Library and Information Sciences

Print ISSN: 1680-9637
Online ISSN: 5977-2676

Investigating the Effectiveness of Semantic Tagging in Sense Disambiguation of Specialized Homographs from the Perspective of Precision in Retrieving Scientific Texts

Mina Rezaei Dinani
*Corresponding Author, Ph.D. Candidate in Knowledge and Information Science, Faculty of Educational Sciences and Psychology, Alzahra University, Tehran, Iran. E-mail: mina.rezaei.d@gmail.com.

Masoumeh Karbala Aghaei Kamran
Associate Professor, Department of Knowledge and Information Science, Faculty of Educational Sciences and Psychology, Alzahra University, Tehran, Iran. E-mail: mkamran@alzahra.ac.ir.

VahidReza Mirzaeian
Assistant Professor, Department of English Language & Literature, Faculty of Literature, Alzahra University, Tehran, Iran. E-mail: mirzaeian@alzahra.ac.ir.

Abstract

Objective: The aim of this study was to explain the application of text corpus tagging method in sense disambiguation from specialized homographs and increasing the retrieval precision of scientific texts containing such homographs.

Methodology: This research was conducted experimentally and it is a supervised method that is one of the three methods of word sense disambiguation. The research sample consisted of 442 scientific articles of two groups of experimental group and control group. The control group had 221 full-text articles without tags and the experimental group had the same 221 tagged articles, which were tested in the information retrieval system to measure the effectiveness of tagging in sense disambiguation from specialized homographs.

Findings: The research findings indicate that while retrieval in the control group due to sense ambiguity of specialized homographs is accompanied with false drop and reduced precision, tagging of specialized homographs in the full text of articles in the experimental group have direct effect in sense disambiguation from specialized homographs. It is possible to retrieve specialized homographs related to each tag, while in retrieval based on the control group, this is not possible. The level of significance of the Wilcoxon signed-rank test (P = 0.0001, Z = -5/909)

shows that the accuracy of retrieval results of specialized homograph after using the tagged text corpus in the information retrieval system is significantly different. Examination of negative and positive rankings shows that the accuracy of the results after using the tagged text corpus has increased significantly and has reached its maximum level of 1.

Conclusion: The rate of precision in retrieving scientific texts in the research findings is evidence of acceptable tagging effectiveness in sense disambiguation of specialized homographs and its effective role in optimizing the information retrieval system. If retrieval system designers focus on optimizing retrieval formulas in search of specialized homograph and empower retrieval systems to search for related documents, researchers with any physiological, experimental, and knowledge characteristics will be able to access related documents. Access their information needs in a short time. In this study, the value of the text corpus as a rich treasure of knowledge-based for information retrieval system was revealed in distinguishing the semantic role of specialized homographs. Although the research was conducted on limited corpus, the researcher believes that because this limited text corpus was designed in a principled way and the texts were consciously selected, the results of the findings can be generalized to all scientific texts in various fields.

Keywords: Specialized homograph, Information retrieval, Information organization, Tagging, Text corpus.

Article type: Research

Publisher: Central Library of Astan Quds Razavi
Library and Information Sciences, 2021, Vol. 24, No.4, pp. 31-59.
Received: 24/06/2021 - Accepted: 08/09/2021

کتابداری و اطلاع رسانی
سازمان کتابخانه‌ها، موزه‌ها و مرکز اسناد آستان قدس رضوی
شاپا چاپی: ۹۶۳۷-۱۶۸۰
شاپا الکترونیکی: ۲۶۷۶-۵۹۷۷

واکاوی اثربخشی برچسب‌گذاری معنایی در رفع ابهام معنایی هم‌نویسه‌های تخصصی از نظر میزان دقت در بازیابی متون علمی

مینا رضایی دینانی
*نویسنده مسئول، دانشجوی دکترای علم اطلاعات و دانش‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه الزهرا، تهران، ایران. رایانامه: mina.rezaei.d@gmail.com

معصومه کربلاآقایی کامران
دانشیار گروه علم اطلاعات و دانش‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه الزهرا، تهران، ایران. رایانامه: mkamran@alzahra.ac.ir

وحیدرضا میرزاییان
استادیار گروه زبان و ادبیات انگلیسی، دانشکده ادبیات، دانشگاه الزهرا، تهران، ایران. رایانامه: mirzaeian@alzahra.ac.ir

چکیده

هدف: تبیین کاربرد روش برچسب‌گذاری پیکره متنی در رفع ابهام معنایی از هم‌نویسه‌های تخصصی از نظر میزان دقت در بازیابی متون علمی حاوی این گونه هم‌نویسه‌ها.

روش: این پژوهش از حیث هدف کاربردی است که به روش تجربی انجام شد و در رفع ابهام معنایی، روشی با نظارت محسوب می‌شود. جامعه پژوهش را ۴۴۲ مقاله علمی در قالب دو گروه گواه و آزمون تشکیل دادند. گروه گواه دارای ۲۲۱ متن کامل مقاله بدون برچسب و گروه تجربی دارای همان ۲۲۱ مقاله اما این بار برچسب‌گذاری شده، بود که در نظام بازیابی اطلاعات برای سنجش کارآیی برچسب‌ها در رفع ابهام معنایی از هم‌نویسه‌های تخصصی مورد آزمون قرار گرفتند.

یافته‌ها: سطح معنی‌داری آزمون رتبه‌های علامت‌دار ویلکاکسون (P = ۰/۰۰۰۱، Z = −۵/۹۰۹) نشان می‌دهد که میزان دقت نتایج بازیابی هم‌نویسه‌های تخصصی بعد از به کارگیری پیکره تخصصی برچسب‌گذاری‌شده در نظام بازیابی اطلاعات نسبت به قبل از آن تفاوت معنی‌داری دارد. بررسی رتبه‌های منفی و مثبت نشان می‌دهد میزان دقت نتایج بعد از به کارگیری پیکره تخصصی برچسب‌گذاری‌شده به میزان معنی‌داری افزایش یافته و به حد بیشینه آن یعنی ۱ رسیده است.

نتیجه‌گیری: اگر طراحان سیستم‌های بازیابی بر بهینه‌سازی فرمول‌های بازیابی متمرکز شوند و نظام‌های بازیابی را برای جستجوی اسناد مرتبط توانمند سازند، پژوهشگران با هر ویژگی فیزیولوژیکی، تجربی و دانشی قادرند به اسناد مرتبط با نیاز اطلاعاتی خود با صرف زمانی اندک دسترسی یابند. در این پژوهش، ارزش پیکره متنی به عنوان گنجینه غنی دانش‌محور، در ایجاد تمایز نقش معنایی هم‌نویسه‌های تخصصی، آشکار شد.

کلیدواژه‌ها: هم‌نویسه تخصصی، بازیابی اطلاعات، سازماندهی اطلاعات، برچسب‌گذاری، پیکره متنی.

نوع مقاله: پژوهشی

ناشر: کتابخانه مرکزی آستان قدس رضوی

کتابداری و اطلاع‌رسانی، ۱۴۰۰، دوره ۲۴، شماره ۴، شماره پیاپی ۹۶، صص. ۳۱-۵۹

تاریخ دریافت: ۱۴۰۰/۴/۳- تاریخ پذیرش: ۱۴۰۰/۶/۱۷

مقدمه

هم‌نویسه‌ها[۱] آن دسته از کلمات چندمعنا هستند که به صورت یکسان نوشته می‌شوند؛ یعنی واژگانی که شکل نوشتاری یکسان اما معنای متفاوتی از همدیگر دارند (هرست[۲]، ۱۹۹۱). معنای این قبیل واژه‌ها، تنها وقتی در متن قرار گیرند، با توجه به بافت موضوعی و زمینه[۳]، قابل تشخیص است. بنابراین نتایج بازیابی هم‌نویسه‌ها معمولاً می‌تواند ضمن پراکندگی موضوع‌های همانند و اجتماع موضوع‌های بی‌ربط، کارایی نظام و سطح رضایت کاربران را تا حد قابل توجهی کاهش دهد.

اصطلاحات تخصصی، برای ارتباط علمی و انتقال صحیح اطلاعات به کار گرفته می‌شود و چنانچه دچار هرج و مرج و نابسامانی شود، زبان تفهیم و تفاهم و جریان درست اطلاعات مختل می‌شود (مرتضایی، ۱۳۸۱)؛ بنابراین دقت و صراحت در ترجمه و گزینش اصطلاح‌های تخصصی برای مفهومی خاص در هر یک از رشته‌های علمی و با هدف جلوگیری از هم‌نویسگی اصطلاح‌ها، امری ضروری است؛ اما با وجود اهمیت این موضوع، تطبیق اصطلاحنامه‌ها نشان می‌دهد که هم‌نویسه‌های تخصصی زیادی در رشته‌های علمی مختلف وجود دارد. هم‌نویسه‌های تخصصی در رشته‌های علمی مختلف، شکل نوشتاری یکسانی دارند اما از مفهوم و تعریف منحصر به آن رشته برخوردارند. در حالی که یک متخصص ممکن است بتواند به لطف سوابق حرفه‌ای خود معنای صحیح هم‌نویسه را در یک مقاله علمی تعیین کند، برای ماشین، شناسایی معنای یک هم‌نویسه تخصصی، به سادگی انسان نیست و روش‌های خودکار بازیابی اغلب بدون داشتن چنین دانشی نمی‌توانند اصطلاحات را به درستی از هم تفکیک کنند (پروکوفوی، دمارتینی، بویاراسکای، روچایسکی و ماروکس[۴]، ۲۰۱۳). در نتیجه این مسئله می‌تواند موجب ابهام زیادی در درک متن و ریزش کاذب شدید به خصوص در جستجوهای تخصصی شود (مینایی بیدگلی، اکبری و حسنی، ۱۳۸۶، نقل در: ستوده و هوشیار، ۱۳۹۷). رچنر[۵] و گوش[۶] (۱۹۹۹) معتقدند حتی اگر تفکیک مدارک مرتبط از غیرمرتبط بیش از یک ثانیه طول نکشد، به طور قابل توجهی گردش کار کاربر را مختل می‌کند و این آزاردهنده است.

اگر طراحان نظام‌های بازیابی بر بهینه‌سازی فرمول‌های بازیابی در جستجوی هم‌نویسه‌های تخصصی متمرکز شوند و نظام‌های بازیابی را برای جستجوی اسناد مرتبط توانمند سازند، پژوهشگران با هر ویژگی فیزیولوژیکی، تجربی و دانشی قادرند به اسناد مرتبط با نیاز اطلاعاتی خود با صرف زمانی اندک دسترسی

---

۱. در این پژوهش، از بین سه واژه مترادف هم‌نویسه، هم‌نگاره و جناس تام، واژه هم‌نویسه برای پرداختن به موضوع استفاده می‌شود.

2. Hearst
3. Context
4. Prokofyev, Demartini, Boyarsky, Mauroux
5. Pretschner
6. Gauch

یابند. ایجاد و توسعه ابزارهای معنایی که مجموعه‌ای از مفاهیم، نمادها و همچنین روابط بین آن‌ها را تبیین می‌کند، از راه‌هایی است که به بازیابی اثربخش کمک می‌کند (هوشیار، ۱۳۹۴).

