دانشکده مهندسی مهندسی کامپیوتر

کاربست یک الگوریتم پیونددهی موجودیت برای مسئله‌ی ابهام‌زدایی معنی واژگان در زبان فارسی

پایان‌نامه یا رساله برای دریافت درجه کارشناسی

در رشته مهندسی کامپیوتر گرایش نرم‌افزار- هوش

نام دانشجو

بنفشه کریمیان

استاد راهنما:

دکتر مینایی

تیر ماه ۱۳۹۸


بسم الله الرحمن الرحيم

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پایان‌نامه/رساله

نام دانشکده: دانشکده کامپیوتر

نام دانشجو: بنفشه کریمیان

عنوان پایان‌نامه یا رساله: کاربست یک الگوریتم پیونددهی موجودیت برای مسئله‌ی ابهام‌زدایی معنی واژگان در زبان فارسی

تاریخ دفاع: ۱۳۹۸/۳/۲۶

رشته: مهندسی کامپیوتر

گرایش: نرم‌افزار- هوش

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما	دکتر بهروز مینائی	دانش‌یار	دانشگاه علم و صنعت	
۲	استاد راهنما				
۳	استاد مشاور				
۴	استاد مشاور				
۵	استاد مدعو خارجی				
۶	استاد مدعو خارجی				
۷	استاد مدعو داخلی	دکتر سید صالح اعتمادی	استادیار	دانشگاه علم و صنعت	
۸	استاد مدعو داخلی				

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب بنفشه کریمیان به شماره دانشجویی ۹۴۵۲۱۱۸۹ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پایان‌نامه/رساله حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: بنفشه کریمیان

امضا و تاریخ: ۱۳۹۸/۳/۲۶

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه/ رساله برای همگان بلامانع است.

✓ بهره‌برداری از این پایان‌نامه/ رساله با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه/ رساله تا تاریخ ..... ممنوع است.

نام استاد یا اساتید راهنما:

دکتر بهروز مینائی

تاریخ:

امضا:

چکیده

ابهام‌زدایی معنای واژگان یکی از مهم‌ترین ارکان استخراج دانش از متن به‌شمار می‌آید، که در آن هر واژه‌ی یک متن به یک هم‌نشیم از یک وردنت پیوند داده می‌شود. این مسئله بسیار مشابه با مسئله‌ی پیونددهی موجودیت‌ها می‌باشد که در آن هر کلمه به یک موجودیت در گراف دانش یا پایگاه دانش پیونددهی می‌شود. در فارسی این دو مسئله با استفاده از وردنت فارسی (فارس‌نت) و گراف دانش فارسی (فارس‌بیس) پوشش داده شده‌اند. با این وجود، در حالی که یک رویکرد متحد برای ابهام‌زدایی معنای واژگان و پیونددهی موجودیت‌ها برای زبان انگلیسی ارائه داده شده است و گراف‌دانش انگلیسی به وردنت انگلیسی نگاشت شده است، در زبان فارسی این رویکرد متحد و نگاشت اتخاذ نشده است.

این پروژه بر آن است که نگاشت بین فارس‌نت و فارس‌بیس که خلأ آن در زبان فارسی موجود است را پوشش دهد. در این راستا، ابتدا بر اساس یک الگوریتم پیونددهی موجودیت (وابسته به فارس‌بیس) یک الگوریتم مشابه برای ابهام‌زدایی معنایی (بر اساس فارس‌نت) ارائه می‌شود. در ادامه، برای گسترش ابهام‌زدایی واژگان در زبان فارسی با الهام‌گیری از پروژه‌ی بیبل‌نت[1] به ترکیب توأمان فارس‌نت و فارس‌بیس می‌پردازیم؛ به این صورت که الگوریتم ابهام‌زدای ارائه شده را به طور توأمان با الگوریتم پیونددهی موجودیت مذکور بر یک مجموعه‌داده اعمال می‌نماییم، تا هر کدام از واژگان مجموعه‌داده به هم‌نشیم‌های فارس‌نت و موجودیت‌های فارس‌بیس پیوند داده شوند. سپس هم‌نشیم‌ها و موجودیت‌های متناظر به یک‌دیگر نگاشت شده و فراوانی این نگاشت اندازه‌گیری می‌شود. بر اساس این فراوانی، و با قراردادن یک آستانه[2] هم‌نشیم‌هایی که تعداد باهم‌آیی آن‌ها از مقدار آستانه بیشتر است را با هم ادغام می‌نماییم. در نهایت، با مقایسه با یک دادگان طلایی[3] معیارهای دقت و بازخوانی خروجی ارائه شده را محاسبه می‌نماییم. شایان ذکر است که دادگان طلایی نگاشت شده توسط یک شخص خبره تولید شده و شامل ۷۶۷ نگاشت است. معیار دقت، بازخوانی و تعداد نگاشت نهایی به‌دست آمده از این پروژه به ترتیب برابر با ۷۰ درصد، ۳۰ درصد و ۵۰۰۰ هم‌نشیم است.

واژه‌های کلیدی: ابهام‌زدایی معنایی، پیونددهی موجودیت، فارس‌نت، فارس‌بیس، بیبل‌نت

[1] Babelnet

[2] Threshold

[3] Gold data

فهرست مطالب

فهرست اشکال

فـصل ۱ :

مقدمه

۱-۱- مقدمه

زبان وسیله‌ی انتقال اطلاعات بین موجودات و نسل‌های مختلف است و این انتقال اطلاعات را می‌توان به دسته‌ی زبانی و غیر زبانی تقسیم کرد. پردازش زبان طبیعی[1] یک روش برای تحلیل و بررسی زبان متن و یکی از حوزه‌های فعال در هوش‌مصنوعی است [1]. در این میان فهم معنای یک متن از اهمیت بسزایی برخوردار است و برای مدت زمانی طولانی از اهداف مهم تحقیق‌های این حوزه و حوزه‌های مشابه بوده‌است [2]. دو مسئله‌ی مهم برای رسیدن به فهم معنای یک متن ابهام‌زدایی معنایی[2] و پیونددهی موجودیت[3]اند. این دو مسئله از مسائل شناخته‌شده در این حوزه‌اند [3].

هر واژه ممکن است معانی متفاوتی با توجه به محتوای مربوطه داشته باشد، که به آن پولیسمی[4] می‌گویند [4]. برای مثال واژه‌ی شیر در سه جمله‌ی زیر معانی متفاوتی دارد:

- شیر سلطان جنگل است.
- شیر آب باز بود.
- شیر در لیوان ریخته‌شده بود.

