یک روش کلی برای معناکردن کلمه

ابهامزدایی در ویکیپدیا

چکیده. در این مقاله ما یک چارچوب کلی برای ابهامزدایی مفهوم کلمه با استفاده از دانش نهفته در ویکیپدیا پیشنهاد می‌کنیم. الی‌الخصوص، از مجموعه متون غنی و در حال رشد ویکیپدیا به منظور دستیابی به مخزن دانش بزرگ و قوی متشکل از عبارات کلیدی ها و مباحث منتخب مرتبط با آن‌ها بهره‌برداری می‌نماییم. عبارات کلیدی عمدتاً از عناوین مقالات ویکیپدیا و متون مرجع مرتبط با لینک‌های ویکی مشتق شده است. ابهامزدایی از عبارات کلیدی هم بر اساس عمومیت موضوع منتخب و هم ارتباط وابسته به متن است که در آن اطلاعات متنی غیرضروری (و به طور بالقوه مختل‌کننده) حذف شده‌اند. ما با ارزیابی‌های گسترده تجربی با استفاده از مقیاس‌های مختلف ارتباطی، نشان می‌دهیم که روش پیشنهادی به دقت ابهامزدایی قابل مقایسه‌ای نسبت به تکنیک های پیشرفته، دست می‌یابد، در حالی که مقدار هزینه محاسبه کمتری را متحمل می‌شود.