فنون رفع ابهام معنایی[1] از هم‌نویسه‌ها، نیاز به منابع ویژه و قابل توجه زبانی دارد و دانش، یک مؤلفه پایه برای رفع ابهام معنایی است. از جمله منابع دانش برای رفع ابهام معنایی، اصطلاحنامه‌ها، سرعنوان‌های موضوعی[2]، وب معنایی[3] و پیکره‌ها[4] هستند. در تعریفی ساده از پیکره می‌توان گفت که مجموعه بزرگی از متون معتبر نوشتاری و یا گفتاری آوانویسی شده است که طبق معیارهای خاصی در قالب الکترونیکی برای هدف مشخصی، جمع‌آوری و ذخیره شدند (بوکر[5]، ۲۰۱۸). برچسب‌زنی موضوعی متون پیکره، امری مهم در حوزه بازیابی اطلاعات و نوعی دسته‌بندی یا طبقه‌بندی در زبان طبیعی است. از طریق این نوع برچسب‌گذاری، طبقه‌ها یا ویژگی هم‌نویسه‌ها مشخص شده و از همدیگر متمایز می‌شوند. این واقعیت که هر کلمه دارای مجموعه معانی خاص خود است، هم مسئله را گسترده می‌کند و هم در نتیجه آن، مقدار داده‌های حاشیه نویسی شده دستی مورد نیاز برای رسیدن به عملکردهای رضایت بخش، افزایش می‌یابد و بنابراین فرایند کلی حاشیه‌نویسی را بسیار سخت می‌کند. بنابراین، تعجب‌آور نیست که حتی زبان‌های با منابع زیاد (مانند انگلیسی)، هنوز از داشتن پیکره‌های بزرگ با برچسب دستی فاصله دارند (باربا، پروکوپیو، لمپلوگنو، پاسینی و نویگلی[6]، ۲۰۲۰).

بنابراین مسئله‌ای که پژوهش حول آن شکل گرفت و به آن پرداخت تعیین میزان تأثیری است که استفاده از پیکره برچسب‌گذاری شده می‌تواند در میزان رفع ابهام معنایی هم‌نویسه‌های تخصصی و دقت نتایج بازیابی حاصل از آن داشته باشد. به عبارت دیگر، پژوهش به منظور پاسخگویی به این سؤال طرح‌ریزی شد که آیا میزان دقت نتایج بازیابی هم‌نویسه‌های تخصصی قبل و بعد از به کارگیری پیکره تخصصی برچسب‌گذاری‌شده در نظام بازیابی اطلاعات تفاوت معناداری وجود دارد یا خیر؟ در همین راستا و به منظور پاسخ به این پرسش، بررسی فرضیه زیر موردنظر است؛ برچسب‌گذاری معنایی هم‌نویسه‌های تخصصی در پیکره متون علمی، دقت بازیابی را افزایش می‌دهد.

از جمله روش‌های رفع ابهام معنایی که در این پژوهش نیز به کار گرفته می‌شود، روش نظارتی[7] است. روش‌های کنونی رفع ابهام معنایی تحت نظارت، معانی را به عنوان برچسب‌های مجزایی تلقی می‌کنند (کومار،

1. Word Sense Disambiguation (WSD)
2. Subject Headings
3. Semantic web
4. Corpus
5. Bowker
6. Barba, Procopio, Campolungo, Pasini & Navigli
7. Supervised

جت، ساکسنا و تلوکدر[1]، ۲۰۱۹). در این پژوهش، مجموعه‌ای از هم‌نویسه‌های تخصصی به همراه برچسب‌هایشان، به عنوان مجموعه آموزش در دسترس است. از دیگر تسهیلاتی که برای رفع ابهام معنایی از هم‌نویسه‌های تخصصی، ضروری به نظر می‌رسد، امکان استفاده از انبوه داده‌های متنی در قالب پیکره است که مجموعه آزمون را تشکیل می‌دهند. در این پژوهش، ارزش پیکره متنی برچسب‌گذاری شده، به عنوان گنجینه غنی دانش‌محور برای نظام بازیابی اطلاعات، در ایجاد تمایز نقش معنایی هم‌نویسه‌های تخصصی، آشکار می‌شود.

پیشینه پژوهش

پیشینه نظری

راه‌حل سنتی و متداول در نظام‌های بازیابی اطلاعات بر جستجوی مبتنی بر کلیدواژه تمرکز دارد (خان ، مک لود و هاوی[2]، ۲۰۰۴). پژوهشگران داخل کشور همچون دلیخون (۱۳۹۵)، شهبازی و شاهینی (۱۳۹۴)، عبدالهی نورعلی (۱۳۸۶)، عبدالهی نورعلی و جوکار (۱۳۸۸)، گل‌تاجی و بذرگر (۱۳۸۹)، نوروزی و هماوندی (۱۳۹۴) و نیز پژوهشگران خارج از کشور مانند لازارینیس[3] (۲۰۰۷)، لواندوفسکی[4] (۲۰۰۸) و ژانگ و لین[5] (۲۰۰۷) بر اهمیت نقش کلیدواژه بر بازیابی اطلاعات تأکید کرده‌اند. پژوهش‌های بسیاری نشان می‌دهد که پژوهشگران نیز به طور معمول، بازیابی مدارک و اسناد علمی مانند مقاله‌های مجلات، همایش‌ها و پایان‌نامه‌ها را در نظام‌های بازیابی خودکار اطلاعات، با درج کلیدواژه انجام می‌دهند و جستجوی کلیدواژه‌ای را به نوع ترکیبی آن و جستجوی ساده را به استفاده از راهبردهای جستجو ترجیح می‌دهند؛ از قبیل این پژوهش می‌توان به پژوهش‌های اسپینک، والفرام، جانسن و ساراسویک[6] (۲۰۰۱)، دلیخون (۱۳۹۴)، شاپوری (۱۳۷۹)، شوتز[7] (۲۰۱۴)، طباطبایی جعفری (۱۳۹۰) و عبدالهی نورعلی و جوکار (۱۳۸۸) اشاره کرد. در برخی از پژوهش‌ها مانند زرداری (۱۳۹۵) و یوسفی‌راد (۱۳۸۸) کلیدواژه‌مدار بودن فرایند بازیابی اطلاعات، مهمترین دلیل نارسایی نظام‌های بازیابی اطلاعات شمرده شده است. رچنر و گوش (۱۹۹۹) نیز بر این عقیده‌اند که کلیدواژه‌ها همیشه وسیله مناسبی برای یافتن اطلاعات مورد علاقه کاربر نیست و معمولاً برای مفهومی‌سازی نیاز و منظور کاربر، کافی نیست. در چنین شرایطی بازنمون اطلاعات در نظام بازیابی با شبکه مفهومی متون

1. Kumar, Jat, Saxena & Talukdar
2. Khan, McLeod & Hovy
3. Lazarinis
4. Lewandowski
5. Zhang & Lin
6. Spink, Wolfram, Jansen & Saracevic
7. Schutze

مورد نیاز کاربران همخوانی ندارد. تکیه ابزارهای جستجو بر شکل کلیدواژه‌های ورودی توسط کاربر، ریزش کاذب و کاهش دقت بازیابی را به ویژه در بازیابی هم‌نویسه‌ها بدیهی می‌سازد. مرور این قبیل پیشینه‌ها نشان می‌دهد وجود چالش‌های ریختی در پایگاه‌های اطلاعاتی و موتورهای جستجوی عمومی و تأثیر آن‌ها بر جامعیت و مانعیت بازیابی اطلاعات، تأیید شده و این چالش‌ها می‌توانند باعث کاهش نمایانی مدارک مرتبط و افزایش نمایانی مدارک غیرمرتبط شوند؛ ابهام معنایی حاصل از وجود هم‌نویسه‌ها نیز از این امر مستثنی نیست و این مسئله می‌تواند بر موفقیت و رضایت کاربر از نتایج بازیابی شده، تأثیر سوء داشته باشد.

پیشینه تجربی

مطالعات در حوزه اثربخشی بازیابی اطلاعات بر سه محور کلی متمرکز است؛ ۱) بررسی دشواری‌های نگارشی در اثربخشی بازیابی اطلاعات ۲) آزمایش تأثیر تکنیک‌ها و یا ابزارهای خاص بر اثربخشی بازیابی ۳) طراحی و آزمایش تکنیک‌ها، الگوریتم‌ها و یا ابزارهای خاص (ستوده و هنرجویان، ۱۳۹۱). این پژوهش در زمره پژوهش‌های محور سوم جای می‌گیرد.

پیشینه‌های مرتبط با طراحی و آزمایش تکنیک‌ها، الگوریتم‌ها و یا ابزارهای خاص

محور طراحی و آزمایش تکنیک‌ها، الگوریتم‌ها و یا ابزارهای خاص در زمره مطالعات تحقیقاتی و عملیاتی قرار می‌گیرد. در این گونه مطالعات دو محور دیگر را نیز مدنظر قرار می‌دهند تا مؤثرترین تکنیک، الگوریتم و یا ابزار برای اثربخشی بازیابی اطلاعات طراحی و آزمایش شود و یا تکنیک‌های موجود بهبود داده شوند.

یوسفان نجف‌آبادی (۱۳۸۲) نظام بازیابی را با استفاده از نمایه‌گذاری معانی و ریشه‌یابی برخی واژه‌های عربی در زبان فارسی طراحی نمودند. نتایج پژوهش وی بهبود کلی در کارایی نظام پیشنهادی را نشان داد.

آزاد (۱۳۸۶) از مدل بازیابی فضای برداری برای طراحی سیستم بازیابی اطلاعات بهره برد و تأثیر سیاست‌های مختلف وزن‌دهی واژه‌ها را در کارایی سیستم مورد بررسی قرار داد.

حسن‌زاده (۱۳۸۹) در پژوهش خود دریافت کارایی مدل نمایه‌گذاری معنایی پنهان از کارایی مدل نمایه‌گذاری مفهومی بیشتر است.

نتایج پژوهش معروفی و پیله‌ور (۱۳۹۰) نشان داد که الگوریتم پیشنهادی در رفع ابهام معنای ۱۵ کلمه با افزودن فاکتور ارتباطی میان دسته‌بندی موضوعی در پیکره، دقت بیشتری نسبت به الگوریتم اولیه دارد.

پژوهش مهرنهاد، قاسم‌زاده و نظارات (۱۳۹۰) نیز نشان داد طراحی و پیاده‌سازی یک سامانه بازیابی اطلاعات دوزبانه با استفاده از پیکره‌های زبانی، دقت بازیابی اطلاعات را در موتورهای جستجو افزایش می‌دهد.

مسعودی و راحتی قوچانی (۱۳۹۴) در مدل پیشنهادی خود، برای رفع ابهام از واژگان مبهم فارسی از روش دسته‌بندی بیشینه بی‌نظمی استفاده کردند و دقت رفع ابهام معنایی واژگان مبهم را ۹۷٪ تخمین زدند.