ابهام‌زدایی معنایی برای تشخیص معنای یک واژه در یک متن با توجه به محتوای متن از میان معانی مختلف آن واژه است و از مسائل سخت در حوزه‌ی هوش‌مصنوعی محسوب می‌شود. روش‌های بسیاری برای ابهام‌زدایی معنایی ارائه شده‌اند ولی با این حال این مسئله هنوز حل‌نشده محسوب می‌شود. با اینکه لغت‌نامه‌ها می‌توانند برای ابهام‌زدایی معنایی مناسب باشند، استفاده از وردنت پیشنهاد شده است [4].

پیونددهی موجودیت نام‌دار همان نگاشت کردن موجودیت‌های موجود در یک متن به موجودیت‌های یک پایگاه‌دانش است [5]. یک موجودیت نام‌دار ممکن است به چند موجودیت در پایگاه‌دانش با یک نام نگاشت شود، که موجب ابهام در پیونددهی می‌شود [6]. برای مثال، شیشه در ویکی‌پدیای فارسی، که نوعی پایگاه‌دانش است، به موجودیت‌های مختلفی نظیر نام فیلم، موادمخدر، یک رمان، به معنای خود شیشه و غیره و در ویکی‌پدیای انگلیسی به خود شیشه، عینک، نام یک روستا، نام یک رود، نام چند فیلم، رمان و غیره نگاشت می‌شود. یک الگوریتم پیونددهی موجودیت می‌بایست با استفاده از محتوای متن و اطلاعات

1 Natural Language Processing (NLP)

2 Word Sense Disambiguation (WSD)

3 Named-Entity Linking

4 Polysemy

موجود در پایگاه‌دانش بتواند این ابهام را حل کند. رفع این ابهام، در صورتی که یک موجودیت از نـام کامـل خود استفاده نکند، بسیار مشکل می‌شود [6].

در پیونددهی موجودیت برخلاف ابهام‌زدایی معنایی یک نگاشت ممکن است نیمه باشد، به این معنا کـه نـام کامل یک موجودیت نیاورده شده باشد ولی با استفاده از محتوای متن بتـوان آن را نگاشـت کـرد. بـا وجـود اختلافات موجود، ابهام‌زدایی معنایی و پیونددهی موجودیت درهم‌تنیده‌اند، چرا کـه هـر دو بـه ابهـام‌زدایـی بخش‌های یک متن می‌پردازند [2]. با اینکه در زبان انگلیسی این دو به هـم پیونـد داده شـده‌انـد، در زبـان فارسی تا کنون مجزا بررسی شده‌اند.

در این پروژه روشی برای نگاشت وردنت فارسی یا همان فارس‌نت به گراف دانش موجودیت‌هـای فارسـی یـا همان فارس‌بیس ارائه می‌دهیم، کـه از نتـایج دو الگـوریتم مشـابه بـرای ابهـام‌زدایـی معنـایی و پیونـددهی موجودیت برای این نگاشت استفاده می‌کند.

در بخش‌های آتی، ابتدا به بررسی کارهای پیشین در این حوزه و سپس، بـه ارائـه‌ی دقیـق روش ارائـه‌شـده می‌پردازیم. در انتها نتایج را بررسی و به جمع‌بندی و ارائه‌ی پیشنهادات می‌پردازیم.

فـصل ۲ :

مروری بر منابع

۲-۱- مقدمه

در این فصل ابتدا به بررسی فارس‌نت و فارس‌بیس و اصول آن دو و سپس، به بررسی و مطالعه‌ی کارهای مشابه انجام شده در این حوزه می‌پردازیم. این مطالعات را می‌توان به سه دسته‌ی مطالعات بر روی ابهام‌زدایی معنایی، مطالعات بر روی پیونددهی موجودیت و مطالعات بر روی نگاشت بین این دو تقسیم‌بندی نمود.

۲-۲- تعاریف، اصول و مبانی نظری

در این بخش ابتدا به بررسی فارس‌نت و اصطلاحات موجود در آن و سپس به بررسی فارس‌بیس می‌پردازیم.

۲-۲-۱- فارس‌نت

فارس‌نت [7] یک پایگاه‌داده‌ی لغوی فارسی و در واقع نسخه‌ی فارسی وردنت [8] است. در فارس‌نت کلمات به صورت گروه‌های هم‌معنایی یا هم‌نشیم‌ها[1] دسته‌بندی شده‌اند. برای هر گروه هم‌نشیم اطلاعاتی نظیر توصیف معنا یا همان گلاس[2]، اطلاحات نحوی، مثالی از کاربرد معنی و غیره نگه‌داری می‌شود. هم‌نشیم‌ها می‌توانند با هم روابط زیر را داشته باشند:

- روابط هایپرنیم/ هایپونیم[3]: برای مثال "رنگ" هایپرنیم برای "قرمز" و "قرمز" هایپونیم برای "رنگ" است.
- روابط مرونیمی[4]: برای مثال دست و چشم برای انسان مرونیم‌اند.
- روابط تضادی[5]

[1] Synsets
[2] Gloss
[3] Hypernymy/ hyponym
[4] Meronymy
[5] Antonymy

هم‌چنین رابطه‌ی خواهری[1] را می‌توان برای هم‌نشیم‌ها تعریف کرد، که به معنای داشتن یک هایپرنیم یکسان بین دو هم‌نشیم است.

۲-۲-۲- فارس‌بیس

گراف‌های دانش، گراف‌های عظیمی از موجودیت‌های با ارتباط داخلی‌اند. فارس‌بیس [9]، [10] یک گراف دانش فارسی است، که از منابع اطلاعاتی نظیر ویکی‌پدیا درست شده است.

۲-۳- مروری بر ادبیات موضوع

در این بخش ابتدا به بررسی مطالعات بر روی ابهام‌زدایی معنایی و سپس به مطالعات بر روی پیونددهی موجودیت می‌پردازیم. در انتها، مطالعات بر روی نگاشت بین این دو را بررسی می‌کنیم.

۲-۳-۱- ابهام‌زدایی معنایی

همان‌گونه که در بخش ۱ بیان کردیم، ابهام‌زدایی معنایی تشخیص معنای درست یک واژه در یک متن با توجه به محتوای متن از میان معانی مختلف آن واژه است. سه روش کلی برای ابهام‌زدایی معنایی وجود دارد که در ادامه به بررسی مطالعات انجام شده بر هر سه روش می‌پردازیم [1].

