DOR: 20.1001.1.2538161.1403.9.3.6.9

DOI: 10.22034/csj.2024.209868

پاییز ۱۴۰۳، دوره ۹، شماره ۳، پیاپی ۳۴

دریافت مقاله: ۱۴۰۳/۰۶/۰۵

پذیرش مقاله: ۱۴۰۳/۰۷/۲۸

نوع مقاله: پژوهشی

بازشناسی و مرتبط‌سازی موجودیت‌های نام‌دار در سند احادیث

محمد ایزدی*

دانشیار، دانشکده مهندسی کامپیوتر- دانشگاه صنعتی شریف- تهران- ایران

پست الکترونیکی: izadi@sharif.edu

عارف صادقیان

دانشجو، دانشکده مهندسی کامپیوتر- دانشگاه صنعتی شریف- تهران- ایران

پست الکترونیکی: aref.sadeghian94@sharif.edu

چکیده

مسئلهٔ بازشناسی موجودیت‌های نام‌دار در متنی از یک زبان طبیعی، این نیازمندی است که درون متن، بخش‌هایی که حاوی اسامی خاص هستند را پیدا کنیم. علاوه بر این، مرتبط‌سازی این موجودیت‌های نام‌دار عبارتست از این که برای هر اسم خاص مشخص شده درون آن متن، مقصود یا مدلول آن اسم در جهان واقع را مشخص کنیم به گونه‌ای که اسامی یکسان یا مشابه اما با مدلول‌های متفاوت، کاملا از یکدیگر متمایز شوند. در این مقاله، شیوه بازشناسی موجودیت‌های نام‌دار در سند احادیث و هم‌چنین مرتبط‌سازی این اسامی با شناسهٔ منحصربه‌فرد افراد تشریح می‌شود. بازشناسی با استفاده از آموزش مدل پیش‌آموختهٔ AraBERTv02 روی دادگان CANERCorpus صورت می‌گیرد که با تخصیص دادگان به نسبت ۰/۸ به ۰/۲ برای دادگان آموزش و آزمون، میزان دقت (accuracy) مساوی ۰/۹۹۰ در دسته‌بندی صحیح کلمات متعلق به دستهٔ اشخاص حاصل می‌شود.

مرتبط‌سازی با استفاده از روش‌های یادگیری مبتنی بر گراف انجام می‌شود. با مدل کردن یک دسته از سند احادیث به صورت یک گراف (در واقع یک جنگل) که سند هر حدیث به صورت یک درخت از گره‌هایی است که متناظر با افراد هستند، می‌توان مسئله را به صورت پیش‌بینی برچسب برای گره‌ها تعریف کرد. به این ترتیب یک شبکهٔ هم‌آمیخت گرافی (GCN) معرفی می‌شود که سعی دارد تا این برچسب را براساس شکل ظاهری افراد پیرامونی هر فرد پیش‌بینی کند. این شبکه می‌تواند با آموزش روی ۰/۸ از دادگان، به دقت میزان دقت (accuracy) مساوی ۰/۸۵۷۰ برسد.

واژه های کلیدی: علوم انسانی محاسباتی، بازشناسی موجودیت‌های نام‌دار، مرتبط‌سازی موجودیت‌های نام‌دار، مدل‌های بزرگ زبانی، شبکه‌های هم‌آمیخت گرافی

۱. مقدمه

این مقاله به بررسی دو مسئلهٔ مهم در حوزه پردازش زبان طبیعی (با تاکید بر زبان‌های فارسی و عربی) یعنی مسائل بازشناسی موجودیت‌های نام‌دار و مرتبط‌سازی آنها با تمرکز بر اسامی موجود در سلسله راویان در اسناد

* نویسندهٔ مسئول

احادیث می‌پردازد. در این بخش، این مسائل و اهمیت آنها بررسی و به اختصار در مورد پیشینهٔ راه‌حل‌ها و ایده‌های پیشنهادی بحث می‌شود.

۲.درآمدی بر موضوع

با داشتن یک متن از یک زبان طبیعی، آن‌چه برای فهم آن متن در یک مطالعهٔ ساختارمند نیاز است، بعد از فهم معانی کلمات عام آن زبان، این است که اسامی خاص را بشناسیم و آن‌ها را مطالعه کنیم. مثلا فرض کنید یک مجموعهٔ حدیثی به زبان عربی در اختیار است و می‌خواهیم آن را مطالعه کنیم. اگر فرض کنیم که زبان عربی را می‌فهمیم، تنها مواردی که می‌تواند در فهم عبارات موجود، اختلال ایجاد کند این است که در این عبارات، دربارهٔ افراد، رویدادها، مکان‌ها و یا سایر اسامی خاصی صحبت به میان آمده باشد که برای خواننده شناخته شده نباشند. به این ترتیب کافی است برای هر اسم خاص، یک بخش توضیحات (یا آن طور که در صفحات ویکی مرسوم است، یک صفحهٔ مجزا) ایجاد کنیم که این ابتدا مستلزم آن است که اسامی خاص را بشناسیم و در مرحلهٔ بعد، نیاز است تا بتوانیم شکل‌های مختلف ظاهر شدن یک اسم خاص واحد در متن را به همان صفحهٔ واحد، مرتبط کنیم. مسئلهٔ اول، در پردازش زبان‌های طبیعی، با نام بازشناسی موجودیت‌های نام‌دار شناخته می‌شود که به‌صورتی که در ادامه می‌آید تعریف می‌شود.

بازشناسی موجودیت‌های نام‌دار: با داشتن یک رشتهٔ متنی از یک زبان طبیعی، نیازمندی این است که درون متن، بخش‌هایی که مربوط به اسامی خاص هستند را پیدا کنیم.

اگر چه که مسئله در صورت‌بندی عمومی آن، قیدی روی نوع اسم خاص نمی‌گذارد، در اغلب راه‌حل‌های جدید که مبتنی بر یادگیری ماشین هستند، این اسامی خاص، از انواعی هستند که دادهٔ برچسب‌خورده برای آن موجود است؛ یعنی چنان‌چه در دادهٔ آموزش، برای مثال اسامی خاص از نوع اسم افراد، اسم شهرها و اسم کشورها برچسب خورده باشد، در دادهٔ آزمون هم مسئله، پیدا کردن همین انواع خواهد بود. مسئلهٔ دوم، در پردازش زبان‌های طبیعی، با عنوان مرتبط‌سازی موجودیت‌های نام‌دار شناخته می‌شود که به صورت زیر، تعریف می‌شود.

مرتبط‌سازی موجودیت‌ها: با داشتن یک رشتهٔ متنی از یک زبان طبیعی که اسامی خاص درون آن معین شده‌اند، نیازمندی این است که مقصود از هر اسم خاص در جهان واقع را مشخص کنیم. به طور مثال، در این پژوهش، به هر موجودیت، یک عدد نسبت می‌دهیم که شناسهٔ منحصربه‌فردی برای رجال موجود در احادیث است. با مشخص بودن این شناسه برای هر موجودیت در دادهٔ آموزش، یک پایگاه دانش کمینه خواهیم داشت که برای هر راوی حدیث، شناسهٔ منحصربه‌فرد و ریخت‌های گوناگون ارجاع به آن در دادهٔ آموزش را در خود دارد. یعنی درایهٔ ازام از این پایگاه‌دانش، حاوی اسامی مختلفی برای فرد واحدی در دنیای واقعی هستند.

۳. پیشینهٔ پژوهش

در خصوص پیشینهٔ راه‌حل‌های مطرح شده برای حل این مسائل، در بخش بعد، با تفصیل بیشتری بحث خواهد شد. در این‌جا به طور خلاصه، روش‌های جدیدتر برای هر کدام از این دو مسئله مرور خواهند شد.

در خصوص بازشناسی موجودیت‌های نام‌دار، تا قبل از شیوع مدل‌های زبانی بزرگ، آموزش مدل‌های ژرف که اغلب به خانوادهٔ شبکه‌های عصبی بازگشتی تعلق داشتند، روی داده‌های هر مسئلهٔ مجزا، بهترین نتیجه را می‌داد. با ظهور مدل‌های زبانی بزرگ و بهبودی که در نصاب‌های موجود در سایر مسائل پردازش زبان طبیعی ایجاد کردند، تلاش برای استفاده از قدرت این مدل‌ها آغاز شد. برای مثال آن طور که در بخش مرور ادبیات خواهد آمد، در ۲۰۲۱، یک مدل ژرف نسبتا پیچیده، روی خروجی مدل BERT زبان انگلیسی، آموزش داده شده و بهترین نتیجه را روی بازشناسی موجودیت‌های نام‌دار در متون حدیثی

به خود اختصاص داده بود. بـا پیش‌رفت مدل‌های زبانی بـزرگ و بهبـودی کـه در نصاب‌های موجود در سـایر مسائل پردازش زبان طبیعی ایجاد کردند، به نظر می‌رسید، fine-tuning مدل‌های پیش آموختهٔ جدید روی زبان عربی، بتواند نتایج بهتر بدهد.

در خصوص مرتبط‌سازی موجودیت‌های نام‌دار، عمدهٔ تمرکز برای حل مسـئله، بر روی پیـدا کردن راه‌حل‌هایی تجربی برای ترکیب ویژگی‌های مربوط به هر موجودیت با موجودیت‌های اطراف آن است، به نحوی که در گام بعدی یک شبکهٔ عصبی نسبتا ساده (یک شبکهٔ Feed Forward) بتواند مرتبط‌سـازی موجودیت‌های یکسان را یاد بگیرد. نزدیک‌تریـن راه‌حل‌هـا که به ایده‌های گرافـی، مربوط به استفاده از PageRank در رتبه‌بندی موجودیت‌های مشابه، برای هرگونه اسـت. به نظر می‌رسد، با توجه به ساختار گرافی ذاتی این مسـئله، یک شبکهٔ عصبی مبتنی بر گراف بتواند نتایج بهتری بگیرد

۴. روش پیشنهادی و نتایج

برای بازشناسی موجودیت‌های نام‌دار، یک مدل مبتنی بر BERT که روی زبان عربی مدرن پیش‌آموزش داده شده، روی یک مجموعـه دادگان عربی کلاسـیک، fine-tuned شـده است و این مدل، روی دادگان آزمون، به میزان دقت مساوی ۰/۹۹۰ رسیده است.