مظفری، تاکی، صباغ جعفری و یوسفیان (۱۳۹۶) در پژوهش خود سامانه‌ای قاعده‌مند برای رفع ابهام معنایی از حروف اضافه (از- در- با- تا) در زبان فارسی پیشنهاد دادند. نتایج آزمایش داده‌ها، دقت بالای عملکرد سامانه۹۹/۱۶ درصد را نشان داد.

دستغیب (۱۳۹۷) برای رفع ابهام واژگان، پیکره تک زبانه، ابزارک تجزیه متن، دیکشنری واژگان و مدل آماری زبان را تهیه کردند تا به عنوان منابع تحقیق در حوزه زبان فارسی مورد بهره‌برداری قرار گیرد.

کامیابی‌گل، اخلاقی باقوجری، عسگریان و حبیبی (۱۳۹۷) در پژوهش خود پیکره‌ای از مقاله‌های علمی-پژوهشی دانشگاه فردوسی مشهد طراحی نموده و برای تحقیقات داده‌کاوی و توصیف‌های داده‌محور در وبگاه کتابخانه مرکزی دانشگاه بارگذاری نمودند.

قیومی (۱۳۹۸) نوعی الگوریتم محاسباتی خوشه‌بندی برای تعیین خودکار معانی واژه‌های هم‌نویسه با توجه به بافت زبانی در یک مدل فضای برداری معرفی کرد. وی دریافت الگوریتم پیشنهادی بسیار خوب توانسته اطلاعات مربوط به واژه را از بافت به دست آورد.

گیل، چرچ و یاروسکای[1] (۱۹۹۲) در پژوهش خود روشی کمّی و تفکیک‌پذیر را برای تفکیک و تمایز معنای اسامی که دارای تعدد معنا هستند، طراحی کردند که میزان دقت آن تا ۹۲ درصد برآورد شده است.

در پژوهش خان، مک لود و هاوی (۲۰۰۴) یک مدل مبتنی بر مفهوم با استفاده از هستی‌شناسی پیشنهاد شده است. در این تحقیق بر داده‌های صوتی تمرکز شده و به صورت تحلیلی و تجربی نشان داده شده است که این مدل، در مقایسه با جستجوی مبتنی بر کلیدواژه، جامعیت و مانعیت بیشتری دارد.

ولت، فرناندز و کستلز[2] (۲۰۰۵) در پژوهش خود، یک مدل مبتنی بر هستی‌شناسی را با استفاده از حاشیه‌نویسی نیمه‌خودکار اسناد در نظام بازیابی، برای بهبود جستجوی مخازن اسناد بزرگ پیشنهاد دادند. آزمایش نمونه، پیشرفت‌هایی را در رابطه با جستجوی کلمات کلیدی نشان می‌دهد و زمینه را برای تحقیق‌ها و بحث‌های بیشتر فراهم می‌کند.

فانژو، کین و تائو[3] (۲۰۰۸) یک الگوریتم هیبرید به نام یادگیری مبتنی بر تبدیل هدایت شده درختی (TTBL) که ترکیبی از درخت تصمیم و یادگیری مبتنی بر تبدیل (TBL) است برای رفع ابهام معنایی هم‌نویسه‌ها معرفی کرده‌اند. آن‌ها دریافتند TTBL به طور قابل توجهی بهتر از درخت تصمیم عمل می‌کند.

1. Gale, Church & Yarowsky
2. Vallet., Fernandez & Castells
3. Fangzhou, Qin & Tao

همو[1] (۲۰۰۹) با استفاده از یک بنیانگر قاعده‌مند و یک پایگاه داده رابطه‌ای معنایی که در یک اصطلاح‌نامه آزمایشی جمع‌آوری شده بود، چارچوبی را برای افزایش کارایی بازیابی موتورهای جستجو برای جستجوی متن عربی پیشنهاد داد. یافته این تحقیق نشان داد کارایی موتورهای کاوش با استفاده از این ابزار افزایش می‌یابد.

نتایج تحقیق شن، وو، ونگ و کای[2]، (۲۰۱۱) نشان داد که الگوریتم‌های نیمه‌نظارتی (SSL) و یادگیری فعال (AL) هنگام رفع ابهام هم‌نویسه‌ها با استفاده از داده‌های بدون برچسب قادرند، ضمن حفظ کارایی، هزینه‌های برچسب‌گذاری دستی را، تا حد زیادی کاهش دهند.

منای[3] (۲۰۱۴) الگوریتم ژنتیک را برای رفع ابهام معنایی واژگان زبان عربی پیشنهاد داد. نتایج پژوهش، دقت پیش‌بینی نتایج را تا ۷۹ درصد نشان داد.

هارادا و سودا[4] (۲۰۱۴) روشی را برای طبقه‌بندی هم‌نویسه‌ها برای جلوگیری از آثار سوء آن‌ها در تحلیل داده‌های متنی رسانه‌های اجتماعی، ارائه دادند. روش ارائه شده ایشان بهبود دقت تا ۸,۵ درصد و کاهش ۱۶,۵ درصد میزان بدتشخیصی را در مقایسه با روش‌های متداول نشان دادند.

محمود، صلاح کریم و الشیشتاوی[5] (۲۰۱۸) استفاده از مدل بولین و تشخیص الگو، سیستمی برای تسهیل بازیابی اسناد بر اساس توسعه پرس و جو و ملزومات استخراج روابط معنایی از اسناد زیست‌شناسی ارائه دادند. نتایج پژوهش مقدار دقت و جامعیت بازیابی را ۱۰۰ درصد نشان داد.

به طور کلی عمده راه‌حل‌ها و پیشنهادات پیشینه‌های فارسی مطالعه شده در این موضوع، بر ضرورت تدوین استاندارد نگارش فارسی، تدوین دستنامه یا راهنمای جستجو، آموزش کاربران و توجه طراحان پایگاه‌های اطلاعاتی به رسم‌الخط فارسی معطوف شده است؛ اما به نظر می‌رسد این راهکارها گرچه در بعد نظری مفید و سازنده هستند ولی تأثیر سوء این چالش‌های زبانی در مرحله اجرا به دلیل عواملی چون نبود وحدت رویَه در تدوین استانداردها، تمایل کاربران به کمترین کوشش و ساده‌نگاری، عدم تمایل کاربران به صرف وقت و هزینه برای آموزش و توجه ناکافی طراحان پایگاه‌های اطلاعاتی به چالش‌های نگارش فارسی، بر جامعیت و مانعیت بازیابی به قوت خود باقی است. برخی دیگر از پیشینه‌های فارسی و غیرفارسی، متناسب با مقتضیات و محدودیت‌های زبانی خود، طراحی نظام‌های بازیابی اطلاعات را بر مبنای مدل‌های مختلف از قبیل فضای برداری، الگوریتم ژنتیک، درخت تصمیم، مدل‌های ترکیبی، روش‌های وزن‌دهی، ریشه‌یابی، نزدیکترین

1. Hammo
2. Shen, Wu, Wang & Cai
3. Menai
4. Harada & Tsuda
5. Mahmoud, Salah Kareem & El-Shishtawy

همسایگی فراوانی وقوع هم‌رخدادی، روابط پنهان کلمات و ... مورد بررسی قرار داده و پس از طراحی، کارایی نظام را در بازیابی اطلاعات متنی آزموده‌اند. پژوهشگران در برخی دیگر از پیشینه‌ها، رفع ابهام از هم‌نویسه‌ها را با روش‌های مختلف مبتنی بر قانون و یا مبتنی بر دانش، روش‌های نظارتی، نیمه‌نظارتی و بدون نظارت مورد مداقه قرار داده‌اند و برخی دیگر، از مجموعه‌های کوچک آموزش یا به عبارتی پیکره‌های برچسب‌گذاری شده و یک مجموعه آزمون برای رفع ابهام معنایی هم‌نویسه‌ها استفاده کرده‌اند.

به طور کلی نتایج حاصل از این پژوهش‌ها حاکی از آن است که ریخت‌شناسی عبارات و واژگان جستجو بر نتایج بازیابی مؤثر است و تکیه ابزارهای جستجو بر شکل کلیدواژه‌های ورودی توسط کاربر است که این می‌تواند بر موفقیت و رضایت کاربر از نتایج بازیابی شده، تأثیر سوء داشته باشد. این پژوهش‌ها عمدتاً به بررسی نقاط ضعف و قوت سامانه‌های جستجو معطوف شده‌اند و با هدف ارائه راهکارهای لازم به شناخت چالش‌ها و دشواری‌های زبانی پرداخته‌اند.

اسکارلینی، پاسینی و ناویگیلی[1] (۲۰۲۰) بر این عقیده‌اند که امروزه داده‌ها به عنوان سوخت عمل می‌کنند و ثابت شده داده‌های آموزشی بیشتر، عملکرد بهتری را باعث می‌شوند. ایجاد پیکره‌های بزرگ حاشیه‌نویسی شده به ویژه وقتی به زبان خاصی مربوط می‌شوند زمان‌بر و گران است. این ابزار برای زبان انگلیسی محدود و برای سایر زبان‌ها تقریباً وجود ندارد. وقتی تمرکز از زبان انگلیسی به سمت زبان‌های با منابع کم تغییر می‌کند، دسترسی به داده‌های حاشیه‌نویسی معنایی به طور فزاینده‌ای، ضرورت می‌یابد. مرور پیشینه‌ها همچنین نشان داد که عمده رویکردهای رفع ابهام معنایی بر مفاهیم اسمی تمرکز دارند و کمتر به رفع ابهام معانی سایر واژگان پرداخته شده این مسئله به ویژه در زبان‌هایی که فقر داده آموزشی دارند از جمله زبان فارسی، محسوس‌تر است.

پس از مرور تحلیلی- انتقادی پیشینه‌های مطالعه شده داخل و خارج از کشور، می‌توان گفت تاکنون پژوهشی که به موضوع واژگان تخصصی متون علمی و به طور ویژه ابهام معنایی هم‌نویسه‌های تخصصی و دانشگاهی و به طور مشخص، هم‌نویسه‌های تخصصی فارسی، رویکرد پیکره‌مدار داشته باشد، مشاهده نشده است و این پژوهش پیکره‌مدار در زمره معدود مطالعات واژه‌پژوهی دانشگاهی محسوب می‌شود. بنابراین به دلیل نقش مهم و تعیین‌کننده هم‌نویسه‌های تخصصی در مسیریابی دقیق و کامل پژوهش‌های علمی، با هدف برون‌رفت از چالش ابهام معنایی هم‌نویسه‌های تخصصی، ارائه راهکاری برای آن ضروری به نظر می‌رسد.

1. Scarlini, Pasini & Navigli

روش پژوهش

اغلب مطالعات پیکره‌ای جایی میان کمیت‌گرایی مطلق و کیفیت‌گرایی مطلق قرار دارند (گریس[1]، ۲۰۰۶ نقل در افراشی، عاصی و جولایی، ۱۳۹۴). ماهیت چندوجهی مسئله، پژوهش حاضر را در زمره تحقیقات تجربی قرار می‌دهد.

