بررسی انواع منابع دارای ساختار و منابع بدون ساختار و پیش پردازش های ابهام زدایی مفهوم کلمات در پردازش زبان طبیعی

علی وحید رودسری [۱]، کبری خوشرفتار [۲]، مرتضی گل پور [۳]، اعظم عندلیب [۴]

۱-کارشناسی ارشد مهندسی فناوری اطلاعات، دانشگاه گیلان، گروه کامپیوتر، رشت، ایران

۲-دانشجوی کارشناسی ارشد مهندسی نرم افزار دانشگاه آزاد اسلامی واحد چالوس، گروه کامپیوتر، مازندران، ایران

۳- دانشجوی کارشناسی ارشد مهندسی نرم افزار ،دانشگاه آزاد اسلامی واحدچالوس، گروه کامپیوتر، مازندران

۴-عضو هیات علمی گروه مهندسی کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، گیلان، ایران

چکیده

دانش مهمترین بخش ابهام زدایی مفهوم کلمات است. این دانش‌ها می‌توانند در شکل های گوناگون و به صورت یک مجموعه از متون باشند که در آن مفهوم کلمه برچسب گزاری شده است. پایگاه دانش یک مجموعه از متن، برچسب ها و توضیحات در جهت تشخیص مفهوم کلمه است. مانند فرهنگ لغت قابل خواندن توسط ماشین، شبکه معنایی، اصطلاحنامه و آنتولوژی. تقریبا از تمام این منابع در ابهام زدایی مفهوم کلمات استفاده می‌شود. کلیه منابع به دو دسته منابع دارای ساختار و منابع بدون ساختار تقسیم می‌شوند. جمله ورودی، یک متن بدون ساختار از اطلاعات است. برای کسب مفهوم صحیح کلمات باید بر روی آن پیش پردازش‌هایی انجام شود تا بتوانیم بستری را فراهم نماییم که بتوان بهترین مفهوم را بدست آورد. در این مقاله، منابع دارای ساختار و منابع بدون ساختار و پیش پردازش های ابهام زدایی مفهوم کلمات در پردازش زبان طبیعی را مورد بررسی قرار می دهیم که بر اساس بررسی های انجام شده، استفاده از وردنت پیشنهاد می شود که یک منبع ضروری برای ابهام زدایی مفهوم کلمات است و یک منبع دارای ساختار می باشد.

کلمات کلیدی: اصطلاحنامه، فرهنگ لغت‌های قابل خواندن توسط ماشین، آنتولوژی، وردنت، Corpora.

۱-مقدمه

یک مسئله ی اصلی در پردازش زبان طبیعی، ابهام زدایی مفهوم کلمات است. وظیفه ی گسترده ای از روش ها برای ابهام زدایی مفهوم کلمات استفاده شده است. با این حال، اغلب به دانش های دستی مانند: فرهنگ لغت قابل خواندن توسط ماشین یا واژفهرست، شبکه های معنایی، نیازمند است (Pedersen and Bruce, ۱۹۹۸). ابهام زدایی مفهوم کلمات وظیفه ی تعیین مفاهیم کلماتی را که دارای معانی زیادی در یک متن خاص هستند را برعهده دارد (Ho et al, ۲۰۱۰). با وجود اهمیت کار ابهام زدایی مفهوم کلمات و محبوبیت آن به عنوان یک موضوع مورد مطالعه، ابزار و منابعی که ابهام زدایی مفهوم کلمات حمایت کنند، تعمیم و استاندارد نسبتا کمی دیده می شود (Miller et al, ۲۰۱۳). ابهام زدایی مفهوم کلمات به طور معمول به عنوان یک مسئله ی طبقه بندی نشان داده می شود که در آن، به هر کلمه‌ی مبهم یک برچسب مفهومی از فهرست مفاهیم از پیش تعریف شده که در طول فرآیند ابهام زدایی مفهوم کلمات وجود دارد اختصاص داده شده است. در اولین گام برای ابهام زدایی مفهوم کلمات باید عبارت را به یک عبارت رسمی براساس قواعد نحوی تعریف شده تبدیل کرد و سپس براساس یک پایگاه دانش معنی درست آن کلمه را بدست آورد. در واقع اسکلت اصلی هر سیستم ابهام زدایی دارای دو قسمت است: پایگاه دانش و الگوریتمی که بتواند معنی صحیح را از پایگاه دانش استخراج نماید. یکی از موانع اصلی برای کارایی بالا در ابهام زدایی مفهوم کلمات، تنگنا و محدودیت در کسب دانش است. در سال‌های اخیر، دو روش اصلی در ابهام زدایی مفهوم کلمات مورد مطالعه قرار گرفته است. مانند روش با ناظر و روش‌های مبتنی بر دانش. بدون داشتن یک پایگاه دانش پیدا کردن معنی کلمه برای انسان و کامپیوتر امری محال می‌باشد. ایجاد پایگاه دانش دشوار است زیرا در هر زمان تغییرات بسیاری در زبان در حال انجام است. این مشکلات، مسائل اصلی در ابهام زدایی مفهوم کلمات هستند (Kilgarriff, ۲۰۰۲). منابع به دو دسته منابع دارای ساختار و منابع بدون ساختار تقسیم بندی می‌شوند که مورد بررسی قرار می گیرند.