کلمات کلیدی: ابهامزدایی مفهوم کلمه، ویکیپدیا، حذف بخش‌های اضافه متن

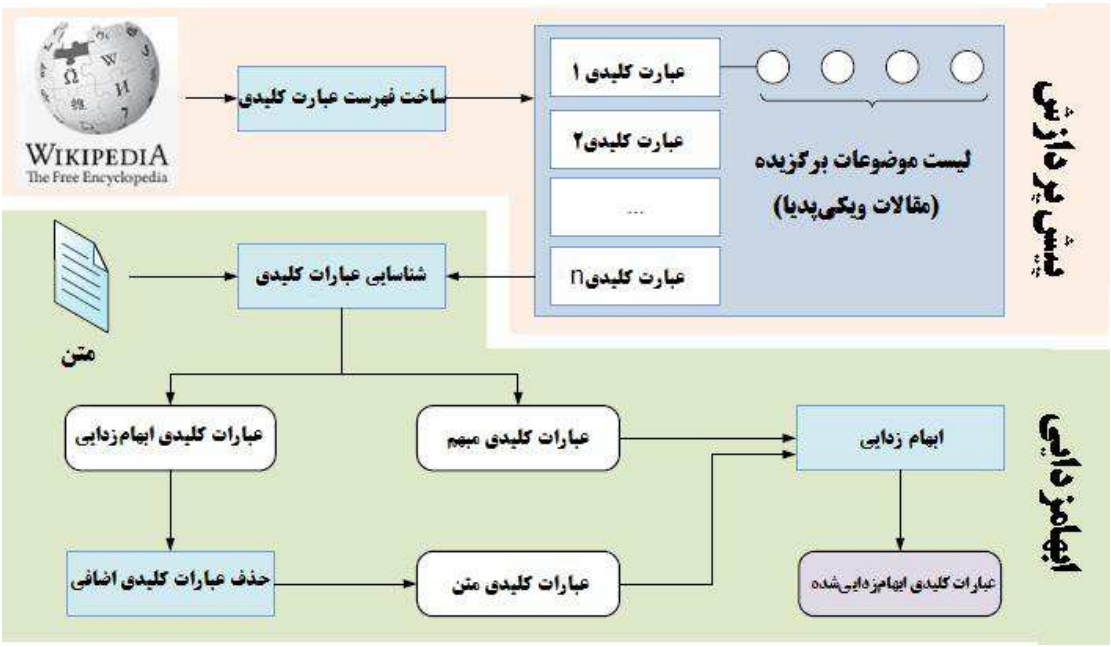
1 مقدمه

ابهامزدایی مفهوم کلمه (WSD) مسئله شناسایی مفهوم (معنی) یک کلمه را در یک متن خاص است. در زندگی روزمره ، ذهن ما ناخودآگاه یک کلمه مبهم را بر اساس متنی که در آن بروز می‌یابد به معنای مناسب ربط می‌دهد. از اینرو در پردازش زبان طبیعی، ابهامزدایی مفهوم کلمه عمل خودکار تعیین معنای یک کلمه با توجه به متن(های) مربوطه است. این یک کار پیچیده اما اساسی در بسیاری از حوزه‌ها از قبیل تشخیص موضوع و نمایه‌سازی، عناصر هم مرجع بین اسناد [2، 18]، و جستجوی وب توسط افراد است. با توجه به رشد روبه افزایش اطلاعات و مضامین آنلاین، روش ابهامزدایی کارآمد و با کیفیت با مقیاس پذیری بالا از اهمیت حیاتی برخوردار است.

دو رویکرد اصلی را می توان در تحقیقات گذشته یافت که درصدد پرداختن به موضوع، یعنی روش‌های مبتنی بر دانش و روش‌های آموزش ماشینی نظارت هستند. رویکرد اول عمدتاً بر فرهنگ‌های لغت، اصطلاحنامه، و یا

پایگاه‌های دانش لغوی، مانند، فهرست مفاهیم متشکل از کلمات / عبارات و تعاریف معـانی ممکـن آن‌هـا متکـی است. الگوریتم لسک[1] یک الگوریتم اصلی از چنین نوعی می‌باشد، با این فرض که واژه‌های اشاره‌کننده به معـانی یکسان با کلمات مجاور خود در یک موضوع مشترک هستند. به دنبال این ایـده، بسـیاری از تحقیقـات درصـدد شناسایی معنای صحیح برای یک کلمه با حداکثر توافق بین تعاریف فرهنگ لغت و اصطلاحات ضـمنی از کلمـه مبهم هستند. در فرایند ابهامزدایی، یک فهرست معانی با کیفیت بالا فاکتور بسیار مهمی اسـت کـه بـر عملکـرد تأثیر می‌گذارد.

با این حال، ساخت چنین منابع لغوی در مقیاس بزرگ، قابل خواندن با ماشین، خسته کننده و پر زحمت است. بنابراین، تنگنای کسب دانش مشکل اصلی در محدود کردن عملکرد چنین سیستمی است.روش دوم مبتنـی بـر تلاش برای یادگیری ماشینی نظارت به منظور استخراج مجموعـه ای از ویژگـی‌هـای متنـی محلـی و جهـانی از مجموعه داده های دستی معانی برچسب گذاشته شده و یکپارچه‌سازی نمونه های آموزشی در یک طبقـه‌بنـدی یادگیری ماشینی است. بسیاری از تکنیک‌های یادگیری ماشینی برای WSD (ابهامزدایی مفهوم کلمه) بـه کـار گرفته شده‌اند، و نشان داده شده که موفق بوده‌اند. با این حال، روش‌های یادگیری ماشینی بیش از حد متحمـل تنگنای کسب دانش زیرا آنها به مقادیر قابل توجهی از نمونه های آموزشی نیاز دارند.



شکل 1.چارچوب ابهامزدایی عبارات کلیدی بر اساس ویکیپدیا.

[1] .Lesk

در این مقاله، ما یک روش کلی برای کاوش در استفاده از ویکیپدیا به عنوان منبع واژگانی به منظور ابهامزدایی مطرح می‌سازیم. ویکیپدیا، بزرگترین دانشنامه و مخزن دانش آنلاین مشارکتی در جهان و دارای بیش از 3.2 مگابایت مقاله صرفاً به زبان انگلیسی است. ویکی‌پدیا با یک گستردگی منطقی شمول جامعی از موضوعات، در مقایسه با بسیاری از دیگر پایگاه‌های دانش فراهم می‌کند. مطالعات قبلی نشان می‌دهد که کیفیت مقاله‌های ویکیپدیا با دانشنامه سردبیر قابل مقایسه است. ویکیپدیا به خاطر مقیاس گسترده همکاری و همچنین کاربرد خود در سال های اخیر به یک منبع مثمر ثمر در بسیاری از زمینه های تحقیقاتی تبدیل شده است.