از دو بعد می‌توان به روش‌شناسی این پژوهش پرداخت؛

۱. بعد جزءمحور روش‌شناسی پژوهش که عمدتاً بر استفاده از پیکره در پژوهش تمرکز دارد.

از این منظر، رویکرد پژوهش، پیکره‌مدار است زیرا به دنبال ارائه و آزمون راهکاری بدیع برای رفع ابهام معنایی هم‌نویسه‌های تخصصی با استفاده از پیکره متنی است. از این بعد، همچنین می‌توان این پژوهش را در زمره پژوهش‌های کاربردشناسی پیکره‌ای یا کاربردشناسی تجربی محسوب کرد.

۲. بعد کلی پژوهش که روش‌شناسی همه مراحل پژوهش را مورد مداقه قرار می‌دهد.

از این بعد پژوهش، ابتدا طی روش کیفی از نوع اسنادی یا کتابخانه‌ای، مشاهده و مطالعه متون انجام و مفاهیم و فرایندها شناسایی و تحلیل شدند. از آنجا که در این پژوهش، تحلیل‌های واژگانی مربوط به پردازش زبان طبیعی مطرح می‌شود، در مراحل ابتدایی تحقیق به منظور شناسایی هم‌نویسه‌ها در متون از روش مشاهده مستقیم و تحلیل واژه بهره گرفته شد. نوع تحلیل در این مرحله از پژوهش، از نوع ریخت‌شناسی هم‌نویسه تخصصی است و در مراحل مختلف شناسایی و تجمیع، ذخیره، پردازش، بازیابی و مقایسه نتایج در نمونه مقاله‌های شش رشته علمی ریاضیات، علوم زمین، علوم زیستی، شیمی، فیزیک و جامعه‌شناسی مورد توجه و استفاده قرار گرفت.

در الگوریتم پیشنهادی پژوهش طبق استاندارد Semeval 2020، سه مرحله مجزا برای رفع ابهام معنایی هم‌نویسه‌های تخصصی فارسی پیش‌بینی شده است؛

۱- در مرحله آموزش، تحلیلگر از مجموعه آموزش استفاده می‌کند تا معانی را از واژه‌های چند معنی استنتاج کند. در این پژوهش، برای تهیه مجموعه آموزش از منابع زیر که از سامانه اصطلاحنامه‌های علمی و فنی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، اخذ شد، استفاده شد؛

- اصطلاحنامه شیمی/ تألیف تقی رجبی، حسین غریبی، ملوک‌السادات حسینی‌بهشتی و مهرداد نوروزی‌اقبالی

1. Gries

- اصطلاح‌نامه علوم زیستی/ تألیف اسماعیل اکبری، ملوک‌السادات حسینی بهشتی و مهرداد نوروزی‌اقبالی
- اصطلاح‌نامه جامعه‌شناسی/ باربارا بوث و میشل بلر
- اصطلاح‌نامه ریاضیات/ ملوک‌السادات حسینی بهشتی، وفایی، سعیده و مهرداد نوروزی‌اقبالی
- اصطلاح‌نامه علوم زمین/ مهری صدیقی، ملوک‌السادات حسینی بهشتی و مهرداد نوروزی‌اقبالی
- اصطلاح‌نامه فیزیک/ مریم نوروزی‌اقبالی، ملوک‌السادات حسینی بهشتی و مهرداد نوروزی‌اقبالی

با استفاده از نرم‌افزار مایکروسافت اکسل[1]، طی فرایند تطبیق واژگان تخصصی اصطلاحنامه‌های مورد نظر، هم‌نویسه‌ها شناسایی و پالایش شد. سپس، این هم‌نویسه‌ها به همراه برچسب موضوعی تعیین‌شده، به عنوان مجموعه آموزش سامانه در مسیر[2] مورد نظر، ذخیره شد.

۲- در مرحله آزمون، تحلیلگر یک مجموعه آزمون را فراهم می‌کند تا با استفاده از فهرست معانی در مرحله آموزش، از معنای واژه موردنظر رفع ابهام کند. مجموعه‌های آزمون ابزار تحقیقاتی هستند که وسیله‌ای برای محققان برای کشف مزایای نسبی استراتژی‌های مختلف بازیابی در یک محیط آزمایشگاهی فراهم می‌کند. به این ترتیب، آن‌ها انتزاعی از یک محیط بازیابی عملیاتی هستند که به آزمایشگران این امکان را می‌دهد تا برخی از متغیرهای تأثیرگذار بر عملکرد بازیابی را کنترل کنند و در نتیجه قدرت آزمایش‌های مقایسه‌ای افزایش می‌یابد (ورهیس، ۲۰۰۲).

استفاده از رویکردهای آماری برای نمونه‌گیری از متون در تولید پیکره -به عنوان مجموعه آزمون- بسیار آرمانی است اما متأسفانه کاربردشان در این زمینه چندان عملی نیست (صفری، ۱۳۹۱). اتکینز، کلیر و استلر[3] (۱۹۹۲) نیز بر این عقیده‌اند که رویکردهای استاندارد برای نمونه‌گیری آماری به سختی در ساخت یک پیکره زبانی کاربرد دارد. در این راستا، جونز و والر[4] (۲۰۱۵) معتقدند که یک پیکره کوچک هم می‌تواند به اندازه یک پیکره بزرگ مؤثر باشد و این به هدف و اصولی که پشتوانه ساخت آن پیکره است، بستگی دارد. در این پژوهش، به دلیل نوع روش جستجو و برچسب‌گذاری دستی و با هدف اطمینان از دقت فرایندها، مجموعه‌ای محدود بررسی شد. این تصمیم همچنین باعث شد که حصول نتیجه، مستلزم صرف وقت و انرژی افزون بر توان پژوهشگر نباشد. به دلیل این که پاسخ به پرسش پژوهش با استفاده از هیچ‌یک از پیکره‌های شناخته‌شده موجود فارسی امکان‌پذیر نیست، پژوهشگر با کمک متخصص نرم‌افزار، ناگزیر به ایجاد پیکره تخصصی بود.

1. Microsoft Excel
2. Directory
3. Atkins, Clear & Ostler
4. Jones & Waller

همان‌طور که از تقسیم‌بندی اتکینز، کلیر و استلر (۱۹۹۲) برمی‌آید، پیکره متنی این پژوهش از نوع نمونه‌ای است و از تمام متون علمی دانشگاهی و علمی- پژوهشی، تنها نمونه‌ای از مقالات علمی شش رشته تخصصی ریاضیات، شیمی، فیزیک، علوم زیستی، جامعه‌شناسی و علوم زمین انتخاب شده است. این پیکره، پیکره‌ای باز است که داده‌های آن بعد از جمع‌آوری قابل افزایش است. همچنین از نوع همزمانی است و در مدت معینی جمع‌آوری می‌شود. پیکره‌ای منفرد و تک‌زبانه است که به زبان فارسی تعلق دارد. سایر تقسیم‌بندی‌های اتکینز، کلیر و استلر (۱۹۹۲) در این پژوهش مصداق نمی‌یابد.

به منظور پاسخ به پرسش پژوهش، از دو گروه آزمودنی شامل گروه تجربی[1] و گروه کنترل[2] استفاده شده است. گروهی که تحت تأثیر متغیر مستقل قرار می‌گیرد، گروه آزمایشی یا تجربی و گروهی که از متغیر مستقل دور هستند را گروه شاهد، گواه یا کنترل می‌گویند. در مرحله بعد، نتیجه آزمون گروه تجربی با نتیجه آزمون گروه گواه یا کنترل مقایسه می‌شود. مقایسه گروه تجربی و کنترل، ارتباط یا همبستگی بین دو متغیر را نشان می‌دهد. پژوهشگر در روش تجربی به استفاده از طرح‌های اجرا شده دیگران، محدود نیست و می‌تواند طرح مناسب پژوهش خود را تهیه کند. در این پژوهش، تعداد مقاله به ازای هر هم‌نویسه به منظور ورود به مجموعه آزمون، در مرحله برنامه‌ریزی حداقل ۵ مقاله تعیین شد؛ اما در مرحله اجرا، تعداد دقیق آن را دو عامل تعیین می‌کرد: ۱- تعداد هم‌نویسه‌های موجود در متن هر مقاله ۲- سهم هر رشته از هم‌نویسه‌های انتخابی و این گریزناپذیر بود. بنابراین از تعداد ۱۵۰ مقاله برنامه‌ریزی شده،

۱۵۰ =(رشته) ۶ * (مقاله) ۵ * (هم‌نویسه) ۵

به تعداد ۲۲۱ مقاله و در مجموع دو گروه آزمون و تجربی به ۴۴۲ مقاله افزایش یافت.

بنا بر آنچه گفته شد، در مرحله عملی این پژوهش با استفاده از روش تجربی، دو گروه وجود داشت؛

۱- گروه گواه، شاهد و یا کنترل، ۲۲۱ مقاله‌ای است که از شش رشته علمی ریاضیات، علوم زمین، علوم زیستی، شیمی، فیزیک و جامعه‌شناسی جمع‌آوری شده و به صورت خام در پیکره قرار گرفته و تحت آزمون اولیه برای ارزیابی نتایج مرحله اول بازیابی قرار گرفتند.

۲- گروه تجربی، همان مقاله‌ها اما این بار با برچسب‌های موضوعی است و برای ارزیابی تأثیر برچسب‌گذاری پیکره متنی به عنوان متغیر مستقل بر میزان دقت مقالات علمی به عنوان متغیر وابسته، تحت آزمون ثانویه برای ارزیابی نتایج مرحله دوم بازیابی قرار گرفتند.

1. Experimental Group
2. Control Group

داده‌های زبانی نمونه پژوهش از وبگاه‌های مجله‌های علمی کشور جمع‌آوری و در پیکره گنجانده شد. در نهایت ارزیابی و مقایسه نتایج حاصل از آزمون دو گروه گواه و تجربی با استفاده از روش کمی پیمایشی انجام شد و داده‌ها مورد تحلیل و تفسیر قرار گرفته، یافته‌ها آشکار شد.

به منظور استفاده از پیکره در این پژوهش کاربردشناسی، باید اطلاعات کاربردشناختی لازم به آن‌ها افزوده شود و بر روی این پیکره‌ها، حاشیه‌نویسی کاربردشناختی[1] انجام گیرد. پیکره‌هایی که به صورت دستی برچسب‌گذاری می‌شوند می‌توانند منبع ارزشمندی برای پژوهش‌های پیکره‌مدار باشند (استفانویچ[2]، ۲۰۰۶). نقش معنایی که در برچسب‌گذاری پیکره این پژوهش استفاده می‌شود، شامل عناوین ۶ رشته علمی است که به هم‌نویسه‌های تخصصی نمونه پژوهش در مجموعه آموزش و آزمون تعلق می‌گیرد و وابستگی علمی آن هم‌نویسه را نسبت به یک رشته خاص نشان می‌دهد.

به منظور تهیه پیکره، تمام اسناد نوشتاری مقاله‌ها در قالب Microsoft Word، PDF و ... به قالب txt تبدیل می‌شوند؛ هم‌نویسه‌های آن‌ها شناسایی و برچسب‌گذاری شده و فایل نهایی با کدگذاری یونیکد[3] (سازگار با زبان فارسی) به عنوان پیکره (مجموعه آزمون) در مسیر داده‌های سامانه مورد استفاده قرار می‌گیرد.

