گسترش پرس‌وجو با سرپرستی ضعیف با استفاده از شبکه سیامی عمیق حافظه کوتاه-مدت طولانی

فاطمه کاوه یزدی[۱]، دانشجوی دکترا؛ علی‌محمّد زارع بیدکی[۲]، دانشیار

۱-گروه مهندسی کامپیوتر - دانشگاه یزد- یزد- ایران/ محقق ارشد در موتور جستجوی پارسی‌جو– fkavehy@stu.yazd.ac.ir
۲- گروه مهندسی کامپیوتر - دانشگاه یزد- یزد - ایران / مدیر عامل موتور جستجوی پارسی‌وجو– Alizareh@yazd.ac.ir

چکیده: عدم‌همخوانی واژگان مهمترین چالش پیش روی سیستم‌های بازیابی اطلاعات از وب هستند. عدم‌همخوانی واژگانی به تفاوت‌های موجود بین پرس‌وجوهای کاربران و محتوای اسناد وب در حالی اطلاق می‌گردد که هر دو به یک موضوع واحد اشاره دارند. روش‌های گسترش پرس‌وجو برای رویارویی با مشکل عدم‌همخوانی واژگانی، پرس‌وجوی کاربر را بازآرایی می‌نمایند تا بدینوسیله همپوشانی بین عبارت‌های موجود در پرس‌وجو و اسناد را افزایش دهند. در این مقاله یک چهارچوب گسترش پرس‌وجوی مبتنی بر شبکه سیامی عمیق حافظه کوتاه-مدت طولانی ارائه شده است. به علاوه، برای نخستین بار وابستگی ارتباطی در این مقاله تعریف شده و برای برچسب‌گذاری جفت‌های متشکل از پرس‌وجوی کاربر و پرس‌وجوی جایگزین مورد استفاده قرار گرفته است. شبکه سیامی آموزش داده شده با استفاده از جفت‌های برچسب‌گذاری شده با نظارت ضعیف، علاوه بر ارائه برچسب برای جفت‌های ورودی، هزینه هم‌سنجی آن‌ها را نیز محاسبه نموده و اعلام می‌کند. پس از برچسب‌گذاری، جفت‌های با کم‌ترین هزینه هم‌سنجی انتخاب و در هم ادغام می‌شوند تا به یک پرس‌وجوی گسترش یافته تبدیل شوند. نتایج آزمایشات نشان‌دهنده برتری روش پیشنهادی بر سایر روش‌های مشابه گسترش پرس‌وجوی مبتنی بر جاسازی کلمات بوده است.

واژه‌های کلیدی: بازیابی اطلاعات، گسترش پرس‌وجو، جاسازی کلمات، وابستگی معنایی، وابستگی ارتباطی، شبکه سیامی عمیق، سلول حافظه کوتاه-مدت طولانی.

Weakly Supervised Query Expansion using Deep Siamese LSTM

Fatemeh Kaveh-Yazdy[1], PhD Student; Ali-Mohammad Zareh-Bidoki[2], Associate Professor

1- Department of Computer Engineering, Yazd University, Yazd, Iran, Email: fkavehy@stu.yazd.ac.ir
Senior Researcher at Parsijoo Persian Search Engine, Email: fkavehy@parsijoo.ir
2- Department of Computer Engineering, Yazd University, Yazd, Iran, Email: alizareh@yazd.ac.ir
CEO of Parsijoo Persian Search Engine, Email: alizareh@parsijoo.ir

Abstract: Term mismatch is the most important challege in web information retrieval. The term mismatch problem is defined as differences between user queries and contents of documents while referring to the same topic. Query expansion methods deal with term mismatch by reformulating the queries to increase their term-overlap with relevant documents. In this paper, we proposed a query expansion framework based on a deep Siamese LSTM neural network. In addition, we defined the relevant relatedness for the first time and used this concept to label pairs made from user query and candidate query. Weakly-supervised labeled pairs are utilized in training of the deep Siamese network. The trained Siamese network provides labels for testset pairs in addition to contrastive loss values. The contrastive loss value reflects the cost of pulling together similar pairs. Pairs with minimum contrastive loss values are selected and merged together to form one expanded query. Results of our tests showed that the proposed framework outperforms similar word embedding based query expansion methods.

Keywords: Information Retrieval, Query Expansion, Word Embedding, Semantic Relatedness, Relevant Relatedness, Deep Siamese Network, LSTM cell.

تاریخ ارسال مقاله: ۱۳۹۷/۳/۱۳
تاریخ اصلاح مقاله: ۱۳۹۷/۰۵/۲۰و ۱۳۹۷/۰۶/۳۰
تاریخ پذیرش مقاله: ۱۳۹۷/۰۷/۲۴
نام نویسنده مسئول: دکتر علی‌محمد زارع بیدکی
نشانی نویسنده مسئول: ایران - یزد – بلوار پژوهش - دانشگاه یزد – گروه مهندسی کامپیوتر.

۱- مقدمه

یکی از مهّم‌ترین چالش‌ها در عرصه بازیابی اطلاعات و موتورهای جستجو، به تفاوت بین دایره واژگان مورد استفاده کاربران در هنگام جستجو و واژگان مورد استفاده در اسناد وب باز می‌گردد. برای رفع این مشکل که "عدم‌همخوانی واژگان"[۱] نامیده می‌شود، مجموعه‌ای از روش‌ها با عنوان روش‌های گسترش پرس‌وجو[۲] مورد استفاده قرار می‌گیرند. این روش‌ها با دریافت یک پرس‌وجو، مانند q ، آنرا به پرس‌وجویی مانند q' تغییر می‌دهند به صورتی که q' بتواند اسناد مرتبط بیشتری را بازیابی کند [۱]. فرایند گسترش پرس‌وجو می‌تواند شامل سه دسته عمل از قبیل حذف، جایگزینی[۳] و یا افزودن یک عبارت باشد. از بین این سه دسته، اَعمال جایگزینی و افزودنِ عبارت بدلیل بروز خطای کمتر بیشتر از سایرین مورد توجه قرارگرفته‌اند.

روش‌های گسترش پرس‌وجو براساس دامنه منابع اطلاعاتی مورد استفاده به دو دسته سراسری[۴] و محلی[۵] تقسیم می‌شوند. روش‌های سراسری از منابع اطلاعات زبانی بیرونی و تمام پیکره برای تأمین نیازهای اطلاعاتی استفاده می‌کنند. در مقابل، روش‌های محلی از اطلاعات مستخرج از اسنادِ مرتبطِ بازیابی شده، بهره می‌گیرند. باید خاطرنشان نمود که در روش‌های گسترش پرس‌وجوی محلی با فرض مرتبط بودن اسناد بازیابی شده برجسته[۶]، از آنها به عنوان منبع اطلاعاتی بهره گرفته می‌شود. در روش‌های سراسری، استفاده از عبارات نامناسب ممکن است منجر به بروز چرخش موضوعی[۷] شود. چرخش موضوعی به تغییر حوزه موضوعی یک پرس‌وجو و در نتیجه بازیابی اسنادی غیرمرتبط با موضوع مورد نظر کاربر اطلاق می‌شود.

با توجه به اینکه روش‌های محلی عبارت‌های مورد نیاز خود را از اسناد بازیابی شده با پرس‌وجوی اوّلیه بدست می‌آورند، عموماً در برابر چرخش موضوعی ایمن بوده و دارای کارایی بالاتری در بازیابی اسناد مرتبط هستند [۲]. در کنار محاسن روش‌های محلی باید خاطر نشان نمود که استفاده از این روش‌ها به معنای یک مرحله بازیابی، یک مرحله پردازش نه چندان سبک برای انتخاب کلمات جایگزین و بعد بازیابی کلی با استفاده از پرس‌وجوی اوّلیه و کاندیداهای گسترش یافته است. اجرای این سه مرحله می‌تواند زمان لازم برای پردازش‌های برخط[۸] و بار کاری جستجوگرها[۹] را افزایش دهد و به همین دلایل این روش‌ها گزینه‌های مناسبی برای پیاده‌سازی در موتورهای جستجو نیستند.

فرایندهای گسترش پرس‌وجو از منابع اطلاعاتی متنوعی استفاده می‌کنند که از آن جمله می‌توان به منابع تهیه شده توسط خبرگان انسانی مانند پایگاه‌های دانش، هستان‌شناسی‌ها و واژه‌نامه‌ها اشاره‌نمود. این منابع که عمدتاً برای یک یا چند دامنه موضوعی محدود آماده می‌شوند، پرهزینه هستند و سیر اعمال تغییرات در آنها کند است. با توجه به این چالش‌ها، محققین به سوی ایده استفاده از منابع داده بیرونی جذب شده‌اند که از آن جمله می‌توان به پیکره‌های متنی، اسناد وب، متن اخبار، تاریخچه پرس‌وجوهای کاربران و پست‌های شبکه‌های اجتماعی اشاره نمود. این دادگان عموماً به صورت آزاد در دسترس هستند و سرعت بروز تغییرات در آنها به مراتب بیشتر از منابع دانش دست‌ساز هستند.

در کنار تنوع منابع داده‌ای، باید به تنوع و تعدّد روش‌های مورد استفاده برای گسترش پرس‌وجوها نیز اشاره نمود که روش‌هایی برای استخراج و استفاده از مشابهت معنایی [۵-۳] ، روش‌های مبتنی بر استفاده از منابع داده بیرونی [۶]، مدل‌های زبانی آماری [۷] و تحلیل‌های دستوری [۸] از مهم‌ترین نمونه‌ها در این عرصه هستند. جدیدترین دسته از روش‌های گسترش پرس‌وجو، روش‌های مبتنی بر جاسازی کلمات هستند که همزمان با معرفی روش‌هایی مانند Word2Vec [۹] و Glove [۱۰] به وجود آمدند. این روش‌ها برای تخمین مشابهت معنایی از فاصله برداری بین بازنمایی‌های برداری عبارت‌ها بهره می‌گیرند.