۲- منابع دارای ساختار

در اینگونه منابع، در صورتی که مفاهیم کلمات شامل رابطه‌ایی با یکدیگر باشند آنها نیز مشخص می شوند. این ساختارها به شرح زیر است:

- اصطلاحنامه
- فرهنگ لغت‌های قابل خواندن توسط ماشین
- آنتولوژی
- وردنت

۲-۱ اصطلاحنامه

این منبع اطلاعاتی درباره‌ی ارتباط بین کلمات فراهم می کند مانند مشابهت، متضاد و همچنین رابطه‌های دیگری که بین کلمات است را نیز نشان می‌دهد (Navigli, ۲۰۰۹).

۲-۲ فرهنگ لغت‌های قابل خواندن توسط ماشین

یکی از مشهور ترین منابعی است که برای پردازش زبان‌های طبیعی استفاده می‌شود. زمانی که این فرهنگ لغت مطرح شد به عنوان اولین فرهنگ لغت الکترونیکی معرفی گردید. این فرهنگ لغت الکترونیکی ترکیبی از فرهنگ لغت‌های CED، OALD، ODOE، LDOCE است. فرهنگ لغت‌های قابل خواندن توسط ماشین برای اولین بار در سال ۱۹۸۰ ساخته شد (Tan, ۲۰۱۳)

و از آن زمان برای تبدیل شدن یک منبع دانش برای مدل سازی زبان انسانی بکار می‌رود. اصطلاحنامه روابط وابسته به فرهنگ نویسی پایه را ارائه می‌دهد. نظیر مترادف و متضاد و احتمالا سایر روابط معنایی مانند: hyponymy، hypernymy، metronymy را فراهم می‌کند. فلباوم[۱] در سال ۱۹۹۸ و میلر[۲] در سال ۱۹۹۰ اصطلاح نامه‌ی بین المللی روگت را که در سال ۱۸۵۲ توسط روگت[۳] ایجاد شده بود، کلمات را بر اساس روابط زیر طبقه بندی می‌کنند:

- مفاهیم انتزاعی
- فضا
- محتوا
- ایده‌های فکری
- اراده و تصمیم
- شهود اجتماعی و عاطفی

آخرین نسخه‌ی آن شامل ۲۵۰۰۰۰ موجودیت است. اصطلاحنامه‌ی مکواری که در سال ۱۹۸۷ توسط برنارد[۴] مورد بررسی قرار گرفت دارای بیش از ۲۰۰۰۰۰ مترادف مبتنی بر انگلیسی استرالیایی از جمله سبک گفتاری عامیانه‌ی استرالیایی است. شکل۱ چهارچوب کلی را برای روش ابهام زدایی مفهوم کلمات مفهومی انطباقی که در این قسمت مطرح می‌شود را نشان می‌دهد. در اینجا فرآیند یادگیری توصیف می‌کند که با یک گام شامل کسب دانش از فرهنگ لغت‌های قابل خواندن توسط ماشین آغاز می‌شود. با استفاده از این کسب دانش، متن ورودی خوانده شده و یک گام ابهام زدایی آزمایشی انجام می‌گیرد. در مرحله‌ی تطابق، بیان می‌شود که ترکیبی از دانش اولیه بر اساس دانش جمع آوری شده از متن تا حدی ابهام زدایی می‌شود. دانش، با توجه به متن در دست تنظیم شده است و سپس برای نتیجه‌ی نهایی ابهام زدایی، دوباره به متن اعمال می‌شود. به عنوان

