان مفعول برگرداند. این مسئله می‌تواند به‌صورت چند جوابی هم وجود داشته باشد. از منظر دیگر، گاهی لازم است چندین جواب محتمل به صورت فهرست ارائه شود و تصمیم نهایی برای اضافه کردن آن قائون به پایگاه‌دانش، توسط بخش دیگری گرفته شود. به این ترتیب، در این پژوهش ما نیز پاسخ‌های نهایی را رتبه‌بندی شده ارائه می‌دهیم.
نمای کلی از سیستم طراحی‌شده، در شکل (۱) آمده‌است. همان‌طور که در شکل مشخص است، فاعل و رابطه‌ی موردنظر وارد سیستم می‌شوند. سپس قطعه‌های متنی محتمل به‌عنوان مفعول در مرحله‌ی اول حدس زده می‌شوند. در مرحله‌ی بعد، سیستم مشخص می‌کند این قطعات متنی اشاره به کدام‌یک از موجودیت‌ها در پایگاه‌دانش می‌کنند.
در ادامه جزئیات هریک از مراحل را شرح می‌دهیم.

۳-۱- گام اول: استخراج متنی پاسخ‌ها

در گام اول باید برای یک فاعل و رابطه‌ی مشخص، مفعول آن پیش‌بینی شود. این پیش‌بینی فقط به صورت متنی است و لازم نیست مشخص شود که مفعول دقیقا معادل با کدام‌یک از موجودیت‌ها در پایگاه‌دانش می‌باشد.
به عنوان مثال، اگر فاعل «حافظ» و رابطه‌ی موردنظر «محل‌تولد» باشد، در این گام باید تعدادی جواب به صورت متنی حدس زده شود و درصورتی که «شیراز» در فهرست جواب‌ها باشد، این مرحله کار خود به خوبی انجام داده است. این درحالی است که ممکن است چندین «شیراز» در پایگاه‌دانش وجود داشته باشد و دقیقاً مشخص نباشد که کدام‌یک از این «شیراز»ها موردنظر است؛ اما حل این مشکل مربوط به گام بعدی می‌شود. در پژوهش انجام‌شده از روش مبتنی بر پرسش از سیستم‌های پرسش‌وپاسخ استفاده شده‌است. این‌گونه سیستم‌ها معمولاً پاسخ را به صورت دقیق‌تر، مشخص‌تر و مختصرتر نسبت به سیستم‌های بازیابی اطلاعات (موتورهای جستجو) برمی‌گردانند. آن‌ها قابلیت دریافت پرسش‌های زبان طبیعی را داشته و ممکن است برای اعلام پاسخ، منابع مختلفی را جستجو کنند. این منابع ممکن است پایگاه‌های دانش مخصوص به خودشان، چندین سند متنی، اینترنت و یا ... باشد.
در روش مبتنی بر پرسش از این سیستم‌ها، می‌توان فاعل و رابطه را در قالب پرسش‌های زبان طبیعی درآورد و آن‌ها را از سیستم پرسید؛ با این امید که مفعول موردنظر در قالب پاسخ برگردانده شود. بنابراین باید با داشتن الگوی مناسب برای طرح پرسش از آن‌ها به گرفتن پاسخ موردنظر امیدوار بود.
سیستم‌های پرسش‌وپاسخی که در این پژوهش به صورت black-box مورد استفاده قرارگرفته‌اند، سیستم‌های پرسش‌وپاسخ START [10,11] و Wolfram|Alpha [12] هستند.
START اولین سیستم پرسش‌وپاسخ مبتنی بر وب در سطح جهان است و برای دریافت پاسخ سوالات منابع مختلفی را در سطح وب جستجو می‌کند. همچنین ممکن است از پایگاه‌دانش خود برای پاسخ‌دهی استفاده کند. این سیستم پرسش‌وپاسخ در دانشگاه MIT طراحی و پیاده‌سازی شده و از سال ۱۹۹۳ به صورت آنلاین در دسترس است. این سیستم در بسیاری از زمینه‌ها پاسخ‌گو است. زمینه‌هایی مثل مکان‌ها، نقشه‌ها، سیستم‌های اقتصادی و سیاسی، آب‌وهوا، فیلم‌ها، بازیگران، افراد، زندگی‌نامه، تعاریف بر طبق لغت‌نامه‌ها و ... .
Wolfram|Alpha یک موتور دانش است که قابلیت انجام محاسبات و تحلیل را دارد و می‌تواند در زمینه‌های کاملا متفاوتی پاسخ‌گو باشد؛ از جمله: مهندسی، شیمی، ریاضی، موزیک، تحصیل، هنر سلامت و ... . این سیستم برای پاسخ‌دهی، پایگاه‌دانش مخصوص به خود را جستجو می‌کند و در صورت