در ادامه این مقاله و در بخش دوم، روش‌های گسترش پرس‌وجوی مبتنی بر جاسازی کلمات معرفی شده‌اند. در بخش سوم، فرایند گسترش پرس‌وجو تعریف شده است. بخش چهارم انواع وابستگی لازم برای تولید یک پرس‌وجوی گسترش‌یافته را مورد بررسی قرار داده است. پنجمین بخش به معرفی چهارچوب پیشنهادی می‌پردازد. جزئیات آزمایشات، مشخصات دادگان و معیارهای ارزیابی در بخش ششم بیان شده است و بخش هفتم به تحلیل نتایج آزمایشات می‌پردازد. هشتمین و آخرین بخش از مقاله به بیان نتیجه‌گیری و مرور اجمالی دستاوردهای روش پیشنهادی می‌پردازد.

۲- مروری بر روش‌های مدرن گسترش پرس‌وجو

در ادامه فراگیر شدن استفاده از جاسازی کلمات[۱۰] در کاربردهای پردازش متن، روش‌های بازیابی اطلاعات نیز از این حوزه متأثّر شده‌اند و استفاده از بردارهای جاسازی کلمات در کاربردهای مختلف از شباهت‌سنجی بین اسناد و پرس‌وجوها تا گسترش پرس‌وجو بسیار متداول شده است. روش‌های گسترش پرس‌وجوی مبتنی بر جاسازی کلمات نیز به دو دسته محلی و سراسری تقسیم می‌گردند. روش‌های محلیِ مبتنی بر استفاده از شبه بازخورد در بین روش‌های گسترش پرس‌وجوی مبتنی بر جاسازی کلمات بیش از سایرین مورد توجه قرار گرفته‌اند. متداول‌ترین دسته از روش‌های محلی با استفاده از یک معیار شباهت‌سنجی معنایی، عبارت های مشابهِ عبارت‌های موجود در پرس‌وجوها را انتخاب نموده و از آنها در بازیابی اوّلیه اسناد استفاده می‌کنند. Roy و همکاران [۱۱] سه روش برای این منظور ارائه نموده‌اند که در دومین موردِ آن، بعد از بازیابی اسناد، عبارت‌های مشابهی را که با استفاده از شباهت کسینوسی انتخاب شده‌اند از محتوای اسناد استخراج نموده و در نهایت از بین مجموعه این عبارت‌ها مرتبط‌ترین عبارت‌ها را انتخاب نموده‌اند.

در تحقیق دیگری، Kuzi و همکاران [۱۲] با استفاده از برآورد درست‌نمایی بیشینه[۱۱]، میزان شباهت یک عبارت جایگزین را به پرس وجوی کاربر تخمین زده و با ترکیب آن با احتمال بازیابی این عبارت از اسناد با بالاترین رتبه، آن را انتخاب نموده‌اند. در نهایت عبارت انتخاب

شده به نوبه خود برای بازیابی و رتبه‌بندی سایر اسناد مورد استفاده قرار گرفته است. Ganguly و همکاران [۱۳] از احتمال انتقال عبارات پرس وجو به عبارات جایگزین برای انتخاب آنها استفاده نموده‌اند که این احتمال به نوبه خود با لحاظ نمودن اسناد بازیابی شده از یک سو و کل اسناد پیکره از سوی دیگر بدست می‌آید. روش‌های مبتنی بر میدان تصادفی مارکف، پیش‌تر نیز در بازیابی اطلاعات مورد استفاده قرار گرفته‌اند امّا Balaneshin-Kordan و Kotov [۱۴] برای اوّلین بار از این چهارچوب در ترکیب با شباهت‌سنجی مبتنی بر بردارهای جاسازی کلمات بهره گرفته و توانسته‌اند با وزن‌دهی مجدد عبارت‌های پرس‌وجو و عبارت‌های جایگزین به نتایج بسیار خوبی در بازیابی اسناد مرتبط دست یابند. Zamani و Croft [۱۵] نشان داده‌اند که استفاده از شباهت کسینوسی به دلیل وضوح کم برای تفکیک بین مقادیر عبارات کاندیدای مشابه مناسب نیست و می‌توان با اعمال توابع Sigmoid و Softmax بر روی حاصل شباهت کسینوسی بر این نقص فائق آمد. در تحقیق دیگری، Zamani و Croft [۱۶] با استفاده از یک شبکه چندلایه پرسپترون به مدلی برای وزن‌دهی عبارت‌ها دست‌یافته‌اند. این مدل به نحوی آموزش می‌بیند که با اختصاص وزن‌های بالاتر به کلمات موجود در اسنادِ مرتبط، احتمال انتخاب آنها را برای گسترش پرس‌وجو افزایش دهد.

روش‌های گسترش پرس‌وجوی سراسری کمتر از روش‌های محلی مبتنی بر شبه بازخورد مورد توجه قرار گرفته‌اند ولی این دسته از روش ها نیز به نوبه خود توانسته‌اند سهمی در تحقیقات معاصر داشته باشند. Zheng و Callan [۱۷] تعدادی ویژگی را از بازنمایی برداری کلمات استخراج نموده و از آنها در چهارچوب یک مدل رگرسیون مبتنی بر چند ویژگی[۱۲] برای وزن‌دهی به کلمات جایگزین بهره گرفته‌اند. Roy و سایرین [۱۱] در یکی از روش‌های پیشنهادی خود از شباهت‌سنجی عبارت‌های جایگزین با بردار بدست آمده از میانگین بردارهای کلمات پرس‌وجو بهره گرفته و با استفاده از هر کدام از کلماتی که به این روش بدست آمده‌اند، گسترش پرس‌وجو را به انجام رسانده‌اند. ALMasri و سایرین [۱۸] با ارائه یک روش ساده مبتنی بر وزن‌دهی استاتیک برای عبارت‌های جایگزین توانسته‌اند نتایج بسیار خوبی ارائه دهند. Zuccon و همکاران [۱۹] از یک مدل زبانی ترجمه[۱۳] که از شباهت برداری جاسازی کلمات پرس‌وجو و کلمات جایگزین بهره می‌گرفته برای گسترش پرس‌وجو استفاده کرده‌اند. یکی از نکات بسیار جالب در این تحقیق، بررسی اثرگذاری تعداد ابعاد بردارهای جاسازی، نوع روش مورد استفاده برای تولید آنها و اندازه و نوع پیکره آموزشی بر میزان دقت روش بوده است. Fernández-Reyes و سایرین [۲۰] دو چهارچوب برای انتخاب و وزن‌دهی عبارات جایگزین را معرفی کرده‌اند.

در روش اوّل، از کلمات پرس‌وجو برای یافتن نزدیکترین جایگزین بهره گرفته می‌شود و در روش دوم از نزدیکترین کلمات جایگزین برای امتیازدهی به کلمات پرس‌وجو (معکوس روش اوّل) استفاده می‌گردد. روش اوّل مشابه تلاش‌های قبلی بوده و تنها در جزئیات و نحوه انتخاب کلمات جایگزین متفاوت است ولی روش دوم یک رویکرد جدید را در پیش گرفته و توانسته به کمک آن نتایج بسیار بهتری را بدست آورد.

در ادامه می‌توان به پژوهش Qian و همکاران [۲۱] اشاره نمود که با استفاده از بردارهای جاسازی کلمات، مشابهت معنایی عبارات جایگزین را بدست آورده و آنها را با استفاده از تابع عضویت فازی وزن‌دهی کرده است. در آخرین مرحله عبارات وزن‌دهی شده را برای رتبه‌بندی اسناد مورد استفاده قرار داده است. Rekabsaz و همکاران [۲۲] برای اوّلین بار با تعریف یک حد آستانه برای مشابهت معنایی و عدم‌اِعمال محدودیت بر روی تعداد شبیه‌ترین کاندیداها، توانسته‌اند کاندیداهای بهتری انتخاب نموده و خطای ناشی از چرخش موضوعی را کنترل نمایند.

علاوه بر دسته‌بندی‌های یاد شده می‌توان از یک دسته‌بندی دیگری نیز نام برد که براساس نوع انتخاب عبارت‌های مورد استفاده در گسترش پرس‌وجو تعیین می‌گردد. این دسته‌بندی بر اساس این حقیقت صورت می‌گیرد که عبارت‌های جایگزین براساس شباهت با یک عبارت از پرس وجو یا کل پرس‌وجو انتخاب شده باشند. از پژوهش‌هایی که قبلاً در این مقاله مورد ارجاع قرار گرفته‌اند، مواردی مانند [۱۴]، [۱۸] و [۲۰] در دسته اوّل قرار می‌گیرند. دسته دوم از روش‌ها، برای انتخاب کلمات جایگزین از مشابهت معنایی آنها با کل پرس‌وجو بهره می‌گیرند که معمولاً در قالب محاسبه بردار میانگین کلمات پرس‌وجو صورت می‌گیرد و از این دسته نیز می‌توان به [۱۳-۱۱]، [۱۶] و [۲۲] اشاره نمود.

۳- معرفی فرایند گسترش پرس‌وجو

برای هر پرس‌وجوی $Q = \{q_1, q_2, \cdots, q_n\}$ شامل n عبارت مانند q_i ، می‌توان پرس‌وجوی گسترش یافته‌ای مانند Q^{exp} تولید نمود که در آن، عبارتی مانند q_i با عبارت t جایگزین شده است، یعنی،

$$Q^{exp} = \{q_1, \cdots, q_{i-1}, t, q_{i+1}, \cdots, q_n\} \quad (۱)$$

پرس‌وجوی Q^{exp} نه تنها باید به لحاظ معنایی با پرس‌وجوی Q مشابهت داشته باشد، بلکه باید بتواند در همراهی با این پرس‌وجو اسناد مرتبط بیشتری را بازیابی کند [۱]. در صورتیکه Rel_Q مجموعه اسناد مرتبط برچسب‌دار برای پرس‌وجوی Q و Rel_Q^{exp} مجموعه اسناد مرتبط برچسب‌دار برای پرس‌وجوی Q^{exp} باشد، آنگاه توانایی پرس‌وجوی گسترش یافته در بازیابی اسناد مرتبط بیشتر با رابطه (۲) قابل تعریف است.

$$\left|Rel_Q^{exp} \cup Rel_Q\right| \geq |Rel_Q| \quad (۲)$$

در این رابطه، تابع $|.|$ نشان‌دهنده اندازه مجموعه است. با توجه به تعداد زیاد متغیرها و توابع مورد استفاده در این مقاله، فهرستی از آنها به همراه محل اوّلین استفاده در جدول شماره (۱) گردآوری شده است.