- روش مبتی بر دانش[2]

در این روش از یک منبع لغوی خارجی مانند لغت‌نامه‌ها استفاده می‌شود که به دلیل نبود نیاز به یادگیری، مجموعه‌داده‌ی برچسب خورده‌ای لازم نیست. با اینکه روش‌های با ناظر ممکن است بهینه‌تر باشند، این روش از مزیت‌های بسیاری مانند پیاده‌سازی آسان و نبود نیاز به مجموعه‌داده‌ی برچسب دار بهره‌مند است [1]، [11]. اولین روش ارائه شده در این دسته، روش لسک[3] است که سعی در

[1] Sisterhood
[2] Knowledge based
[3] Lesk

بیشینه‌کردن هم‌پوشانی لغوی بین معانی مختلف کلمات درون محتوا دارد [12]. در این روش معانی کلمات درون یک محتوا به صورت هم‌زمان ابهام‌زدایی می‌شوند. کیلگاریف و روزنویگ [13] روش لسک ساده‌شده[1] را ارائه دادند، که در هر زمان تنها یک کلمه را ابهام‌زدایی می‌کند. این روش فضای جست‌وجو را کاهش بسزایی داد.

- روش با ناظر[2]

در روش‌های با ناظر یک طبقه‌بند[3] با استفاده از یک مجموعه‌داده‌ی برچسب‌زده‌شده، که در اینجا یک کلمه و معنای متناظر آن در یک محتوا است، آموزش داده می‌شود. سپس از مدل یافت شده برای یافتن معنای متناظر کلمات در متن‌های بدون برچسب استفاده می‌شود [11]. در [14] روشی با استفاده از طبقه‌بند بیز ساده[4] ارائه‌شده‌است. لی، نگ و چیا [15] با استفاده از اجزای کلام[5]، لغات اطراف در محتوا، همایندها[6] و روابط نحوی[7] و یک ماشین بردار پشتیبانی[8] ابهام‌زدایی معنایی را انجام دادند. هم‌چنین در [16] نیز از بیز ساده برای ابهام‌زدایی استفاده شده‌است. در [17] و [18] به ترتیب از شبکه‌ی عصبی[9] و لیست تصمیم[10] به عنوان طبقه‌بند برای ابهام‌زدایی استفاده شده‌است. لیست تصمیم ارائه‌شده برای زبان فرانسه و اسپانیایی است و از ویژگی‌های اجزای کلام و فاصله‌ی کلمات در متن استفاده می‌کند. علاوه بر این مطالعات، چترجی و میسرا [19] یک مدل قابل آموزش ارائه داده‌اند، که با گرفتن مجموعه‌داده‌ی برچسب‌زده‌شده یادگیری را انجام داده و برای یک مجوعه‌داده‌ی دیده‌نشده معنای کلمات را می‌یابد. در [20] روشی مبتنی بر شباهت کسینوسی ارائه‌شده که از دو ویژگی کلمات متواتر متن و کلمات اطراف کلمه‌ی مبهم برای ابهام‌زدایی استفاده می‌کند.

- روش بدون ناظر[11]

برخلاف روش‌های با ناظر، این روش به مجموعه‌داده‌ی برچسب‌دار نیازی ندارد و بر مبنای این کار می‌کند که کلمات با معانی یکسان، در یک محتوای یکسان ظاهر می‌شوند[11]. چونهی ژانگ [21]

1 Simplified Lesk
2 Supervised
3 Classifier
4 Naïve Bayes
5 Part of speech (POS)
6 Collocation
7 Syntactic Relations
8 Support-vector machine
9 Neural Network
10 Decision list
11 Unsupervised

روشی مبتنی بر ژنتیک الگوریتم به نام ابهام‌زدایی معنایی ژنتیکی[1] ارائه داده است که ابتدا از وردنت تمام هم‌نشیم‌های مورد نظر را استخراج و سپس از الگوریتم ژنتیک ساده و وزن‌دار برای بیشینه‌کردن میزان شباهت معنایی استفاده می‌کند. پدرسن در [22] با استفاده از برداری از ویژگی‌ها برای ابهام‌زدایی، خوشه بندی[2] ارائه داده است. ورونیس در [23] برای برنامه‌های بازیابی اطلاعات[3] یک الگوریتم بدون ناظر مبتنی بر گراف برای ابهام‌زدایی معنایی به نام هایپرلکس[4] ارائه داده است.

لازم به ذکر است که با استفاده از الگوریتم بوت‌استرپ[5] و یادگیری ترارسانی[6] می‌توان الگوریتم‌هایی شبه با ناظر برای ابهام‌زدایی معنایی پیاده‌سازی نمود که میزان کمی داده‌ی برچسب‌گذاری‌شده نیاز دارند [24].

از میان روش‌های ارائه‌شده در زبان فارسی می‌توان به دو مطالعه‌ی زیر اشاره کرد:

- روش ارائه شده در [25] گراف وابستگی معنایی[7] است، که گره‌ها معانی مختلف کلمات درون یک جمله‌اند و با دنبال کردن یک مسیر معانی برای هر کلمه یک معنا انتخاب می‌شود. در این روش وزن هر گره توسط الگوریتم مرکزیت[8] و وزن یال‌ها از سه روش جی-پی‌ام‌آی[9]، پی‌ام‌آی[10] و اس-پی‌ام‌آی[11] محاسبه می‌شوند. معنای هر کلمه توسط دو روش جی‌ای‌دبلیو[12] که مسیری با بیش‌ترین جمع وزن یال‌ها و جی‌ان‌دبلیو[13] که گره‌های با بیش‌ترین وزن را به عنوان معنی انتخاب می‌کنند، محاسبه شده است.
- در [26] دو روش با ناظر کی‌ان‌ان[14] و بیز ساده مقایسه شده‌اند. برای این دو روش مجموعه‌داده‌ی دارای برچسب نیاز بوده‌است، که به صورت دستی مجموعه‌داده‌ی همشهری برچسب‌گزاری شده‌است. معیار شباهت کسینوسی نرمال‌شده در نظر گرفته شده و در طبقه‌بندها از برداری از ویژگی‌ها استفاده شده است.