نیاز، تحلیل‌هایی انجام می‌دهد تا جواب نهایی را اعلام کند. از جمله مواردی که این سیستم قادر به پاسخ‌گویی در آن زمینه است، افراد و تاریخچه‌ی آن‌هاست. بنابراین درصورتی که دانش کافی درباره‌ی افراد در آن وجود داشته باشد، می‌تواند در نقش سیستم پرسش‌وپاسخ برای سیستم ما عمل کند.

در روش پیشنهادی برای تکمیل پایگاه‌دانش، این قابلیت وجود دارد که پرسش زبان طبیعی به یکی از این سیستم‌های پرسش‌وپاسخ ارسال می‌گردد و از عملکرد آن برای استخراج پاسخ استفاده می‌شود.

۳-۱-۱- استفاده از سیستم START:

این سیستم به دلیل اینکه منابع ساختاریافته یا بدون ساختار مختلفی را در سطح وب جستجو می‌کند، ممکن است پاسخ را در قالب‌های متفاوتی نیز برگرداند. مثلا پاراگرافی که جواب با احتمال زیادی در آن است و یا جواب دقیقی که در مقابل برچسبی نوشته شده است. این سیستم اگر به ابهام برسد، همه‌ی حالت‌های مبهم را به عنوان پاسخ برمی‌گرداند. همچنین درصورتی که پاسخ را بداند، مشخص می‌کند که از چه منبعی این پاسخ را استخراج نموده است. وقتی که پرسش در رابطه با اشخاص باشد، اغلب به جستجوی آن‌ها در صفحات ویکی‌پدیا و IMDB[۱۰] می‌گردد و طبق شباهت ظاهری نام شخص مورد جستجو و کل افراد ذخیره شده در آن منابع، همه‌ی گزینه‌های محتمل را برمی‌گرداند.

سیستم ما هر زوج فاعل و رابطه را در قالب پرسش زبان طبیعی از START می‌پرسد. پاسخ‌های بدون منبع و بدون برچسب را دور می‌ریزد و بقیه‌ی جواب‌ها را که می‌تواند جداکننده‌های متفاوتی بین‌شان باشد، استخراج می‌کند.

۳-۱-۲- استفاده از سیستم Wolfram|Alpha:

زمانی که پرسشی از این سیستم پرسیده می‌شود، پاسخ در result field نمایش داده می‌شود. درباره‌ی افراد و پرسش‌هایی که اطلاعات خاصی درباره‌ی آن‌ها را جستجو می‌کنند، پاسخ‌های Wolfram|Alpha به صورت دقیق هستند و نیاز به تحلیل خاصی ندارند. البته این به این معنا نیست که پاسخ‌هایی که مطمئن هستیم اشتباه هستند را دور نریزیم. زمانی که ظاهر پاسخ کاملا نشان می‌دهد این یک پاسخ صحیح نیست، آن را دور میریزیم. مثلا در جستجوی یک اسم خاص بوده ایم، درحالی که سیستم پاسخی را برگردانده که با حرف بزرگ شروع نمی‌شود. حالت‌های دیگری که عدم پاسخ‌دهی Wolfram|Alpha را نشان می‌دهد این است که عبارت « data not available» در فیلد نتایج داشته باشیم و یا اینکه این فیلد به‌طورکلی نمایش داده نشود. بنابراین سیستم ما هر زوج فاعل و رابطه را در قالب پرسش زبان طبیعی از سیستم Wolfram|Alpha می‌پرسد و از پاسخ آن به عنوان کاندیدای متنی مفعول استفاده می‌کند.