چارچوب ابهامزدایی مطرح شده در شکل 1 نشان داده شده است. سه مؤلفه اصلی، فهرست ویکیپدیا، شناسایی عبارات کلیدی و از بین بردن عبارات کلیدی اضافی و ابهام‌زدایی ویکی‌پدیا، در تحقیق ما برای ابهامزدایی شرح و بسط داده شده‌اند. به طور خاص، ما یک فهرست از مفهوم کلمه با استخراج کلمات چند معنایی، مترادف و فراپیوند کد گذاری شده در ویکیپدیا می‌سازیم. هر مدخل در فهرست یک عبارت کلیدی است که حداقل به یک مقاله ویکیپدیا اشاره دارد. در بخش 3.1 به تفضیل، عبارت کلیدی هم یک عنوان مقاله در ویکیپدیا هستند، و هم به صورت ظاهری (یا متون مرجع) از لینک ویکی‌پدیا آمده‌اند. این عبارات کلیدی، که هر یک دقیقا به یک مقاله ویکیپدیا اشاره دارد، عبارات کلیدی بدون ابهام هستند. بعضی عبارات کلیدی مبهم هستند که هر یک از آنها به مقاله‌های چندگانه ویکیپدیا اشاره دارند (یعنی، موضوعات / مفاهیم منتخب، که در شکل 1 نشان داده شده است).

با توجه به یک متن، عبارات کلیدی بدون ابهام شناخته شده از متن به عنوان اطلاعات متنی برای ابهام‌زدایی از عبارات کلیدی مبهم هستند. در این میان، از بین بردن عبارات کلیدی اضافه به شناسایی عبارات کلیدی مهم در متن که به صورت عبارت کلیدی مبهم معین وقوع یافته به ابهامزدایی کمک می کند ، و تا حد زیادی موارد مختل کننده را فیلتر نموده و کارایی سیستم را بهبود می‌بخشد. این ابهام‌زدایی جزء اصلی چارچوب ماست. هدف از آن تعادل توافق بین مضمون عبارت کلیدی مبهم و مضمون هر مفهوم منتخب است.

ارزیابی تجربی بر اساس مجموعه داده‌های مبتنی برحقیقت نشان می‌دهد که روش ما هم از نظر اثربخشی و هم بهره‌وری بهتر از روش‌های پیشرفته دیگر، است. علاوه بر این، چون فهرست ویکی‌پدیا که ما ایجاد می‌کنیم متکی بر اطلاعات غنی معنایی موجود در ویکیپدیا است، رویکرد ما تنگنای کسب دانش سنتی اجتناب نموده و برای هر

دامنه در اندازه های مختلف قابل اجرا است. این روش می‌تواند به تحقیقات موجود که به بررسی ابهامزدایی مفهوم کلمه و همچنین کاربردهای بالقوه نیاز دارد، مرتبط شود.

رویکرد ما در چندین مفهوم به طور کلی کافی است: با توجه به شمول جامع‌تر مباحث ویکیپدیا، فهرست ویکیپدیا دارای دامنه مستقلی است، و با توجه به محبوبیت رو به رشد ویکیپدیا در زبان‌های دیگر، رویکرد ما می‌تواند به آسانی در همه زبان‌های مختلف مورد استفاده مجدد قرار گیرد. و در نهایت، چارچوب های مدولار امکان استفاده از مقیاس‌های ارتباطی مختلف متناسب با نیازهای کاربردی مختلف را فراهم می‌آورد.

سایر مطالب این مقاله، به شرح زیر است: بخش 2 پژوهش‌های مرتبط را بررسی می‌کند. بخش 3 رویکرد مارا همراه با اجزای منحصر به فرد در چارچوب پیشنهادی معرفی می‌نماید. در بخش 4، ما نتایج تجربی را ارائه داده و مورد بحث قرار می‌دهیم. در نهایت، ما بخش 5 نتیجه‌گیری می‌نماییم.

2 پژوهش‌های مرتبط

بسیاری از پژوهش‌های اخیر ویکیپدیا را برای افزایش وظایف استخراج متن، مانند مقیاس ارتباط معنایی، طبقه بندی و دسته‌بندی متن، و تشخیص موضوع مورد کاوش قرار داده‌اند. در میان این پژوهش‌ها، آثار مربوط که شامل ابهامزدایی مفهوم کلمه و مقیاس‌های ارتباط معنایی می‌باشند را، بررسی می‌کنیم.