---

1 genetic word sense disambiguation (GWSD)
2 Clustering
3 Information Retrieval applications
4 HyperLex
5 Bootstrap
6 Transductive learning
7 Semantic Dependency graph
8 Centrality
9 J-PMI
10 PMI
11 S-PMI
12 GEW
13 GNW
14 kNN

۲-۳-۲- پیونددهی موجودیت

همان‌طور که در بخش ۱ بیان کردیم، پیونددهی موجودیت، نگاشت کردن موجودیت‌های موجود در یک متن به موجودیت‌های یک پایگاه‌دانش است. گلو[1] یک سیستم پیونددهی موجودیت است، که توسط راتینوو در [27] ارائه شده‌است. این سیستم از پایگاه‌دانش ویکی‌پدیا استفاده می‌کند و ابتدا چند کاندید برای یک موجودیت انتخاب و آن‌ها را رتبه‌بندی می‌کند. سپس تصمیم می‌گیرد که آیا موجودیت با بالاترین رتبه می‌بایست به موجودیت مورد نظر نگاشت شود یا نگاشتی برای این موجودیت وجود ندارد. هوفارت در [28] سیستمی به نام آیدا[2] ارائه داده‌است که با ساخت گرافی جدید از پایگاه‌دانش و با امتیازدهی با استفاده از احتمال و شباهت کاندیدها عمل پیونددهی را انجام می‌دهد. در مطالعه‌ای دیگر گوین به همراه هوفارت و همکاران [29] نسخه‌ای دیگر از آیدا به نام آیدا-لایت[3] ارائه‌داده‌اند. هر دو آیدا و آیدا-لایت از یاگو[4] به عنوان پایگاه‌دانش استفاده می‌کنند. آیدا-لایت از یک روش دولایه برای ابهام‌زدایی پیونددهی استفاده می‌کند. به این گونه که ابتدا موجودیت‌هایی که ابهام‌زدایی پیونددهی آن‌ها کم هزینه‌تر و ساده‌تر است، پیونددهی شده و سپس با داشتن این پیونددهی‌ها و محتوای مربوطه در لایه‌بعد به ابهام‌زدایی پیونددهی‌های پرهزینه‌تر و سخت‌تر پرداخته می‌شود. پی‌پی‌آرسیم[5] سیستم ارائه‌شده در [30] است، که از تلفیق اطلاعات محلی و سراسری استفاده می‌کند و الگوریتمی بر پایه‌ی گراف برای پیونددهی موجودیت است. در این سیستم از سرتیتر ویکی‌پدیا برای تولید کاندیدها استفاده می‌شود. پیونددهی و شناخت مشترک موجودیت‌های نام‌دار[6] که در [31] ارائه شده‌است، از اولین مطالعاتی است که شناخت و پیونددهی موجودیت نام‌دار را به عنوان یک کار مشترک انجام داده است. در [32] از سی‌ان‌ان[7] برای استخراج اطلاعات در راستای پیونددهی موجودیت استفاده‌شده‌است. در [33] از روش جاگذاری[8] به منظور ابهام‌زدایی پیونددهی موجودیت استفاده‌شده‌است. در [10] روشی برای پیونددهی موجودیت فارسی با استفاده از فارس‌بیس ارائه داده شده‌است. در این روش ابتدا تشابه کسینوسی میان محتوای متن و محتوای متنی مقاله‌ی ویکی‌پدیا محاسبه و با استفاده از فرمولی با شباهت بین محتوای متن و ابرپیوندهای مقاله ترکیب و امتیاز نهایی هر کاندید را مشخص می‌کنند.

---

1 GLOW
2 AIDA
3 AIDA-light
4 YAGO
5 PPRSim
6 Joint named entity recognition and disambiguation (JERL)
7 Convolutional neural networks (CNN)
8 Embedding method

۲-۳-۳- نگشات میان ابهام‌زدایی معنایی و پیونددهی موجودیت

ناویگلی و همکاران ابتدا در [34] پایگاه‌دانش ویکی‌پدیا را به وردنت انگلیسی نگاشت کردند و بیبل‌نت[1] را ارائه‌دادند و سپس، در [35] الگوریتمی به نام بیبلیفای[2] برای ابهام‌زدایی معرفی کردند. در نگاشت ارائه‌شده از هم‌معنی‌ها، هایپرنیم‌ها، هیپونیم‌ها، خواهر و گلاس مربوطه برای هر هم‌نشیم وردنت و برچسب داخل پرانتز، لینک‌ها و دسته‌بندی‌های ویکی‌پدیا استفاده می‌شود.

۲-۴- نتیجه‌گیری

در این بخش مطالعات پیشین مربوط به حوزه‌ی مدنظر بررسی شدند. همان‌گونه که مطالعه شد، با وجود پیشرفت ابهام‌زدایی معنایی و پیونددهی موجودیت به صورت جداگانه و وجود نگاشت میان پایگاه‌دانش و وردنت انگلیسی، برای زبان فارسی این نگاشت وجود نداشته و نیاز به آن در این حوزه احساس می‌شود. در مطالعه‌ی پیش رو سعی در نگاشت فارس‌نت و فارس‌بیس شده است.

[1] BabelNet

[2] Babelify

فـصل ۳ :

روش تحقیق

۳-۱- مقدمه

در این بخش به بررسی روش ارائه‌شده در این پروژه برای نگاشت فارس‌نت به فارس‌بیس می‌پردازیم.

۳-۲- تشریح کامل روش تحقیق

در ابتدا قصد بر انجام روشی همانند روش ارائه‌شده برای ساخت بیبل‌نت [34] کردیم. به این معنا که از فارس‌بیس موارد زیر:

- سرتیترهای داخل پرانتز در صورت موجود
- پیوندها
- دسته‌بندی‌ها

و از فارس‌نت موارد زیر را جمع‌آوری کردیم.

- کلمات درون یک هم‌نشیم
- هاپرنیم و پیپونیم‌ها
- خواهرها
- گلاس مربوطه

سپس با استفاده از فرمول زیر که در مقاله‌ی بیبل‌نت آورده شده‌بود امتیاز هر هم‌نشیم را محاسبه می‌کنیم.

$$score\left(s، w\right) = |Ctx(s) \cap Ctx(w)| + 1 \qquad (۱-۳)$$

$$p\left(s، w\right) = \frac{score(s، w)}{\sum_{s0 \in SensesWN(w) w0 \in SensesWiki(w)، } score(s0، w0)} \qquad (۲-۳)$$

اما به دلیل پایین بودن دقت این روش تصمیم به تعویض روش و یافتن روشی جدید برای نگاشت فارس‌نت به فارس گرفتیم.