۳-۲- گام دوم: نگاشت موجودیت

تا این‌جا همه‌ی عملکرد سیستم پیرامون یافتن نسخه‌ی متنی برای مقدار مفعول مجهول بود. در این قسمت نحوه‌ی نگاشت این مقادیر به موجودیت‌های پایگاه‌دانش شرح داده می‌شود.

نگاشت متن به موجودیت از چندین جنبه می‌تواند همراه با مشکلاتی باشد؛ از جمله آن موجودیت‌های نامداری که نام یکسان دارند اما خودشان یکسان نیستند. اشخاص زیادی سرتاسر دنیا می‌توانند نام مشترکی داشته باشند و نباید آن‌ها را با هم اشتباه گرفت. دلیل دیگری که می‌تواند عملکرد نگاشت را مشکل کند، عدم استفاده از یک متن کمکی برای رفع ابهام می‌باشد. در روش‌های معمول برای نگاشت از یک متن کمکی هم استفاده می‌شود تا این نگاشت اشتباه رخ ندهد. معمولا متن اطلاعات جانبی راجع‌به فاعل وجود دارد و از طریق روش‌های پردازش زبان طبیعی سعی می‌شود نزدیک‌ترین موجودیت کاندیدا در پایگاه‌دانش را به‌عنوان مفعول درنظرگرفت. اما در روش استفاده‌شده در طراحی این سیستم، فرض شده که هیچ متن کمکی وجود ندارد و با این فرض کارایی سیستم در بدترین شرایط اطلاعاتی تخمین زده شده است. به این دلیل که این شرایط بسیار محتمل است که حتی اگر به عنوان مثال متنی هم در اختیار باشد، در آن متن پارامترهایی که کمک‌کننده برای نگاشت هستند موجود نیست. این مسئله حداقل در حوزه‌ی تعریف‌شده برای این پژوهش زیاد رخ می‌دهد. اشخاص مشهوری را فرض کنید که اطلاعات متنی که از آن‌ها در صفحات موجود است اغلب پیرامون همان دلیل شهرت‌شان است. مثلا صفحه‌ی یک بازیگر بیشتر پیشینه‌ی هنری وی را پوشش می‌دهد تا اینکه به تشریح زندگی شخصی‌اش بپردازد و زندگی شخصی وی را به صورت مختصر شرح می‌دهد. همین مسئله باعث می‌شود زمانی که به دنبال اطلاعات شخصی افراد هستیم، آن صفحه برای رفع ابهام از اطلاعات شخصی کارا نباشد. البته این اختصار در اطلاعات در همه‌ی موارد رخ نمی‌دهد اما به دلایلی که ذکر شد، در این پژوهش فرض را بر این گذاشتیم که در همه‌ی حالات از متن کمکی بهره نگیرد. موضوع دیگری که نگاشت متن به موجودیت در این سیستم با آن مواجه است، نتیجه‌ی پذیرش هر قطعه‌ی متنی کوچک اطلاعاتی است که از مرحله‌ی قبل به این‌جا رسیده است. در مرحله‌ی استخراج متنی پاسخ‌ها، هر متن کوچکی را پذیرفتیم و از این مرحله می‌خواهیم تا جایی که ممکن است معادل‌هایی برایشان بیابد و آن‌ها را دور نریزد؛ در حالی‌که در روش‌های متداول چنین نیست. صفحه‌ی فوتبالیستی را در ویکی‌پدیا فرض کنید که کاملا درباره‌ی زندگی حرفه‌ای وی می‌باشد و تنها چیزی که درباره‌ی فرزندان و همسر او گفته نام کوچک آن‌ها است. سیستم طراحی‌شده سعی می‌کند از همین اطلاعات کم هم برای نگاشت استفاده کند و آن‌ها را اطلاعات بی ارزش نخواند و کنار نزند.