---

۱ Fellbaum
۲ Miller
۳ Roget
۴ Bernard

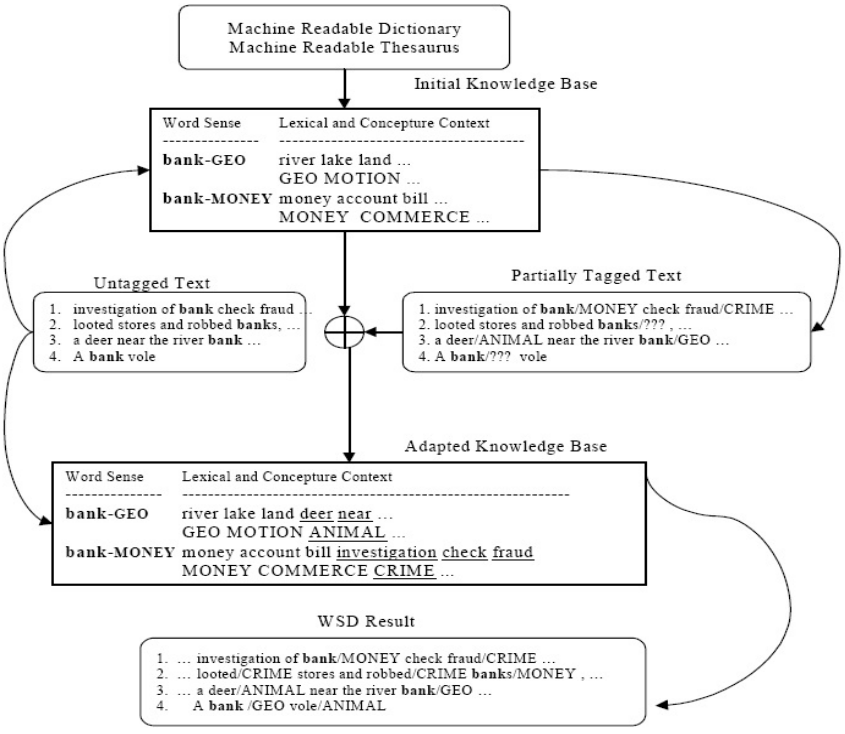
مثال: نمایش متنی اولیه از فرهنگ لغت انگلیسی معاصر لانگمن استخراج شده است. مفهوم bank_GEO شامل هر دو اطلاعات لغوی و مفهومی است:

{land, river, lake, ⋯} ∪ {GEO, MOTION, ⋯}

نمایش متنی اولیه، حاوی اطلاعات کافی برای ابهام زدایی عبارت زیر در متن ورودی است.

"A deer near the river bank"

در این گام ابهام زدایی آزمایشی، برچسب‌های مفهومی از deer/ANIMAL (آهو) و bank/GEO را تولید می‌کند. اما در بعضی از موارد bank به دلیل عدم آگاهی از دانش ابهام زدایی مفهوم کلمات، بدون برچسب هستند. ما مشاهده می‌کنیم که مفهوم bank_GEO در زمینه‌ی Vole (موش صحرایی) از آنجایی که بین ANIMAL و GEOGRAPHY ارتباطی وجود ندارد، حل نشده است. پس از آن مرحله‌ی انطباق، deer (آهو) و ANIMAL را به نمایش متنی برای bank_GEO اضافه می‌کند. تطابق نمایش متنی اولیه در حال حاضر با اطلاعات موجود بسیار غنی است و قادر به ابهام زدایی مثال bank در زمینه‌ی Vole(موش صحرایی) برای تولید نتیجه‌ی ابهام زدایی نهایی است (Chen, ۲۰۰۰).



شکل ۱- ابهام زدایی مفهوم کلمات تطبیقی با استفاده از فرهنگ لغت‌های قابل خواندن توسط ماشین (Chen, ۲۰۰۰)

CR ها برای ۹ مفهوم اسمی از bank در فرهنگ لغت انگلیسی معاصر لانگمن (LDOCE) در زیر فهرست شده است. که در جدول۱ نشان داده شده است:

bank.۱.n.۱ land along the side of a river, lake, etc.;

bank.۱.n.۲ earth which is heaped up in a field or garden, often making a border or division;

bank.۱.n.۳ a mass of snow, clouds, mud, etc.;

bank.۱.n.۴ a slope made at bends in a road or race-track, so that they are safer for cars to go round;

bank.۱.n.۵ a high underwater of bank in a river, harbour, etc.;

bank.۳.n.۱ a row, esp. of OAR in an ancient boat or KEY on a TYPEWRITER;

bank.۴.n.۱ a place in which money is kept and paid out on demand, and where related activities go on;

bank.۴.n.۲ a place where something is held ready for use, esp. ORGANIC products of human origin for medical use;

bank.۴.n.۳ a person who keeps a supply of money or pieces for payment or use in a game of chance.

جدول۱- نمایش متنی واژگان برای مفهوم bank (Chen, ۲۰۰۰)

Sense ID	Sense Label S	Lexical Context Representation $LCR(D_{bank,s})$
bank.4.n.1	MONEY	{place, money, keep, pay, demand, activity}
bank.1.n.1	RIVER	{land, lake, river}
bank.1.n.5	SANDBANK	{underwater, sand, harbour}
bank.1.n.2	EARTH	{earth, heap, field, garden, boarder, division}
bank.1.n.3	PILE	{mass, snow, cloud, mud}
bank.1.n.4	ROAD	{car, aircraft, move, side, turn}
bank.3.n.1	ROW	{row, oar, boat, key, typewriter}
bank.4.n.2	MEDICINE	{place, hold, use, organic, product, human, origin, medical}
bank.4.n.3	GAMBLE	{person, keep, supply, money, payment, game, chance}