در روش جدید بنا بر آن شد که ابتدا با استفاده از کد [10] الگوریتم پیونددهی موجودیت معرفی شده در

این مقاله را بر روی مجموعه‌داده اجرا کرده و سپس الگوریتم ابهام‌زدایی معنایی مشابه با این مقاله پیاده‌سازی کرده و روی مجموعه‌داده اجرا کنیم.

در این روش برای هر جمله‌ی ورودی ابتدا اجزای نحوی آن و سپس هم‌نشیم‌های کاندید برای هر کلمه پس از پیش پردازش استخراج می‌شوند. سپس هم‌نشیم‌هایی که با اجزای نحوی هم‌خوانی ندارند حذف شده و برای هم‌نشیم‌های باقی‌مانده موارد زیر استخراج می‌شوند.

- کلمات درون یک هم‌نشیم
- هاپرنیم و پیپونیم‌ها
- خواهرها
- گلاس مربوطه
- مثال مربوطه

پس از بدست آوردن این موارد امتیاز با استفاده از جمع تشابهات میان کلمات درون این موارد و کلمات محتوای مربوطه بدست می‌آید و ابهام‌زدایی معنایی انجام می‌شود.

پس از انجام ابهام‌زدایی معنایی بر روی مجموعه‌داده‌ی مورد نظر، نتایج پیونددهی و ابهام‌زدایی با یک‌دیگر مقایسه شده و در صورت نگاشت به بخش مشترک به یک‌دیگر نگاشت می‌شوند و تواتر این نگاشت‌ها محاسبه می‌شود. در آخر با استفاده از یک مجموعه‌داده‌ی برچسب‌دار از نگاشت بین فارس‌نت و فارس‌بیس یک آستانه‌ی مناسب برای قبول و یا عدم قبول نگاشت انتخاب می‌شود.

فـصل ۴ :

نتایج و تفسیر آنها

۴-۱- مقدمه

در این بخش نتایج حاصل از نگاشت بین فارس‌نت و فارس‌بیس با استفاده از روش مطرح شده در بخش ۳ را بررسی می‌کنیم.

۴-۲- محتوا

۴-۲-۱- مجموعه‌داده‌ها

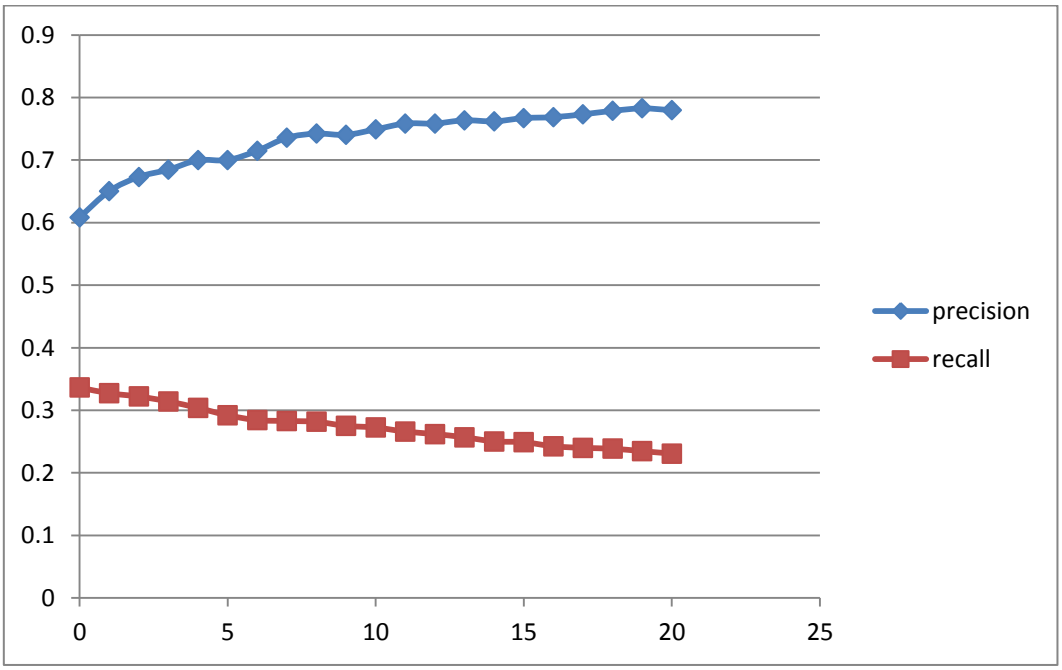
به منظور ارزیابی الگوریتم ارائه شده و به دلیل جدید بودن نگاشت بین فارس‌نت و فارس‌بیس مجموعه‌داده‌ای برای ارزیابی و یا الگوریتمی برای مقایسه‌ی نتایج وجود نداشت. به همین دلیل، از مجموعه‌داده‌ای که توسط یک شخص خبره تولید و شامل ۷۶۷ نگاشت بین فارس‌نت وفارس‌بیس بود، برای ارزیابی استفاده‌کردیم. همچنین، الگوریتم‌های اابهام‌زدایی معنایی و پیونددهی موجودیت بر روی مجموعه‌داده‌ی همشهری اجرا شده‌اند.