برای مرتبط‌سازی موجودیت‌های نام‌دار هم، سلسلهٔ راویان هر حدیث، به‌صورت یک گراف (در واقع درخت) مدل شـده است و گره‌ها، که هنگام آموزش از روی نام ظاهری و برچسـب موجود برای گره به دست می‌آیند، در صورتی به هم متصل هستند که رابطهٔ نقل بلافصل بین آن‌ها حاکم باشـد. به این ترتیب مسـئله به‌صورت پیش‌بینی برچسب در یک شـبکهٔ گرافی در می‌آید. حل این مسئله با اسـتفاده از شبکهٔ گراف هم‌آمیختی منجر به میزان دقت مسـاوی ۰/۸۵۷۰ روی داده‌های آزمون می‌شود.

۵. ساختار مقاله

در ادامه، در بخش دوم مقاله، ادبیات موضوع برای هر کدام از این دو مسـئله به تفکیک مرور می‌شود. سپس در بخش سـه روش‌های پیشنهادی که در بالا به‌طور خلاصه بیان شـد، ارائه می‌شـوند. پـس از آن در بخش چهارم، جزئیات پیاده‌سازی و نتایج ارزیابی‌ها تشریح می‌شود و در آخر، در بخش جمع‌بندی و نتیجه‌گیری، نتایج حاصل از پژوهش مرور می‌شوند.

مرور کارهای پیشین

بازشناسی موجودیت‌های نام‌دار

در ایـن بخش، مـرور پژوهش‌های مرتبط با مسـئلهٔ بازشناسـی موجودیت‌های نام‌دار و مسئلهٔ مرتبط‌سازی موجودیت‌ها را به طور مجزا بررسی می‌کنم.

دادگان مورد استفاده

پیش از مرور سـیر پیشرفت راه‌حل‌های مسئله، دادگان موجود برای آموزش مدل‌های مبتنی بر یادگیری ماشـین را مـرور می‌کنیم. جدول (۱)، دادگان موجود برای مسـئلهٔ بازشناسی موجودیت‌های نام‌دار در زبان عربی را به همراه سـال انتشار نشـان می‌دهد. همان‌طور که مشخص است، بیشـتر این داده‌ها، براساس برچسـب‌زنی اخبار به دست آمده‌اند. از بین ایـن مجموعـه دادگان، CANERCorpus با موضوع مورد علاقهٔ ما بیشـتر مرتبط است. این مجموعه چیزی بالـغ بر ۷۰۰۰ حدیث از متـن کتاب صحیح بخاری است که سه کارشـناس با بیش از ۵ سال سابقه در حوزهٔ زبان‌شناسی محاسباتی، آن را برچسب زده‌اند. تا پیش از این دادگان، آن‌طور که در [۱] ادعا شده است، هیچ دادگانی برای بازشناسی موجودیت‌های نام‌دار در عربی کلاسیک وجود نداشته و همین موضوع، سبب محبوبیت آن شده است.

کارهای پیشین

توسعهٔ راه‌حل‌ها برای مسئلهٔ بازشناسی موجودیت‌های

نام‌دار در زبان عربی، به تبع پیشـــرفت الگوریتم‌ها و توان محاسباتی در طول زمان رشد کرده است.

اولین تلاش‌ها برای حل این مســـئله، آن طور که در [۱] تشریح شده، سیستم‌هایی بوده‌اند که براساس قواعد دستی کار می‌کردند. پس از آن‌ها روش‌هایی جایگزین شـــدند که سعی داشـــتند تا با اســـتفاده از ویژگی‌های مشخصی که دوباره توسط متخصص و به صورت دستی تعیین می‌شد، مسئله را به یک مسئلهٔ دسته‌بندی در یادگیری ماشین تبدیل کنند. در نهایت روش‌های مبتنی بر یادگیری ژرف جایگزین شدند.

جدول (۱): دادگان بازشناسی موجودیت‌های نام‌دار در عربی

دادگان	سال	منبع	تعداد موجودیت
ANERcorp	۲۰۰۷	درگاه، مجلات، اخبار	۴
ACE 2003	۲۰۰۴	اخبار	۷
ACE 2004	۲۰۰۴	اخبار	۷
ACE 2005	۲۰۰۵	اخبار	۷
ACE2007	۲۰۰۷	اخبار	۷
REFLEX	۲۰۰۹	اخبار رویترز	۴
AQMAR	۲۰۱۲	ویکی‌پدیای عربی	۷
OntoNotes 5.0	۲۰۱۳	اخبار	۱۸
WDC2014	۲۰۱۴	ویکی‌پدیا	۴
DAWT	۲۰۱۷	ویکی‌پدیا	۴
CANERCorpus	۲۰۱۸	دینی	۲۰
Wojood	۲۰۲۲	ویکی‌پدیا	۲۱

روش‌های حل مسئلهٔ بازشناسی موجودیت‌های نام‌دار را می‌توان به طورکلی به چهار دستهٔ زیر تقسیم کرد:

روش‌های مبتنی بر قاعده

روش‌های مبتنی بر یادگیری ماشین

روش‌های مبتنی بر یادگیری ژرف

روش‌های مبتنی بر مدل‌های زبانی پیش‌آموخته

در ادامـــه هر کدام از این روش‌ها و پژوهش‌هایی که از آن‌ها استفاده کرده‌اند بررسی خواهد شد.

روش‌های مبتنی بر قاعده

روش‌های مبتنی بر قاعده در [۲,۳,٤] مطالعه شده‌اند. در این پژوهش‌ها، قواعد دستور زبانی برای مکان‌هایی که نام گونه‌ها در آن ظاهر می‌شوند، توسط متخصصان پیشنهاد شده‌اند. در همان ابتدای سیر تکامل روش‌های بازشناسی موجودیت‌های نـــام‌دار برای عربـــی، [٥] یک روش مبتنی بر قاعده پیشـــنهاد کرد که براســـاس ویژگی‌های صرفی در زبـــان عربی، به دنبال الگوها می‌گشـــت. [٦] پس از آن، روشی پیشنهاد کرد که در آن، با استفاده از تحلیل صرفی و یک مجموعـــه از کلیدواژه‌ها که احتمال نام‌دار بودن یک موجودیت را بیشتر می‌کردند، توانست به امتیاز F1 =۰/۸۹ برای تشخیص نام افراد برسد.

روش‌های مبتنی بر ویژگی

روش‌هـــای مبتنـــی بـــر ویژگـــی [۷] از الگوریتم‌های یادگیری ماشین برای تصمیم‌گیری در خصوص برچسب موجودیت‌ها اســـتفاده می‌کنند. کارها در این پژوهش‌ها با [۸] شروع شـــد که براساس الگوریتم آنتروپی بیشینه کار می‌کـــرد اما در پیدا کـــردن موجودیت‌هایی که بیش از یک کلمه طول داشتند، عملکرد خوبی نداشت. سپس، به منظور بهبود عملکرد، [۹] پیشـــنهاد شـــد که ویژگی‌های صرفی، برچســـب‌های اجزای ســـخن و تعداد زیادی ویژگی از این دست را برای دســـته‌بندی اســـتفاده می‌کرد اما هم‌چنان، روش همـــان آنتروپی بیشـــینه بود. اولین اســـتفاده‌ها از الگوریتم CRF در همان حوالی در [۱۰] رخ داد که دو درصد بهتر از نتیجهٔ قبل خروجی می‌داد.

روش‌های مبتنی بر یادگیری ژرف

برای رفع مشـــکل غیرخودکار بودن انتخاب ویژگی‌ها، مدل‌های یادگیری ژرف به تدریج پیشنهاد شدند. این مدل‌ها، اغلب به گونه‌ی واحدی در پژوهش‌ها استفاده شده‌اند: ابتدا در خصـــوص نمایش ورودی تصمیم‌گیری شـــده اســـت، سپس معماری مدل تعیین شده است و در نهایت، برچسب موجودیت، از خروجی مدل حســـاب شده است. گزینه‌های قابل انتخاب برای نمایش برداری، درونه‌ســـازی در سطح کلمه یا حرف بوده اســـت اما در مراحـــل دیگر، انتخاب‌ها

متنوع بوده است. برای به کار بستن مدل‌های ژرف، [۱۱] نمایش Wrod2Vec را برای ورودی به مدل داده است. مدل طراحی شده، سه لایه دارد. در لایه ورودی آن، نمایش برداری هر کلمه، با همسایگان مجاور بلافصل آن ترکیب شده است. لایهٔ بعدی، به منظور یادگیری ویژگی‌های ترکیبی از نمایش برداری در لایهٔ قبل در مدل تعبیه شده است. لایهٔ خروجی هم نمایش لایهٔ میانی (مخفی) را به برچسب گونه‌های موجود تبدیل می‌کند. این ساختار در بسیاری از پژوهش‌ها دیده می‌شود که در آن یک شبکهٔ عصبی نه‌چندان ژرف، با سه لایه استفاده می‌شود:

لایهٔ ورودی: این لایه، نقش گرفتن نمایش ورودی‌ها و ترکیب آن‌ها را دارد.

لایهٔ سیاق: این لایه، ویژگی‌های مهم را در متن یاد می‌گیرد.

لایهٔ خروجی: این لایه نقش دسته‌بند را ایفا می‌کند و برچسب را حدس می‌زند.