جدول ۱: جدول نمادهای مورد استفاده در روابط و متن مقاله شامل متغیرها و توابع.

#	نماد	توصیف	محل	#	نماد	توصیف	محل
۱	Q	پرس‌وجوی کاربر	(۱)	۱۹	tr	آستانه برش امتیاز وابستگی ارتباطی	(۶)
۲	Q^{exp}	پرس‌وجوی گسترش یافته	(۱)	۲۰	$\vec{X}_i$	بردار هر شی ورودی شبکه سیامی	(۷)
۳	q_i	کلمه i-ام پرس‌وجوی Q	(۱)	۲۱	W	وزن بردارهای ورودی شبکه سیامی	(۷)
۴	V	مجموعه واژگان پیکره	متن	۲۲	$O_w(.)$	تابع پارامتری بازنمایی اشیا	(۷)
۵	t	یک عبارت از مجموعه واژگان	(۱)	۲۳	D_w	فاصله وزن‌دار دو شی	(۷)
۶	Rel_Q	اسناد مرتبط پرس‌وجوی Q	(۲)	۲۴	$\|.\|$	نرم مرتبه ۲	(۷)
۷	Rel_Q^{exp}	اسناد مرتبط پرس‌وجوی Q^{exp}	(۲)	۲۵	Y	برچسب دودویی خروجی شبکه سیامی	(۸)
۸	t_c	یک عبارت جایگزین	(۳)	۲۶	m	شعاع حاشیه اطراف تابع $O_w(.)$	(۸)
۹	$\overrightarrow{v_{t_c}}$	بردار جاسازی عبارت t_c	(۴)	۲۷	k_t	تعداد عبارت جایگزین	متن
۱۰	$\overrightarrow{v_{q_i}}$	بردار جاسازی عبارت q_i	(۴)	۲۸	k_c	تعداد جایگزین‌ها با کمترین هزینه هم‌سنجی	متن
۱۱	$Sim(.,.)$	تابع شباهت معنایی یک عبارت با عبارت جایگزین	(۴)	۲۹	w_t	وزن هر عبارت در پرس‌وجوی گسترش‌یافته	(۹)
۱۲	$Cos(.,.)$	تابع محاسبه کسینوس زاویه بین دو بردار	(۴)	۳۰	γ_t	پارامتر تنظیم وزن عبارت‌ها	(۹)
۱۳	D	مجموعه اسناد نمایه (دامنه)	(۵)	۳۱	C_i	هزینه هم‌سنجی پرس‌وجوی کاندیدای i-ام	(۹)
۱۴	T	اسناد بازیابی‌شده با ارسال Q	(۵)	۳۲	$E(.,.)$	نشانگر وجود یک عبارت در یک پرس‌وجوی	(۹)
۱۵	T^{exp}	اسناد بازیابی‌شده با ارسال Q^{exp}	(۵)	۳۳	$W_{Q_i^{exp}}$	وزن پرس‌وجوی گسترش‌یافته i- ام	(۱۰)
۱۶	$\mathcal{R}(.,.)$	وابستگی ارتباطی دو پرس‌وجو	(۵)	۳۴	γ_Q	پارامتر تنظیم وزن پرس‌وجوی گسترش‌یافته i- ام	(۱۰)
۱۷	$\mathbb{R}$	مجموعه اعداد حقیقی	(۵)	۳۵	rel_i	نشانگر مرتبط بودن نتیجه i-ام	(۱۱)
۱۸	$f(.,.)$	تابع نگاشت اشتراک اسناد مرتبط با اجتماع اسناد بازیابی شده $f: D \times D \to \mathbb{R}$	(۵)	۳۶	r	رتبه اوّلین سند مرتبط با پرس‌وجو	(۱۳)

با توجه به این نکات باید از دو نوع معیار برای تعیین میزان وابستگی معنایی[۱۴] و وابستگی ارتباطی[۱۵] جفت پرس‌وجوها بهره گرفت تا به کمک آنها مقیاسی برای مناسب بودن پرس‌وجوی گسترش یافته به دست آورد. به همین دلیل در ادامه این مقاله دو نوع وابستگی بین پرس‌وجوی کاربر و پرس‌وجوی گسترش‌یافته معرفی خواهند شد.

۴- وابستگی جفت پرس‌وجو

تعیین مشابه بودن/نبودن یک جفت شی در کاربردهایی مانند شباهت سنجی تصاویر و یا عبارات کوتاه چندان پیچیده نیست و با کمک ناظر انسانی می‌توان جفت دادگان آموزشی را برچسب‌دار نمود. امّا در حوزه گسترش پرس‌وجو، تعریف دو پرس‌وجوی مشابه بسیار پیچیده است، زیرا علاوه بر مشابهت معنایی، پرس‌وجوی جایگزین باید بتواند اسناد مرتبط بیشتری را بازیابی کند و به همین دلیل عملکرد آن به صورت غیرقابل انکاری به محتوای متنی مجموعه اسناد موجود در نمایه وابسته است. بنابراین یک ناظر انسانی نمی‌تواند با دریافت دو پرس‌وجو به تنهایی در مورد مناسب بودن آنها هیچ اظهارنظری بکند؛ بلکه باید تمام نتایج هر دو پرس‌وجو بازیابی و بررسی شده و در صورتیکه پرس وجوی جدید توانسته باشد اسناد مرتبط بیشتری را بازیابی کند از آن به عنوان نسخه گسترش‌یافته مناسب نام برد.

۴-۱- تعریف وابستگی معنایی

وابستگی معنایی پرس‌وجوی کاربر به پرس‌وجوی گسترش‌یافته اوّلین شرط برای یک پرس‌وجوی جایگزین است. تعریف وابستگی معنایی دو عبارت عموماً برگرفته از تعریف مشابهت معنایی بین آن دو پرس‌وجو است [۵]. مشابهت معنایی را می‌توان براساس مشابهت اجزا یا کل دو پرس‌وجو در نظر گرفت. فرایند تخمین شباهت‌سنجی یکپارچه برای دو پرس‌وجو عمدتاً براساس ترکیب نمودن شباهت اجزای آنها و تعیین یک مقدار واحد صورت گرفته است. در این تحقیق، برای تولید پرس‌وجوهای مشابه، از مشابهت معنایی جزئی مشابه رابطه (۳) بهره گرفته شده است.

$$t_c = \text{Argmax}\, Sim(t_c, q_i) \qquad \forall q_i \in Q, \forall\, t \in V \qquad (۳)$$

در این رابطه t_c (برگرفته از واژه‌نامه V[۱۶]) یک عبارت جایگزین برای کلمه ای مانند q_i است و $Sim(.,.)$ نشان‌دهنده شباهت دو کلمه با عبارت جایگزین است. در این روش، عبارت‌های جایگزین برای هر کلمه براساس شباهت معنایی رتبه‌بندی می‌گردند و شبیه‌ترین آنها برای جایگزینی انتخاب می‌شوند. در این مقاله بنا به استفاده از بازنمایی‌های برداری از معیار فاصله کسینوسی به عنوان معیار شباهت بهره گرفته شده است (رابطه (۴)).

$$Sim(t_c, q_i) = Cos(\overrightarrow{v_{t_c}}, \overrightarrow{v_{q_i}}) \qquad (۴)$$

در این رابطه، $\overrightarrow{v_{t_c}}$ و $\overrightarrow{v_{q_i}}$ به ترتیب بردارهای عبارت‌های t_c و q_i هستند و $Cos(.,.)$ تابع کسینوس برداری است. تا این مرحله، پرس‌وجوهای جایگزین با استفاده از مشابهت معنایی جزئی با پرس‌وجوی کاربر انتخاب

شده‌اند، امّا در خصوص میزان اثرگذاری آنها در بهبود بازیابی نمی‌توان اظهارنظر نمود و به همین دلیل در مرحله بعد از وابستگی ارتباطی بهره خواهیم گرفت. از آنجا که در این مقاله از مشابهت کسینوسی برای وزن‌دهی عبارات بهره گرفته نمی‌شود و تنها کاربرد آن مرتب‌سازی عبارت‌های جایگزین و انتخاب بهترین‌ها است بنابراین نیازی به استفاده از توابع sigmoid و softmax برای افزایش وضوح نیست.

۴-۲- تعریف وابستگی ارتباطی[۱۷]

فرایند بازیابی بر روی مجموعه اسناد موجود در نمایه صورت می‌گیرد که از این به بعد D نامیده می‌شود. با ارسال دو پرس‌وجوی Q و Q^{exp} به نمایه، به ترتیب، مجموعه اسناد T و T^{exp} بازیابی می‌شوند که $T \subset D, T^{exp} \subset D$. در صورتی که برای پرس‌وجوی Q مجموعه‌ای از اسناد مرتبط برچسب‌دار مانند Rel_Q، موجود باشد، آنگاه وابستگی ارتباطی دو پرس‌وجو به شکل مندرج در رابطه‌ی (۵) تعریف می‌شود.