۲-۳ آنتولوژی

آنتولوژی(هستان شناسی)، یک مدل خاص از دامنه‌های ویژه است که عمدتا شامل طبقه بندی و مجموعه ای از ارتباطات دارای

مفهوم است. داده‌هایی که در اینجا دارای رابطه هستند به صورت لینک‌هایی به هم وصل می‌شوند و هر یک از این لینک‌ها، یک نوع از رابطه را توصیف می‌کنند (Philpot et al, ۲۰۰۵). آنتولوژی، شامل بیش از ۸۰ هزار روابط بین مفهوم کلمات است. به عنوان مثال: synonym، hypernym، abbreviation (Tanaka et al, ۲۰۰۷).

آنتولوژی مدلی انتزاعی از جهان واقع است که مفاهیم و روابط میان آن را در قلمروی مورد بحث نمایش می‌دهد. آنتولوژی‌ها که پایگاه دانش مفهومی هستند و در محدوده‌ی وسیعی از قلمروها کاربرد دارند که برای نمونه می‌توان به شبکه‌های جهان گستر معنایی، موتورهای جستجو، تجارت الکترونیک، پردازش زبان طبیعی، مهندسی دانش، استخراج و بازیابی اطلاعات، سیستم‌های چندعاملی، مدل سازی کیفی از سیستم‌های فیزیکی، طراحی پایگاه داده، سیستم‌های اطلاعات جغرافیایی و کتابخانه‌های رقمی اشاره نمود. در قلمروی کامپیوتر، آنتولوژی را می توان با یک چهارتایی (C,R,F,A) تعریف کرد که در آن:

- C مجموعه مفاهیم موجود در جهان مدل شده است.
- R مجموعه روابط میان مفاهیم است و خود به دو زیرمجموعه مجزای R_t و R_N افراز می‌شود:
- R_t مجموعه روابط طبقه ای میان مفاهیم است که سلسله مراتب مشمول را ایجاد می‌کند و دودویی است.
- R_N مجموعه روابط غیرطبقه ای است که ممکن است n تایی نیز باشد. ($n \geq 1$)
- F مجموعه تصریحات آنتولوژی در مورد مفاهیم و روابط آنها است و خود به دو زیرمجموعه F_t و F_N افراز می‌شود:
- F_t مجموعه تصریحات آنتولوژی در باره روابط طبقه ای مفاهیم است. به عبارت دیگر، سلسله مراتب شمول را نشان می‌دهد.
- F_N مجموعه اصول بدیهی آنتولوژی درباره‌ی روابط غیر طبقه ای مفاهیم است.
- A مجموعه اصول بدیهی آنتولوژی است. که به زبان صوری، مثل منطق بیان می‌شود (شمس فرد و عبداله زاده بارفروش، ۱۳۸۱).

تفاوت‌های اصلی بین آنتولوژی انسان و ماشین عبارتند از:

آنتولوژی انسان بسیار وسیع تر از آنتولوژی ماشین است. اما آنتولوژی ماشین باید رسمی باشد. یعنی با یک زبان قابل فهم توسط ماشین بیان شود. همچنین باید شفاف و دقیق باشد. یعنی تمام جزئیات به صورت کامل و غیر مبهم توصیف شده باشد. اما همان طور که ذکر شد، آنتولوژی انسان به صورت ناگهانی و تلویحی شکل می‌گیرد (نوروزی و طاهریان، ۱۳۹۰).

۲-٤ وردنت

وردنت، فرهنگ لغت الکترونیکی با قابلیت دسترسی آزاد است که شامل اسم، صفت و قید است و در دانشگاه پرینستون، توسعه داده شده است (Patwardhan et al, ۲۰۰۳). وردنت یک واژه نامه محاسباتی زبان انگلیسی در پردازش زبان طبیعی

است که به صورت گسترده مورد استفاده قرار می‌گیرد. وردنت یک منبع ضروری برای ابهام زدایی مفهوم کلمات است و وردنت یک منبع دارای ساختار است. در وردنت، همه‌ی مفاهیم مترادف کلمات در یک مجموعه قرار دارد که آن مجموعه‌ی مترادف نامیده می‌شود. آخرین نسخه‌ی این فرهنگ لغت ۳,۰ است و دارای ۱۵۵۰۰۰ کلمه است که این کلمات در ۱۱۷۰۰۰ گروه معنایی قرار گرفته اند. برای هر مجموعه مترادف یک مجموعه اطلاعات وجود دارد یعنی رابطه‌هایی وجود دارد که اطلاعاتی را برای کامل کردن این داده در اختیار کاربرها قرار می‌دهد.