۳- در سومین مرحله از رفع ابهام معنایی واژه‌ها یعنی در مرحله ارزیابی، پاسخ‌های حاصل از مرحله آزمون در یک چارچوب نظارتی یا بدون نظارت ارزیابی می‌شوند.

به منظور پاسخ به پرسش پژوهش قضاوت ربط نتایج بازیابی‌شده توسط شش متخصص موضوعی از شش رشته علمی موردنظر به صورت دودویی انجام گرفت و سپس توسط پژوهشگر صحت‌سنجی شد. برای هر رشته علمی حداقل ۵ هم‌نویسه‌تخصصی و در مجموع ۴۶ واژه تخصصی در قالب ۱۶ هم‌نویسه تخصصی به عنوان کلیدواژه‌های جستجوی متون علمی در شش رشته مورد نظر، تعیین شدند. این تعداد هم‌نویسه از میان ۸۹۹ هم‌نویسه تک‌واژه‌ای، گزینش شدند که از مقایسه و تطبیق اصطلاحنامه‌های شش رشته موردنظر، به دست آمد.

مبنای جستجو برای هر رشته، تعداد ۵ هم‌نویسه بود، اما به دلیل خاصیت هم‌نویسگی و نوعی تکرار در ظاهر واژه، از ۵ تا ۱۲ واژه در بین رشته‌های مختلف متغیر بود و در مجموع ۴۶ کلیدواژه جستجو را تشکیل دادند.

1. Pragmatic Annotation
2. Stefanowitsch
3. Unicode

جدول ۱ هم‌نویسه‌های مورد نظر پژوهش را به همراه رشته‌های دربردارنده آن نشان می‌دهد.

جدول ۱. هم‌نویسه‌های انتخابی پژوهش

جرم	صفحه	پیچش	پروانه	آرایه	چشمه	برگ	حلال
فیزیک شیمی علوم زمین جامعه‌شناسی	ریاضی شیمی علوم زمین جامعه‌شناسی	ریاضی فیزیک علوم زمین علوم زیستی	ریاضی فیزیک علوم زیستی	ریاضی فیزیک علوم زمین	فیزیک شیمی علوم زمین	ریاضی علوم زمین علوم زیستی	ریاضی شیمی علوم زمین
درخت	**قطر**	**بازیابی**	**تورم**	**تقلب**	**دوران**	**یال**	**پلاسما**
ریاضی علوم زمین علوم زیستی	ریاضی علوم زمین جامعه‌شناسی	فیزیک شیمی علوم زمین	فیزیک جامعه‌شناسی	فیزیک جامعه‌شناسی	ریاضی علوم زمین	ریاضی علوم زمین	فیزیک علوم زیستی

جدول ۲ تعداد هم‌نویسه‌های تخصصی منتخب پژوهش را به تفکیک رشته‌های علمی نشان می‌دهد.

جدول ۲. تعداد هم‌نویسه‌های تخصصی منتخب پژوهش در شش رشته علمی

رشته‌ها	علوم زمین		ریاضیات		فیزیک		شیمی		علوم زیستی		جامعه‌شناسی		جمع	
	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد
مقادیر	۱۲	۲۶/۱	۱۰	۲۱/۷	۹	۱۹/۵	۵	۱۰/۹	۵	۱۰/۹	۵	۱۰/۹	۴۶	۱۰۰

همان‌طور که جدول ۲ نشان می‌دهد، مجموع واژگان تخصصی انتخاب شده در شش رشته علمی مورد نظر، ۴۶ واژه است؛ اما به دلیل همپوشانی و به عبارت بهتر هم‌نویسگی آن‌ها در دو رشته، سه رشته یا چهار رشته، همچنان که جدول ۱ نشان می‌دهد، ۱۶ واژه منحصر به فرد محسوب می‌شوند.

یافته‌های پژوهش

پاسخ به پرسش پژوهش شامل یک مرحله ارزیابی سه بخشی است و توسط پژوهشگر انجام شد؛
در بخش اول و دوم ارزیابی، دقت نتایج بازیابی هم‌نویسه‌های تخصصی رشته‌های علمی موردنظر در دو گروه گواه و تجربی، محاسبه شد.

مانعیت (دقت) عبارت است از؛ تعداد کل مقالات مرتبط بازیابی شده تقسیم بر کل مقالات بازیابی شده. این معیار، میزان مرتبط بودن مدارک بازیابی شده با مورد جستجو را محاسبه می‌کند.

رابطه ۱)

$$\text{دقت} = \frac{\text{تعداد مدارک بازیابی شده مرتبط}}{\text{کل مدارک بازیابی شده}}$$

زمانی که فقط اسناد مرتبط بازیابی شوند، دقت به مقدار حداکثری خود یعنی عدد ۱,۰ می‌رسد.

جداول ۳ تا ۸ میزان دقت نتایج بازیابی هم‌نویسه‌های موردنظر پژوهش را به تفکیک شش رشته علمی در نظام بازیابی اولیه و نظام بازیابی پیشنهادی مورد مقایسه قرار می‌دهد.

جدول ۳ میزان دقت نتایج بازیابی هم‌نویسه‌های موردنظر رشته زیست‌شناسی را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ۳. میزان دقت نتایج بازیابی شده در نمونه پژوهش رشته زیست‌شناسی در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه				نظام بازیابی پیشنهادی			
	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت
پیچش	۲۱	۵	۱۶	۰/۲۴	۵	۵	۰	۱
پروانه	۱۵	۵	۱۰	۰/۳۳	۵	۵	۰	۱
برگ	۴۴	۱۶	۲۸	۰/۳۶	۱۶	۱۶	۰	۱
درخت	۲۰	۸	۱۲	۰/۴	۸	۸	۰	۱
پلاسما	۱۲	۵	۷	۰/۴۲	۵	۵	۰	۱
مجموع	۱۱۲	۳۹	۷۳	۰/۳۵	۳۹	۳۹	۰	۱

همچنان که جدول ۳ نشان می‌دهد میانگین دقت بازیابی ۵ هم‌نویسه موردنظر در متون زیست‌شناسی از ۰/۳۵ در نظام بازیابی اولیه به حد بیشینه آن یعنی ۱ در نظام پیشنهادی افزایش یافته است.

جدول ۴ میزان دقت نتایج بازیابی هم‌نویسه‌های موردنظر رشته ریاضیات را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ٤. میزان دقت نتایج بازیابی شده در نمونه پژوهش رشته ریاضیات در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه				نظام بازیابی پیشنهادی			
	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت
صفحه	۲۶	۷	۱۹	۰/۲۷	۷	۷	۰	۱
پیچش	۲۱	۵	۱۶	۰/۲۴	۵	۵	۰	۱
پروانه	۱۵	۵	۱۰	۰/۳۳	۵	۵	۰	۱
آرایه	۱۵	۵	۱۰	۰/۳۳	۵	۵	۰	۱
برگ	۴۴	۵	۳۹	۰/۱۱	۵	۵	۰	۱
حلال	۲۹	۵	۲۴	۰/۱۷	۵	۵	۰	۱
درخت	۲۰	۶	۱۴	۰/۳	۶	۶	۰	۱
قطر	۳۵	۶	۲۹	۰/۱۷	۶	۶	۰	۱
دوران	۱۵	۶	۹	۰/۴	۶	۶	۰	۱
یال	۳۴	۶	۲۸	۰/۱۸	۶	۶	۰	۱
مجموع	۲۵۴	۵۶	۱۹۸	۰/۲۲	۵۶	۵۶	۰	۱

همچنان که جدول ۴ نشان می‌دهد میانگین دقت بازیابی ۱۰ هم‌نویسه مورد نظر در متون ریاضیات از ۰/۲۲ در نظام بازیابی اولیه به حداکثر مقدار آن یعنی ۱ در نظام پیشنهادی افزایش یافته است.

جدول ۵ میزان دقت نتایج بازیابی هم‌نویسه‌های موردنظر رشته علوم زمین را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ٥. میزان دقت نتایج بازیابی شده در نمونه پژوهش رشته علوم زمین در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه				نظام بازیابی پیشنهادی			
	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت
جرم	۳۳	۷	۲۶	۰/۲۱	۷	۷	۰	۱
صفحه	۲۶	۶	۲۰	۰/۲۳	۶	۶	۰	۱
پیچش	۲۱	۵	۱۶	۰/۲۴	۵	۵	۰	۱
آرایه	۱۵	۵	۱۰	۰/۳۳	۵	۵	۰	۱

چشمه	۲۳	۹	۱۴	۰/۳۹	۹	۹	۰	۱
برگ	۴۴	۵	۳۹	۰/۱۱	۵	۵	۰	۱
حلال	۲۹	۵	۲۴	۰/۱۷	۵	۵	۰	۱
درخت	۲۰	۵	۱۵	۰/۲۵	۵	۵	۰	۱
قطر	۳۵	۶	۲۹	۰/۱۷	۶	۶	۰	۱
بازیابی	۲۰	۶	۱۴	۰/۳	۶	۶	۰	۱
دوران	۱۵	۵	۱۰	۰/۳۳	۵	۵	۰	۱
یال	۳۴	۱۵	۱۹	۰/۴۴	۱۵	۱۵	۰	۱
مجموع	۳۱۵	۷۹	۲۳۶	۰/۲۵	۷۹	۷۹	۰	۱

جدول ۵ نشان می‌دهد میانگین دقت بازیابی ۱۲ هم‌نویسه موردنظر در متون علوم زمین از ۰/۲۵ در نظام بازیابی اولیه به حداکثر مقدار آن یعنی ۱ در نظام پیشنهادی افزایش یافته است.

جدول ۶ میزان دقت نتایج بازیابی هم‌نویسه‌های مورد نظر رشته جامعه‌شناسی را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ٦. میزان دقت نتایج بازیابی شده در نمونه پژوهش رشته جامعه‌شناسی در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	**نظام بازیابی اولیه**				**نظام بازیابی پیشنهادی**			
	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت
جرم	۳۳	۱۰	۲۳	۰/۳	۱۰	۱۰	۰	۱
صفحه	۲۶	۵	۲۱	۰/۱۹	۵	۵	۰	۱
قطر	۳۵	۵	۳۰	۰/۱۴	۵	۵	۰	۱
تورم	۱۳	۶	۷	۰/۴۶	۶	۶	۰	۱
تقلب	۱۰	۵	۵	۰/۵	۵	۵	۰	۱
مجموع	۱۱۷	۳۱	۸۶	۰/۲۶	۳۱	۳۱	۰	۱

همان‌طور که جدول ۶ نشان می‌دهد میانگین دقت بازیابی ۵ هم‌نویسه مورد نظر در متون جامعه‌شناسی از ۰/۲۶ در نظام بازیابی اولیه به حداکثر مقدار آن یعنی ۱ در نظام پیشنهادی افزایش یافته است.