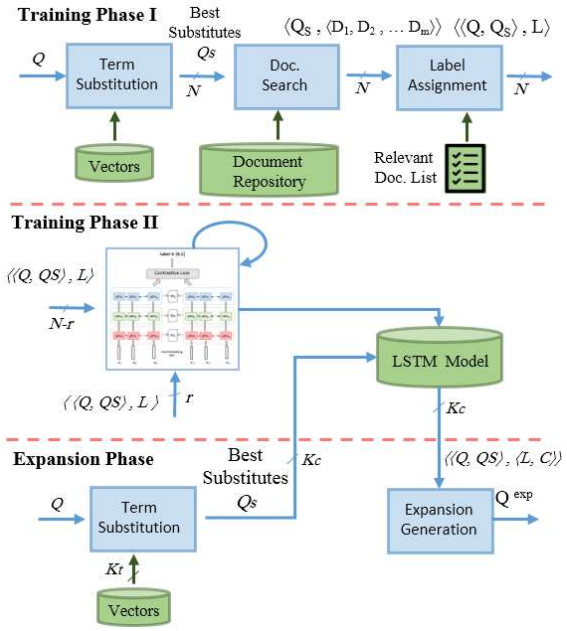
$$\mathcal{R}(Q, Q^{exp}) \propto \frac{f(T \cup T^{exp}, Rel_Q)}{f(T, Rel_Q)} \qquad (۵)$$
$$f: D \times D \to \mathbb{R}$$

$\mathcal{R}(.,.)$ نشان‌دهنده وابستگی ارتباطی دو پرس‌وجو است که با استفاده از تابع f محاسبه می‌شود. تابع $f(.,.)$ با دریافت یک مجموعه از اسناد و مقایسه آنها با مجموعه اسناد مرتبط Rel_Q، امتیازی حقیقی ($\mathbb{R}$) را به آنها اختصاص می‌دهد. برای تعریف ضابطه تابع f می‌توان از توابع مختلفی که همه دارای این شروط باشند استفاده نمود. پس از تعیین امتیاز میزان وابستگی ارتباطی دو پرس‌وجو، باید جفت‌هایی که دارای وابستگی کافی بوده و برای بازیابی مناسب باشند را به عنوان جفت‌های قابل گسترش برچسب‌گذاری نمود. برای این منظور از یک آستانه برش[۱۸] بهره خواهیم گرفت، که بر طبق آن جفت‌های دارای امتیاز بالاتر از آستانه برش، tr، به عنوان جفت قابل گسترش برچسب "1" دریافت می‌نمایند و سایر جفت‌ها برچسب "0" (مطابق رابطه (۶)).

$$Label(Q, Q^{exp}) = \begin{cases} 1 & R(Q, Q^{exp}) \geq tr \\ 0 & o.w. \end{cases} \qquad (۶)$$

۵- چهارچوب پیشنهادی برای گسترش پرس‌وجو

این پژوهش برای اوّلین بار مسئله گسترش پرس‌وجو را از دیدگاه متفاوتی مورد بررسی قرار داده است. رویکرد این مقاله قابلیت اجرا در یک موتور جستجوی بزرگ-مقیاس را دارد و می‌تواند بر مشکل عدم وجود انبوه دادگان برچسب‌دار نیز غلبه کند. چهارچوب پیشنهادی مزبور دارای دو مرحله یادگیری و گسترش پرس‌وجو است. مرحله یادگیری نیز به نوبه خود به دو زیرمرحله تقسیم می‌شود. برای آموزش و راه‌اندازی این چهارچوب در ابتدای امر نیازمند یک مجموعه کوچک از پرس‌وجوها، فهرست اسناد برچسب‌دار برای هر پرس‌وجو و بردارهای کلمات خواهیم بود.



شکل ۱: شمای کلی چهارچوب پیشنهادی برای گسترش پرس‌وجوی مبتنی بر شبکه سیامی عمیق.

در ادامه این مقاله، هر یک از مراحل آموزش و گسترش پرس‌وجو به تفکیک معرفی می‌گردند. شمای ساده‌ای از این چهارچوب در شکل (۱) قابل مشاهده است.

۵-۱- مرحله آموزش مدل وابستگی

در فاز اوّل، با استفاده از بردارهای کلمات و با احتساب وابستگی معنایی، بهترین کلمات کاندیدا با کلمات موجود در پرس‌وجو جایگزین می‌شوند تا تعدادی پرس‌وجوی جایگزین تولید شوند. پرس‌وجوهای جایگزین به یک جستجوگرِ متصل به نمایه اسناد ارسال می‌شوند تا برای هر یک از آنها فهرستی از اسناد بازیابی شده، فراهم گردد. در گام بعدی، اسناد به بخش برچسب‌گذاری وارد می‌شوند. این بخش فهرست اسناد مرتبط برای هر پرس‌وجو را در اختیار دارد و براساس معیار وابستگی ارتباطی به هر یک از جفت پرس‌وجوهای اصلی و جایگزین یک برچسب مبنی بر قابل جایگزین بودن (1) و یا نبودن (0) را اختصاص می‌دهد.

تا این مرحله، تعیین مناسب بودن/نبودن جفت گسترش یافته به صورت نسبی و بر اساس میزان مشابهت مجموع بازیابی پرس‌وجوی کاندیدا با پرس‌وجوی ورودی صورت می‌گیرد. امّا این روش را نمی‌توان برای هر پرس‌وجوی جدید ورودی به کار گرفت که مجموعه اسناد مرتبط آن معین نشده باشند. به همین دلیل در گام دوم، از یک الگوریتم کلاسه سازی دودویی برای تعیین برچسب جفت‌های جدید بهره خواهیم گرفت.

مدل کلاسه‌سازی انتخاب شده که یک شبکه سیامی حافظه کوتاه-مدت طولانی[۱۹] است نیازمند دادگان برچسب‌دار است. با توجه به اینکه تولید دادگان برچسب‌دار برای سیستم‌های بازیابی اطلاعات بسیار پرهزینه است، در این مقاله از یادگیری ضعیف برای غلبه بر این کمبود بهره گرفته‌ایم. مسائلی که از راهکار یادگیری صعیف بهره می‌گیرند در

واقع دارای همان خصوصیات و چینش مسائل یادگیری با سرپرست هستند، با این تفاوت که در مسائل با سرپرست، برچسب‌های قطعی و مطمئنی برای دادگان موجود هستند. در مقابل، در روش‌های با سرپرستی ضعیف از یک مدل ساده و نه چندان دقیق که از منابعی مانند یک مجموعه دادگان برچسب‌دار کوچک یا پایگاه دانش استفاده می‌کند برای برچسب‌گذاری مجموعه بزرگتری از دادگان بهره گرفته می شود [۲۳]. در واقع یک مدل با سرپرست، دانشی که فقط از بخش کوچکی از دادگان استخراج شده را به حجم بیشتری از دادگان تعمیم می‌دهد تا آنها را برچسب‌دار کند. وظیفه یادگیری ضعیف در چهارچوب پیشنهادی توسط بخش برچسب‌گذاری به اجرا در آمده است.

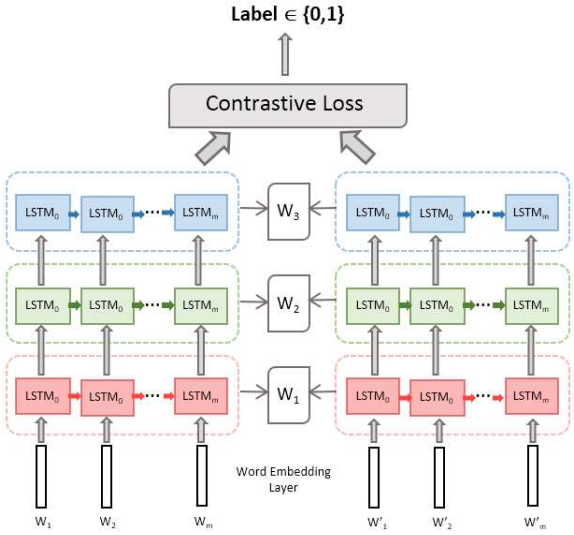
جفت‌های برچسب‌دار شده با نظارت ضعیف در فاز دوم به دو بخش آموزش و آزمون تقسیم شده و با استفاده از آنها یک شبکه سیامی آموزش داده می‌شود. ساختار پیشنهادی در این تحقیق، از یک شبکه سیامی با سه لایه همسان LSTM در دو زیرشبکه و یک لایه مشترک برای اعمال تابع هزینه هم‌سنجی[۲۰] تشکیل شده است. شمای کلی این ساختار در شکل شماره (۲) قابل مشاهده است. ورودی لایه اوّل در این شبکه بردارهای جاسازی کلمات است. فرایند تولید این بردارها به صورت جداگانه و با استفاده از داده‌های هر یک از پیکره‌های متنی به اجرا در آمده است. بردارهای کلمات هر یک از دو پرس‌وجو در لایه ورودی به شبکه وارد می‌شوند.

هر یک از لایه‌های درونی شبکه مزبور از سلول‌های حافظه کوتاه-مدت طولانی تشکیل شده است که تعداد آنها در جریان آزمایشات و با براساس بهترین مقدار برای معیارهای ارزیابی تعیین گردیده است. اتصال سلول‌ها در درون هر لایه به صورت یک‌جهته و پیش‌رو بوده و خروجی های آخرین لایه در هر دو زیرشبکه به یک تابع هزینه هم‌سنجی ارسال می‌گردند.

تابع تخمین هزینه هم‌سنجی را می‌توان از بین توابع خطای مرتبه یک یا دو انتخاب کرد؛ هر چند تجربیات منعکس شده در [۲۴] و [۲۵] نشان داده‌اند که استفاده از یک تابع مرتبه یک نتایج بهتری بدنبال دارد امّا در آزمایشات مربوط به این مقاله مشاهده گردید محاسبه هزینه هم‌سنجی با استفاده از یک تابع مرتبه دو نتایج بهتری را به همراه خواهد داشت و به همین دلیل از تابع هزینه اقلیدسی به شکل مندرج در رابطه (۷) استفاده می‌گردد (مشابه رابطه پیشنهادی [۲۶]):

$$D_w(\vec{X}_1,\vec{X}_2) = \left\| O_w(\vec{X}_1) - O_w(\vec{X}_2) \right\|_2 \qquad (۷)$$

در این رابطه، $\vec{\boldsymbol{X}}_{\boldsymbol{1}}$ و $\vec{\boldsymbol{X}}_{\boldsymbol{2}}$ بردارهای دو شی مورد نظر و $\boldsymbol{O_w}(.)$ تابع پارامتری بازنمایی اشیا می‌باشد. با بازآرایی رابطه (۷) می‌توان به رابطه (۸) رسید.