به عنوان مثال در اینجا وقتی کلمه automobile را جستجو می کنیم تنها گروه معنایی که مشاهده می شود را نشان داده ایم.

$$\{car_n^1, auto_n^1, automobile_n^1, machine_n^4, motorcar_n^1\}$$

شماره بالا هر کلمه نشان دهنده این است در صورتی این کلمه را جستجو کنیم این مجموعه مترادف چندمین گروه مفهومی این کلمه است و شماره پایین نشان دهنده نقش کلمه در جمله است. مانند: اسم، فعل. پس همانطور که مشاهده کردید مجموعه مترادف یک مجموعه از مفهوم‌های کلمه است که کلمات هر دسته دارای معناهای تقریبا مشابهی هستند. بر اساس رابطه‌ایی که در قسمت‌های قبلی درباره پیدا کردن مفهوم درست کلمه‌ها براساس نقشی را که دارا هستند توضیح دادیم، اکنون آن رابطه را برای یافتن مفهوم در وردنت به صورت زیر تغییر می‌دهیم.

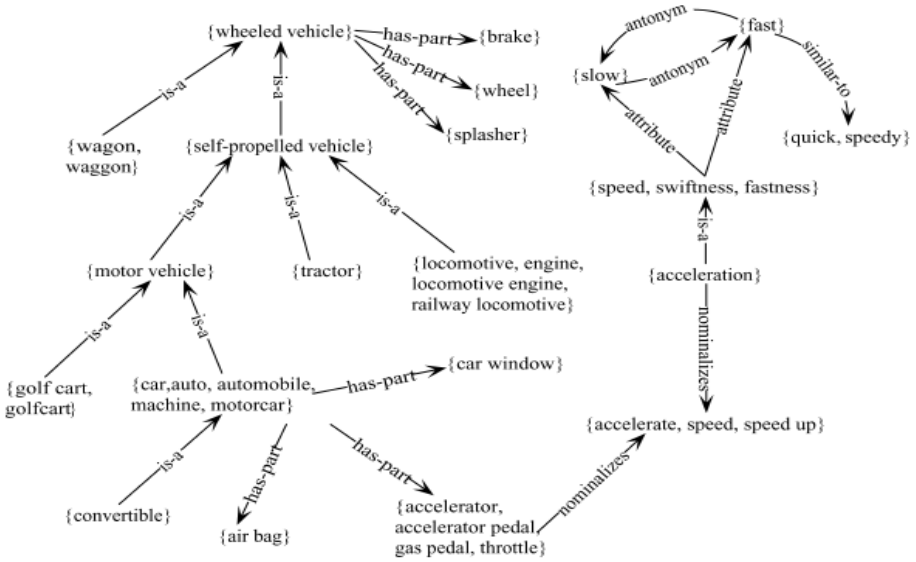
$$SensesWN : L \times POS \rightarrow ۲^{SYNSETS} \qquad (۱)$$

$$Senses_{WN}(car_n) = \begin{Bmatrix} car_n^1, auto_n^1, automobile_n^1, machine_n^4, motorcar_n^1 \\ car_n^2, rail\ car_n^1, rail\ way\ car_n^1, rail\ road\ car_n^1 \\ cable\ car_n^1, car_n^3 \\ car_n^4, gondola_n^3 \\ car_n^3, elevator\ car_n^1 \end{Bmatrix}$$

در صورتی که هر کلمه بدون ابهام باشد فقط باید یک مجموعه مترادف داشته باشد. به عنوان مثال car_n^1 فقط با یک دسته معنایی مشخص ارتباط دارد.

$$\{car_n^1, auto_n^1, automobile_n^1, machine_n^4, motorcar_n^1\}$$

در شکل ۲ بخشی از شبکه معنایی وردنت که از مجموعه مترادف ، car_n^1 بدست آمده را نشان می دهد.



شکل۲- ساختار داخلی وردنت همراه با رابطه‌ها (Navigli, ۲۰۰۹)

مثال ۲)

$$\{\ \text{pop}^2_n,\ \text{soda}^2_n,\ \text{soda pop}^1_n,\ \text{soda water}^2_n,\ \text{tonic}^2_n\ \}$$

عددهای بالای کلمات بیان کننده‌ی این است که در صورتی که کلمه‌ی مورد نظر را جستجو کنیم. آن مجموعه مترادف چندمین گروه دارای مفهوم کلمه‌ی مورد نظر است و عددهای پایین بیان کننده‌ی نقش آن کلمه است و n آخر کلمات مخفف اسم است و وردنت معنی کلمات را فراهم می‌کند. برای هر مجموعه مترادف یک مجموعه اطلاعات وجود دارد. یعنی رابطه‌هایی وجود دارد که اطلاعاتی را برای تکمیل کردن این داده در اختیار کاربرها قرار می‌دهد. تعدادی از این رابطه‌ها به شرح زیر است:

- Gloss

متن‌هایی است که برای بیان کردن بعضی از کارکردها و مثال‌هایی از این مجموعه مترادف می باشد.