۴-۲-۲- ارزیابی نتایج

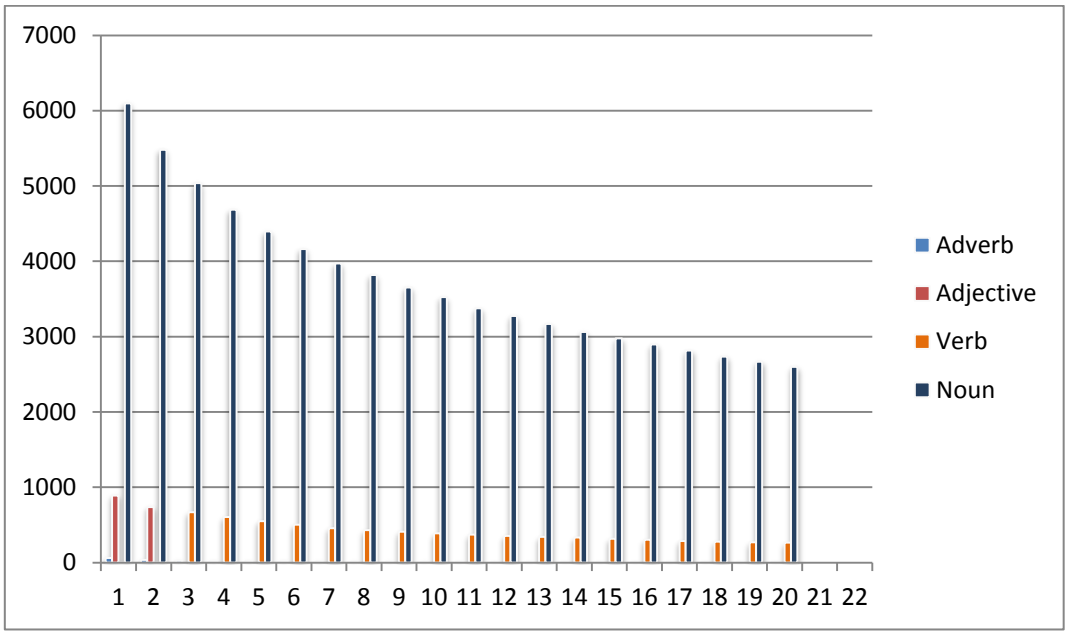
نتایج به دست آمده از روش ارائه‌شده را پس از اعمال آستانه‌های مختلف برای قبولی با مجموعه‌داده‌ی ارزیابی مقایسه کردیم. این مقایسه با دو معیار دقت و بازخوانی و به منظور یافتن آستانه‌ی بهینه انجام‌شده است.

شکل (۴-۱) افزایش و کاهش معیار دقت و بازخوانی با افزایش آستانه را نمایش می‌دهد. همان‌گونه که قابل انتظار است، با افزایش آستانه‌ی تواتر نگاشت قابل قبول معیار دقت افزایش و معیار بارخوانی کاهش پیدا می‌کند. همچنین، همان‌طور که در شکل (۴-۳) قابل مشاهده است، افزایش آستانه تعداد نگاشت‌ها را کاهش می‌دهد. در شکل (۴-۲) تعداد هر یک از اجزای سخن تشخیص داده‌شده در این نگاشت را مشاهده می‌کنیم. تعداد اسم‌های تشخیص داده‌شده از سایر اجزای سخن بیش‌تر است و بعد از اسم‌ها، فعل‌ها

بیش‌ترین‌اند.

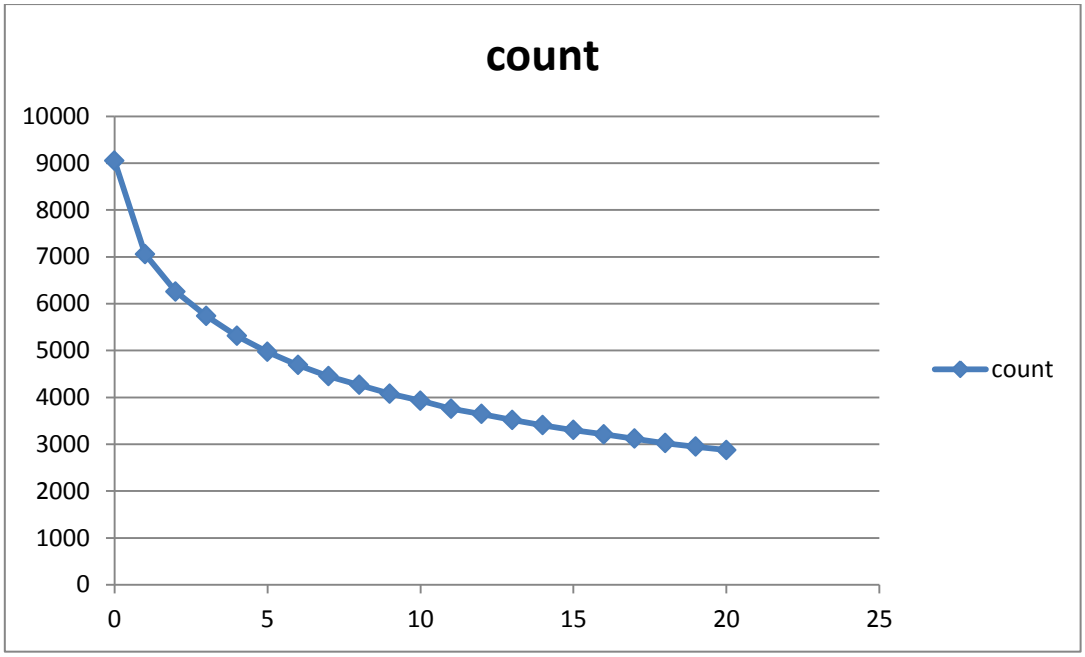


شکل (۴-۱) مقدار معیار دقت و بازخوانی برای آستانه‌های متفاوت



شکل (۴-۲) تعداد اجزای سخن در میان نگاشت‌ها برای هر آستانه

در شکل (۴-۳) تعداد نگاشت‌ها برای هر آستانه را نشان می‌دهد. برای تواتر حداقل یکبار تعداد ۹۰۰۰ نگاشت و برای آستانه‌ی حداقل ۲۰ تعداد ۳۰۰۰ نگاشت انجام شده است.



شکل (۴-۳) تعداد نگاشت‌ها در هر آستانه

با توجه به سه نمودار نمایش‌داده‌شده، آستانه‌ی ۵ به نظر مناسب‌تر از بقیه‌ی آستانه‌ها می‌آید. با گرفتن آستانه‌ی ۵، معیار دقت برابر ۷۰، بازخوانی برابر ۳۰ و تعداد نگاشت‌ها برابر ۵۰۰۰ است.

فـصل ۵ :

جمع‌بندی و پیشنهادها

۵-۱- مقدمه

ابهام‌زدایی معنایی و پیونددهی موجودیت از موضوعات مطرح در حوزه‌ی پردازش زبان طبیعی‌اند، که با استفاده از فارس‌نت و فارس‌بیس انجام می‌گیرند. در این مطالعه سعی در پیوند این دو حوزه و نگشات بین فارس‌نت و فارس‌بیس شده‌است. در این بخش به معرفی کلی مطالعه‌ی انجام شده و جمع‌بندی می‌پردازیم. در انتها نیز درباره‌ی نوآوری و پیشنهادات آتی برای این مطالعه بحث می‌کنیم.