شکل ۲: شمای کلی شبکه سیامی عمیق حافظه کوتاه-مدت طولانی مورد استفاده برای مشابهت‌سنجی پرس‌وجوهای کاندیدا و کاربر

$$L(W,Y,\vec{X}_1,\vec{X}_2) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{\max(0,m-D_w)\}^2 \qquad (۸)$$

در این رابطه Y برچسب دودویی جفت اشیا، W مجموعه وزن‌های مورد استفاده بین دو شی X_1 و X_2 است، متغیر D_w نماینده فاصله وزن‌دار بین دو شی است و $m > 0$ شعاع حاشیه‌ای اطراف تابع $\boldsymbol{O_w}(.)$ است. برای آموزش این شبکه باید دادگانی شامل جفت پرس‌وجوی کاربر-جایگزین که دارای برچسب باشند را تأمین نمود. در این راستا، جفت پرس‌وجوهای دارای وابستگی بیشتر از آستانه برش توسط بخش برچسب گذاری با برچسب "1" و سایرین با برچسب "0" برچسب‌دار می‌شوند تا برای آموزش در اختیار شبکه سیامی قرار گیرند.

۵-۲- گسترش پرس‌وجو با استفاده از جفت‌های مرتبط

مرحله گسترش پرس‌وجو شامل تولید پرس‌وجوهای جایگزین، برچسب دار نمودن و ادغام آنها برای تولید پرس‌وجوی گسترش‌یافته است. در اوّلین گام، برای تولید پرس‌جوهای جایگزین برای هر عبارت غیرتوقفی[۲۱] در پرس‌وجوی کاربر، تعداد k_t عبارت جایگزین با بیشترین شباهت معنایی از مجموعه کل عبارات موجود در واژه‌نامه انتخاب و در پرس وجوی کاربر جایگزین می‌شوند تا تعدادی پرس‌وجوی جایگزین به دست آید. پرس‌وجوهای جایگزین در اختیار شبکه سیامی آموزش دیده قرار می‌گیرند و این شبکه به آنها مقدار هزینه هم‌سنجی و یک برچسب را اختصاص می‌دهد.

سپس از بین مجموعه کاندیداهایی که دارای وابستگی ارتباطی با پرس‌وجوی کاربر باشند، تعداد k_c جایگزین که دارای کمترین هزینه هم سنجی (بیشترین وابستگی ارتباطی) هستند انتخاب و از آنها در گسترش پرس‌وجوی اوّلیه بهره گرفته می‌شود. تولید پرس‌وجوی جدید به دو

شکل مبتنی بر عبارت[۲۲] و مبتنی بر پرس‌وجو[۲۳] صورت می‌گیرد. در روش مبتنی بر عبارت، کلمات موجود در پرس‌وجوهای جایگزین انتخاب شده و براساس رابطه (۹) وزن‌دهی می‌شوند. این پرس‌وجوها در قالب یک پرس‌وجوی واحد در کنار پرس‌وجوی کاربر به نمایه ارسال شده و لیست اسناد بازیابی شده برای ارزیابی مورد استفاده قرار می‌گیرند.

$$W_t = \gamma_t \frac{\sum_{i=1}^{k_c}(1-C_i)\,E(t,Q_i^{exp})}{\sum_{i=1}^{k_c} E(t,Q_i^{exp})} \qquad (۹)$$

$$E\left(t,Q_i^{exp}\right) = \begin{cases} 1 & \forall\, t \in Q_i^{exp} \\ 0 & o.w. \end{cases}$$

در این رابطه، وزن هر عبارت t با W_t نمایش داده می‌شود. پارامتر γ_t در بازه $[0\,,1]$ برای تنظیم وزن عبارت‌ها مورد استفاده قرار می‌گیرد و C_i نشان‌دهنده مقدار هزینه هم‌سنجی خروجی شبکه برای پرس‌وجوی جایگزین i-ام است. تابع $E(.,.)$ نشانگر وجود عبارت t در Q_i^{exp} است.

در روش مبتنی بر پرس‌وجو، به جای ارسال هر یک از عبارت‌ها به صورت جداگانه، مجموعه k_c پرس‌وجوی جایگزین به صورت جداگانه با وزن‌هایی به صورت رابطه (۱۰) به نمایه ارسال شده و مجموعه اسناد بازیابی شده توسط آنها ادغام شده و رتبه‌بندی می‌گردند. در این رابطه، $W_{Q_i^{exp}}$ نشان‌دهنده وزن هر پرس‌وجوی گسترش‌یافته، C_i مقدار هزینه هم‌سنجی و γ_Q در بازه $[0\,,1]$ پارامتر تنظیم‌کننده وزن پرس‌وجوها است.

$$W_{Q_i^{exp}} = \gamma_Q(1-C_i) \qquad (۱۰)$$

۶- آزمایشات

در این بخش، جزئیات پیاده‌سازی روش پیشنهادی به همراه ویژگی‌های دادگان مورد استفاده برای آموزش و آزمون معرفی خواهند شد. به علاوه، تعدادی از روش‌های گسترش پرس‌وجوی مبتنی بر جاسازی کلمات که برای مقایسه با روش پیشنهادی انتخاب و پیاده‌سازی شده‌اند نیز در این بخش معرفی خواهند شد.

۶-۱- معرفی دادگان

برای انجام آزمایشات از دادگان دو پیکره مشهور و متداول برای زبان فارسی بهره گرفته‌ایم که عبارتند از پیکره محک وب DotIR [۲۷] و پیکره همشهری ویرایش ۲ [۲۸]. پیکره‌های مزبور در دانشگاه تهران برای کاربردهای بازیابی اطلاعات و کلاسه‌سازی موضوعی تدوین شده‌اند. این دو پیکره هر کدام دارای ۵۰ پرس‌وجو هستند که مجموعه اسناد مرتبط با آنها نیز برچسب‌گذاری شده‌اند و بنابراین می‌توان از آنها در آموزش و آزمون روش‌های گسترش پرس‌وجو بهره برد. آماده‌سازی این دادگان شامل یکسان‌سازی نویسه‌ها، حذف برچسب‌های XML و استخراج محتوای صفحات و خصوصیات آنها است. دادگان اسناد هر دو پیکره با استفاده از کتابخانه متن‌باز Lucene V6.5.1 به صورت مجزا نمایه‌سازی شده‌اند. در آزمایشات مشابه روش‌های مورد مطالعه از مدل زبانی با هموارسازی Jelinek-Mercer بهره گرفته شده است و به همین دلیل ما نیز از همین مدل زبانی بهره گرفته‌ایم. مدل زبانی مزبور دارای یک پارامتر قابل تنظیم به نام لاندا ($0 < \lambda \leq 1$) است که با استفاده از الگوی اعتبارسنجی چند دسته‌ای[۲۴] تعیین می‌شود. در این پژوهش نیز با تغییر این پارامتر در بازه بین صفر تا یک مقدار مناسب را به صورت $\lambda = 0.4$ بدست آورده‌ایم. از آنجا که تشخیص کلمات توقف در روش‌های مقایسه شده نقش اساسی ایفا می‌کند و در عین حال برای کاربردهای مبتنی بر پیکره‌های متنوع نمی‌توان از یک لیست واحد استفاده نمود، یک چهارچوب تعیین کلمات توقف براساس روش پیشنهادی [۲۹] پیاده‌سازی شده و مورد استفاده قرار گرفته است.

۶-۲- پیاده‌سازی روش پیشنهادی

در اوّلین گام برای پیاده‌سازی چهارچوب نرم‌افزاری روش پیشنهادی، متن اسناد، عناوین، برچسب‌ها و انکورها برای تولید بازنمایی‌های جاسازی کلمات هر یک از پیکره‌ها استخراج شده و بردارهای جاسازی به روش FastText با ۳۰۰ بعد تولید شده‌اند. در ادامه، برای انتخاب عبارت‌های جایگزین از معیار شباهت کسینوسی (مشابه رابطه (۴)) بهره گرفته شده است. تعداد عبارت‌های جایگزین، k_t، با استفاده از اعتبارسنجی چند دسته‌ای برابر ۷ تعیین شده است. برای انتخاب تابع $f(.,.)$ مناسب جهت تخمین وابستگی ارتباطی، باید به این نکته توجه نمود که این تابع باید بتواند جفت‌هایی را متمایز کند که بیشترین تعداد از اسناد مرتبط متفاوت با اسناد مرتبط بازیابی شده توسط پرس‌وجوی اوّل را بازیابی کنند. توابع Jaccard، Odds Ratio، Support، Piatetsky-Shapiro و Information Gain معرفی شده در [۳۰] برای این منظور مورد ارزیابی قرار گرفته‌اند که از بین آنها بهترین نتایج به ترتیب توسط توابع Jaccard و Piatetsky-Shapiro بدست آمد که با توجه به سادگی و برد محدود، تابع Jaccard مورد استفاده قرار گرفت. درنهایت برای پیکره همشهری۲، تعداد ۱۵۰۳۵۵ جفت و برای پیکره محک وب ۱۷۰۴۰۶ جفت پرس‌وجوی برچسب‌دار تولید شده‌اند. شبکه سیامی مورد استفاده دارای ۳ لایه حافظه کوتاه-مدت طولانی است که هر لایه شامل ۷۸ سلول می‌باشد. در جریان آموزش این شبکه احتمال ماندگاری[۲۵] برابر ۰/۹۵ بیشترین دقت را به ارمغان آورده که با توجه به کمبود دادگان آموزش به نظر منطقی می‌رسد. تعداد دورهای آموزش[۲۶] برای این مدل در بهترین حالت برای پیکره همشهری ۲ برابر ۱۰۵ دور و برای پیکره محک وب برابر ۹۳ دور بوده است. پس از آموزش این شبکه، دو نوع آزمایش تشخیص جفت‌های با وابستگی ارتباطی و گسترش پرس‌وجو بر روی نتایج انجام گرفته است. ساختار آزمونها از الگوی اعتبارسنجی ۵ دسته‌ای[۲۷] تبعیت نموده است.