- ارتباطات لغوی و معنایی

این قسمت برای برقرار کردن ارتباط بین مفاهیم و مجموعه مترادف‌ها است. تعدادی از ارتباطات معنایی و واژه‌ایی که در وردنت استفاده شده اند به شرح زیر است:

Antonymy: X با Y دارای رابطه است اگر و فقط اگر این دو نسبت به یکدیگر دارای مفهوم متضاد باشند.

Pertainymy: X یک صفت است که می‌تواند بیان کند به صورتی به Y وابسته است. Y می‌تواند اسم یا یک صفت دیگر باشد.

Nominalization: این رابطه تبدیل اسم X به فعل Y را نمایش می‌دهد.

Hypernymy: این مورد، به نام‌های is-a یا kind-of نیز شناخته می‌شود. در این رابطه می‌گوییم Y دارای رابطه Hypernymy با X است اگر و فقط اگر X یک قسمتی از Y باشد.
hyponymy و troponymy: این رابطه‌ها بر عکس رابطه Hypernymy مورد استفاده قرار می‌گیرند.

Meronymy: این رابطه به نام part of شناخته می‌شود. Y دارای رابطه Meronymy با X است اگر و فقط اگر Y یک بخشی از X باشد.

Holonymy: X دارای رابطه Holonymy با Y است اگر و فقط اگر X یک بخشی از Y باشد. این رابطه برعکس رابطه meronymy است.

Entailment: فعل Y برای اجرا شدن مستلزم اجرا شدن فعل X است. اگر می خواهید X را انجام دهید باید Y را انجام دهید.

Similarity: صفت X شبیه صفت Y است.
Attribute: اسم X یک خصوصیت است که صفت Y برای بیان کردن ارزشی برای اسم مورد استفاده قرار می‌گیرد.
See also: این یک ارتباط برای برقراری ارتباطات بین صفت‌ها مورد استفاده قرار می‌گیرد.

۲-٤-۱ توسعه‌ی وردنت

رویکردی که مطرح است متشکل از دو فاز اصلی است:

- نگاشت است. که به طور خودکار بین ویکیپدیا و وردنت انتشار می‌یابد.
- روابط اتصال صفحات ویکیپدیا است که به وردنت متصل می‌شود.

۲-٤-۲ نگاشت ویکیپدیا به وردنت

(۲)

$$\mu : Senses_{\text{Wiki}} \rightarrow Senses_{\text{WN}},$$

طوری که، برای هر Wikipage :

$$w \in Senses_{\text{Wiki}} :$$

$$\mu(w) = \begin{cases} s \in Senses_{\text{WN}}(w) & \text{اگر یک پیوند را بنوان انتشار داد} \\ \epsilon & \text{در غیر این صورت} \end{cases}$$

الگوریتم نگاشت: به منظور پیوند هر صفحه و ویکیپدیا به یک مفهوم وردنت، طبق مقاله (Ponzetto and Navigli, ۲۰۱۰) یک الگوریتم جدید توسعه داده شده است که شبه کد آن در الگوریتم ۱ نشان داده شده است و در مراحل زیر انجام می‌شود:

Input: $Senses_{\text{Wiki}}$, $Senses_{\text{WN}}$
Output: a mapping $\mu : Senses_{\text{Wiki}} \rightarrow Senses_{\text{WN}}$

for each $w \in Senses_{\text{Wiki}}$
  $\mu(w) := \epsilon$
for each $w \in Senses_{\text{Wiki}}$
  if $|Senses_{\text{Wiki}}(w)| = |Senses_{\text{WN}}(w)| = 1$ **then**
    $\mu(w) := w_n^1$
for each $w \in Senses_{\text{Wiki}}$
  if $\mu(w) = \epsilon$ **then**
    for each $d \in Senses_{\text{Wiki}}$ s.t. d redirects to w
      if $\mu(d) \neq \epsilon$ and $\mu(d)$ is in a synset of w **then**
        $\mu(w) :=$ sense of w in synset of $\mu(d)$; **break**
for each $w \in Senses_{\text{Wiki}}$
  if $\mu(w) = \epsilon$ **then**
    if no tie occurs **then**
      $\mu(w) := \underset{s \in Senses_{\text{WN}}(w)}{\text{argmax}} \; p(s|w)$
return μ

الگوریتم۱- الگوریتم نگاشت (Ponzetto and Navigli, ۲۰۱۰)

در این الگوریتم ، بهترین مفهوم S، به وسیله محاسبه‌ی بخش‌هایی از زمینه‌ی ابهام زدایی از S و w و نرمال سازی توسط امتیازات خلاصه بر روی همه‌ی مفاهیم از w در ویکیپدیا و وردنت، بدست می‌آید. نتایج نشان داده است که بهبود روش‌های مطرح شده بر پایه‌ی حاشیه نویسی بزرگ و دارای عملکرد بالاتر، توسط استفاده از اطلاعات ابهام زدایی بیشتر، می‌تواند بدست آید و با استفاده از زمینه ابهام زدایی غنی کمک می‌کند تا مناسبترین مفهوم وردنت را برای یک صفحه ویکیپدیا انتخاب کنیم.