استراب و پونزتو ویکیپدیا را برای سنجش ارتباط معنایی مورد استفاده قرار دادند. روش آنها در مقالات ویکیپدیا که شامل کلمه خاص در عناوین خود بودند جستجو می‌نمود و با در نظر گرفتن مقیاس طول مسیر در سلسله مراتب رده‌بندی ویکیپدیا، همپوشانی متن، و همچنین احتمال بروز آنها . میلنه و ویتن یک مقیاس وزنی سبک را بر اساس ارتباط معنایی لینک‌های ویکیپدیا ایجاد نمودند، که مقیاس مبتنی بر لینک ویکیپدیا (WLM) نامیده می‌شد. اولاً آن‌ها مقالات ویکیپدیا را که با عبارت مورد نظر ارتباط داشت را شناسایی نموده و پس از آن، ارتباط دو عبارت را با مقالات ذکر شده در ویکیپدیا به شرح زیر است، محاسبه نمودند:

$$relatedness(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \quad (1)$$

که در آن a و b دو مقاله ویکیپدیا هستند، A و B مجموعه‌ای از همه مقالات ویکیپدیا که به ترتیب به a و b مرتبط هستند، و W مجموعه‌ای از کل مقالات ویکیپدیا است. با توجه به دقت بالا و هزینه کم این روش، معمولاً

در پژوهش‌های موجود استفاده می‌شود. در پژوهش ما، WLM را به عنوان یک گزینه برای محاسبه ارتباط معنایی کارامد بین دو مقاله ویکیپدیا به کار می‌گیریم.

چون این روش بر فراپیوند بین مقالات ویکی‌پدیا متمرکز است، ما همچنین مقیاس‌های دیک و جاکارد را در فراپیوند پژوهش خود بررسی می کنیم.

ویکی‌فای! [۲] در صدد تفسیر عبارات کلیدی در یک متن با موضوعات ویکیپدیا است که در آن ابهامزدایی عبارات کلیدی یک گام اساسی است. هم الگوریتم های مبتنی بر دانش و هم الگوریتم‌های مشتق شده از داده‌ها در ویکی‌فای مورد استفاده قرار می‌گیرند. روش مبتنی بر دانش، از الگوریتم لسک، با بهره گیری از ظهور عبارات کلیدی مبهم و اطلاعات متنی الهام گرفته است. با این حال، یک روش مستقل از روش پایه مورد استفاده برای رایج ترین مفاهیم بدتر عمل می‌کند. روش مبتنی بر داده به طبقه بندی‌کننده تعدادی ویژگی می‌آموزد، از جمله به عنوان ادات سخن و کلمات محلی متنی. آنها سپس هر دو الگوریتم را با استفاده از طرح رای‌گیری ترکیب می‌نمایند. به حائز اهمیت است که، روش از نظر محاسباتی گران است زیرا یک بردار ویژگی آموزشی برای هر عبارت کلیدی مبهم از بین همه عبارات وقوع یافته در کل ویکیپدیا استخراج می‌نماید.

مدلیان و همکارانش. هر دو مقیاس ارتباطی و دارای عمومیت را مورد استفاده قرار دادند. برای یک متن خاص، همه عبارات کلیدی، که هر یک به صورت منحصر به فرد برای یک موضوع از ویکیپدیا نشان داده شده است، به عنوان بافت شناخته می‌شوند. سپس بافت برای ابهام زدایی از عبارات کلیدی که هر یک می‌تواند در بیش از یک موضوع ویکیپدیا وجود داشته باشد، استفاده می‌شود. در پژوهش آن‌ها، ارتباط با بافت برای هر موضوع منتخب از عبارت کلیدی مبهم توسط WLM محاسبه شده است. برای موضوع منتخب t، عمومیت برای عبارت کلیدی معین k احتمال پیشینی عبارت کلیدی k با اشاره به موضوع منتخب t است، یعنی، $P(t|k)$. با این دو مقیاس، یک امتیاز برای هر موضوع منتخب t برای عبارت کلیدی معین k با استفاده از معادله زیر محاسبه می شود.

$$Score(t,k) = \frac{\sum_{c \in C} relatedness(t,c)}{|C|} \times P(t|k) \quad (2)$$

[2] . Wikify!