برای ارزیابی، پرس‌وجوهای مجموعه آزمون با تعداد ۵ عبارت جایگزین تولید و برای هم‌سنجی در اختیار شبکه سیامی قرار داده شده‌اند. در پایان ارزیابی، جفت‌های ورودی یک برچسب مبنی بر همسانی/عدم‌همسانی به همراه هزینه هم‌سنجی دریافت نموده‌اند. سپس با استفاده از k_c پرس‌وجوی جایگزین، یک ویرایش مبتنی بر عبارت و k_c ویک یرایش مبتنی بر پرس‌وجو برای هر پرس‌وجو تولید شده است. برای هر یک از دو ویرایش، دو برابر تعداد اسناد مورد نیاز جهت ارزیابی،

بازیابی شده و با ادغام و رتبه‌بندی مجدد، یک لیست واحد به عنوان خروجی تولید گردیده است.

۶-۳- روش‌های مورد مقایسه

برای ارزیابی روش پیشنهادی، تعدادی از روش‌های گسترش پرس‌وجوی مبتنی بر جاسازی کلمات از هر دو گروهِ روش‌های با مشابهت معنایی جزئی و کلی انتخاب و پیاده‌سازی شده‌اند. این روش‌ها عبارتند از:

- روش Q2V پیشنهادی توسط Fernández-Reyes و همکاران[۲۰]
- روش VEXPپیشنهادی توسطALMasri و همکاران [۱۸]
- روش Pre-retrieval KNN پیشنهادی توسط Roy و همکاران [۱۱]
- روش پیشنهادی Zuccon و همکاران [۱۹]

علاوه بر این روش‌ها، کلیه اسناد یکبار با استفاده از الگوریتم بازیابی با مدل زبانی Jelinek-Mercer و $\lambda = 0.4$ بدون هیچ نوع پردازش اضافی (بجز حذف کلمات توقف) بازیابی شده‌اند.این روش‌ها برای اختصار به ترتیب PG-Q2V ، VEXP، ROY ، ZCCN و JM LM نامیده می‌شوند و همچنین برای روش پیشنهادی در دو حالت مبتنی بر عبارت و مبتنی بر پرس‌وجو از عبارات اختصاری DS_Term و DS_Query بهره گرفته خواهد شد.

پارامتر α برای روش VEXP عددی در بازه $[0\ ,1]$ است که برای تعیین مقدار دقیق آن از آزمایشات متوالی در قالب اعتبارسنجی پنج-دسته‌ای با بازه‌هایی به طول ۰/۰۵ بهره گرفته شده است. این پارامتر برای اسناد همشهری۲ برابر ۰/۳ و برای اسناد پیکره محک وب DotIR برابر ۰/۴ تعیین شده است. پارامترهای γ_Q و γ_t نیز با روشی مشابه و با استفاده از اعتبارسنجی پنج-دسته‌ای محاسبه شده‌اند. این مقادیر برای پیکره همشهری به ترتیب برابر ۰/۰۵ و۰/۱۰ تعیین شده و همین پارامترها برای پیکره محک وب به ترتیب برابر ۰/۱۵ و ۰/۰۵ تعیین شده‌اند.

۶-۴- معیارهای ارزیابی

برای اندازه‌گیری میزان تأثیر هر یک از روش‌های توصیف شده بر عملکرد فرایند بازیابی اسناد و برچسب‌گذاری جفت پرس‌وجوها از چهار معیار متداول بهره گرفته شده است که عبارتند از:

- دقت نقطه‌ای درN[۲۸]
- میانگین متوسط دقت[۲۹]
- میانگین رتبه متقابل[۳۰]
- دقت برچسب‌گذاری[۳۱]

دقت نقطه‌ای در N از محاسبه نسبت تعداد اسناد مرتبط تا رتبه Nام بر N بدست می‌آید (مشابه رابطه (۱۱)) و نشان‌دهنده توان روش در گنجاندن اسناد مرتبط در فاصله N رتبه بالای لیست بازیابی است. در این رابطه rel_i نشانگر مرتبط بودن سند i ام است که برای سند مرتبط برابر 1 و برای سند غیرمرتبط برابر 0 است. برای محاسبه‌ی میانگین متوسط دقت برای تعداد $|Q|$ پرس‌وجو تا رتبه N، ابتدا برای هر پرس‌وجو متوسط مقادیر دقت نقطه‌ای برای همه نقاط قبل از N محاسبه شده و سپس از مقادیر بدست‌آمده برای همه پرس‌وجوها میانگین‌گیری خواهد شد (مشابه رابطه (۱۲)).

$$P@N = \frac{\sum_{i=1}^{N} rel_i}{N} \qquad (۱۱)$$

$$MAP = \frac{1}{|Q|}\sum_{Q_i \in Q}\frac{1}{N}\sum_{i=1}^{N} P@i \qquad (۱۲)$$

رتبه متقابل از معکوس رتبه اوّلین سند مرتبط بدست می‌آید و با میانگین‌گیری از این معیار (مشابه رابطه (۱۳)) برای تمام پرس‌وجوها می‌توان از آن به عنوان معیاری برای ارزیابی توانایی روش در بازیابی اسناد مرتبط در بالاترین رتبه ممکن بهره گرفت. در این رابطه r نشان دهنده رتبه اولین سند مرتبط است.

$$RR = \begin{cases} \frac{1}{r} & r: rank\ of\ the\ 1st\ relevant\ doc. \\ 0 & No\ relevant\ result \end{cases} \qquad (۱۳)$$

$$MRR = \frac{1}{|Q|}\sum_{Q_i \in Q} RR_{Q_i}$$

آخرین معیار ارزیابی با سایر معیارهای بالا متفاوت بوده و برای کاربردی غیر از بازیابی اطلاعات مورد استفاده قرار می‌گیرد. با توجه به نیاز مبرم به ارزیابی دقت شبکه سیامی در برچسب‌گذاری جفت پرس‌وجوها، باید از یک معیار برای بررسی دقت این الگوریتم بهره برد. دقت برچسب‌گذاری براساس نسبت تعداد برچسب‌های درست مثبت به مجموع برچسب‌های درست مثبت و نادرست مثبت محاسبه می‌گردد (مشابه رابطه (۱۴)).

$$precision = \frac{TP}{TP + FP} \qquad (۱۴)$$

۷- نتایج آزمایشات

در روش پیشنهادی از یک شبکه سیامی برای برچسب‌گذاری جفت پرس وجوی کاربر و پرس‌وجوی گسترش یافته بهره گرفته شده است. به همین دلیل در اوّلین بخش از آزمایشات، دقت این شبکه در برچسب‌گذاری زوج‌های آزمون ارزیابی می‌شود و در گام دوم توانایی چهارچوب پیشنهادی در بهبود بازیابی اسناد مرتبط ارزیابی می‌شود.

۷-۱- ارزیابی دقت برچسب‌گذاری شبکه سیامی

پرس‌وجوهای هر دو مجموعه در ۵ دسته تقسیم شده و موارد جایگزین برای مجموعه‌های آموزش و آزمون تولید شده و برای تمام جفت‌های بدست آمده، اسناد مرتبط بازیابی شده و در صورت احراز شرط وابستگی ارتباطی به آنها برچسب مناسب اختصاص یافته است. در هر نوبت، شبکه سیامی آموزش داده شده برچسب جفت‌های آزمون را حدس زده که در نهایت توانسته است به دقت برچسب‌گذاری برابر ۸۴/۸ درصد با دادگان پیکره همشهری۲ و ۹۰/۴۵ درصد با دادگان DotIR دست‌یابد.

۷-۲- ارزیابی دقت بازیابی

برای ارزیابی کیفیت فرایند بازیابی پس از اعمال فرایند گسترش پرس وجو از معیارهای MAP، MRR، P@5 و P@10 بهره گرفته شده است. نماد JM LM نشان‌دهنده بازیابی اسناد تنها با استفاده مدل زبانی با هموارسازی Jelinek-Mercer و بدون بهره‌گیری از یک روش گسترش پرس‌وجو است. سایر روش‌های معرفی شده در بخش ۶-۳ همه از یک الگوی گسترش پرس‌وجو بهره می‌گیرند. چنانکه در توصیف روش پیشنهادی ذکر شد این روش در زمان بازیابی در دو حالت مبتنی بر عبارت و مبتنی بر پرس‌وجو قابل استفاده است که برای اشاره به آنها از عبارات اختصاری DS_Term و DS_Query بهره گرفته شده است. نتایج این آزمایشات در جدول شماره (۲) قابل مشاهده است.

۷-۳- تحلیل نتایج بازیابی

یافته‌ها نشان می‌دهد، روش پیشنهادی در هر یک از دو حالت توانسته امتیازات بهتری را نسبت به سایر روش‌ها بدست آورد. روش پیشنهادی با بازیابی مبتنی بر عبارت در حالت کلی (با لحاظ MAP) بهتر از حالت مبتنی بر پرس‌وجو عمل می‌کند. روش‌های ROY و ZCCN نسبت به سایر روش‌ها نتایج بهتری را به ارمغان آورده‌اند که دلیل آنرا باید در تفاوت تعریف وابستگی معنایی در آنها جستجو کرد. در هر دو روش ROY و ZCCN برای انتخاب عبارت‌های گسترش پرس‌وجو از فاصله بردار کلمه جایگزین با بردار کلِ پرس‌وجو یا فاصله آن با تمام بردارهای کلمات موجود در پرس‌وجو بهره گرفته می‌شود. همین ویژگی در روش پیشنهادی نیز دیده می‌شود و از نقاط قوت آن محسوب می‌گردد.

دقت بازیابی در پیکره محک وب DotIR نسبت به پیکره همشهری پایین‌تر است که دلیل آن را باید در تمایز ماهیت اسناد دو پیکره جستجو نمود.

پیکره‌های مستخرج از وب معمولاً اسنادی با پراکندگی موضوعی بیشتر دارند. در مقام مقایسه باید توجه نمود که در پیکره محک وب، کلمات دارای بردارهای هم‌وقوعی تنک‌تر هستند. فرایند آموزش با مقادیر اوّلیه تصادفی به علت داشتن همسایه‌های کمتر، تغییرات اندکی نسبت به مقادیر اوّلیه خواهند داشت و چندان تمایزپذیر نخواهند بود.

همین مشکل، امکان انتخاب همسایه‌های نویزی را فراهم آورده که به نوبه خود به بازیابی اسناد غیرمرتبط و کاهش کارایی منجر می‌شود.