با این ساختار توضیح داده شده در بالا، پژوهش‌های زیادی انجام شده است. به طور خاص، در [۱۲،۱۳] از RNN برای یادگیری ویژگی‌ها در لایهٔ سیاق استفاده شده است. در [۱۴] از یک BLSTM برای همین منظور استفاده شده است. BLSTM، به طور موثری در استفاده از ویژگی‌های گذشته و آینده در ورودی توان‌مند است [۱۵] به طور خاص، در همین مرجع اخیر، ابتدا درونه‌سازی در سطح کلمه و همین‌طور درونه‌سازی در سطح حروف، به عنوان ورودی مدل به کار گرفته می‌شود. سپس یک لایه برای یک‌پارچه کردن این دو نوع ویژگی به مدل اضافه می‌شود که در حقیقت سازوکار توجه[1] را پیاده‌سازی می‌کند. خروجی این لایه، در واقع پردازش شدهٔ ورودی توسط مدل است. این ورودی پردازش شده، سپس به یک BLSTM داده می‌شود که در نوبت بعدی، خروجی‌های آن به لایهٔ دسته‌بندی داده می‌شود. وقتی باب استفاده از شبکه‌های بازگشتی در این مسئله باز شد، قطعا پژوهش‌هایی که انواع دیگری از این مدل‌ها را هم امتحان کنند، پیشنهاد می‌شوند. به دلیل وجود ساختارهای ریختی[2] غنی در زبان عربی، این ایده به‌طور طبیعی به ذهن می‌رسد که درونه‌سازی در سطح زیرکلمه توانایی این را دارد تا شباهت ظاهری بین لغات را بهتر مدل کند. به همین خاطر، این موارد در پژوهش‌های اخیر بررسی شده است. برای مثال در [۱۶] ایده این بوده است که از یک BLSTM برای کد کردن حروف و سپس الصاق نمایش به‌دست آمده برای اولین و آخرین حرف هر کلمه، برای نمایش برداری برای هر کلمه استفاده شود. واضح است که با توجه به ساختار BLSTM که در فصل قبل توضیح داده شد، آموزش توصیف شده به روش مذکور، اثری از هر حرف، در حروف مجاورش خواهد گذاشت و لذا الصاق بردار اولین و آخرین حرف کلمه بعد از آموزش، ردّی از سایر حروف کلمه را هم در خود دارد. این پژوهش، سپس خروجی‌های به‌دست آمده برای هر کلمه را به یک دسته‌بند می‌دهد و تاثیر استفاده از درونه‌سازی در سطح کلمه را در حل مسئلهٔ بازشناسی موجودیت‌های نام‌دار روی دادگان بازشناسی موجودیت‌های نام‌دار توئیتر [۱۷] بررسی می‌کند.

روش‌های مبتنی بر مدل زبانی پیش‌آموخته

در سال‌های اخیر، مدل‌های زبانی پیش‌آموخته (PLM[3]) توانایی کشف اطلاعات مربوط به سیاق در خصوص متن را نشان داده‌اند و از این رو، در بسیاری از مسائل حوزهٔ پردازش زبان‌های طبیعی جا باز کرده‌اند. بیشتر PLMها معماری مبتنی بر تبدیل‌گرها را برای آموزش استفاده می‌کنند. BERT یکی از معروف‌ترین مدل‌های توسعه داده شده در این حوزه است که مدل‌های زیادی براساس معماری آن برای سایر زبان‌ها نیز توسعه داده شده است. اولین تلاش‌ها برای آموزش چنین مدلی برای زبان عربی، توسط [۱۸] انجام شد. این مدل، نقش زیادی در پژهش‌های بعدی خود ایفا کرده است. در بین این پژهش‌ها، [۱۹] بسیار به پژوهشی که در این نوشتار انجام شده است، نزدیک است. در [۱۹]، به منظور حل مسئلهٔ بازشناسی

1-Attention Mechanism

2-Morphological

3-Pre-trained Language Model

موجودیت‌های نـــام‌دار روی دادگان CANERCorpus روی خروجـــی BERT، یک مـــدل BLSTM/BGRU-CRF آموزش داده شـــده اســـت و بهترین نتیجه‌ای که گرفته شده است، F1= ۰/۹۴۷۶ بـــوده اســـت. این احتمـــالا بهترین نتیجه‌ای اســـت که تا کنون روی CANERCorpus منتشر شده است. پژوهش‌ها برای استفاده از مدل‌های زبانی از پیش‌آموخته در حل مسئله بازشناســـی موجودیت‌های نام‌دار در زبان عربی، به این‌جا خلاصه نمی‌شـــود و [۱]، موارد بســـیاری از چنین پژوهش‌هایی را ذکر کرده است؛ اما، پژوهش‌های ذکر شـــده اغلـــب ایدهٔ جدیدی کـــه در انجام بازشناســـی موجودیت‌هـــای نـــام‌دار روی دادگان مربـــوط بـــه عربی کلاسیک ندارند.

مرتبط‌سازی موجودیت‌های نام‌دار

کارهای پیشین

روش‌های ســـنتی مبتنی بر یادگیری ماشـــین در حل مســـئلهٔ مرتبط‌ســـازی موجودیت‌هـــا، به خوبـــی در [۲۰] گردآوری شـــده‌اند. لذا در مرور ادبیـــات، از این پژوهش استفاده خواهیم کرد. اگر چه که برخی پژوهش‌های منتشر شده بعد از آن را که به موضوع پژوهش ما مرتبط بوده‌اند، نیز بررسی خواهیم کرد.

وجوه تفاوت در روش‌ها

چنان‌چه بخواهیم یک دورنما از دسته‌بندی این روش‌ها داشته باشیم، مشـــابه روش‌های موجود برای بازشناسی موجودیت‌های نام‌دار، ســـیر روش‌های اســـتفاده شده در پژوهش‌ها برای معرفی موجودیت‌های نام‌دار هم بســـیار متاثر از پیش‌رفت‌های جدید در توسعهٔ مدل‌های ژرف بوده است. چنان که می‌توان به دو دستهٔ کلی زیر تقسیم کرد:

- روش‌های مبتنی بر یادگیری ماشین سنتی
- روش‌های مبتنی بر یادگیری ژرف

در ادامه سعی می‌شود تا این روش‌ها به تفکیک مرور شـــوند. اگرچه که تمرکز اصلی، روی مرور روش‌های مبتنی بر یادگیری ژرف خواهد بود و روش‌های مبتنی بر یادگیری ماشـــین صرفا در حد اشاره مرور خواهند شد.

روش‌های مبتنی بر یادگیری ماشین

این روش‌ها، اساســـا بر ویژگی‌های طراحی شـــده به صورت دســـتی اعم از محبوبیت موجودیت، سازگاری با سیاق محلی و وابستگی‌های سراسری در سطح متن، مبتنی هســـتند و مســـئلهٔ پیدا کردن موجودیت مرتبط با نام ذکر شده را به یک مسئلهٔ رتبه‌بندی تقلیل می‌دهند. الگوریتم‌های رتبه‌بنـــدی در ایـــن روش‌هـــا، در دو دســـتهٔ بانظـــارت و بدون‌نظـــارت قابـــل اشـــاره‌اند. روش‌هـــای بدون‌نظارت، شـــامل روش‌های مبتنی بر مدل‌های فضای برداری [۲۱] و روش‌های مبتنـــی بر بازیابی اطلاعات [۲۲] می‌شـــوند. روش‌هـــای با نظـــارت، به‌طـــور عمده شـــامل روش‌های دســـته‌بندی دودویـــی [۲۳]، روش‌هـــای یادگیـــری برای رتبه‌بنـــدی[4] [۲۴]، روش‌های احتمالاتـــی [۲۵] و روش‌های مبتنی بر گراف [۲۶] می‌شـــود. اغلب ایـــن روش‌ها، فرایند مرســـومی را طی می‌کنند که طـــی آن، ابتدا در یک مرحله، تعدادی ویژگی به‌صورت دستی استخراج می‌شود و سپس این ویژگی‌ها به روشی که برای رتبه‌بندی موجودیت‌های مشـــابه، انتخاب شـــده اســـت، ورودی داده می‌شوند و به همین خاطر، دو محدودیت اساسی دارند:

ویژگی‌هایـــی که منجر بـــه عملکرد خوبـــی برای مدل می‌شـــوند، اغلب دقت و مهارت بالایی در مرحلهٔ استخراج ویژگی‌ها می‌طلبند.

مدل‌هایی که به این شـــیوه‌ها آموزش داده می‌شـــوند، اغلب بسیار به پایگاه‌دانش یا متنی که برحسب آن آموزش داده شـــده‌اند، وابسته می‌شـــوند. به طوری که استفاده از آن‌هـــا روی پایگاه دانش‌هایـــی در همان حـــوزه، معمولا نتایجی به مراتب ضعیف‌تر از آن‌چه گزارش شده، می‌دهد.

روش‌های مبتنی بر یادگیری ژرف

در مقایســـه با روش‌های سنتی، این روش‌ها، می‌توانند

4-Learning to Rank

به طـور خودکار، ویژگی‌های مهم را یـاد بگیرند و لذا، با سـهولت به مراتب بیشـتری قابل انتقال به مسائل مشابه دیگرنـد. برای مثال، روش‌های مبتنـی بر یادگیری عمیق، قادرند تا یک نمایش برداری برای کلمات از طریق مدل‌های پیش‌آموختـه، نظیـر Word2Vec [۲۷] و GloVe [۲۸] بـه دسـت بیاورند. این نمایش می‌تواند همزمان، دربردارندهٔ ویژگی‌های معنایی و نحوی برای نام‌ها ذکر شده باشد که دسته‌بند ورودی داده می‌شود.

بعـد از بازشناسـی موجودیت‌هـای نـام‌دار، مرحلهٔ تشـخیص موجودیت‌های نام‌دار در سـه مرحلـه دنبال می‌شود:

تولیـد موجودیت‌هـای نامزد: این مرحلـه، به تولید مجموعهٔ موجودیت‌هـای محتمل در پایگاه دانش برای هر موجودیت ذکر شده در متن اختصاص دارد.