۳- منابع بدون ساختار

منابع بدون ساختار به گروه های زیر تقسیم می‌شود:

۳-۱ Corpora

یک مجموعه از متون است که برای مدل‌های یادگیری استفاده می‌شود و شامل ۲ مدل دیگر می باشد که برای ابهام زدایی مفهوم کلمات با روش‌های باناظر و روش‌های بدون ناظر کاربرد دارند.

۳-۱-۱ Raw corpora

مجموعه براون[۵] در سال ۱۹۶۷ ارائه شد و دارای یک میلیون کلمه در قالب مجموعه متن‌های متوازن است که در سال ۱۹۶۱ در آمریکا منتشر شده است. در حال حاضر، این مجموعه بالغ بر ۳۰ میلیون کلمه را شامل می‌شود.

۳-۱-۲ Sense-Annotated Corpora

منظور از Sense Annotated یعنی کلیه مفاهیمی که از یک کلمه برداشت می‌شود را دارا است. این منبع برای بهبود در انجام ابهام زدایی مفهوم کلمات مورد استفاده قرار می‌گیرد. یکی از معروفترین منابعی که از این ساختار استفاده می‌کند

[۵] Brown

Semcor است.

۳-۱-۲-۱ SemCor

Semcor زیرمجموعه‌ایی از مجموعه براون است که شامل کلماتی است که به طور دستی دارای ضمیمه‌های نقش کلمات در جمله، ریشه کلمه و مفهوم کلمه از دیدگاه وردنت است. Semcor دارای ۳۵۲ متن است که در ۱۸۶ تا از این متن‌ها، نقش کلمات مشخص شده است ولی در ۱۶۶ تا باقیمانده فقط کلمات فعل مشخص شده اند و مفهومی را که براساس نقش آن در جمله دارند نیز مشخص گردیده است. علاوه بر این، این مجموعه دارای ۲۳۴۰۰۰ تفسیر درباره‌ی مفهوم کلمات است. بنابراین یک مجموعه بزرگ از برچسب‌های مفهومی است که می‌توان آن را برای آموزش دادن روش‌هایی که برپایه باناظر برای ابهام زدایی هستند استفاده کرد. در شکل ۳ یک قسمت کوچک از متن‌هایی که در این مجموعه وجود دارد و برچسب گذاری شده است نشان داده شده است. همان طور که مشاهده می‌کنید همه‌ی کلمات بر اساس وردنت برچسب گذاری شده اند. به عنوان مثال کلمه‌ی word در جمله اول براساس وردنت دارای مفهوم "عبارت مختصر" و در جمله دوم دارای مفهوم "یک بخشی از زبان است که مردم محلی امکان تشخیص آن را دارند." است (حسامی و محمودی، ۱۳۸۹).

As of Sunday$_n^1$ night$_n^1$ there was$_v^4$ no word$_n^2$ of a resolution$_n^1$ being offered$_v^2$ there$_r^1$ to rescind$_v^1$ the action$_n^1$. Pelham pointed out$_v^1$ that Georgia$_n^1$ voters$_n^1$ last$_r^1$ November$_n^1$ rejected$_v^2$ a constitutional$_a^1$ amendment$_n^1$ to allow$_v^2$ legislators$_n^1$ to vote$_n^1$ on pay$_n^1$ raises$_n^1$ for future$_a^1$ Legislature$_n^1$ sessions$_n^2$.

شکل۳- ساختار داخلی Semcor همراه با برچسب گذاری کلمات (Navigli, ۲۰۰۹)

Semcor براساس نسخه وردنت ۱,۵ است ولی نسخه جاری از نسخه ۲,۱ استفاده کرده است. براساس semcor یک مجموعه دو زبانی به وسیله پیانتا[۶] و همکارانش به وجود آمده است که نام آن را Multisemcor نامیده‌اند. این مجموعه دارای دو مجموعه موازی به زبان‌های ایتالیایی و انگلیسی است. این مجموعه دارای برچسب‌های نقش کلمه، ریشه آنها و همچنین مفهوم آنها در ایتالیایی و انگلیسی است که برای مفاهیم از نسخه انگلیسی و ایتالیای وردنت استفاده کرده اند. برچسب‌های مفهومی کلمات اصلی متن semcor به ایتالیایی ترجمه می‌شوند.