زمانی که سیستم در جستجوی «همسر» یا «فرزندان» شخصی است، «اشخاص» موجود در پایگاه‌دانش را بررسی می‌کند که آیا حالت‌هایی که در ادامه شرح داده شده برای آن‌ها صادق است یا خیر. اگر روش‌های ذکرشده نتیجه دهند، تعدادی موجودیت از بین «اشخاص» را خواهیم داشت که کاندیدای مفعول بودن هستند.

۳-۲-۱- روش مبتنی‌بر استنتاج از روابط موجود دیگر

در این روش سیستم از روابط دیگر موجود در پایگاه‌دانش برای یافتن موجودیت موردنظر استفاده می‌کند. برای این منظور ارتباط‌های خانوادگی، فاعل و قطعات متن دریافتی کاندیدا استفاده می‌شوند. درصورتی‌که سیستم بتواند از این طریق موجودیت‌ها را بیابد که همان به عنوان پاسخ برگشت داده می‌شود؛ درغیراین‌صورت هردو روشی که در ادامه آمده‌اند، اجرا می‌شوند.

۳-۲-۲- روش تطابق کامل

اگر در پایگاه‌دانش دقیقاً یک شخص وجودداشته‌باشد که قطعه‌پاسخ متنی با نام وی برابر باشد، آن موجودیت به مجموعه‌ی پاسخ‌ها اضافه می‌شود.

۳-۲-۳- روش نام‌خانوادگی

از آن‌جایی که در بسیاری از موارد، نام‌خانوادگی همسر و فرزندان یک فاعل با نام‌خانوادگی وی برابر است، اگر در پایگاه‌دانش شخصی داشته باشیم که نام وی با قطعه‌متن دریافتی و نام‌خانوادگی آن با نام‌خانوادگیِ فاعل برابر باشد، موجودیت به مجموعه‌ی پاسخ‌ها اضافه می‌شود.