نکته دیگری که باید بدان توجه نمود تفاوت روش پیشنهادی و سایر روش‌های مورد مقایسه در نگاه به بازنمایی کل پرس‌وجو می‌باشد. در برخی از مقالات از بردار میانگین بردارهای کلمات به عنوان بردار بازنمایی پرس‌وجو بهره گرفته شده است. در حالی که این ایده در هنگام مواجهه با پرس‌وجوهای حاوی کلمات دارای چند معنا و یا کلمات با تعداد تکرار بسیار کم مشکل ساز خواهد شد؛ زیرا پراکندگی بردارهای کلمات می تواند بردار میانگین را از محل تجمع اکثر کلمات دور کند. در مقابل رویه بالا، در روش پیشنهادی از یافتن یک نمایش برای یک پرس‌وجو اجتناب شده و به جای آن تلاش‌ها بر یافتن دو بازنمایی متقابلاً مشابه (و نه لزوماً میانگین هریک از عبارات) برای دو پرس‌وجوی مرتبط، استوار بوده است.

در آزمایشات بازیابی، در وهله اوّل تمام روش‌ها با k برابر ۱۰ مورد ارزیابی قرارگرفته‌اند، انتخاب مقدار ۱۰ برای k بدین منظور صورت گرفته است که کمترین مقدار k برای رسیدن دقت روش‌های PG-Q2V و VEXP به مقدار MAP بالای ۵۰٪ مقدار MAP بازیابی بدون گسترش پرس‌وجو مقادیر ۷ و ۱۰ بودند، در حالیکه سایر روش‌ها با مقادیر کوچکتری از k نیز قادر به دست یافتن به این محدوده بودند.

در این زمان باید به این نکته توجه نمود که انتخاب k برابر ۱۰ لزوماً به معنای افزودن ۱۰ عبارت به پرس‌وجو نیست، زیرا در روش‌هایی مانند PG-Q2V، VEXP در واقع برای هر عبارت غیرتوقفی ۱۰ عبارت جایگزین تولید و در گسترش مورد استفاده قرار می‌گیرد. از طرفی باید توجه نمود که برخی از کاندیداها بیش از یکبار در جریان گسترش تولید می‌شوند و در این صورت تنها یکبار با امتیازی حاصل از مجموع امتیازهای جزئی در بازیابی مورد استفاده قرار می‌گیرند.

جدول ۲: نتایج ارزیابی کارایی بازیابی اسناد بدون استفاده از هیچ روش گسترش پرس‌وجو در مقایسه با استفاده از روش‌های مختلف

#	پیکره	متریک	JM LM	PG-Q2V	VEXP	ZCCN	ROY	DS_Term	DS-Query
۱	همشهری ۲	MAP	۰/۵۷۹	۰/۵۵۲	۰/۶۲۲	۰/۵۸۶	۰/۵۸۸	**۰/۶۷۶**	۰/۶۵۵
۲		MRR	۰/۶۹۸	۰/۷۱۲	۰/۸۰۹	۰/۷۲۲	۰/۷۱۷	**۰/۸۱۵**	۰/۷۹۶
۳		P@5	۰/۶۴۰	۰/۶۰۰	۰/۶۴۸	۰/۵۸۰	۰/۶۱۶	۰/۶۷۲	**۰/۶۸۸**
۴		P@10	۰/۵۸۸	۰/۵۴۴	۰/۶۲۲	۰/۵۴۸	۰/۵۸۰	۰/۶۴۲	**۰/۶۶۴**
۵	محک وب DotIR	MAP	۰/۴۴۴	۰/۴۰۴	۰/۳۸۵	۰/۴۱۷	۰/۴۵۰	**۰/۴۷۲**	۰/۴۶۹
۶		MRR	۰/۶۲۴	۰/۵۹۶	۰/۵۹۳	۰/۴۹۵	۰/۵۴۰	۰/۶۱۷	**۰/۶۵۷**
۷		P@5	۰/۴۵۲	۰/۴۱۲	۰/۴۰۸	۰/۴۳۶	۰/۴۶۸	۰/۴۶۴	**۰/۴۸۸**
۸		P@10	۰/۴۶۲	۰/۴۲۰	۰/۳۹۲	۰/۴۲۸	۰/۴۶۴	۰/۴۶۲	**۰/۴۹۲**

جدول شماره (۳) نشان‌دهنده میانگین تعداد عبارت‌های موجود در پرس‌وجوها، میانگین عبارات یکتای افزوده شده به آنها در جریان گسترش و نسبت این دو مقدار است. در این جدول، تعداد متوسط عبارت ها در پرس‌وجوها با نماد $\bar{L}$، متوسط تعداد عبارات افزوده با $\overline{Uex}$ مشخص شده و برای سهولت در مقایسه روش‌ها، نسبت این دو پارامتر نیز در آخرین سطر هر بخش گنجانده شده است.

چنانکه نتایج مندرج در جدول (۳) نشان می‌دهد برای پرس‌وجوهای مستخرج از پیکره محک وب، هر دو مقدار کمینه اوّل و دوم متعلق به روش پیشنهادی هستند. در دادگان پیکره همشهری۲ اوّلین مقدار کمینه متعلق به روش پیشنهادی در حالت مبتنی بر پرس‌وجو است ولی دومین کمترین مقدار به صورت مشترک به روش‌های ROY و PG-Q2V اختصاص دارد که توانسته‌اند با اختلاف ۰/۴۶ متوسط عبارت کمتری نسبت به روش پیشنهادی در حالت مبتنی بر عبارت را ارائه دهند.

با توجه به برتری روش پیشنهادی بر سایر روش‌ها در دادگان این پیکره و همچنین اختلاف کمتر از یک واحد باز هم می‌توان گفت استفاده از روش پیشنهادی حتی در حالت مبتنی بر عبارت نیز نسبت به سایر روش‌ها مقرون به صرفه‌تر است؛ زیرا کاهش تعداد عبارت‌های افزوده به صورت مستقیم به کاهش بارِکاری جستجوگرها و در نتیجه کاهش زمان لازم برای پاسخگویی به کاربران می‌انجامد.

۸- نتیجه‌گیری

در این مقاله یک چهارچوب نرم‌افزاری برای گسترش پرس‌وجوی مبتنی بر شبکه سیامی عمیق حافظه کوتاه-مدت طولانی ارائه شده است. برای توسعه این روش در ابتدا دو نوع وابستگی معنایی و ارتباطی تعریف شده است که وابستگی ارتباطی در این تحقیق برای اوّلین بار معرفی شده است.

در تحقیقات مشابه از وابستگی معنایی بین عبارت‌های موجود در پرس‌وجو (به صورت جداگانه) و یا وابستگی معنایی به کل یک پرس‌وجو برای انتخاب عبارت‌های کاندیدای گسترش بهره گرفته می‌شده است. این نوع وابستگی در کنار فراهم آوردن امکان بازیابی اسنادِ مشابهِ مرتبط باعث بازیابی اسناد غیرمرتبط با مشابهت موضوعی نیز می‌گردید؛ در حالیکه وابستگی ارتباطی می‌تواند احتمال بازیابی چنین اسنادی را کاهش دهد.

از آنجا که اسناد مرتبط تنها برای پرس‌وجوهای اصلی برچسب‌دار شده و برای پرس‌وجوهای جایگزین هیچ برچسبی وجود ندارد، برای غلبه بر این مشکل از یک روش با نظارت ضعیف بهره گرفته شده است. مهّم ترین مزیت استفاده از این تکنیک فراهم آوردن حجم انبوهی دادگان برچسب‌دار برای آموزش شبکه عصبی تنها با استفاده از دادگان برچسب‌دار ۵۰ پرس‌وجو است. با آموزش شبکه‌عصبی می‌توان پرس‌وجوهای جایگزین گسترش پرس‌وجو را برای ارزیابی به این شبکه سپرد.

یکی دیگر از ویژگی‌های روش پیشنهادی، عدم اتّکا بر یک بازنمایی نه‌چندان دقیق است؛ زیرا در این روش، بردارهای میانگین برای هر یک پرس‌وجوها (به عنوان بردار نمایش پرس‌وجو) به کار برده نمی‌شوند. بردارهای میانگین به دلیل اینکه می‌توانند تحت تأثیر وجود کلمات با بردارهای پرت باشند، امکان بروز چرخش موضوعی را افزایش می‌دهند. در روش پیشنهادی، در واقع پرس‌وجوها براساس مشابهت خروجی‌های دو زیرشبکه که براساس تعاریف این مقاله حاکی از وابستگی ارتباطی آنها هستند، برچسب‌گذاری می‌شوند. در کنار این مزیت، روش مزبور تضمین می‌کند عبارت‌های مورد استفاده در بازیابی نه تنها با تک تک عبارت‌های موجود در پرس‌وجوی کاربر مشابهت دارند بلکه با کل پرس‌وجو در هماهنگی هستند.

جدول ۳: مقایسه تعداد متوسط عبارت‌ها در پرس‌وجوها ($\bar{L}$)، متوسط تعداد عبارات افزوده ($\overline{Uex}$) و نسبت این متغیر که نشان‌دهنده میزان بار تحمیلی روش به جستجوگرها است.

#		متریک	JM LM	PG-Q2V	VEXP	ZCCN	ROY	DSi_Term	DSi-Query
۱	همشهری ۲	$\bar{L}$	۴/۵۴	۴/۵۴	۴/۵۴	۴/۵۴	۴/۵۴	۴/۵۴	۴/۵۴
۲		$\overline{Uex}$	صفر	۱۰	۳۵/۶	۳۵	۱۰	۱۲/۰۸	۸/۰۶
۳		$\frac{\overline{Uex}}{\bar{L}}$	صفر	**۲/۲۰**	۷/۸۴	۷/۷۱	**۲/۲۰**	۲/۶۶	**۱/۷۸**
۴	محک وب DotIR	$\bar{L}$	۳/۹۶	۳/۹۶	۳/۹۶	۳/۹۶	۳/۹۶	۳/۹۶	۳/۹۶
۵		$\overline{Uex}$	صفر	۱۰	۳۳/۶	۳۲/۶۶	۱۰	۶/۶۶	۷/۴۴
۶		$\frac{\overline{Uex}}{\bar{L}}$	صفر	۲/۵۳	۴/۴۹	۸/۲۵	۲/۵۳	**۱/۶۸**	**۱/۸۸**

علاوه براین موارد، آزمایشات نشان داده است که روش پیشنهادی می‌تواند با استفاده از تعداد کمتری عبارت افزوده به کارایی بالاتری دست‌یافته و در کاهش بار کاری جستجوگرها بر سایر روش‌ها غلبه کند. علاوه بر محاسن بالا، با توجه به بازنمایی کل پرس‌وجو در یک لایه شبکه سیامی امکان تعیین اثر توأم تغییر چند عبارت بر بهبود کارایی بازیابی اطلاعات نیز در روش پیشنهادی این مقاله میسر است و همچنین می توان اثر حذف یک عبارت را نیز در نظر گرفت.