در این معادله C نشاندهنده بافت عبارت کلیدیk *است*. مشاهده شده است که همه عبارات کلیدی بافت در پژوهش آن‌ها به طور یکسان عمل می‌کنند. با ارزیابی100 مقاله ویکیپدیا مشخص شد، روش پیشنهادی که رایج‌ترین روش استخراج مفهوم با معناداری 2.4 درصدی در مقیاس-F است، عملکرد بهتری دارد.

به طور طبیعی، برخی از عبارات کلیدی بیشتر به مضمون متن مربوط هستند تا موارد دیگر، به ویژه هنگامی که یک متن مباحث متعددی را پوشش می‌دهد. میلن و ویتن سنجش عبارات کلیدی مضمون متن را بر اساس ارتباط آن‌ها با یکدیگر و همچنین عبارت کلیدی بودن آن‌ها مطرح ساختند. به طور خاص، اگر مضمون متن منسجم باشد، و پس مقیاس ارتباط مناسب‌تر می‌شود، در حالی که عمومیت هنگامی که مضمون متن متنوع‌تر باشد، مفیدتر است. مطالعه تجربی آنها نشان می‌دهد که طبقه بندی C4.5 عملکرد بهتری نسبت به روش مدلیان و همکارانش به دست می‌دهد.

در حالی که پژوهش‌های مدلیان و همکارانش، میلنه و ویتن تا به امروز به عملکرد امیدبخشی در میان روش‌های موجود دست یافته‌اند، آنها بر ارتباط مضمون متن با در نظر گرفتن تمام عبارات کلیدی بدون ابهام در متن مورد بررسی، که کارایی ندارد، تکیه می‌کنند. همانگونه که یک متن اغلب شامل برخی موارد مختل‌کننده، یعنی، صفحات وب است، تمام عبارات کلیدی بدون ابهام برای بیان این موضوع از متن به یک اندازه مفید نیستند ، و برخی از آنها حتی ممکن است دقت کمتری علاوه بر اتلاف منابع محاسباتی داشته باشند. اگر چه میلن و ویتن یک طرح سنجش را برای برجسته ساختن عبارات کلیدی مبتنی بر معنا به کار برده‌اند، لاجرم متحمل هزینه‌های اضافی می‌باشند.

در این پژوهش، ما طرح حذف موارد اضافه با برداشتن مهم‌ترین عبارات کلیدی برای پردازش بیشتر را به کار برده‌ایم. این مرحله غیربدیهی عبارات کلیدی سطحی را فیلتر نموده و به طور قابل توجهی موارد اضافی مختل‌کننده را کاهش می‌دهد، وکه هم منجر به بهره وری و هم دقت بهتر می‌شود. علاوه بر این، روش‌های موجود با استفاده از مقیاس ارتباطی خاص تعریف شده و مورد بررسی قرار گرفته است. در اینجا، ما یک الگوریتم کلی را مورد بحث قرار می‌دهیم که می تواند با مقیاس‌های مختلف ارتباطی تطبیق یابد.

3. چارچوب ابهامزدایی

در این بخش، ما توصیف دقیقی از سه جزء اصلی تحقق را به منظور دستیابی به ابهامزدایی، یعنی فهرست ویکیپدیا، شناسایی عبارات کلیدی و حذف موارد اضافه عبارات کلیدی، و ابهام زدایی به ترتیب در آن توالی ارائه می‌دهیم، زیرا متناسب با ترتیب کاربرد آنها در چارچوب ما هستند.

1.3 فهرست ویکیپدیا

فهرست ویکیپدیا شامل عبارات کلیدی ها و مباحث منتخب مرتبط با آن‌ها می‌باشد. عبارات کلیدی از دو منبع سرچشمه می‌گیرند، یعنی، عنوان مقاله ویکیپدیا و متون مرجع لینک‌های ویکی‌پدیا.