رتبه‌بنـدی موجودیت‌هـای نامزد: در ایـن مرحله برحسب شواهد متنوعی، نامزدهای تعیین شده در مرحلهٔ قبل، رتبه‌بندی می‌شوند.

پیش‌بینی موجودیت‌های غیرقابل مرتبط‌سازی: این مرحله تعیین می‌کند که چه زمان یک موجودیت باید به NIL نگاشته شود.

تکنیک‌هـای مرحلهٔ اول و سـوم، چنـدان تغییری در سال‌های اخیر نداشته‌اند و عمدهٔ تغییرات مربوط به مرحلهٔ رتبه‌بندی موجودیت‌های نام‌دار است. آن چه در خصوص کاربـرد شـبکه‌های عمیق در حل مسـئلهٔ مرتبط‌سـازی موجودیت‌هـای نام‌دار اهمیـت دارد، این اسـت که آن‌ها می‌توانند به طور خودکار، سـطوح مختلفی از نمایش‌های توزیع‌شده را برای اسـناد و پایگاه‌دانش به‌دست بیاورند که در ابهام‌زدایی بدون مداخلهٔ انسان می‌تواند بسیار موثر باشد. در بین همهٔ پژوهش‌هایی که از مدل‌های عمیق برای حل مسـئلهٔ مرتبط‌سازی موجودیت‌ها اسـتفاده می‌کنند، می‌توان سـه تکنیک کلی برای رتبه‌بنـدی موجودیت‌های نامزد را مشاهده کرد:

درونه‌سـازی[۵]: از آن‌جا که یک درونه‌سـازی، بردار چگال و کم‌بُعدی اسـت که هم نحـو و هم معنا را در خود دارد، مدل‌های مبتنی بر یادگیری عمیق، به عنوان ورودی از آن‌ها اسـتفاده می‌کنند. در چهار مـورد مختلف هنگام حل مسـئلهٔ معرفی موجودیت‌های نام‌دار، نیاز به محاسبهٔ درونه‌سـازی می‌تواند بـروز کند که همین مـوارد، وجه تفاوت بین پژوهش‌ها بوده‌اند:

- درونه‌سازی کلمه[۶]
- درونه‌سازی موجودیت‌های ذکر شده[۷]
- درونه‌سازی موجودیت‌های نام‌دار[۸]
- درونه‌سازی تراز [۹]

ویژگی‌ها: ویژگی‌ها، به منظور اندازه‌گیری شباهت بین موجودیت‌های ذکر شـده در متن و موجودیت‌های نامزد برای آن‌ها در پایگاه‌دانش تعریف می‌شوند. این ویژگی‌ها، به‌طور خودکار توسـط مدل‌های عمیق برای متن محاسبه می‌شوند. در پژوهش‌های مرتبط، معمولا پنج دسته ویژگی زیر موردنظر پژوهش‌گران بوده‌اند:

- محبوبیت پیشین[۱۰]
- شباهت ریختی
- شباهت گونه
- شباهت سیاق
- انطباق موضوعی

الگوریتم: آخرین وجه تفاوت پژوهش‌های موردنظر، الگوریتم اسـتفاده شـده برای رتبه‌بندی نامزدها با فرض معلوم بودن ویژگی‌های آن‌هاسـت. الگوریتم‌های استفاده شده، شـامل PageRank، MLP و یادگیری تقویتی هستند. خروجی این الگوریتم‌ها، یک رتبه‌بندی روی نامزدها برای هر موجودیت ذکر شـده در متن اسـت. پس به طورکلی، پژوهش‌ها برحسـب نوع الگوریتمی کـه برای انتخاب بین نامزدها استفاده کرده‌اند، در سه دستهٔ زیر جا می‌گیرند:

- MLP
- الگوریتم‌های مبتنی بر گراف

5-Embedding

6-Word embedding
7-Surface form embedding
8-Entity embedding
9-Alignment embedding
10-Prior popularity

– الگوریتم‌های تقویتی

در ادامه، سعی می‌شود تا در بخش‌ها، این جنبه‌های تفاوت و آثاری که هر یک بر نتایج پژوهش‌ها گذاشته‌اند، مرور شوند.

تفاوت در درونه‌سازی

همان‌طور که در فصل قبل هم مرور شد، ایدهٔ اصلی در پس درونه‌سازی، نمایش معنای یک تکه از کلام با یک بردار کم‌بُعد و چگال است به گونه‌ای که حافظ فاصلهٔ معنایی باشد؛ یعنی ورودی‌ها با معنای شبیه به هم را به بردارهایی با فاصله‌ی نزدیک به هم بنگارد. در پژوهش‌ها، درونه‌سازی در سطوح مختلفی استفاده شده است که در ادامه می‌خواهیم انواع آن را بررسی کنیم.

درونه‌سازی کلمه

درونه‌سازی کلمه، آن‌طور که از نام آن مشخص است، به هر کلمه، یک بردار حقیقی نسبت می‌دهد. به دو صورت می‌توان درونه‌سازی کلمات را به دست آورد. یکی این‌که از مدل‌های پیش‌آموخته استفاده شود و دیگری این‌که یک مدل، آموزش داده شود که بردارهایی با چنان خصوصیاتی بدهد. در بین پژوهش‌هایی که از مدل‌های پیش‌آموخته برای محاسبه درونه‌سازی کلمه استفاده شده است، [۲۹,۱۵,۳۰,۳۱,۳۲] مواردی هستند که از Word2Vec استفاده کرده‌اند و [۳۳,۳٤] مثال‌هایی از استفادهٔ GloVe به عنوان مدل پیش‌آموخته‌ای که درونه‌سازی کلمه با استفاده از آن استفاده می‌شود، هستند. اگر چه که استفاده از مدل‌های پیش‌آموخته به خاطر حذف هزینه‌های آموزش داده بسیار محبوب است، اما پژوهش‌های زیادی هم وجود دارند که به دلیل نبود مدل‌های پیش‌آموختهٔ مناسب، مجبور به آموزش مدل‌ها با دادهٔ خود شده‌اند. این پژوهش‌ها، اغلب مربوط به حوزه‌های خاصی بوده‌اند. برای مثال، [۳۵] برای تولید نمایش برداری برای کلمات، از آموزش Sent2Vec که یک توسعه از CBoW است استفاده کرده است.

درونه‌سازی نام ذکر شده

نام‌های ذکر شده در متن هم نیاز است که به منظور محاسبهٔ شباهت با نامزدها، به صورت برداری نمایش داده شوند و لذا باید برای آن‌ها هم درونه‌سازی محاسبه کرد و از این وجه، پژوهش‌ها، رویکردهای مختلفی اتخاذ کرده‌اند. ساده‌ترین رویکرد ممکن، آن‌طور که توسط [۳٦,۳۷] اتخاذ شده است، این بوده که صرفا درونه‌سازی کلمات برای کلمات تشکیل‌دهندهٔ نام ذکر شده در متن را میانگین بگیرند. رویکردی که برخی تفاوت‌های مهم در سطح کلمات یا زیر آن را محو می‌کند و از اینرو چندان جالب نیست. روش‌های پیچیده‌تر و به تبع آن، موثرتر در خصوص شیوهٔ محاسبهٔ درونه‌سازی نام‌های ذکر شده از روی درونه‌سازی کلمات تشکیل‌دهندهٔ آن پیشنهاد شده‌اند. یک دسته از روش‌ها که به‌عنوان نمونه در [۳۸,۳۹,٤۰] به آن پرداخته شده و به نسبت نتایج جالب‌تری داده است، این بوده که به جای آن‌که به طور ایستا، نحوهٔ ترکیب درونه‌سازی کلمات تشکیل‌دهندهٔ یک موجودیت ذکر شده در متن را برای به‌دست آوردن درونه‌سازی نهایی آن تعیین کنیم، اجازه بدهیم تا یک مدل هوشمند، خود با استفاده از داده‌ها، تصمیم بگیرد که چگونه این بردارها را برای رسیدن به بردار نهایی با هم ترکیب کند. مثلا در مرجع اخیر، درونه‌سازی کلمات تشکیل‌دهندهٔ نام ذکر شده در متن، به یک مدل SLP[11] داده می‌شود و برحسب داده‌های آموزش، پارامترهای مدل تعیین می‌شوند.

درونه‌سازی گونه‌ها

نیازمندی محاسبهٔ درونه‌سازی گونه‌های موجود در پایگاه دانش، به منظور محاسبهٔ میزان شباهت آن‌ها با نام‌های ذکر شده در متن، در مرحلهٔ تولید نامزدها یا انتخاب آن‌ها بروز می‌کند. پژوهش‌ها در خصوص درونه‌سازی گونه‌ها به فراخور غنای پایگاه‌دانشی که در اختیار داشتند، روش‌های مختلفی برای محاسبهٔ درونه‌سازی گونه‌ها استفاده کرده‌اند. در غنی‌ترین حالت ممکن، از شکل‌های

11- Single Layer Perceptron

مختلف نام‌های آن گونه، توضیحات موجود، سیاق گونه و اطلاعات سلسله‌مراتبی که گونه متعلق به آن است در محاسبهٔ درونه‌سازی گونه استفاده شده است؛ اما در پژوهشی که ما انجام داده‌ایم، پایگاه‌دانشی که در اختیار بوده، صرفا شامل یک شناسهٔ عددی برای افراد است که برچسب‌ها بر اساس آن زده شده‌اند. در واقع، آن‌چه پایگاه‌دانش را تشکیل داده است، سطرهایی است که هر کدام از آن‌ها یک شناسهٔ عددی به همراه شکل‌های مختلف ذکر شدن گونه‌ها در داده‌های برچسب‌خورده هستند. در بین پژوهش‌ها، به عنوان نمونه، [۳۷] درونه‌سازی گونه‌ها را براساس میانگین نمایش برداری شکل‌های مختلف آن به‌دست آورده است. [۴۱] با استفاده از CNN و [۴۲] با LSTM سعی کرده‌اند تا درونه‌سازی گونه را از روی درونه‌سازی شکل‌های مختلف آن به طور پویا براساس داده‌های برچسب‌خورده به‌دست بیاورند.