جدول ۷ میزان دقت نتایج بازیابی هم‌نویسه‌های موردنظر رشته شیمی را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ۷. میزان دقت نتایج بازیابی شده در نمونه پژوهش رشته شیمی در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه		نظام بازیابی اولیه			نظام بازیابی پیشنهادی			
	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت
جرم	۳۳	۵	۲۸	۰/۱۵	۵	۵	۰	۱
صفحه	۲۶	۵	۲۱	۰/۱۹	۵	۵	۰	۱
چشمه	۲۳	۵	۱۸	۰/۲۲	۵	۵	۰	۱
حلال	۲۹	۱۱	۱۸	۰/۳۸	۱۱	۱۱	۰	۱
بازیابی	۲۰	۸	۱۲	۰/۴	۸	۸	۰	۱
مجموع	۱۳۱	۳۴	۹۷	۰/۲۶	۳۴	۳۴	۰	۱

همچنان که از داده‌های جدول ۷ برمی‌آید میانگین دقت بازیابی ۵ هم‌نویسه مورد نظر در متون شیمی از ۰/۲۶ در نظام بازیابی اولیه به حداکثر مقدار آن یعنی ۱ در نظام پیشنهادی افزایش یافته است.

جدول ۸ میزان دقت نتایج بازیابی هم‌نویسه‌های مورد نظر رشته فیزیک را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ۸. میزان دقت نتایج بازیابی شده در نمونه پژوهش رشته فیزیک در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه				نظام بازیابی پیشنهادی			
	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت	کل مدارک بازیابی شده	تعداد مدارک بازیابی شده مرتبط	تعداد مدارک بازیابی شده نامرتبط	میزان دقت
جرم	۳۳	۷	۲۶	۰/۲۱	۷	۷	۰	۱
پیچش	۲۱	۶	۱۵	۰/۲۹	۶	۶	۰	۱
پروانه	۱۵	۵	۱۰	۰/۳۳	۵	۵	۰	۱
آرایه	۱۵	۵	۱۰	۰/۳۳	۵	۵	۰	۱
چشمه	۲۳	۶	۱۷	۰/۲۶	۶	۶	۰	۱

بازیابی	۲۰	۵	۱۵	۰/۲۵	۵	۵	۰	۱
تورم	۱۳	۵	۸	۰/۳۸	۵	۵	۰	۱
تقلب	۱۰	۵	۵	۰/۵	۵	۵	۰	۱
پلاسما	۱۲	۶	۶	۰/۵	۶	۶	۰	۱
مجموع	۱۶۲	۵۰	۱۱۲	۰/۳۱	۵۰	۵۰	۰	۱

نتایج جدول ۸ نشان می‌دهد که میانگین دقت بازیابی ۹ هم‌نویسه مورد نظر در متون فیزیک از ۰/۳۱ در نظام بازیابی اولیه به حداکثر مقدار آن یعنی ۱ در نظام پیشنهادی افزایش یافته است.

در بخش سوم ارزیابی، نتیجه حاصل از میانگین دو بخش پیشین ارزیابی، مقایسه و به پرسش پژوهش پاسخ داده شد. برای پاسخ به پرسش پژوهش از آزمون‌های آماری مقایسه‌ای استفاده شده است. از آنجا که در داده‌های حاصل از به کارگیری نظام بازیابی پیشنهادی واریانس وجود ندارد و بنابراین داده‌ها دارای توزیع نرمال نیستند، برای مقایسه نتایج حاصل از به کارگیری نظام بازیابی اولیه و نظام بازیابی پیشنهادی، نمی‌توان از آزمون t جفتی استفاده کرد لذا برای مقایسه از آزمون ناپارامتریک رتبه‌های علامت‌دار ویلکاکسون[1] استفاده شده است. این آزمون به جای مقایسه میانگین مقادیر، با محاسبه رتبه‌های مقادیر، بر اساس تعداد رتبه‌های منفی و مثبت در مقایسه‌های زوجی، نتیجه به کارگیری دو نظام بازیابی را مورد بررسی قرار می‌دهد.

برای مقایسه دقت نتایج بازیابی هم‌نویسه‌های تخصصی در رشته‌های علمی موردنظر، نتایج آزمون ویلکاکسون در به کارگیری دو نظام بازیابی اولیه و پیشنهادی در جدول ۹ ارائه شده است.

جدول ۹. رتبه‌ها: مقایسه دقت نتایج دو نظام بازیابی

		تعداد	میانگین رتبه‌ها	جمع رتبه‌ها
نظام بازیابی پیشنهادی - نظام بازیابی اولیه	**رتبه‌های منفی***	۰	۰/۰۰	۰/۰۰
	رتبه‌های مثبت**	۴۶	۲۳/۵۰	۱۰۸۱/۰۰
	گره‌ها***	۰		
	جمع	۴۶		
سطح معنی‌داری، ۰/۰۰۰۱ = Z = −۵/۹۰۹				
* نظام بازیابی پیشنهادی > نظام بازیابی اولیه				
** نظام بازیابی پیشنهادی < نظام بازیابی اولیه				
*** نظام بازیابی پیشنهادی = نظام بازیابی اولیه				

1. Wilcoxon Signed Ranks Test

سطح معنی‌داری آزمون رتبه‌های علامت‌دار ویلکاکسون (P = ۰/۰۰۰۱، Z = -۵/۹۰۹) در جدول ۹ نشان می‌دهد که میزان دقت نتایج بازیابی هم‌نویسه‌های تخصصی بعد از به کارگیری پیکره تخصصی برچسب‌گذاری‌شده در نظام بازیابی اطلاعات نسبت به قبل از آن تفاوت معنی‌داری دارد. بررسی رتبه‌های منفی و مثبت نشان می‌دهد میزان دقت نتایج بعد از به کارگیری پیکره تخصصی برچسب‌گذاری شده به میزان معنی‌داری افزایش یافته است.

نتیجه‌گیری

در امر بازیابی اطلاعات، نگاه علمی به پدیده‌های زبانی و بررسی آن‌ها بر پایه یافته‌های معتبر، راهگشای چالش‌های پردازش زبان طبیعی است. مطالعه پژوهش‌های داخلی و خارجی نشان می‌دهد به طور معمول تکیه نظام‌های بازیابی بر شکل واژه ورودی است. در این صورت نتیجه جستجوی کلیدواژه‌ها به بازیابی نتایج نامرتبط و کاهش دقت منجر خواهد شد. از آنجا که بیشتر پژوهش‌های رفع ابهام معنایی به زبان‌های غیرفارسی و عموماً به زبان انگلیسی معطوف می‌شود، پرداختن به ابهام معنایی هم‌نویسه‌ها در زبان فارسی فوریت و اهمیت بیشتری می‌یابد و به ویژه اگر مسئله هم‌نویسه‌های تخصصی مطرح باشد، به سبب حساسیت موضوع، اهمیت آن دوچندان می‌شود. در این پژوهش استفاده از پیکره متنی و دستاوردهای زبان‌شناسی پیکره‌ای، به عنوان راهکاری مناسب برای مطالعه بر روی داده‌های واقعی و رفع ابهام معنایی از هم‌نویسه‌های تخصصی شناسایی شد. با توجه به اهمیت رفع ابهام معنایی از هم‌نویسه‌های تخصصی در متون علمی و در راستای دستیابی به هدف پژوهش، عملکرد نظام بازیابی متون علمی در دو وضعیت برچسب‌گذاری شده و برچسب‌گذاری نشده بررسی شد و میزان دقت بازیابی در این دو حالت محاسبه شده و مورد مقایسه قرار گرفت. ارزیابی میزان دقت بازیابی متون علمی شش رشته موردنظر نشان داد که میزان دقت در نظام بازیابی پیشنهادی به میزان معناداری بهبود یافته و به حد بیشینه ان یعنی ۱ رسیده است. به عبارت دیگر برچسب‌گذاری معنایی پیکره متون تخصصی موجب بهبود بازیابی و افزایش دسترس‌پذیری متون حاوی هم نویسه‌های تخصصی می‌شود. بر اساس روش مورد استفاده در پژوهش حاضر و نتایج به دست آمده از آزمون فرضیه می‌توان نتیجه گرفت که امکان تعمیم یافته‌ها در مورد رفع ابهام معنایی هم‌نویسه‌های تخصصی برای متون همه رشته‌های علمی وجود دارد. مقایسه یافته‌های پژوهش حاضر با پژوهش‌های پیشین نشان می‌دهد که روش مورد استفاده در این پژوهش، گامی رو به جلو در پیدانمایی و دسترس‌پذیری بیشتر متون حاوی هم نویسه‌های تخصصی خواهد بود و با توجه به تکیه نظام‌های بازیابی اطلاعات بر شکل ظاهری کلیدواژه‌های ورودی، این دستاورد حائز اهمیت است. بنابراین طراحی نظام بازیابی تمام متن از مقالات علمی با قابلیت رفع

ابهام معنایی هم‌نویسه‌های تخصصی با رویکرد پیکره‌مدار شایان توجه است. همچنین رویکرد پیکره‌مدار نظام بازیابی اطلاعات، فرصت‌های بسیاری را برای مطالعه سایر چالش‌های پردازش زبان طبیعی بر روی داده‌های عینی و واقعی فراهم می‌نماید.

پیشنهادها

با توجه به این که فرایندها، فنون و ابزارهای به کار رفته در این روش به متون علمی رشته‌ی خاصی محدود نیست، ارزیابی الگوی پیشنهادی برای متون علمی همه رشته‌ها امکان‌پذیر است. پژوهش در زمینه موضوع‌های زیر به عنوان پژوهش‌های آینده مطلوب خواهد بود؛

- روش پیشنهادی برای متون علمی سایر زبان‌ها از جمله زبان انگلیسی بررسی و امکان‌سنجی شود.
- در مرحله آزمون، به جای استفاده از پیکره خام از دیگر رویکردهای رفع ابهام معنایی به ویژه رویکرد با ناظر برای رفع ابهام معنایی از هم‌نویسه‌ها استفاده شود و سپس نتایج آن با نتایج بازیابی در پیکره برچسب‌گذاری شده، مقایسه و معناداری اختلاف نتایج بازیابی این دو روش، بررسی شود.
- اگرچه این پژوهش به رشته‌های علمی معینی محدود شده است، اما نتایج رویکردهای به کار گرفته شده آن، می‌تواند نقشه راهی برای بازیابی هم‌نویسه‌های تخصصی در سایر رشته‌های علمی باشد و در ایجاد نظام کارآمد بازیابی متون و مدارک علمی، راهگشا باشد. به این منظور تهیه پیکره بزرگ متون دانشگاهی پیشنهاد می‌شود. نهادهایی همچون فرهنگستان زبان و ادب فارسی، پژوهشگاه علوم و فناوری اطلاعات ایران (ایران‌داک)، پایگاه اطلاعات علمی جهاد دانشگاهی، بانک اطلاعات نشریات کشور، کتابخانه‌های دانشگاهی و عمومی بزرگ و بسیاری از نهادهای دیگر می‌توانند با ایجاد هماهنگی لازم، زمینه ایجاد پیکره و تحقیقات وسیع در زمینه واژگان و متون علمی را فراهم کنند و پیکره متون فارسی دانشگاهی را ایجاد کنند.