به صورت خلاصه می‌توان گفت، روش‌های مبتنی بر شباهت‌سنجی معنایی با کل پرس‌وجو بهتر از روش‌های مبتنی بر شباهت جزئی عمل می‌کنند. با همین دلیل روش پیشنهادی و روش‌های ROY و ZCCN نتایجی بهتر از سایرین ارائه داده‌اند. سپس باید به این نکته اشاره نمود که روش پیشنهادی توانسته است با ایجاد کمترین بار کاری برای جستجوگرها به نتایجی بهتر از سایرین دست یابد.

نویسندگان این مقاله در نظر دارند در راستای ادامه این مسیر، از روش‌های انتخاب جایگزین مشابه ROY و ZCCN برای ترکیب با چهارچوب پیشنهادی بهره ببرند. به علاوه، استفاده از یک ساختار امتیازدهی مبتنی بر چندویژگی[۳۲] (به جای روابط ۹ و ۱۰) نیز می‌تواند گام مهمی در بهبود این روش باشد. همچین استفاده از این الگو به همراه شبکه سیامی در ساختارهایی نظیر بخش تصحیح املایی معرفی شده در [۳۱] نیز در دستور کارهای آتی قرار خواهد داشت.

سپاسگزاری

این مقاله با پشتیبانی مالی و علمی موتور جستجوی پارسی‌جو به انجام رسیده است و تمام محاسبات مربوط به پردازش دادگان و آموزش شبکه‌های عصبی در این مقاله بر روی بستر محاسباتی توزیع شده این پروژه به اجرا درآمده است. نویسندگان مایلند مراتب قدردانی خود را از اعضای تیم پارسی‌جو و به طوراخص خانم‌ها مهدیه فلاح، سودابه زارع‌زاده، شادی عطارها و آقای امین رئیس‌زاده ابراز کنند.

مراجع

[1] C. Carpineto and G. Romano, “A Survey of Automatic Query Expansion in Information Retrieval,” ACM Comput. Surv., vol. 44, no. 1, pp. 1:1–1:50, Jan. 2012.

[۲] رضا خدایی، محمدعلی بالافر، سیدناصر رضوی، «اثربخشی بسط پرس‌وجو مبتنی بر خوشه‌بندی اسناد شبه بازخورد با الگوریتم K-NN»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۶، شماره ۱، صفحات ۱۴۳-۱۵۱، ۱۳۹۵.

[3] S. A. Takale and S. S. Nandgaonkar, “Measuring Semantic Similarity Between Words Using Web Search Engines,” in Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, 2007, pp. 757–766.

[4] K. Gulordava and M. Baroni, “A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus,” in Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, Stroudsburg, PA, USA, 2011, pp. 67–71.

[۵] فاطمه کاوه یزدی، علی‌محمّد زارع بیدکی، محمدرضا پژوهان، «تعیین مشابهت معنایی به روش بدون سرپرست با استفاده از قدم‌زنی تصادفی بر گراف جایگزینی زبانی»، دوره ۴۸، شماره ۱، صفحات ۲۳۷-۲۴۹، ۱۳۹۷.

[6] H. Bast, B. Buchhold, and E. Haussmann, “Semantic Search on Text and Knowledge Bases,” Found. Trends Inf. Retr., vol. 10, no. 1, pp. 119–271, 2016.

[7] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, “Probabilistic Query Expansion Using Query Logs,” in Proceedings of the 11th International Conference on World Wide Web, Honolulu, Hawaii, USA, 2002, pp. 325–332.

[8] S. V. Pantazi, “Unsupervised grammar induction and similarity retrieval in medical language processing using the Deterministic Dynamic Associative Memory (DDAM) model,” J. Biomed. Inform., vol. 43, no. 5, pp. 844–857, Oct. 2010.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in Advances in Neural Information Processing Systems (NIPS’ 13), Lake Tahoe, Nevada, 2013, pp. 3111–3119.

[10] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global Vectors for Word Representation,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14), Doha, Qatar, 2014, pp. 1532–1543.

[11] D. Roy, D. Paul, M. Mitra, and U. Garain, “Using Word Embeddings for Automatic Query Expansion,” in SIGIR Workshop on Neural Information Retrieval, Pisa, Italy, 2016, pp. 1–5.

[12] S. Kuzi, A. Shtok, and O. Kurland, “Query Expansion Using Word Embeddings,” in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM ’16), Indianapolis, Indiana, USA, 2016, pp. 1929–1932.

[13] D. Ganguly, D. Roy, M. Mitra, and G. J. F. Jones, “Word Embedding based Generalized Language Model for Information Retrieval,” in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’15), Santiago, Chile, 2015, pp. 795–798.

[14] S. Balaneshin-kordan and A. Kotov, “Embedding-based Query Expansion for Weighted Sequential Dependence Retrieval Model,” in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’17), Shinjuku, Tokyo, Japan, 2017, pp. 1213–1216.

[15] H. Zamani and W. B. Croft, “Estimating Embedding Vectors for Queries,” in Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR ’16), Newark, Delaware, USA, 2016, pp. 123–132.

[16] H. Zamani and W. B. Croft, “Relevance-based Word Embedding,” in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’17), Shinjuku, Tokyo, Japan, 2017, pp. 505–514.

[17] G. Zheng and J. Callan, “Learning to Reweight Terms with Distributed Representations,” in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 2015, pp. 575–584.

[18] M. ALMasri, C. Berrut, and J.-P. Chevallet, “A Comparison of Deep Learning Based Query Expansion with Pseudo-Relevance Feedback and Mutual Information,” in Advances in Information Retrieval, Padua, Italy, 2016, pp. 709–715.

[19] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi, “Integrating and Evaluating Neural Word Embeddings in Information Retrieval,” in Proceedings of the 20th Australasian Document Computing Symposium (ADCS ’15), Sydney, Australia, 2015, pp. 12:1–12:8.

[20] F. C. Fernández-Reyes, J. Hermosillo-Valadez, and M. Montes-y-Gómez, “A Prospect-Guided global query expansion strategy using word embeddings,” Inf. Process. Manag., vol. 54, no. 1, pp. 1–13, Jan. 2018.

[21] Q. Liu, H. Huang, J. Lut, Y. Gao, and G. Zhang, “Enhanced word embedding similarity measures using fuzzy rules for query expansion,” in IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 2017, pp. 1–6.

[22] N. Rekabsaz, M. Lupu, A. Hanbury, and H. Zamani, “Word Embedding Causes Topic Shifting; Exploit Global Context!,” in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’17), Shinjuku, Tokyo, Japan, 2017, pp. 1105–1108.

[23] Z.-H. Zhou, “A brief introduction to weakly supervised learning,” Natl. Sci. Rev., vol. 5, no. 1, pp. 44–53, 2018.

[24] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a Similarity Metric Discriminatively, with Application to Face Verification,” in Proceedings of the IEEE Computer Society Conference on Computer

[25] J. Mueller and A. Thyagarajan, "Siamese Recurrent Architectures for Learning Sentence Similarity," in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Terrigal, NSW, 2016, pp. 2786–2792.

[26] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006, vol. 2, pp. 1735–1742.

[27] E. Darrudi, H. Baradaran-Hashemi, A. AleAhmad, A.-M. Zareh-Bidoki, A.-H. Habibian, F. Mahdikhani, A. Shakery, and M. Rahgozar, "dorIR collection for Persian web retrieval," DBRG-TR-138702, Tech. Report for Iran Telecommunication Research Center, Tehran, Iran, 2008.

[28] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," Knowledge-Based Syst., vol. 22, no. 5, pp. 382–387, Jul. 2009.

[29] R. T.-W. Lo, B. He, and I. Ounis, "Automatically Building a Stopword List for an Information Retrieval System," J. Digit. Inf. Manag., vol. 3, no. 1, pp. 3–8, 2005.

[30] L. Geng and H. J. Hamilton, "Interestingness Measures for Data Mining: A Survey," ACM Comput. Surv., vol. 38, no. 3, 2006.

[31] F. Kaveh-Yazdy and A.-M. Zareh-Bidoki, "Aleph or Aleph-Maddah, That is the Question! Spelling Correction for Search Engine Autocomplete Service," in The 4th International eConference on Computer and Knowledge Engineering (ICCKE'14), Mashhad, Iran, 2014, pp. 1-10.

زیرنویس‌ها

[1] Term Mismatch
[2] Query Expansion (QE)
[3] Term Substitution
[4] Global QE
[5] Local QE
[6] Top Retrieved Documents
[7] Topic Drift
[8] Online
[9] Searchers
[10] Word Embedding
[11] Maximum Likelihood Estimation (MLE)
[12] Multi-feature Regression Model
[13] Translation Language Model
[14] Semantic Relatedness
[15] Relevance Relatedness
[16] Vocabulary
[17] Relevance Relatedness
[18] Cut Threshold
[19] Long Short-Time Memory (LSTM)
[20] Contrastive loss
[21] Non-stopword Term
[22] Term-based QE
[23] Query-based QE
[24] Cross-fold Validation
[25] Keeping Ratio
[26] Training Epochs
[27] Five-Fold Cross Validation
[28] Precision at N
[29] Mean Average Precision (MAP)
[30] Mean Reciprocal Rank (MRR)
[31] Labeling Precision
[32] Multi-Feature