تفاوت در ویژگی‌ها

به لطف ابزارهای فراهم شده توسط یادگیری ژرف و به طور خاص، به لطف درونه‌سازی‌ها، اغلب راه‌حل‌ها، از درونه‌سازی به طور مستقیم و یا برای طراحی سایر ویژگی‌ها استفاده می‌کنند. با این وجود، هنوز هم برخی از ویژگی‌هایی که در ادامه بحث خواهند شد، به طور دستی طراحی می‌شوند.

محبوبیت پیشین

محبوبیت پیشین، احتمال وقوع یک گونه در متن با فرض داشتن نام ذکر شده در متن، بدون در نظر گرفتن سیاق است. این ویژگی، اگرچه که خیلی ساده است، اما با این حال، بسیار قدرتمند است و تقریبا در همهٔ پژوهش‌های اخیر استفاده شده است. به طور دقیق‌تر، محبوبیت پیشین یک گونه برای یک نام ذکر شده، نسبت تعداد مواردی است که آن نام ذکر شده، مربوط به آن گونه بوده، به کل تعداد دفعاتی که آن نام، در متن ذکر شده است. دقت کنید که این مقدار، روی دادهٔ برچسب خورده محاسبه می‌شود.

شباهت شکل حروف

به طور شهودی، احتمال مربوط بودن یک نام ظاهر شده در متن، به گونه‌ای که از نظر ظاهری بیشتر به آن شبیه است بیشتر از گونه‌ای است که از نظر ظاهری کم‌تر به آن شبیه است. البته که این شهود، در اغلب موارد درست است و یک قاعدهٔ دقیق و کلی نیست. برای تعیین شباهت ظاهری، معیارهای متعددی وجود دارد؛ فاصله‌ی ویرایشی، امتیاز دایس، SBD و فاصلهٔ همینگ از جملهٔ این متریک‌ها هستند. در پژوهش‌های اخیر، روش‌های مبتنی بر یادگیری ژرف برای محاسبهٔ شباهت ظاهری ابداع شده‌اند. به‌طور خاص، در [۴۱,۴۳,۴۴] فاصله کسینوسی درونه‌سازی نام‌ها برای محاسبهٔ شباهت بین آن‌ها استفاده شده است.

شباهت نوع

در پژوهش‌ها، یک ویژگی دیگر که در محاسبهٔ شباهت نام‌ها استفاده می‌شود، توجه به نوع یک کلمه است. مثلا این که یک نام ذکر شده در متن، نام یک شهر است یا یک فرد، یا مثلا «صفر» نام یک فرد است، نام یک تاریخ است یا یک عدد. از آن‌جا که در پژوهش پیش رو، فقط به دنبال معرفی افراد هستیم، این ویژگی‌ها، چندان در رفع ابهام به کار ما نمی‌آیند.

شباهت سیاق

شباهت سیاق، اصلی‌ترین ویژگی مورد استفاده الگوریتم‌های موجود برای پیدا کردن گونه مرتبط با یک موجودیت است. هسته اصلی نوآوری در پژوهش‌ها، در ابداع الگوریتم‌هایی است که برمبنای شباهت سیاق، گونه مرتبط با نام ذکر شده را پیدا می‌کنند. منظور از سیاق یک عبارت، لغت‌های اطراف آن هستند. یک راه‌حل ساده برای محاسبهٔ سیاق یک کلمه، استفاده از رویکرد BoW [۱۲] است چنان که تا یک فاصلهٔ طبیعی ثابت، هر لغتی که در همسایگی یک کلمه هست به عنوان مجموعه سیاق آن کلمه در نظر گرفته می‌شود و سپس برای محاسبهٔ شباهت

12-Bag of Words

سیاق‌ها، تعداد کلمات مشـــترک در سیاق دو کلمه شمرده می‌شـــوند. این روش به‌عنوان مثال در [٤٥] استفاده شده اســـت. در حال حاضر، دیگر از چنیـــن پژوهش‌هایی برای محاســـبهٔ ویژگی‌های مبتنی بر سیاق اســـتفاده نمی‌شود. بلکه اغلب روش‌های ژرف جایگزین شده‌اند. این روش‌ها، عمدتا همان‌هایی هستند که در مرور روش‌های حل مسئله بازشناسی موجودیت‌های نام‌دار گذشت. به طور مختصر، استفاده از مدل‌های مبتنی بر معماری GRU، LSTM، RNN، CNN و نهایتا مدل‌های مبتنی بر تبدیل‌گرها، روش‌هایی است که از آن برای به‌دست آوردن سیاق یک کلمه استفاده می‌شود. پس از پیدا کردن نمایش برداری یک کلمه مبتنی بر ســـیاق آن، معمولا شباهت کسینوسی بین بردار نمایش‌ها، مدنظر محاسبهٔ سیاق است.

شباهت موضوعی

این نوع شـــباهت، که از آن به عنوان شباهت سراسری هم یاد می‌شود، چنین تعریف می‌شود که با داشتن یک سند، شـــباهت موضوعی یک گونه برای آن‌که به عنوان یک نام ذکر شده در متن این سند معرفی شود، به صورت میانگین شـــباهت گونه موردنظر، با همه نام‌های ذکر شـــده در آن سند تعریف می‌شـــود. به این ترتیب، مشخص می‌شود که یک گونه، چقدر با یک ســـند شبیه اســـت و از این شباهت، برای رتبه‌بندی نامزدها می‌توان اســـتفاده کرد. با همه این توضیحات، در مسئله‌ای که قصد حل آن را داریم، موضوع همه سندها، به هم شبیه است و لذا این ویژگی، چندان موثر نخواهد بود.

تفاوت در الگوریتم

منظور از الگوریتم، رویه‌ای است که در پژوهش برای استفاده از ویژگی‌ها برای تعیین گونه‌ی مربوط به نام ذکر شده استفاده می‌شود. در پژوهش‌ها، برای این کار، از سه دسته از الگوریتم‌ها اســـتفاده شده است که ادامه بررسی می‌شوند.

MLP

برای هر گونهٔ نامزد و هر نام ذکر شـــده در متن، MLP ویژگی‌هـــا را به یک امتیاز تبدیل می‌کند و بر اســـاس این امتیاز، گونهٔ پیشنهادی خود را معرفی می‌کند. ایده‌ای اصلی این اســـت که تابع امتیازدهی با استفاده از یک مدل عصبی تخمین زده شـــود. برخی از پژوهش‌ها، نظیر [٤٦,٤٧,٤٨] از یک مدل پرســـپترون یک لایه، استفاده کرده‌اند تا امتیاز گونه‌ها را به کمک آن حســـاب کنند ســـپس گونه با امتیاز بیشتر را گزارش کرده‌اند. برخی دیگر از پژوهش‌ها، نظیر [٤٢] نیز از مدل‌هایی با چندلایه استفاده کرده‌اند.

الگوریتم‌های گرافی

کاربـــرد الگوریتم‌هـــای مبتنی بر گراف، به نســـبت جدیدتر از سایر حوزه‌هاست. مسئله در این الگوریتم‌ها، به گونه‌های مختلفی مدل می‌شود. از موارد اولیهٔ استفاده این گونه الگوریتم‌ها برای کاربرد موجودیت‌های نام‌دار، می‌توان [۳۰] را نام برد که مســـئله در آن این‌گونه مدل شده است که همه نامزدها برای همهٔ نام‌های ذکر شده در متن، گره‌های گراف‌اند و شـــباهت بین موجودیت‌ها هم به یال‌های وزن‌داری بین این گره‌ها تبدیل می‌شود. حالا برای هر گونهٔ نام‌دار یک رتبه فرض می‌شـــود که با استفاده از الگوریتم PageRank، سعی می‌شود تا این رتبه تخمین زده شـــود. بجز این پژوهش، پژوهش‌های دیگری هـــم بوده‌اند که از ایده‌هـــای قدم‌زنی تصادفی استفاده کرده‌اند. مثلا [٤٣] یکی از آن‌هاست که باز هم گونه‌های نام‌دار را به عنوان گره می‌گیرد و سعی می‌کند با تعریف ماتریس گذار بر اساس شباهت موجودیت‌ها، رتبهٔ نامزدها را تخمین بزنند.

۶. روش‌های پیشنهادی

بازشناسی موجودیت‌های نام‌دار

برای تشـــریح توســـعهٔ انجام شـــده در بازشناســـی موجودیت‌های نام‌دار ابتدا دادگان مورد اســـتفاده و سپس

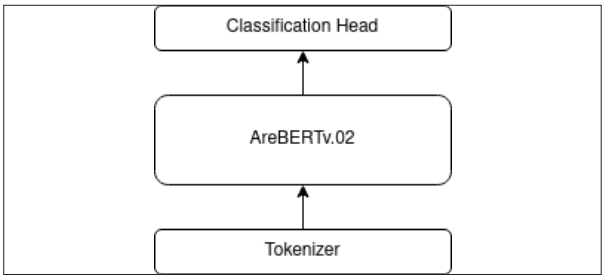
مدل را بررسی می‌کنیم. تشریح پیاده‌سازی و نتایج در بخش بعد انجام خواهد شد.

دادگان

دادگان مورد استفاده در این مسئله، همان CANERCorpus است که پیشتر گذشت. این مجموعه داده، نمونه خوبی برای پژوهش در علوم حدیثی از جهت نزدیکی سیاق است؛ یعنی سایر متون مذهبی هم، از حیث سیاق کلمات، تفاوت چندانی با آن ندارند و امید می‌رود که پژوهش‌ها را بتوان بدون کاهش معنی‌دار نتایج، به آن‌ها منتقل کرد.