۵-۲- جمع‌بندی

در این مطالعه، سعی در نگاشت فارس‌نت و فارس‌بیس که تاکنون برای زبان فارسی انجام نشده‌است، کرده‌ایم. برای این کار، ابتدا یک الگوریتم پیونددهی موجودیت مناسب یافته و سپس یک الگوریتم برای ابهام‌زدایی معنایی مشابه آن پیاده‌سازی نمودیم. سپس اشتراک خروجی این دو الگوریتم را گرفته و تواتر آن‌ها را حساب کردیم. در آخر با استفاده از معیار دقت، بازخوانی و تعداد نگاشت‌ها آستانه‌ای برای نگاشت با تواتر قابل قبول در نظر گرفتیم. معیار دقت،بازخوانی و تعداد نگاشت نهایی برابر با ۷۰ درصد، ۳۰ درصد و ۵۰۰۰ نگاشت بوده‌است.

۵-۳- نوآوری

نگاشت میان فارس‌نت وفارس‌بیس به منظور اجرای هم‌زمان و بهینه‌ی ابهام‌زدایی معنایی و پیونددهی موجودیت برای زبان فارسی تاکنون انجام‌نشده‌است. در این مطالعه، سعی در نگاشت این دو منبع مهم در حوزه‌ی پردازش زبان طبیعی کرده‌ایم، که خود نوآوری محسوب می‌شود.

۵-۴- پیشنهادها

برای کارهای آتی می‌توان الگوریتم را بر روی مجموعه‌داده‌های متنوع و بیش‌تری اجرا کرد تا نگاشت‌های

فعل‌ها و تواتر نگاشت‌ها بهبود پیدا کند. هم‌چنین می‌توان با نگاشت‌های یافت شده، یک الگوریتم برای ابهام‌زدایی و پیونددهی موجودیت هم‌زمان، همانند بیبل‌نت [34] ، ارائه داد.

مراجع

مراجع

[1] J. Sreedhar، S. Raju، A. Babu، A. Shaik، and P. Kumar، "Word Sense Disambiguation: An Empirical Survey،" International Journal of Soft Computing and Engineering،Volume-2، Issue-2، pp. 494-503، May 2012

[2] A. Moro، A. Raganato، and R. Navigli، "Entity Linking meets Word Sense Disambiguation: a Unified Approach،" Trans. Assoc. Comput. Linguist.، vol. 2، pp. 231–244، 2018.

[3] A. Moro، F. Cecconi، and R. Navigli، "Multilingual word sense disambiguation and entity linking for everybody،" CEUR Workshop Proc.، vol. 1272، pp. 25–28، 2014.

[4] S. Kumar، N. Sharma، and S. Niranjan، "Word Sense Disambiguation Using Association Rules : A Survey،" Int. J. Comput. Technol. Electron. Eng.، vol. 2، no. 2، pp. 93–98، 2012.

[5] B. Hachey، W. Radford، and J. R. Curran، "Graph-based named entity linking with Wikipedia،" Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)، vol. 6997 LNCS، pp. 213–226، 2011.

[6] T. Huber and P. Definition، "Entity Linking - A Survey of Recent Approaches،" Department of Computer Science، Humboldt-Universität zu Berlin pp. 1–5، 2012.

[7] M. Shamsfard et al.، "Semi automatic development of farsnet; the persian wordnet،" Proc. 5th Glob. WordNet Conf. Mumbai، India، vol. 29، 2010.

[8] A. M. George، "WordNet: A Lexical Database for English،" Commun. ACM، vol. 38، pp. 39–41، 1995.

[9] M. B. Sajadi، B. M. Bidgoli، and A. Hadian، "FarsBase: A cross-domain farsi knowledge graph،" CEUR Workshop Proc.، vol. 2198، 2018.

[10] M. Asgari، A. Hadian، and B. Minaei-Bidgoli، "FarsBase: The Persian Knowledge Graph،" Semnatic Web J.، vol. 1، no. 0، pp. 1–5، 2018.

[11] P. P. Borah، G. Talukdar، and A. Baruah، "Approaches for Word Sense Disambiguation – A Survey،" Int. J. Recent Technol. Eng.، vol. 3، no. 1، pp. 35–38، 2014.

[12] M. Lesk، "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone،" Proc. 5th Annu. Int. Conf. Syst. Doc.، pp. 24–26، 1986.

[13] A. Kilgarriff، "{E}nglish {SENSEVAL}: Report and Results،" Proc. Int. Conf. Lang. Resour. Eval.، 1998.

[14] W. A. Gale، K. W. Church، and D. Yarowsky، "A method for disambiguating word senses in a large corpus،" Comput. Hum.، vol. 26، no. 5–6، pp. 415–439، 1992.

[15] Y. K. Lee، H. T. Ng، and T. K. Chia، "Supervised word sense disambiguation with support vector

machines and multiple knowledge sources،" Senseval-3 Third Int. Work. Eval. Syst. Semant. Anal. Text، no. July، pp. 137–140، 2004.

[16] R. Bruce and J. Wiebe، "Word-sense disambiguation using decomposable models،" In 32nd Annual Meeting of the Association for Computational Linguistics، Las Cruces، NM، pp. 139–146، 2007.

[17] G. Towell، "Disambiguating Highly Ambiguous Words،" In: Computational Linguistics، March 1998، Vol. 24، Number 1، vol. 20899، pp. 1–22، 2002.

[18] D. Yarowsky، "Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French،" Natural Language Engineering، 2002.

[19] N. Chatterjee and R. Misra، "Word-Sense Disambiguation using maximum entropy model،" in 2009 Proceeding of International Conference on Methods and Models in Computer Science (ICM2CS).

[20] M. Nameh، S. M. Fakhrahmad، and M. Z. Jahromi، "A New Approach to Word Sense Disambiguation Based on Context Similarity،" Proc. World Congr. Eng.، vol. I، pp. 8–11، 2011.

[21] C. H. Zhang، Y. Zhou، and T. Martin، "Genetic word sense disambiguation algorithm،" Proc. - 2008 2nd Int. Symp. Intell. Inf. Technol. Appl. IITA 2008، vol. 1، pp. 123–127، 2008.

[22] T. Pedersen and R. Bruce، "Unsupervised Text Mining،"، Dallas TX،Departement of computer scienceand engineering، southern methodist university 1997.

[23] J. Véronis، "HyperLex: Lexical cartography for information retrieval،" Comput. Speech Lang.، vol. 18، no. 3 SPEC. ISS.، pp. 223–252، 2004.