در ویکیپدیا، هر مقاله به شرح یک موضوع واحد می‌پردازد و با استفاده از نامی است اغلب برای اشاره به موضوع 1 مورد استفاده قرار می‌گیرد، عنوان‌گذاری می‌شود. از این رو، عناوین مقالات ویکیپدیا شامل فهرست موجودی ویکیپدیای ما مانند عبارات کلیدی است، که هر یک از آنها به مقاله مربوط به ویکیپدیا به عنوان موضوع 2 منتخب آن اشاره دارد. توجه داشته باشید که، صفحات ویکیپدیا برای اهداف اجرایی و یا پشتیبانی (به عنوان مثال، بحث، گفتگو، صفحات کاربری)، حذف شده، اما صفحات تغییر مسیر در آن گنجانده می شوند. صفحه تغییر مسیر در ویکیپدیا عنوان صفحه مقاله مورد نظر را با عنوان ارجح‌تر، با توجه به دو عنوان اشاره کننده به موضوع یکسان عوض می‌نماید. چنین تغییر مسیری می تواند ما را در پرداختن به کلمات مترادف (نام‌های جایگزین)، کلمات مخفف، کلمات دارای تنوع املایی، و غلط املائی کمک نماید. به طور طبیعی، مقاله تغییر مسیر مورد نظر موضوع منتخب برای عنوان صفحه تغییر مسیر به عنوان یک عبارت کلیدی در فهرست ویکی‌پدیا است.

بر اساس سیاست ویکیپدیا، لینک‌های ویکی (یا فراپیوند) باید در ویکیپدیا برای مباحث مربوط به مقاله، اصطلاحات فنی ذکر شده، و یا برای نام‌های خاص که به احتمال زیاد برای خوانندگان 3 ناآشنا هستند، ایجاد شوند. به این ترتیب، متون مرجع و مقالات لینک‌شده فراپیوند ارتباطات معنایی ساخته شده توسط خرد جمعی مشارکت‌کنندگان در ویکیپدیا هستند. توجه داشته باشید که، متن مرجع شکل ظاهری یک لینک که همیشه ممکن است با عنوان مقاله مرتبط نباشد.

از این رو متون مرجع فهرست عبارات کلیدی را تا حد زیادی توسط کلمات چندمعنایی، و روابط مشارکتی و اجتماعی منعکس شده توسط آنها، غنی می‌سازند. متن مرجع و مقاله مرتبط با آن در فهرست ویکیپدیای ما به ترتیب به عنوان عبارت کلیدی منتخب و موضوع آن اضافه شده است.

صفحات ابهامزدایی ویکیپدیا برای ابهام‌زدایی از تعدادی موضوعات مشابه که ممکن است با یک اصطلاح مبهم به آن اشاره شده باشد، طراحی شده‌اند. عناوین چنین صفحاتی به طور معمول یکی از اصطلاحات مبهم است، که با ابهامزدایی از برچسب آن دنبال می‌شود. مباحث منتخب در یک صفحه و هر کدام با یک توضیح کوتاه در مورد آن ذکر شده است. ما روش اکتشافی توراداکف و ولیخف را برای استخراج موضوعات منتخب از هر صفحه ابهامزدایی اتخاذ نموده‌ایم. هنگامی که یک واژه مبهم قبلاً در فهرست به عنوان یک عبارت کلیدی وجود داشته باشد، ما لیست مووضوعات منتخب آن‌را با آنهایی که از صفحه ابهامزدایی مربوطه استخراج شده به روز رسانی می‌کنیم.

به طور خلاصه، فهرست عبارات کلیدی ویکیپدیا با در نظر گرفتن عنوان مقاله ویکیپدیا، پردازش صفحات تغییر مسیر داده، تجزیه صفحات ابهامزدایی و استخراج لینک‌های مافوق ایجاد شده است. در این فهرست، اگر یک عبارت کلیدی دقیقا با یک موضوع (یا مقاله) در ارتباط باشد، ما آن را عبارت کلیدی بدون ابهام می‌نامیم. یک عبارت کلیدی مبهم با بیش از یک موضوع در ارتباط است.