مدل

مدلی که برای بازشناسی موجودیت‌های نام‌دار از آن استفاده کردیم، مدلی به نام AraBERTv02 است که [۱۸] معرفی کرده است. این مدل، همان‌طور که از نام آن مشخص است، یک مدل زبانی پیش‌آموخته مبتنی بر معماری BERT است که روی متون زبان عربی، آموزش داده شده است. رویهٔ استفاده از این مدل، یا هر مدل پیش‌آموخته دیگری برای استفاده برای تشخیص موجودیت‌های نام‌دار، به این شکل است که ابتدا باید مدل را روی دادگان CANERCorpus، آموزش دهیم تا مدل با متن آشنا شود. این از آن حیث ضروری است که این مدل، برای زبان عربی مدرن استاندارد پیش‌آموخته شده است و چنانچه بدون آموزش ثانویهٔ آن روی داده‌هایی در عربی کلاسیک از آن استفاده شود، نتیجه کار چندان موثر نخواهد شد. پس از آموزش، با استفاده باید متن را تکه‌تکه[13] کرد که برای این کار، خود مدل زبانی، دارای چنین واحدی هست. سپس، از آن‌جا که بازشناسی موجودیت‌های نام‌دار یک مسئلهٔ برچسب‌زنی دنباله[14] است، کافی است مدل روی داده‌های برچسب‌خورده آموزش داده شود. در هنگام انجام مسئلهٔ برچسب‌زنی، یک سر دسته‌بند در انتهای BERT قرار می‌گیرد که سعی دارد با استفاده از نمایش بردای کلمات، تشخیص دهد که چه هر کلمه به چه موجودیتی تعلق دارد و سعی می‌شود که یک تابع ضرر، اغلب آنتروپی متقابل، نسبت به داده‌های آموزش کمینه شود.



شکل (۱): معماری شبکه برای بازشناسی

پس از آموزش مدل، می‌توان از آن برای برچسب‌زنی روی داده‌های دیده نشده استفاده کرد. به طور شهودی، مدلی که برای بارشناسی موجودیت‌های نامدار استفاده می‌شود، ساختاری مشابه شکل (۱) دارد. در این شکل، لایهٔ توکنایزر، یک پیش‌پردازش انجام می‌دهد که طی آن متن ورودی تکه‌تکه می‌شود و به شکلی در می‌آید که قابل فهم برای مدل زبانی است. سپس در لایهٔ بعدی این متن تکه‌تکه شده، به مدل زبانی داده می‌شود. خود این لایه، ساختار درونی پیچیده‌ای کاملا مشابه ساختار BERT دارد که مشتمل بر بلوک‌های تبدیل‌گر و لایه‌های دیگر است. در نهایت، در لایهٔ آخر، یک دسته‌بند قرار دارد که نمایش برداری تکه‌ها را دریافت و براساس آن دسته‌بندی را انجام می‌دهد. آن‌چه اهمیت دارد این است که این لایه، یک لایهٔ خطی با تابع فعال‌سازی است.

مرتبط‌سازی موجودیت‌ها

برای مرتبط‌سازی موجودیت‌های نام‌دار، ایده‌های نسبتا جدیدی به کار گرفته شده است. به منظور مدل کردن ارجاعات در سند احادیث، از مدل شبکهٔ عصبی گرافی، به طور دقیق‌تر، نوع پیچشی آن، شبکهٔ هم‌آمیخت گرافی استفاده شده است. در این بخش، به معرفی این ایده‌ها و در فصل بعد، به تشریح پیاده‌سازی و نتایج پرداخته خواهد شد.

13-Tokenize
14-Sequence Labeling

شبکه‌های عصبی گرافی

در مسئله مرتبط‌سازی موجودیت‌های نام‌دار، آن‌چه هوش طبیعی هنگام انجام چنین کاری انجام می‌دهد، این است که هنگامی که یک موجودیت نام‌دار، مثلا «زید» می‌بیند، در گام اول همهٔ موجودیت‌هایی که نام آن‌ها، با این نام شباهت دارد را پیدا می‌کند و سپس در صورت وجود چندین نامزد محتمل برای مرتبط بودن، به روابط روایی بین رجال موجود در سلسله اهمیت می‌دهد. مثلا اگر یک «زید بن عمرو» و یک «زید بن بکر» در بین نامزدهای احتمالی حاضر باشند، نامزدی انتخاب می‌شود که رجال موجود در سلسلهٔ راویان احادیثی که در آن‌ها ظاهر شده باشد، با رجال موجود در این سلسله سازگارتر باشد؛ آن چیزی هم که از سازگاری مدنظر است، در وهلهٔ اول، همسایگان بلافصل موجودیت است، و بعد همسایگان بعدی با درجهٔ اهمیت کمتر. این نکته اخیر، تناسب مسئله با ساختار شبکه‌های هم‌آمیختی روی گراف‌ها را واضح می‌کند. تنها اختیار عمل، در این خواهد بود که نمایش متنی اسامی، چگونه به ویژگی‌هایی برای هر گره تبدیل شوند. این تبدیل، هر چه باشد، خوب است که به ساختارهای نحوی توجه بیشتری از معنا داشته باشد. زیرا متن‌های موردنظر در این مسئله، چندان دارای معانی نیستند و اگر هم باشند در حل این مسئله با کمک هوش طبیعی هم کمک‌کننده نیست.

نمایش برداری اسامی

برای استفاده از ظرفیت‌های شبکه‌های هم‌آمیختی روی گراف‌ها، لازم است تا برای نام‌های ذکر شده برای هر فرد، یک نمایش برداری به‌دست بیاوریم که خوشبختانه، کار مسبوق به سابقه‌ای در مطالعه زبان‌های طبیعی است. از بین روش‌های مرسوم، آن‌هایی که نمایشی برای هر اسم فراهم می‌کنند که اسامی شبیه به هم (به خصوص از حیث نحو) را به بردارهای نزدیک به هم نگاشت کند. به این منظور، در این پژوهش، از همان مدلی که روی مجموعه دادهٔ CANERCorpus آموزش دادیم، استفاده خواهیم کرد. حسن این کار این است که مدلی که آموزش داده شده، با متون حدیثی آشناست و لذا نمایش‌های برداری به دست آمده از آن، نمایش‌های دقیقی خواهند بود.

ویژگی‌ها

به منظور تعیین شباهت نام‌های ذکر شده و گونه‌های نام‌دار، از شباهت سیاق به‌دست آمده توسط نمایش برداری آن‌ها، به همراه محبوبیت پیشین استفاده شده است. به این شکل که در هنگام دسته‌بندی هر نام، امتیاز هر دسته، در محبوبیت پیشین آن دسته برای آن نام ضرب شده است. به بیان دیگر، هر نام، یک بردار از مقادیر محبوبیت‌های پیشین روی همه نام‌ها تعریف می‌کند که این بردار، در بردار امتیاز دسته‌ها برای آن نام، ضرب داخلی شده است.

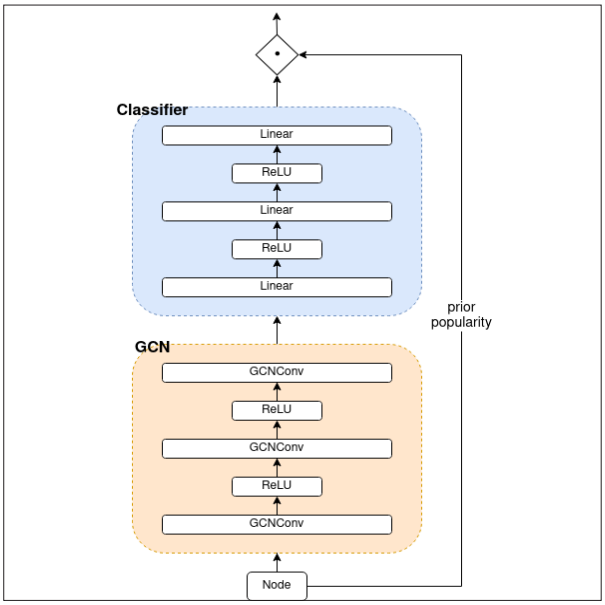
داده و الگوریتم

در خصوص مرتبط‌سازی موجودیت‌های نام‌دار، برخلاف بازشناسی آن‌ها، خیلی دادهٔ برچسب خورده در دسترس نیست. به همین خاطر، پژوهش‌هایی برای تولید داده‌های مصنوعی، به نحوی که به طور قابل اتکایی بتوان از آن‌ها برای آموزش و ارزیابی استفاده کرد صورت گرفته است. به طور مثال، برای زبان عربی، [۴۹] تلاش کرده است تا ۱۸۲۹۸ راوی واقعی حدیث را به صورتی کنار هم بیاورد که هر فرد در کنار افرادی باشد که در احادیث واقعی هم ممکن است. به عبارت دیگر، یک سند ساختی در این مجموعه‌داده، اگر چه که ممکن است هیچ‌گاه در واقعیت دیده نشده باشد، اما هر همسایگی از آن سند، در سندهای واقعی هم دیده شده است. به این ترتیب چیزی در حدود ۲۸۰ هزار سند حدیث تولید شده است که در این پژوهش هم از آن استفاده شده است. مسئله به این صورت مدل شده است که برای هر زوج مرتب نام ذکر شده و گونهٔ مرتبط با آن در پایگاه‌دانش یک گره در نظر گرفته شده است. برای نمونه اگر یک گونهٔ مشخص در پایگاه‌دانش، با اشکال ظاهری متعددی در دادهٔ برچسب‌خورده ظاهر شده است، برای هر یک، یک گره جداگانه در نظر گرفته