[24] T. P. Pham، H. T. Ng، and W. S. Lee، "Word sense disambiguation with semisupervised learning،" Proc. 20th Natl. Conf. Artif. Intell. - Vol. 3، pp. 1093–1098، 2005.

[25] M. Soltani and H. Faili، "A statistical approach on Persian word sense disambiguation،" 7th Int. Conf. Informatics Syst.، pp. 1–6، 2010.

[26] M. Hamidi، A. Borji، and S. S. Ghidary، "Persian Word Sense Disambiguation،" 15th Iran. Conf. Electr. Electron. Eng. Proc.، pp. 114–118، 2007.

[27] L. Ratinov، D. Roth، D. Downey، and M. Anderson، "Local and Global Algorithms for Disambiguation to Wikipedia،" Annu. Meet. Assoc. Comput. Linguist.، pp. 1375–1384، 2011.

[28] J. Hoffart et al.، "Robust Disambiguation of Named Entities in Text Johannes،" Proc. 2011 Conf. Empir. Methods Nat. Lang.، pp. 782–792، 2011.

[29] D. B. Nguyen، J. Hoffart، M. Theobald، and G. Weikum، "AIDA-light: High-throughput named-entity disambiguation،" CEUR Workshop Proc.، vol. 1184، 2014.

[30] M. Pershina، Y. He، and R. Grishman، "Personalized Page Rank for Named Entity Disambiguation،" Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies، pp. 238–243، 2015.

[31] T. Ishizaki، K. Yokochi، K. Chiba، T. Tabuchi، and T. Wagatsuma، "Joint Named Entity Recognition and Disambiguation،" Pediatr. Pharmacol. (New York).، vol. 1، no. 4، pp. 291–303، 1981.

[32] M. Francis-Landau، G. Durrett، and D. Klein، "Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks،" arXiv preprint arXiv:1604.00734، 2016.

[33] I. Yamada، H. Shindo، H. Takeda، and Y. Takefuji، "Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation،" The SIGNLL Conference on Computational Natural Language Learning (CoNLL) 2016.

[34] R. Navigli and S. Ponzetto، "BabelNet: Building a very large multilingual semantic network،" Proc. 48th Annu. Meet. …، no. July، pp. 216–225، 2010.

[35] T. Flati and R. Navigli، "Three birds (in the LLOD cloud) with one stone: BabelNet، Babelfy and the Wikipedia Bitaxonomy،" Proc. Semant.، pp. 4–5، 2014.

فهرست واژگان فارسی به انگلیسی

آستانه..... Threshold
ابهام‌زدایی معنایی ..... Word Sense Disambiguation (WSD)
ابهام‌زدایی معنایی ژنتیکی..... Genetic word sense disambiguation (GWSD)
اجزای کلام..... Part of speech (POS)
با ناظر..... Supervised
بی‌ناظر..... Unsupervised
برنامه‌های بازیابی اطلاعات..... Information Retrieval applications
بیز ساده..... Naïve Bayes
پردازش زبان طبیعی ..... Natural Language Processing (NLP)
پیونددهی موجودیت ..... Named-Entity Linking
تضاد..... Antonymy
خوشه‌بندی..... Clustering
دادگان طلایی..... Gold dataset
روابط نحوی..... Syntactic Relations
روش جایگذاری..... Embedding method
شبکه‌ی عصبی..... Neural Network
رده‌بند..... Classifier
گراف وابستگی معنایی..... Semantic Dependency graph
لیست تصمیم..... Decision list
ماشین بردار پشتیبانی..... Support-vector machine
مبتنی بر دانش..... Knowledge based
مرکزیت..... Centrality
همایند..... Collocation
هم‌نشیم‌ها..... Synsets
یادگیری ترارسانی..... Transductive learning

فهرست واژگان انگلیسی به فارسی

تضاد..... Antonymy

مرکزیت..... Centrality

رده‌بند..... Classifier

خوشه‌بندی..... Clustering

همایند..... Collocation

لیست تصمیم..... Decision list

روش جایگذاری..... Embedding method

ابهام‌زدایی معنایی ژنتیک..... Genetic word sense disambiguation (GWSD)

دادگان طلایی..... Gold dataset

برنامه‌های بازیابی اطلاعات..... Information Retrieval applications

مبتنی بر دانش..... Knowledge based

پیونددهی موجودیت ..... Named-Entity Linking

پردازش زبان طبیعی ..... Natural Language Processing (NLP)

بیز ساده..... Naïve Bayes

شبکه‌ی عصبی..... Neural Network

اجزای کلام..... Part of speech (POS)

گراف وابستگی معنایی..... Semantic Dependency graph

با ناظر..... Supervised

ماشین بردار پشتیبانی..... Support-vector machine

هم‌نشیم‌ها..... Synsets

روابط نحوی..... Syntactic Relations

یادگیری ترارسانی..... Transductive learning

آستانه..... Threshold

بی‌ناظر..... Unsupervised

ابهام‌زدایی معنایی ..... Word Sense Disambiguation (WSD)

Abstract:

Word sense disambiguation، which maps each word to a Synset of a Wordnet، is one important problem in the field of text knowledge extraction. This problem is related to the problem of Entity linking، which links words to corresponding entities of a knowledge graph. In Persian، these two problems can be solved using Persian Wordnet (Farsnet) and Persian knowledge graph (FarsBase). Although there exists a uniform combination of these two problems for English language، there are no such work for Persian language.

The main aim of this project is to represent a mapping between Farsnet and Farsbase. First of all، our proposed method applies a Persian entity linker on a corpus and then، uses a word sense disambiguation algorithm similar to the entity linker for the same corpus. Then، we check the result of the entity linker and the word sense disambiguation algorithm for similarities، in order to map linked entities to word senses (Synsets of FarsNet). At the end، we count the number of occurrence of each mapped pairs and look for an appropriate threshold for accepting a mapped pair. The final mapping is evaluated using a dataset of 767 manually mapped entity-Synset pairs. The result had a precision and recall of 70 and 30 presents and contains 5000 mapped pairs.

Keywords: Word Sense Disambiguation- Entity Likning- FarsNet- FarsBase

Iran University of Science and Technology
Computer engineering Department

Entity linking algorithm for Word sense disambiguation of Persian language

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Bachelors in Computer engineering

By:
Banafsheh Karimian

Supervisor:
Dr. Minaei

June 2019