٤- مجموعه‌ی داده

ارزیابی سیستم از طریق یک مجموعه‌ی داده‌ی تست انجام شد. از آن‌جایی که مجموعه‌ی داده‌ای که متناسب با موضوع ما باشد در دسترس نبود، مجموعه‌ای جدید از داده‌ها را ایجاد کردیم. برای اطمینان از اینکه یک مجموعه‌ی داده‌ی قابل اتکا ایجاد شود، روشی را که آقای West و همکاران [9] در ساخت مجموعه‌ی داده استفاده کردند به کار بردیم. هر رکورد از داده‌ها یک دوتایی است که باید عنصر سوم آن به عنوان مفعول توسط سیستم ما پیش‌بینی گردد. این دوتایی برای اینکه از فرضیات مسئله‌ی ما و مجموعه‌ی داده‌ی آقای West پیروی کنند باید چند ویژگی داشته‌باشند. عنصر دوم که معادل با رابطه می‌باشد در نهایت باید یکی از حالت‌های «همسر» یا «فرزندان» داشته باشد و فاعل آن «اشخاصی» باشند که آقای West استفاده کرده است و یا حداقل اینکه «اشخاصی» باشند که به روشی مشابه با روش آقای West و از منابع مشابه‌ای انتخاب شده باشند. آقای West و همکارانشان برای انتخاب فاعل‌ها ابتدا ۱۰۰ هزار شخصی را که بیش‌ترین جستجو درباره‌شان در سطح وب صورت گرفته را به عنوان اشخاص پایه درنظر گرفتند. اما اینکه چه کسانی و با چه رتبه‌ای در این فهرست هستند را هم به صورت عمومی در اختیار نگذاشته‌اند. بنابراین اولین مرحله، یافتن ۱۰۰ هزار شخص است که از مهم‌ترین افراد باشند. سیستمی که [13] طراحی کردند، جستجو در پایگاه‌دانشِ Freebase راحت و ساده می‌کند. علاوه‌براین، آن‌ها به هر موجودیت در Freebase یک امتیاز اختصاص داده‌اند که شیوه‌ی امتیازدهی آن‌ها برای رتبه‌دهی به اشخاص استفاده شد و با کمک آن امتیازها ۱۰۰ هزار شخص مهم و برجسته را یافته و برای ساخت مجموعه‌ی داده از آن‌ها استفاده کردیم. در ادامه‌ی ساخت مجموعه‌ی داده به روش آقای West برمی‌گردیم. آن‌ها برای ساخت هریک از مجموعه‌های داده‌شان به ازای هر رابطه از نمونه‌گیری طبقه‌ای که در ادامه توضیح می‌دهیم، استفاده کردند: برای هر رابطه، از بین آن ۱۰۰ هزار نفر، تنها آن فاعل‌هایی را درنظرگرفتند که مقدار مفعول آن مشخص بود و مجهول نبود. سپس این فاعل‌ها طبق درجه اهمیت‌شان به ۱۰۰ دسته تقسیم شدند. سپس از هر دسته به صورت تصادفی ۱۰ نمونه انتخاب کردند که در نهایت ۱۰۰۰ نمونه از فاعل به‌ازای هر رابطه بدست آمد. به این ترتیب سعی شد نزدیک‌ترین داده با توجه به کار آقای West ایجاد شود تا از این طریق امکان مقایسه با آن تا حدودی فراهم گردد؛ هرچند این مقایسه به صورت تقریبی باشد و دقیق نباشد. البته از آن‌جایی که مجموعه‌ی داده‌ی دیگری هم به همین روش و با نمونه‌های تصادفی دیگری تشکیل شد و عملکرد سیستم در قالب معیارهای ارزیابی مشابه با عملکرد سیستم روی مجموعه‌ی داده‌ی اول مشاهده شد، این مسئله می‌تواند نتایج دریافتی از سیستم را برای مقایسه با سیستم West اعتبار بیشتری ببخشد.

جدول (۲) : مقایسه‌ی عملکرد سیستم طراحی شده با سیستم پیشین

روش	همسر		فرزندان	
	MRR	MAP	MRR	MAP
West	۰.۵۴	۰.۵۰	۰.۲۵	۰.۱۸
START	۰.۴۸	۰.۴۱	۰.۱۹	۰.۱۶
Wolfram|Alpha	**۰.٦۱**	**۰.۵۵**	**۰.٦۸**	**۰.٦۳**

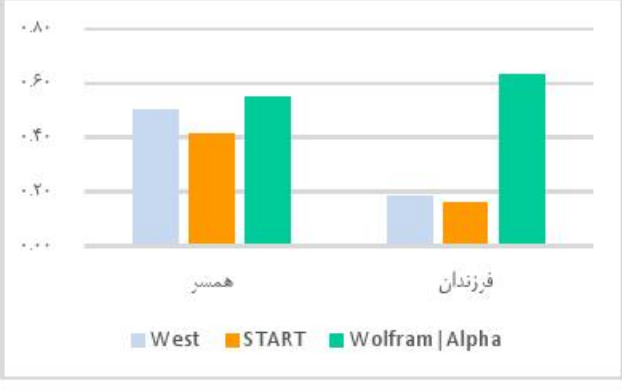
۱- معیارها و نتایج ارزیابی

در این بخش، ابتدا کیفیت پاسخ‌ها را از نظر دقت رتبه‌بندی در چیدمان پاسخ‌های صحیح بررسی می‌کنیم. از آن‌جایی که هم تعداد پاسخ‌های صحیح اهمیت دارد و هم جایگاه آن‌ها در بین کل پاسخ‌های برگشتی مهم است، از معیارهای MRR و MAP استفاده شده است. مقدار RR یک موجودیت در یک لیست رتبه‌بندی‌شده برابر با معکوسِ رتبه‌ی بالاترین پاسخ صحیح است. وقتی‌که با چند لیست رتبه‌بندی شده روبه‌رو باشیم و از RR آن‌ها میانگین بگیریم، به آن MRR گفته می‌شود. معیار MAP علاوه‌بر اینکه پاسخ صحیح اول در فهرست را درنظر می‌گیرد، رتبه‌ی همه‌ی پاسخ‌های صحیح را نیز به‌کار می‌برد. فرمول این دو معیار در [9] ذکر شده‌اند.