شـــده اســـت. هم‌چنین با دخالت دادن برچسب هر داده در ســـاختن گره در این زوج مرتب، اطمینان حاصل می‌شود که گره‌ها صرفا بر اساس نام‌های ذکر شده در متن ساخته نمی‌شوند که منجر به یکی شدن اسامی یکسان اما مربوط به افراد متفاوت شـــود. یعنی اگر یک نام ذکر شده، در دو جای مختلف به افراد مختلفی مربوط شـــده باشـــد، با این روش مدل‌سازی، این اطمینان وجود دارد که گره‌های آن‌ها از هم جـــدا خواهند بود. یال‌ها هم بین راس‌ها، در صورت وجود رابطه‌ی نقل بلافصل ایجاد می‌شود. یعنی دو گره با یال به هم متصل خواهند شـــد اگر و فقط اگر در یک ســـند حدیث، بلافاصله در کنار هم دیده شده باشند. برچسب هر گره هم شناسهٔ عددی آن در پایگاه‌دانش خواهد بود. با این پیکربندی، مسئله مرتبط‌سازی موجودیت‌های نام‌دار، تبدیل به مسئله پیش‌بینی برچســـب گره در شبکهٔ عصبی گرافی خواهد شد. مدل به صورت یک GCN با سه لایهٔ هم‌آمیختی و یک مولفهٔ دســـته‌بندی در انتها ساخته شـــده است. این یعنی، هر گره، اثری از سه گره پس و پیش خود در هنگام تصمیم‌گیری در خصوص برچسب آن دارد. بین این لایه‌ها، تابع فعال‌ساز ReLU استفاده می‌شود. به طور شهودی، در هنگام آموزش، مدل یاد می‌گیرد که چگونه با اســـتفاده از نمایش برداری به‌دست آمده از همسایه‌های هر گره، بتواند برچسب آن را حدس بزند. با آموزش مدل، آن‌چه به دست می‌آید، ضرایبی است که تعیین می‌کند که بردار ویژگی هر فرد، چگونه باید با بردار ویژگی اطرافیانش ترکیب شود تا در هنگام مقایسه با بردار ویژگی موجودیت‌های نام‌دار، به موجودیت مرتبط با خود بیشتر شبیه شود. معماری مدل در شـــکل (۲) نشان داده شده است. در این شکل، هر راس ابتدا توســـط شـــبکهٔ هم‌آمیختی، با همسایگان خود ترکیب می‌شـــود. این مولفه، همان‌طور که گفته شد مشتمل بر سه لایهٔ هم‌آمیختی اســـت که تابع فعال‌سازی ReLU رابط بین آن‌هاســـت. سپس یک مولفهٔ دسته‌بندی مشتمل بر سه لایهٔ خطـــی که بین آن‌ها هم تابع فعال‌ســـازی ReLU قرار دارد، برای هر راس، احتمال قرارگیری در هر دسته را مشخص می‌کند. در این مرحله ایـــن بردار احتمال در بردار مربوط به محبوبیت پیشین دسته‌ها برای نام موجود در آن راس، ضرب داخلی می‌شود و حاصل آن به عنوان خروجی مدل در نظر گرفته می‌شود.

پیاده‌سازی و ارزیابی

در این بخش، به منظور ارزیابی روش‌هایی که در بخش قبل پیشنهاد شدند، نتایج حاصل‌شـــده از پیاده‌سازی این روش‌هـــا و اجرای آنها بر مجموعه دادگان موردنظر برای ارزیابی ارائه می‌شوند. ابتدا نتایج برای مسئلهٔ بازشناسی موجودیت‌های نام‌دار و ســـپس، برای مسئلهٔ مرتبط‌سازی آن‌ها ارائه و بررسی می‌شوند.



شکل (۲): معماری شبکه برای بازشناسی بازشناسی موجودیت‌های نام‌دار

پـــس از مختصـــری توضیـــح در خصـــوص نحـــوه پیاده‌ســـازی و هم‌چنین بررســـی نتایج پرداخته می‌شود. برای آماده‌سازی داده‌ها، نیاز بود تا مجموعه دادهٔ برچسب خورده به قالب BIO[۱۵] تبدیل شوند.

پس از فراهم‌ســـازی داده در این قالب، ابتدا با استفاده از Tokenizer داده‌ها را به نحوی که برای مدل زبانی قابل فهم باشد تکه‌تکه کرده و سپس برای ارزیابی مدل، آموزش را روی هشتاد درصد داده‌ها انجام دادیم و بیست درصد

15-Begin-In-Out

آن برای آزمون نگه داشته شد. مدل در ۱۰ دوره، با نرخ یادگیری ۰/۰۰۰۲ روی این داده، آموزش داده شده است. در هر پایان هر دوره، ارزیابی روی دادهٔ آزمون انجام شده و نتیجه مطابق آن چیزی است که در جدول (۲) قابل مشاهده است. این جدول، معیارها را به صورت سراسری در هر دوره نشان می‌دهد. یعنی مثلا، دقت، دقت سراسری برای همهٔ موجودیت‌های برچسب‌خورده است. این مدل به صورت عمومی نیز در دسترس قرار گرفته است و قابل امتحان کردن است؛ به این شکل که با دادن یک متن حدیثی، موجودیت‌های نام‌دار آن در ۲۱ دستهٔ مشخص را برمی‌گرداند.

جدول (۲): ارزیابی بازشناسی

دوره	ضرر آموزش	ضرر آزمون	Percision	بازآوری	F1	دقت Accuracy
۱	۰.۱۶۶	۰.۱۲۶	۰.۹۳۸	۰.۸۹۸	۰.۹۱۷	۰.۹۷۵
۲	۰.۰۸۱	۰.۰۶۸	۰.۹۴۵	۰.۹۴۹	۰.۹۴۷	۰.۹۸۶
۳	۰.۰۶۲	۰.۰۵۳	۰.۹۵۹	۰.۹۵۹	۰.۹۶۱	۰.۹۸۹
۴	۰.۰۴۵	۰.۰۴۸	۰.۹۶۰	۰.۹۶۰	۰.۹۶۴	۰.۹۸۹
۵	۰.۰۳۴	۰.۰۴۶	۰.۹۵۷	۰.۹۵۷	۰.۹۶۳	۰.۹۸۹
۶	۰.۰۳۲	۰.۰۴۴	۰.۹۶۳	۰.۹۶۳	۰.۹۶۷	۰.۹۹۰
۷	۰.۰۲۸	۰.۰۴۴	۰.۹۶۲	۰.۹۶۲	۰.۹۶۷	۰.۹۹۰
۸	۰.۰۲۶	۰.۰۴۴	۰.۹۶۱	۰.۹۶۱	۰.۹۶۷	۰.۹۹۰
۹	۰.۰۲۴	۰.۰۴۴	۰.۹۶۳	۰.۹۶۳	۰.۹۶۷	۰.۹۹۰
۱۰	۰.۰۲۵	۰.۰۴۳	۰.۹۶۴	۰.۹۶۴	۰.۹۶۸	۰.۹۹۰

مرتبط‌سازی موجودیت‌های نام‌دار

همان‌طور که گفته شد، به منظور مرتبط‌سازی موجودیت‌های نام‌دار، از شبکهٔ گراف هم‌آمیختی روی یک مجموعه دادهٔ ترکیب شده استفاده شده است. پیاده‌سازی به این شکل انجام گرفته که ابتدا با یک دور پیمایش داده‌های برچسب خورده، محبوبیت پیشین برای نامزدهای محتمل هر موجودیت محاسبه شده است؛ به این ترتیب که، برای هر نام ذکر شده در متن، همهٔ برچسب‌هایی که به آن‌ها نگاشته شده است را جمع‌آوری کرده‌ایم. مثلا برای نام «علی» همه برچسب‌های مربوط به افرادی واقعی که در جایی از متن نام آن‌ها به همین شکل ذکر شده است را برحسب تعداد دفعاتی که این نام، مربوط به هر برچسب بوده را به‌دست آورده‌ایم. به منظور نرمال‌سازی، امتیاز همه موجودیت‌ها را در ابتدا ۱ فرض کردیم تا نام‌های جدید دیده نشده هم بخت انتساب داشته باشند. سپس به منظور کارایی بیشتر در پیاده‌سازی، یک بار برای همیشه، درونه‌سازی همه نام‌های ذکر شده در متن را به‌دست آورده و ذخیره کرده‌ایم. بعد از تعریف مجموعه‌داده هر زنجیرهٔ حدیثی را به صورت یک واحد مجزا تعریف کرده‌ایم که خواص راسی آن‌ها همان درونه‌سازی نام ذکر شده است و یال‌ها هم که همان رابطه روایی بین افراد در یک حدیث است. در این مرحله، به همراه هر واحد داده که یک سند حدیث است، نامزدهای احتمالی هر راس را هم در ساختارداده نگه می‌داریم. بعد مدل همان‌طور که در فصل گذشته آمد، تعریف شده است. آن‌چه در این‌جا اشاره به آن اهمیت دارد، این است که تاثیر محبوبیت پیشین، در خروجی همین مدل دیده شده است؛ یعنی، بردار امتیازی که دسته‌بند در آخرین لایهٔ مدل حساب می‌کند، در بردار محبوبیت پیشین ضرب داخلی می‌شود و به این‌ترتیب، برچسب‌های با محبوبیت بیشتر، شانس بالاتری در خروجی نهایی پیدا خواهند کرد. آموزش در ۳ دوره انجام شده است که از نظر اندازهٔ دسته با هم متفاوت هستند. در دوره اول، احادیث به صورت دسته‌های ۲۵۶تایی به مدل داده شده‌اند. در دوره بعد در دسته‌های ۱۲۸تایی و در دوره سوم در دسته‌های ٦٤تایی. نرخ یادگیری در همه‌ی دوره‌ها ۰/۰۰۱ بوده است. معیار بهینه‌سازی در آموزش شبکهٔ عصبی، آنتروپی متقابل بوده است و الگوریتم مورد استفاده برای رسیدن به آن، الگوریتم آدام بوده است. نتیجه ارزیابی مشابه چیزی است که در جدول (۳) آورده شده است. پس از آموزش، این مدل توانست روی مجموعهٔ دادهٔ آزمون، به میزان دقت[16] ۰/۸۵۷۰ دست یابد.