انجام چنین پژوهش‌هایی نه تنها منجر به مقایسه و بازآزمون روش پیشنهادی این پژوهش می‌شود، همچنین زمینه غنای ادبیات نظری و پژوهش‌های عملیاتی بیشتر در این حوزه را با رویکرد بین‌رشته‌ای فراهم می‌کند.

سپاسگزاری

نگارندگان بر خود لازم می‌دانند از معاونت محترم پژوهشی دانشگاه الزهرا (س) به خاطر حمایت معنوی در اجرای پژوهش و همچنین داوران محترم جهت ارائه نظرات ارزشمند، سپاسگزاری نمایند.

منابع

آزاد، احسان (۱۳۸۶). *طراحی و پیاده‌سازی یک سیستم بازیابی اطلاعات متنی جدید برای زبان فارسی*. پایان‌نامه کارشناسی ارشد. گروه مهندسی کامپیوتر- هوش مصنوعی. دانشکده فنی-مهندسی. دانشگاه شیراز، شیراز.

افراشی، آزیتا؛ عاصی، مصطفی؛ جولایی، کامیار (۱۳۹۵). استعاره‌های مفهومی در زبان فارسی؛ تحلیلی شناختی و پیکره‌مدار. *زبان‌شناخت*، ۶ (۱۲)، ۳۹-۶۱.

اکبری، اسماعیل؛ حسینی بهشتی، ملوک‌السادات؛ نوروزی‌اقبالی، مهرداد (۱۳۸۴). *اصطلاح‌نامه علوم زیستی*. تهران: مرکز اطلاعات و مدارک علمی ایران.

بوث، باربارا؛ بلر، میشل (۱۳۸۲). *اصطلاح‌نامه جامعه‌شناسی*. ترجمه مهوش معترف. تهران: مرکز اطلاعات و مدارک علمی ایران.

حسن‌زاده، شیرین (۱۳۸۹). یک سیستم بازیابی اطلاعات متنی برای زبان فارسی. *چهارمین کنفرانس داده‌کاوی ایران. تهران*، دانشگاه صنعتی شریف.

حسینی‌بهشتی، ملوک‌السادات؛ وفایی، سعیده؛ نوروزی‌اقبالی، مهرداد (۱۳۹۳). *اصطلاح‌نامه ریاضیات*. تهران: مرکز اطلاعات و مدارک علمی ایران.

دستغیب، محمدباقر (۱۳۹۷). تولید پیکره متنی برای زبان فارسی با استفاده از راه‌حل‌های مبتنی بر دانش. *چهارمین کنفرانس ملی دستاوردهای نوین در برق و کامپیوتر و صنایع. اسفراین*، مجتمع آموزش عالی فنی و مهندسی اسفراین.

دلیخون، لیلا (۱۳۹۵). *بررسی راه‌های گسترش پرسش کاربران در موتورهای جستجو و پایگاه داده‌های تخصصی: مطالعه موردی دانشجویان کارشناسی ارشد فنی و مهندسی دانشگاه الزهرا (س)*. پایان‌نامه کارشناسی ارشد. دانشکده علوم تربیتی و روانشناسی. دانشگاه الزهرا، تهران.

رجبی، تقی؛ غریبی، حسین؛ حسینی‌بهشتی، ملوک‌السادات؛ نوروزی‌اقبالی، مهرداد (۱۳۸۳) *اصطلاح‌نامه شیمی*. تهران: مرکز اطلاعات و مدارک علمی ایران.

زرداری، سولماز (۱۳۹۵). *مهندسی هستی‌نگاری علم اطلاعات و دانش‌شناسی بر اساس دائرةالمعارف کتابداری و اطلاع‌رسانی*. پایان‌نامه دکتری، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه شهید چمران اهواز، اهواز.

ستوده، هاجر؛ هنرجویان، زهره (۱۳۹۱). مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آن‌ها بر اثر بخشی پردازش خودکار متن و بازیابی اطلاعات. *کتابداری و اطلاع‌رسانی*، ۱۵(۴)، ۵۹-۹۲.

ستوده، هاجر؛ هوشیار، مژگان (۱۳۹۷). بررسی نقش انواع بافتار هم نویسه‌ها در تعیین شباهت بین مدارک. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۳(۳)، ۱۱۸۳-۱۲۰۶.

شاپوری، سودابه (۱۳۷۹). مشکلات جستجوی موضوعی استفاده‌کنندگان از فهرست رایانه‌ای کتابخانه مرکزی دانشگاه فردوسی مشهد. *کتابداری و اطلاع‌رسانی*، ۳(۲)، ۴۹-۶۸أ.

شهبازی، مهری و شاهینی، شبنم (۱۳۹۴). بررسی میزان کارآیی پایگاه‌های اطلاعاتی مگ ایران، نورمگز و اس.آی.دی. در بازیابی و ربط مباحث علم اطلاعات و دانش‌شناسی با استفاده از کلیدواژه‌های آزاد و مقایسه آن‌ها از نظر میزان استفاده از کلیدواژه‌های مهارشده. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۱(۲)،۴۳۱-۴۵۴.

صدیقی، مهری؛ حسینی‌بهشتی، ملوک‌السادات؛ نوروزی‌اقبالی، مهرداد (۱۳۸۴). *اصطلاح‌نامه علوم زمین*. تهران: مرکز اطلاعات و مدارک علمی ایران.

صفری، سعید (۱۳۹۱). *طراحی و ایجاد پیکره‌ی تولیدی زبان‌آموز فارسی*. پایان‌نامه کارشناسی ارشد. دانشکده ادبیات و علوم‌انسانی. دانشگاه علامه طباطبائی، تهران.

طباطبایی جعفری، زهرا (۱۳۹۰). *بررسی شیوه‌های بسط پرسش در رفتار جستجوی اطلاعاتی کاربران در موتورهای جستجو: مطالعه در میان دانشجویان تحصیلات تکمیلی علوم کتابداری و اطلاع‌رسانی دانشگاه‌های سراسری شهر تهران*. پایان‌نامه کارشناسی ارشد. دانشکده ادبیات و علوم انسانی. دانشگاه قم، قم.

عبدالهی نورعلی، محمدصادق (۱۳۸۶). *کندوکاو مسائل ریخت شناسی زبان فارسی در بازیابی اطلاعات از جستجوگرهای وب*. پایان‌نامه کارشناسی ارشد، گروه علم اطلاعات و دانش‌شناسی. دانشکده علوم تربیتی و روان‌شناسی. دانشگاه شیراز، شیراز.

عبدالهی نورعلی، محمدصادق؛ جوکار، عبدالرسول (۱۳۸۸). چالش‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب. *مطالعات تربیتی و روانشناسی دانشگاه فردوسی مشهد*. ۱۰(۲)، ۶۷-۹۰.

قیومی، مسعود (۱۳۹۸). تعیین خودکار معنای واژه‌های فارسی با استفاده از تعبیه معنایی واژه. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۵(۱)، ۲۵-۵۰.

کامیابی‌گل، عطیه؛ اخلاقی باقوجری، الهام؛ عسگریان، احسان؛ حبیبی، هانیه (۱۳۹۷). استخراج اطلاعات از پیکره زبانی؛ معرفی پیکره مقاله‌های علمی پژوهشی دانشگاه فردوسی مشهد. *کتابداری و اطلاع‌رسانی*. ۲۱(۲)، ۳-۲۵.

گل تاجی، مرضیه؛ بذرگر، سعیده (۱۳۸۹). بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی. *کتابداری و اطلاع‌رسانی*، ۱۳(۲)، ۱۹۱-۲۱۴.

مرتضایی، لیلا (۱۳۸۱). مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. *فصلنامه اطلاع‌رسانی*، ۱۷(۱-۲)، ۱-۷.

مسعودی، بابک؛ راحتی قوچانی، سعید (۱۳۹۴). رفع ابهام معنایی واژگان مبهم فارسی با مدل موضوعی LDA. *پردازش علائم و داده‌ها*، ۴(۲۶)، ۱۱۷-۱۲۵.

مظفری، زهرا؛ تاکی، گیتی؛ صباغ جعفری، مجتبی؛ یوسفان، پاکزاد (۱۳۹۷). سامانه رفع ابهام معنایی از حروف اضافه در زبان فارسی با استفاده از قالب‌های معنایی. *پژوهش‌های زبانی*، ۹(۱)، ۹۹-۱۱۷.

معروفی، افسانه؛ پیله‌ور، عبدالحمید (۱۳۹۲). رفع ابهام از معنی کلمه مبهم فارسی با استفاده از روش‌ها مبتنی بر پیکره و قاموس. *اولین همایش منطقه‌ای رویکردهای نوین در مهندسی کامپیوتر و فناوری اطلاعات*. رودسر، دانشگاه آزاد اسلامی.

مهرنهاد، زینب؛ قاسم‌زاده، محمد؛ نظارات، امین (۱۳۹۶). به کارگیری پیکره‌های زبانی در طراحی یک سامانه بازیابی اطلاعات دوزبانه. *دومین کنفرانس ملی محاسبات نرم*. دانشکده فنی و مهندسی شرق گیلان، دانشگاه گیلان.

نوروزی، یعقوب؛ هماوندی، هدی (۱۳۹۴). بررسی مشکلات جستجو و بازیابی تصاویر در موتورهای کاوش برگزیده مبتنی بر ویژگی‌های نگارشی زبان فارسی. *کتابداری و اطلاع‌رسانی*، ۵(۲)، ۲۰۶-۲۲۲.

نوروزی اقبالی، مریم؛ حسینی بهشتی، ملوک‌السادات؛ نوروزی اقبالی، مهرداد (۱۳۸۵). *اصطلاح‌نامه فیزیک*. تهران: مرکز اطلاعات و مدارک علمی ایران.

هوشیار، مژگان (۱۳۹۴). *مقایسه قدرت انواع بافتار متن در ابهام‌زدایی معنایی از هم‌نویسه‌های انگلیسی*. پایان‌نامه کارشناسی ارشد. گروه علم اطلاعات و دانش‌شناسی. دانشکده علوم تربیتی و روانشناسی. دانشگاه شیراز، شیراز.

یوسفان نجف‌آبادی، احمد (۱۳۸۲). *یک سیستم بازیابی اطلاعات متنی برای زبان فارسی بر پایه نمایه‌گذاری معانی پنهان*. پایان‌نامه کارشناسی ارشد، مهندسی کامپیوتر، دانشگاه شیراز، شیراز.

یوسفی‌راد، ابراهیم (۱۳۸۸). آر. دی. اف: الگویی برای توصیف منابع در وب معنایی. *فصلنامه کتاب*. ۷۹، ۹– ۲۲.