2.3 شناسایی و حذف عبارات کلیدی اضافه

ما متن ورودی را تجزیه کرده و تمام عبارات کلیدی که در فهرست نیز موجودند ، با اولویت دادن به آن‌هایی که قدیمی‌ترند استخراج می‌نماییم. به عنوان مثال، با توجه به جمله‌ی" دریای جاوه..... است"، ما دریای جاوه را به جای جاوه استخراج می‌کنیم. برای از عبارات کلیدی بدون ابهام استخراج شده، مباحث ویکیپدیا مرتبط با آن‌ها به طور مستقیم از فهرست به دست آمده است. این موضوعات ویکیپدیا به درک با موضوعات تحت پوشش متن کمک نموده، و زمینه برای تعیین عبارات کلیدی مبهم استخراج شده، فراهم می‌سازد.

با این حال، یک متن ممکن است موضوعات بسیار متنوعی را پوشش دهد. بنابراین، همه عبارات کلیدی بدون ابهام اهمیت یکسانی برای ابهامزدایی ندارند. در حالی که عبارات کلیدی مرتبط می‌توانند به شناسایی مفهوم صحیح یک عبارت کلیدی مبهم، و نامربوط که ممکن است به دقت و صحت ابهامزدایی صدمه زده و متحمل

هزینه‌های محاسباتی بیشتری شوند، کمک نماید. این امر نیازمندیک طرح حذف عبارات اضافی مناسب برای اثر بخشی و بهره وری است.

ما از مقیاس عبارات کلیدی برای تعیین کمیت اهمیت یک عبارت کلیدی استفاده می‌کنیم. برای یک عبارت کلیدی بدون ابهام مشخص، عبارت کلیدی بودن احتمال قیاسی است که در آن یک عبارت کلیدی به عنوان متن مرجع استفاده می‌شود، بدون توجه به اینکه آن کلمه در کجا بروز یافته باشد. بر اساس این مقیاس، ما عبارات کلیدی بالای M را انتخاب می‌کنیم که بالاترین ارزش عبارت کلیدی بودن را برای شکل عبارات کلیدی مضمون متن دارند. سپس عبارات کلیدی مبهم مشخص شده از متن با استفاده از عبارات کلیدی مضمون متن ابهامزدایی می‌شوند. ما در آزمایش‌های خود، باید تأثیر M را در مورد اثربخشی و بهره‌وری ابهامزدایی مورد ارزیابی قرار دهیم.

3.3 ابهامزدایی

برای یک عبارت کلیدی k مبهم داده شده، همه عبارات کلیدی مضمون متن دارای اهمیت یکسانی نیستند، همانگونه که برخی از آنها از نظر معنایی به k بیش دیگران مربوط هستند. به عنوان مثال، عبارت کلیدی آلبرت اینشتین در جستجوی گوگل 4 در مقاله ویکیپدیا به عنوان نمونه‌ای برای معرفی ویژگی‌های *گوگل دودل*[3] به نظر می رسد.

بدیهی است که ارتباط بسیار کمی (در صورت وجود) در بین نبوغ در علم و بزرگی موتور جستجو وجود دارد. با این حال، با توجه به ارزش بالای عبارت کلیدی بودن، آلبرت انیشتین اغلب به عنوان یکی از عبارات کلیدی متن زمینه انتخاب شده است.

از آنجا که هر عبارت کلیدی (یا یکی از موضوعات منتخب آن) به یک مقاله ویکیپدیا اشاره دارد، بنابراین محاسبه ارتباط بین دو عبارات کلیدی (یا موضوعات منتخب) می‌تواند به مسئله محاسبه ارتباط بین مقالات ویکیپدیای وابسته تنزل یابد. مقیاس‌های کمی در تحقیقات گذشته برای سنجش ارتباط معنایی بین دو مقاله ویکیپدیا گزارش داده شده‌اند، که به طور عمده مبتنی بر لینک‌های ویکی، مانند مقیاس‌های دیک، جاکارد، و WLM هستند (به بخش 2نگاه کنید). به عنوان یک چارچوب عمومی، روش پیشنهادی ما می‌تواند از هر مقیاسی

[3] . Google Doodle

استفاده نماید و در بحث زیر ما از ارتباط $(k; k0)$ برای اختصاص ارتباط بین دو عبارات کلیدی k و $k0$ (و یا موضوع منتخب t) استفاده می‌کنیم.