16-Accuracy

جدول (۳): ارزیابی مرتبط‌سازی

دوره	ضرر	Accuracy
۱	۰.۰۰۱۸	۰.۷۹۱۲
۲	۰.۰۰۱۶	۰.۸۵۴۸
۳	۰.۰۰۲۷	۰.۸۶۳۲

۷. جمع‌بندی

در ایـــن پژوهـــش، به منظـــور تشـــخیص و معرفی یا مرتبط‌سازی موجودیت‌های نام‌دار در احادیث، تلاش‌هایی برای حل دو زیرمسئلهٔ مرتبط انجام شد.

ابتدا سعی شـــد تا با استفاده از آموزش دادن یک مدل پیش‌آموخته بر روی داده‌های برچسب‌خورده یک مجموعهٔ روایی، یک مدل زبانی منطبق بر سیاق احادیث که به زبان عربی کلاســـیک نوشته شده‌اند به‌دســـت آید که از آن، هم به عنوان مدل اصلی در بازشناسی موجودیت‌های نام‌دار اســـتفاده شـــود و هم بـــرای به‌دســـت آوردن نهفت برای کارهای بعدی؛ از جمله مرتبط‌سازی موجودیت‌های نام‌دار. این مدل در بازشناسی موجودیت‌های نام‌دار، توانست پس از ۱۰ دوره آموزش به امتیاز F1 =۰/۹۶۸ برســـد. سپس با استفاده از مدل به دست آمده، سعی شد تا برای حل مسئلهٔ مرتبط‌ســـازی موجودیت نام‌دار، ابتدا یک نهفت محاســـبه شـــود و سپس با مدل کردن مســـئله به صورت یک شبکهٔ عصبی گرافی، و تبدیل مســـئلهٔ مرتبط‌سازی موجودیت‌ها به یک مســـئلهٔ پیش‌بینی برچســـب در این گراف، از نهفت محاســـبه شده هم در آموزش شـــبکهٔ هم‌آمیختی گرافی و هم در تعیین نامزدهای موجود برای هر نام ذکر شـــده در متن، استفاده شـــد. به این ترتیب، پس از ۳ دوره آموزش، به میزان دقت (accuracy) ۰/۸۶۳۲ به دســـت آمد که روی داده‌های آزمون، ۰/۸۵۷ حاصل شد.

مراجع

1. X. Qu, Y. Gu, Q. Xia, Z. Li, Z. Wang, and B. Huai. A survey on arabic named entity recognition: Past, recent advances, and future trends. IEEE Transactions on Knowledge and Data Engineering, pages 1–18, 2023.
2. K. Shaalan and H. Raza. Nera: Named entity recognition for arabic. Journal of the American Society for Information Science and Technology, 60(8):1652–1663, 2009
3. S. Abdallah, K. Shaalan, and M. Shoaib. Integrating rule-based system with clas sification for arabic named entity recognition. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 311–322, 2012.
4. S. Abuleil and M. Evens. Extracting names from arabic text for question-answering systems. In RIAO, pages 638–647, 2004.
5. J. Maloney and M. Niv. Tagarab: a fast, accurate arabic name recognizer using high-precision morphological analysis. In Computational approaches to Semitic languages, 1998.
6. T. Buckwalter. Issues in arabic orthography and morphology analysis. In Proceed ings of the workshop on computational approaches to Arabic script-based languages, pages 31–34, 2004.
7. M. Oudah and K. Shaalan. A pipeline arabic named entity recognition using a hybrid approach. In Proceedings of COLING 2012, pages 2159–2176, 2012.
8. Y. Benajiba, P. Rosso, and J. M. Bened´ıruiz. Anersys: An arabic named entity recognition system based on maximum entropy. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 143–153, 2007
9. Y. Benajiba and P. Rosso. Arabic named entity recognition using conditional random fields. In Proc. of Workshop on HLT & NLP within the Arabic World, LREC, volume 8, pages 143–153, 2008.
10. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer, 2009
11. M. Gridach. Deep learning approach for arabic named entity recognition. Inter national Journal of Computer Applications, 2016.
12. Y. Zhao, Y. Shen, and J. Yao. Recurrent neural network for text classification with hierarchical multiscale dense connections. In IJCAI, pages 5450–5456, 2019.
13. J. C.-W. Lin, Y. Shao, Y. Djenouri, and U. Yun. Asrnn: A recurrent neural network with an attention model for sequence labeling. Knowledge-Based Systems, 212:106548, 2021.
14. M. N. A. Ali, G. Tan, and A. Hussain. Bidirectional recurrent neural network approach for arabic named entity recognition. Future Internet, 10(12), 2018.
15. Z. Huang, W. Xu, and K. Yu. Bidirectional lstm-crf models for sequence tagging. In arXiv preprint arXiv:1508.01991, 2015.
16. M. Gridach. Character-aware neural networks for arabic named entity recognition for social media. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), pages 23–32, 2016.

Joint entity linking with deep reinforcement learning. In WWW, pages 438–447, 2019.
35. O. Adjali, R. Besancon, O. Ferret, H. Le Borgne, and B. Grau. Multi-modal entity linking for tweets. In ECIR, 2020
36. L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In EMNLP, pages 6397–6407, 2020.
37. Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji, and X. Wang. Modeling mention, context and entity with neural networks for entity disambiguation. In IJCAI, pages 1333 1339, 2015.
38. M. Francis-Landau, G. Durrett, and D. Klein. Capturing semantic similarity for entity linking with convolutional neural networks. In NAACL, pages 1256–1261, 2016
39. N. Kolitsas, O.-E. Ganea, and T. Hofmann. End-to-end neural entity linking. In CoNLL, pages 519–529, 2018.
40. Yamada, K. Washio, H. Shindo, and Y. Matsumoto. Global entity disam biguation with pretrained contextualized embeddings of words and entities. arXiv preprint arXiv:1909.00426, 2020.
41. X. Yang, X. Gu, S. Lin, S. Tang, Y. Zhuang, F. Wu, Z. Chen, G. Hu, and X. Ren. Learning dynamic context augmentation for global entity linking. In EMNLP IJCNLP, pages 271–281, 2019.
42. L. Chen, G. Varoquaux, and F. M. Suchanek. A lightweight neural model for biomedical entity linking. In AAAI, pages 12 657–12 665, 2021.
43. M. Xue, W. Cai, J. Su, L. Song, Y. Ge, Y. Liu, and B. Wang. Neural collective entity linking based on recurrent random walk network learning. In IJCAI, pages 5327–5333, 2019.
44. T. H. Nguyen, N. R. Fauceglia, M. R. Muro, O. Hassanzadeh, A. Gliozzo, and M. Sadoghi. Joint learning of local and global features for entity linking via neural networks. In COLING, pages 2310–2320, 2016.
45. S. Guo, M.-W. Chang, and E. Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In NAACL, pages 1020–1030, 2013.
46. P. Le and I. Titov. Boosting entity linking performance by leveraging unlabeled documents. In ACL, pages 1935–1945, 2019.
47. P. Le and I. Titov. Distant learning for entity linking with automatic noise detec tion. In ACL, pages 4081–4090, 2019.
48. Y. Cao, L. Hou, J. Li, and Z. Liu. Neural collective entity linking. In COLING, pages 675–686, 2018.
49. S. Mahmoud, O. Saif, E. Nabil, M. Abdeen, M. ElNainay, and M. Torki. Ar-sanad 280k: A novel 280k artificial sanads dataset for hadith narrator disambiguation. Information, 13:55, 2022.
17. K. Darwish. Named entity recognition using cross-lingual resources: Arabic as an example. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1558–1567, 2013.
18. W. Antoun, F. Baly, and H. Hajj. Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104, 2020.
19. N. Alsaaran and M. AlRabiah. Classical arabic named entity recognition using variant deep neural network architectures and bert. IEEE Access, 9:91 537–91 547, 2021.
20. W. Shen, Y. Yin, Y. Yang, J. Han, J. Wang, and X. Yuan. Toward tweet entity linking with heterogeneous information networks. TKDE, 2021.
21. S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In EMNLP-CoNLL, pages 708–716, 2007
22. V. Varma, P. Pingali, R. Katragadda, S. Krishna, S. Ganesh, K. Sarvabhotla, H. Garapati, H. Gopisetty, V. B. Reddy, K. Reddy, and et al. Iiit hyderabad at tac 2009. In Text Analysis Conference 2009 Workshop, 2009.
23. Z. Chen and H. Ji. Collaborative ranking: A case study on entity linking. In EMNLP, pages 771–781, 2011.
24. L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In ACL, pages 1375–1384, 2011.
25. W. Shen, J. Han, J. Wang, X. Yuan, and Z. Yang. Shine+: A general framework for domain-specific entity linking with heterogeneous information networks. TKDE, 30(2):353–366, 2017
26. J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named enti ties in text. In EMNLP, pages 782–792, 2011
27. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In ICLR, 2013
28. J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In EMNLP, pages 1532–1543, 2014.
29. O.-E. Ganea and T. Hofmann. Deep joint entity disambiguation with local neural attention. In EMNLP, pages 2619–2629, 2017.
30. S. Zwicklbauer, C. Seifert, and M. Granitzer. Robust and collective entity disam biguation through semantic embeddings. In SIGIR, pages 425–434, 2016.
31. F. Hou, R. Wang, J. He, and Y. Zhou. Improving entity linking through semantic reinforced entity embeddings. In ACL, pages 6843–6848, 2020.
32. L. Hu, J. Ding, C. Shi, C. Shao, and S. Li. Graph neural entity disambiguation. KBS, 195:105620, 2020.
33. D. Gillick, S. Kulkarni, L. Lansing, A. Presta, J. Baldridge, E. Ie, and D. Garcia Olano. Learning dense representations for entity retrieval. In CoNLL, pages 528 537, 2019
34. Z. Fang, Y. Cao, Q. Li, D. Zhang, Z. Zhang, and Y. Liu.