References

Abdollahi NorAli, M. S. (2007). *Survey on morphological difficulties of Persian language in information retrieval from web search tools*. M.S thesis, Library and Information science. Faculty of Education and Psychology, Shiraz University, Shiraz. (in Persian)

Abdollahi NorAli, M. S., & Jokar, A. (2009). Survey on morphological difficulties of Persian language in information retrieval from Web Search Engines. *Educational and Psychological Studies*, 10(2), 67-90. (in Persian)

Afrashi, A., Asi, S. M., & Joulaei, K. (2016). Conceptual metaphors in Persian: A cognitive perspective and a corpus driven analysis. *Zabanshenakht (Language studies)*, 6(2), 39-61. (in Persian)

Akbari, E., Hosseini Beheshti, M., & Noroozi Eghbali, M. (2005). *Thesaurus of Biological Science*. Tehran, Iranian Research Institute for Information Science and Technology. (in Persian)

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1-16.

Azad, E. (2007). *Design and implementation of a modern information retrieval system for Farsi language*. M. S thesis. Computer engineering artificial intelligence. Faculty of Engineering, Shiraz University, Shiraz. (in Persian)

Barba, E., Procopio, L., Campolungo, N., Pasini, T., & Navigli, R. (2020). MuLaN*:* Multilingual label propagation for word sense disambiguation. *IJCAI*, 3837-3844.

Booth, B., & Blair, M. (1992). *Thesaurus of sociological indexing terms*. Tehran, Iranian Research Institute for Information Science and Technology. (in Persian)

Bowker, L. (2018). Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research. *Library Hi Tech*, 36(4), 1-28. DOI: 10.1108/LHT-12-2017-0271

Dastghaib, M. B. (2018). Constructing Persian text corpus by using some knowledge-based approaches. *Fourth National Conference on New Achievements in Electrical and Computer and Industries*. Esfarayen, university of Technology. (in Persian)

Delikhoun, L. (2016). *A survey of query expansion (QE) of Users in Search Engines and Specialized Databases: A Case Study of Engineering Graduate Student at Alzahra University.* Thesis for Master, Knowledge and Information Science. Faculty of Education and Psychology, Alzahra University, Tehran. (in Persian)

Fangzhou L., Qin, S. & Tao, J. (2008). Tree-guided transformation-based homograph disambiguation in Mandarin TTS system. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA. 4657-4660

Gale, K. W., Church, W. A., & Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computer and the Humanities*, 26(5-6), 415-439.

Ghayoomi, M. (2019). Identifying Persian Words' Senses Automatically by Utilizing the Word Embedding Method. *Iranian Journal of Information Processing & Management*; 35(1), 25-50. (in Persian)

Goltaji, M., & Bazregar, S. (2010). Investigating the morphological problems of Persian language in three databases of ISC, Irandoc & Jahad Institute. *Library and Information Sciences*, 13(2), 191-214. (in Persian)

Hammo, B. H. (2009). Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. *Information Retrieval*, 12(3), 300-323.

Harada, T., & Tsuda, K. (2014). Classifying homographs in Japanese social media texts using a user Interest model. *Procedia Computer Science*, (35), 929-936.

Hasanzadeh, S. (2010). A text information retrieval system for Persian language. The *Fourth Iran Data Mining Conference/ IDMC 10*. Tehran, Sharif University of Technology. (in Persian)

Hearst, M. A. (1991). Noun homograph disambiguation using local context in large text corpora. *Proceedings of the 7st Annual conference of the University of Waterloo Centre for the new OED and text research*, Berkeley, 185-188.

Hoseini Beheshti, M. S., Vafaei, S., & Noroozi Eghbali, M. (2015). *Thesaurus of mathematic*. Tehran, Iranian Research Institute for Information Science and Technology. (in Persian)

Houshyar, M. (2016). *The comparison of powers of different kinds of text contexts in sense disambiguation of English homographs*. M. A. Thesis, Knowledge and Information Science, Information Management. Faculty of Educational Sciences and Psychology, Shiraz University, Shiraz. (in Persian)

Jones, C., & Waller, D. (2015). *Corpus linguistics for grammar, A guide for research*. London, Routledge.

Kamyabi Gol, A., Akhlaghi Baghujeri, E., Asgarian, E., & Habibi, H. (2018). Extracting information from language corpus: introducing the corpus of scientific articles of Ferdowsi University of Mashhad. *Library and Information Sciences*, 21(2), 3-25. (in Persian)

Khan, L., McLeod, D., & Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. The VLDB Journal- *The International Journal on Very Large Data Bases*, 13(1), 71-85.

Kumar, S., Jat, S., Saxena, K., & Talukdar, P. (2019). Zero-shot word sense disambiguation using sense definition embedding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Florence, Italy, 5670–5681.

Lazarinis, F. (2007). Evaluating the searching capabilities of e-commerce web sites in a non-English language: A Greek case study, *Online Information Review*, 31(6), 881-891.

Lewandowski, D. (2008). Problems with the use of web search engines to find results in foreign languages. *Online Information Review*, 32(4), 668-672.

Maaroufi, A., & Pilehvar, A. (2013). Word sense disambiguation of Persian words using methods based on corpus and dictionary. *First National Conference on Advances in computer science and information retrieval approaches*. Rudsar, Azad University, Guilan. (in Persian)

Mahmoud, H., Salah Kareem, S., & El-Shishtawy, T. (2018). A semantic retrieval system for extracting relationships from biological corpus. *International Journal of Computer Science & Information Technology (IJCSIT)*, 10(1), 43-53.

Masoudi, B., & Rahati Ghochani, S. (2016). Farsi word sense disambiguation with LDA topic model. *JSDP*. 12(4), 117-125. (in Persian)

Mehrnahad, Z., Ghasemzadeh, M. & Nazarat, A. (2017) .Using language corpuses in designing a bilingual information retrieval system. *2rd International Conference on Soft Computing*. Rudsar, Guilan University. (in Persian)

Menai, M. B. (2014). Word sense disambiguation using an evolutionary approach. *Informatica*, 38(3), 155-169.

Mortezai, L. (2001). Persian language and orthography for Information storage and retrieval. *Iranian Journal of Information Processing & Management*, 17(1-2), 9-26. (in Persian)

Mozaffari, Z., Taki, G., Sabbagh Jaffari, M., & Yusefian, P. (2018). Preposition sense disambiguation in Persian using semantic frames. *Language Research*, 9(1), 99-117. (in Persian)

Norouzi, Y., & Homavandi, H. (2015). Survey of Image Search and Retrieval Problems in Selected Search Engines based on the Persian Writing Styles. *Library and Information Sciences*, 5(2), 206-222. (in Persian)

Noroozi Eghbali, M., Hoseini Beheshti, M. S., & Noroozi Eghbali, M. (2007). *Thesaurus of physics*. Tehran, Iranian Research Institute for Information Science and Technology. (in Persian)

Pretschner, A., & Gauch, S. (1999). Ontology based personalized search. In Proceedings of the 11th IEEE, *International Conference on Tools with Artificial Intelligence*. Chicago, IL, USA.

Prokofyev, R., Demartini, G., Boyarsky, A., Ruchayskiy, O., & Cudré-Mauroux, P. (2013). Ontology-based word sense disambiguation for scientific literature. Advances in information retrieval: *35th European conference on IR research, ECIR*, Berlin, Germany: Springer. 594-605.

Rajabi, T., Gharibi, H., Hosseini Beheshti, M. S. & Noroozi Eghbali, M. (2004). *Thesaurus of Chemistry*. Tehran, Iranian Research Institute for Information Science and Technology. (in Persian)

Safari, S. (2011). *Designing and Developing a FFL Learner Corpus*. Thesis for Degree of Master of ART (M.A*)*, Department of English Language & Literature, Faculty of Literature, Allameh Tabatabaei University, Tehran. (in Persian)

Scarlini, B., Pasini, T., & Navigli, R. (2020). Sense-Annotated Corpora for Word Sense Disambiguation in Multiple Languages and Domains. *LREC*. Marseille, France, 5905-5911.

Schutze, H. (2014). *Introduction to Information Retrieval: Relevance Feedback and Query Expansion*. Retrieved from http://www.cis.uni-muenchen.de/~hs/teach/13s/ir/pdf/09expand.pdf

Sedighi, M., Hoseini Beheshti, M. S., & Noroozi Eghbali, M. (2004). *Thesaurus of geosciences*. Tehran, Iranian Research Institute for Information Science and Technology. (in Persian)

Semeval website (2020). *International workshop on semantic evaluation*. Retrieved 9 December from https://semeval.github.io/

Shahbazi, M. & Shahini, S. (2016). Study of the the efficacy Magiran, Noormags and SID database in retrieval and relevance of Information Science and Knowledge subject by free keywords and Compare them in terms of the use of controlled keywords. *Iranian Journal of Information Processing & Management*, 31(2), 431-454. (in Persian)

Shapoori, S. (2000). Problems of subject search for users of the computer catalog of the Central Library of Ferdowsi University of Mashhad. *Library and Information Sciences*, 3(2), 49-68. (in Persian)

Shen, B., Wu, Z., Wang, Y., & Cai, L. (2011). Combining Active and Semi-Supervised Learning for Homograph Disambiguation in Mandarin Text-to-Speech Synthesis. *12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 27-31.

Sotoudeh, H., & Honarjooyan, Z. (2012). A review of the difficulties of the Persian language in the digital environment and their effects on the effectiveness of automatic text processing and information retrieval. *Library and Information science*, 15(4), 59-92. (in Persian)

Sotoudeh, H. & Houshyar, M. (2018).The Role of Different Types of Homograph Contexts in Measuring Documents Similarities, *Iranian Journal of Information Processing & Management*, 33(3), 1195-1220. (in Persian)

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*. 52(3), 226-234.

Stefanowitsch, A. (2006). *Corpus-based approaches to metaphor and metonymy. In Corpus-Based Approaches to Metaphor and Metonymy*. Edited by Anatol Stefanowitsch and Stefan Th. Gries, Berlin, New York: De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110199895.1

Tabatabaie Jafari, Z. (2011). *A Survey to Query Expansion (QE) in Information Searching Behavior in Search Engines: A study of LIS graduate student Tehran states university.*

Thesis for degree of master of ART (MA), Library & Information science. Faculty of Humanities, University of Qom, Qom. (in Persian)

Vallet, D., Fernandez, M. & Castells, P. (2005). An ontology-based information retrieval model. *In European Semantic Web Conference*. Springer, Berlin, Heidelberg. 455-470.

Voorhees, E. M. (2002). The philosophy of information retrieval evaluation. *In CLEF '01*: *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, London, UK. Springer-Verlag, 355-370.

Yousefan Najafabadi, A. (2004). *A Farsi text Information retrieval system based on latent semantic indexing.* M. A. Thesis. School of Electrical and Computer Engineering, Shiraz University, Shiraz. (in Persian)

Yousefi Rad, E. (2009). R.D.F.: A model for resource description in semantic web. *National Studies on Librarianship and Information Organization*, 20(3), 9-22. (in Persian)

Zardary, S. (2016). *Ontology engineering of knowledge and information science based on Encyclopedia of Library and Information Science*. Ph. D. degree, Knowledge and Information Science Department. Faculty of Education and Psychology, Shahid Chamran University of Ahwaz, Ahwaz. (in Persian)

Zhang, J., & Lin, S. (2007), Multiple language supports in search engines, *Online Information Review*, 31(4), 516-532.