به یاد داشته باشید که یک متن ممکن است بسیاری از موضوعات متنوع را پوشش دهد، که اغلب عبارات زمینه متن M منعکس شده‌اند. به این معنا که برخی از عبارات متن از M ممکن است تا حد زیادی به عبارات دیگر مضمون متن مرتبط باشند. عبارات کلیدی مضمون متن با روابط آن‌ها با دیگر عبارات کلیدی سنجیده می‌شوند، همانگونه که در معادله زیر نشان داده شده است. ، در این معادله C بیانگر مجموعه‌ای از عبارات کلیدی زمینه متن و $|C| \leq M$ است.

$$Weight(k, C) = \frac{\sum_{k' \in C \setminus k} Relatedness(k, k')}{|C| - 1} \tag{3}$$

با معیار سنجش تعریف شده، ارتباط بین موضوع منتخب t با کل متن زمینه C در معادله 4 محاسبه شده است. شباهت متنی مشابهی در سایر روش‌ها اقتباس شده است.

$$Relatedness(t, C) = \frac{\sum_{k \in C} Weight(k, C) \times Relatedness(t, k)}{\sum_{k \in C} Weight(k, C)} \tag{4}$$

همانگونه که در بخش 2 بحث شد، عمومیت، احتمال قیاس یک عبارت کلیدی با اشاره به یک موضوع خاص است. تحقیقات موجود قبلاً کارایی مقیاس عمومیت را نشان داده‌اند. ما در چارچوب خود، ارتباط و عمومیت را با استفاده از عامل نمایی C متعادل می‌سازیم. با توجه به عبارت کلیدیk برای ابهامزدایی، باید Ck مجموعه‌ای از موضوعات منتخب k باشد. ما موضوع t را به عنوان موضوع ابهامزدایی به k اختصاص می‌دهیم که هم ارتباط و هم عمومیت را با پارامتر C که از قبل مشخص شده حداکثر می‌سازد، همانگونه که در معادله 5نشان داده شده است.

$$t_o = \arg\max_{t \in C_k} \left(Relatedness(t, C)^c \times P(t|k) \right) \tag{5}$$

بنابراین، چارچوب ما مستلزم دو پارامتر می‌باشد: M برای به اندازه چهار چوب و c برای ایجاد تعادل در ارتباط و عمومیت.M کوچکتر موضوعات مفیدتری را برای ابهامزدایی حفظ می‌کند و همچنین با خطر فیلتر مباحث

مفید بهره‌وری را بهبود می‌بخشد. ، از سوی دیگر، M بزرگتر ممکن است موضوعات مفیدتر و موارد مختل کننده بیشتری را بروز دهد، و قطعاً از نظر محاسباتی پرهزینه‌تر است. همانند عامل مقیاس‌گذاریc، این مقیاس موجب انعطاف پذیری تنظیم اثر مقیاس ارتباط بر اساس تعاریف ارتباطی مختلف می‌شود (به عنوان مثال، جاکارد و WLM). در بخش زیر، ما تاثیر دو پارامتر تجربی را نشان می‌دهیم.

5 نتیجه‌گیری

ابهامزدایی مفهوم کلمه یک مسئله اساسی برای پرداختن به بسیاری از کاربردها در زمینه‌های پردازش زبان طبیعی و بازیابی اطلاعات، و غیره است.

دانش گسترده و دارای کیفیت بالا در ویکیپدیا یک مخزن دانش با حوزه مستقل برای ابهامزدایی مفهوم کلمه را فراهم می‌سازد. در این مقاله، ما یک چارچوب کلی (که می‌تواند مقیاس‌های ارتباطی متنوع را تطبیق داده، و مستقل از حوزه، به طور بالقوه در زبان‌های دیگر استفاده شود) را برای استفاده از ویکیپدیا به منظور ابهامزدایی مفهوم کلمه/ عبارت کلیدی با استفاده از مقیاس‌های عمومیت و ارتباط مطرح ساخته‌ایم. ما نشان دادیم که حذف موارد غیرضروری یا مضامین مختل کننده بالقوه مرتبه بزرگی روند ابهامزدایی را سریع تر از روش‌های موجود می‌سازد، در حالی که به دقت ابهامزدایی نظیری (اگر نه بهتر) دست می‌یابد.