به ازای هریک از ۱۰۰۰ فاعلی که برای هر رابطه داریم یک لیست پاسخ رتبه‌بندی شده وجود دارد و طبق آن محاسبات لازم انجام شده است. برای ارزیابی، یک بار سیستم را با تنظیم سیستم پرسش‌وپاسخ START و بار دیگر با Wolfram|Alpha برای مرحله‌ی استخراج متنی پاسخ‌ها اجرا کردیم و عملکرد هریک را بررسی نمودیم. جدول (۲) و شکل‌های (۲) و (۳)،

شکل (۲) : عملکرد سیستم با توجه به معیار MRR.

شکل (۳) : عملکرد سیستم با توجه به معیار MAP.



سیستم طراحی‌شده را از نظر این دو معیار ارزیابی، با سیستم آقای West مقایسه کرده و مشاهده می‌شود زمانی که از سیستم پرسش‌وپاسخ Wolfram|Alpha برای مرحله‌ی استخراج پاسخ‌ها استفاده می‌شود نتیجه بهبود می‌یابد. برای رابطه‌ی «فرزندان»، این بهبود به شکل قابل‌توجهی محسوس می‌باشد. دلیل تفاوت در نتایج بین این دو رابطه می‌تواند به ماهیت رابطه و نوع نگاشت موجودیت‌ها برگردد. اینکه معمولا «فرزندان» بیشتر از «همسر» می‌توانند از روش سوم در نگاشت موجودیت پیروی کنند.

اگر بخواهیم علت کم‌تر بودن مقادیر معیارهای ارزیابی را در حالت استخراج پاسخ بر اساس START بررسی کنیم، عامل آن می‌تواند ضعیف‌تر بودن سیستم پرسش‌وپاسخ START نسبت به سیستم Wolfram|Alpha باشد. زیرا START پاسخ‌ها را از کل وب باید جستجو کند درحالی‌که Wolfram|Alpha پایگاه‌دانش خود را که بسیار غنی است جستجو می‌کند.

اگر بخواهیم روش Wolfram|Alpha را به عنوان نماینده‌ای موفق از سیستم طراحی‌شده با سیستم آقای West مقایسه کنیم و علت این تفاوت در نتایج را جستجو کنیم می‌توان به درنظر گرفتن همه‌ی قطعه‌های کوچک از پاسخ‌ها اشاره کرد. در سیستم جدید طراحی شده، اگر به عنوان مثال تنها نام کوچک یک شخص به عنوان پاسخ در دسترس باشد و استخراج شده باشد، همین اطلاعات کم توسط سیستم مورد استفاده قرار می‌گیرد و سعی می‌شود با استفاده از آن موجودیت کاندید حدس زده شود. این درحالی است که روش‌های مشابه ممکن است قطعه متن‌هایی که خود به تنهایی اطلاعات تعیین‌کننده‌ای ندارند را نادیده بگیرند. همین مسئله باعث می‌شود امکان پیش‌بینی‌های بیشتر برای سیستم ما فراهم شود. این امکان درصورتی می‌تواند به جواب‌های صحیح بیشتر منجر شود که مرحله‌ی استخراج متنی پاسخ‌ها از طریق منابع غنی اتفاق بیفتد.

دلیل دیگری که در درنتیجه می‌تواند موثر باشد تفاوت عملکرد سیستم West برای استخراج اطلاعات می‌باشد. آن سیستم کل وب را برای یافتن پاسخ با استفاده از یک سیستم پرسش‌وپاسخ خودساخته جستجو می‌کند و این جستجو از طریق چندین پرس‌وجو صورت می‌گیرد. همین مسئله می‌تواند درعین حال که پاسخ‌های صحیح بیشتری را منجر شود، پاسخ‌های اشتباه دیگری را هم درصورت برخورد نادقیق با جواب‌های استخراجی و به دلیل وسعت اطلاعات ایجاد کند. بنابراین باید مکانیزم دقیقی برای استخراج نهایی پاسخ از بین آن‌ها داشته باشد تا از اشتباه‌ها جلوگیری شود و در رتبه‌بندی نهایی بیشتر شاهد پاسخ‌های صحیح باشیم.

به عنوان معیار ارزیابی دیگر برای مقایسه‌ی عملکرد سیستم در دو رابطه‌ی «همسر» و «فرزندان» می‌توان از precision و recall کمک گرفت. معیار precision نشان می‌دهد از بین همه‌ی موجودیت‌هایی که توسط سیستم پیش‌بینی شده، چند مورد از آن‌ها صحیح پیش‌بینی شده‌اند. تفاوت عملکرد این معیار برای هردو روش استخراج متنی پاسخ‌ها و برای هر دو رابطه‌ی «همسر» و «فرزندان» در شکل (۴) قابل مشاهده می‌باشد. همان‌طور که این نمودار نشان می‌دهد در حالت Wolfram|Alpha سیستم برای رابطه‌ی «همسر» موفق شده است از بین کل موجودیت‌هایی که برای آن رابطه پیش‌بینی کرده است، پیش‌بینی‌های صحیح‌تری داشته باشد. از زاویه‌ی دیگر، اگر از کل موجودیت‌هایی که باید سیستم پیش‌بینی می‌کرد، تعداد آن‌هایی که توانسته پیش‌بینی کند را بدست آوریم؛ درواقع recall آن را یافته‌ایم. نتیجه‌ی

شکل (٤) : عملکرد سیستم با توجه به معیار precision.

شکل (٥) : عملکرد سیستم با توجه به معیار recall.

بررسی این معیار ارزیابی برای رابطه‌های «همسر» و «فرزندان» با استفاده از Wolfram|Alpha و START در شکل (۵) نشان داده شده است.

همان‌طور که در نمودار مشخص است برای رابطه‌ی «فرزندان» در این زمینه موفقیت بیشتری کسب شده است. به طور کلی با درنظرگرفتن هر دو معیار precision و recall می‌توان چنین نتیجه گرفت که اگرچه Wolfram|Alpha برای رابطه‌ی «فرزندان» موجودیت‌های مورد انتظار بیشتری را پیش‌بینی می‌کند اما در این پیش‌بینی، به‌صورت نسبی اشتباه‌های بیشتری مرتکب شده است؛ هرچند این مقدار زیاد محسوس نیست. سیستم طراحی‌شده اگر مبتنی‌بر START اجرا شود نتیجه در قالب precision و recall نیز در مقایسه با Wolfram|Alpha ضعیف‌تر است. این تنظیم از سیستم برای رابطه‌ی «همسر» نسبت به رابطه‌ی «فرزندان» هم حدس‌های صحیح‌تری زده و هم موفق شده بیشتر از «فرزندان» نسبت به حالت موردانتظار پیش‌بینی داشته‌باشد.

در کل با توجه به همه‌ی معیارهای ارزیابی، تنظیم Wolfram|Alpha به عنوان استخراج‌کننده‌ی متنی پاسخ‌ها می‌تواند نماینده‌ی موفقی از سیستم باشد زیرا در همه‌ی آن معیارها به نتایج قابل قبولی رسیده است.

۲- بحث و نتیجه‌گیری

در این پژوهش راهکاری برای تکمیل اطلاعات موجود در پایگاه‌های دانش ارائه شد. این راهکار برای اشخاصی که اطلاعات همسر و فرزندان‌شان در Freebase موجود نیست پیشنهاد شده و آزمایش گردیده است. روش پیشنهادی از قابلیت دریافت پرسش‌های زبان طبیعی سیستم‌های پرسش-

وپاسخ استفاده کرده و سعی می‌کند از این طریق، بخش‌های مجهول از پایگاه‌دانش را به‌صورت متنی بیابد. از آن‌جایی‌که در سیستم طراحی‌شده فرض شده که هیچ اطلاعات جانبی دیگری در دسترس نیست که با تکنیک‌های متن‌کاوی و شباهت‌یابی پاسخ‌های متنی را به موجودیت‌های صحیح نگاشت دهد، سعی می‌کند با استفاده از رابطه‌های مبتنی‌بر پایگاه‌دانش، روابط خانوادگی، ماهیت رابطه و موجودیت مورد جستجو نگاشت موجودیت را با حداقل اطلاعاتِ دردسترس انجام دهد. آن راه‌حل پیشنهادی که مبتنی‌بر سیستم پرسش‌وپاسخ Wolfram|Alpha است، در مقایسه با کار مشابه در گذشته موفق شده از نظر معیارهای MRR و MAP بهبود یابد.

به‌عنوان کارهایی که در آینده می‌توان برای سیستم انجام داد، بهبود آن در رابطه‌های مهم دیگری است. در صورتی که ماهیت موجودیت‌های مورد جستجو امکان ارتباط با منابع معنایی مانند WordNet [14] را فراهم کند، از این امکان استفاده شود تا قوانین بیشتری استخراج شود. هم‌چنین می‌توان مرحله‌ی استخراج اطلاعات متنی را با استفاده از منابع متنی یا ساختاریافته‌ی قابل اتکا مانند صفحات بیوگرافی افراد بهبود بخشید.

مراجع

[1] Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge." In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 1247-1250. AcM, 2008.

[2] Vrandečić, Denny. "Wikidata: A new platform for collaborative data collection." In Proceedings of the 21st International Conference on World Wide Web, pp. 1063-1064. ACM, 2012.

[3] Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." In Proceedings of the 16th international conference on World Wide Web, pp. 697-706. ACM, 2007.

[4] Wu, Wentao, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. "Probase: A probabilistic taxonomy for text understanding." In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 481-492. ACM, 2012.

[5] Rodriguez, Miguel, Sean Goldberg, and Daisy Zhe Wang. "SigmaKB: multiple probabilistic knowledge base fusion." Proceedings of the VLDB Endowment 9, no. 13 (2016): 1577-1580.

[6] Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. "Dbpedia: A nucleus for a web of open data." In The semantic web, pp. 722-735. Springer Berlin Heidelberg, 2007.

[7] Zhu, Jun, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. "StatSnowball: a statistical approach to extracting entity relationships." In Proceedings of the 18th international conference on World Wide Web, pp. 101-110. ACM, 2009.

[8] Etzioni, Oren, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. "Web-scale information extraction in knowitall:(preliminary results)." In Proceedings of the 13th international conference on World Wide Web, pp. 100-110. ACM, 2004.

[9] West, Robert, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. "Knowledge base completion via search-based question answering." In Proceedings of the 23rd international conference on World Wide Web, pp. 515-526. ACM, 2014.

[10] Katz, Boris. "Annotating the World Wide Web using natural language." In Computer-Assisted Information Searching on Internet, pp. 136-155. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1997.

[11] MIT InfoLab. START. http://start.csail.mit.edu/index.php.

[12] Wolfram Alpha LLC. Wolfram|Alpha. http://www.wolframalpha.com.

[13] Bast, Hannah, Florian Bäurle, Björn Buchhold, and Elmar Haußmann. "Easy access to the freebase dataset." In Proceedings of the 23rd International Conference on World Wide Web, pp. 95-98. ACM, 2014.

[14] Fellbaum, Christiane. WordNet. Blackwell Publishing Ltd, 1998.

زیرنویس‌ها

1 Databases
2 Entity
3 Lexical Analysis
4 Resource Description Framework
5 Query
6 Search Engines
7 Knowledge-Sharing Communities
8 Facts
9 Question Answering (QA)
10 http://www.imdb.com/