٤- پیش‌پردازش‌های ابهام زدایی مفهوم کلمات

پیش پردازش‌های ابهام زدایی مفهوم کلمات به شرح زیر است:

٤-۱) تفکیک کردن

تفکیک کردن به این معنی است که یک مجموعه را به توکن‌هایی که سازنده این مجموعه است تقسیم بندی کنیم.

٤-۲ برچسب زنی نقش کلمات

[۶] Pianta

در اینجا باید نقشی را که هر کلمه در این عبارت بازی می‌کند را تعیین کنیم. هر کلمه که در جمله قرار می‌گیرد دارای نقشی مانند اسم، فعل، صفت یا غیره است. همانطور که می‌دانیم کلمه در هر نقشی که قرار می‌گیرد مفهوم خاصی را پیدا می‌کند.

٤-٣ ریشه یابی

در این گام ریشه هر کلمه را بدست می‌آوریم. این کار به علت آنکه هر کلمه در هر موقعیت زمانی از لحاظ گرامر قرار می‌گیرد به آن پیشوند و پسوندهایی می‌چسبد و حتی در مواقعی نیز موجب تغییر در ریشه کلمه می‌شود، اما همه‌ی این تغییرات باز هم آن کلمه دارای مفهوم ریشه خودش است. برای اینکه کار شناسایی مفهوم کلمه آسان تر شود باید تمام کلمه‌هایی که در جمله قرار دارند را ریشه یابی نماییم. برای اینکار روش‌های مختلفی مانند ریشه یاب پورتر برای زبان انگلیسی و ریشه یاب کاظم تقوی برای زبان فارسی ارائه شده‌اند.

٤-٤ قطعه بندی

براساس قواعد زبان شناسی، زمانی که دو یا چند کلمه در کنار یکدیگر قرار می‌گیرند موجب ایجاد یک عبارت قاعده ایی خاص می‌شوند. به عنوان مثال عبارت اسمی یا عبارت فعلی. این کار یکی از کارها در پردازش زبان‌های طبیعی است که در آن برای بدست آوردن نوع عبارت مورد استفاده قرار می‌گیرد.

٤-٥ تجزیه کردن

تعیین ساختار نحوی جمله که به طور متداول به کمک رسم درخت پارسر است. که در آن درختی را براساس داده‌هایی که داریم می‌سازیم تا ببینیم که آیا ساختار این جمله درست است یا خیر.

٥- انتخاب متد کلاس بندی

آخرین مرحله‌ایی که در ابهام زدایی مفهوم کلمات مطرح است انتخاب کردن یکی از روش‌های کلاس‌بندی است. بیشترین روش‌هایی که در این مرحله استفاده می‌شود روش‌های یادگیری ماشین است. همچنین از روش‌های شناسایی الگو نیز می‌توان استفاده کرد ولی استقبال چندانی از آنها نشده است. به طور کلی دو دسته‌بندی برای روش‌های ابهام زدایی مفهوم کلمات وجود دارد روش‌های باناظر که در این روش ها ابتدا از یک مجموعه برای آموزش استفاده می‌کنیم و سپس براساس اطلاعات بدست آمده کار دسته بندی‌های بعدی را انجام می‌دهیم. روش‌هایی که از بدون ناظر استفاده می‌کنند آموزشی درکار نیست، بلکه در این روش‌ها از یک منبع دانشی استفاده اطلاعات مناسب را از داخل منبع دانش استخراج می‌کنند. همچنین پایگاه‌های دانشی که مورد استفاده قرار می‌گیرند را هم می‌توان به دو دسته تقسیم بندی کرد، پایگاه دانش قوی و ضعیف. در پایگاه‌های دانش قوی از ساختارهایی مانند آنتولوژی استفاده می‌شود که در آنها تمام رابطه‌های بین واژگان در نظر گرفته شده است. از این نوع پایگاه‌های دانش می‌توان وردنت را نام برد. اما پایگاه‌های دانش ضعیف مانند براون که مجموعه از تمامی کلمات موجود است که همه آنها را برچسب گذاری کرده‌اند و ارتباطات بین کلمه‌ها مشخص نشده است. این ساختار دارای یک

قسمت به نام حذف کلمات اضافه است. این قسمت مربوط به کلماتی در جمله است که در اصل دارای بار معنایی خاصی نیست و فقط برای اینکه خواننده راحتتر بتواند متن را درک نماید مورد استفاده قرار می‌گیرند. به عنوان مثال بعضی از این کلمات در زبان انگلیسی عبارتند از the, a, it, that و غیره. در ابهام‌زدایی مفهوم کلمات این کلمات در نظر گرفته نمی‌شوند تا بتوان سریع‌تر و راحتتر مفهوم کلمات اصلی را بدست آورد.

٦- معرفی چهار دسته از روش‌های یادگیری در ابهام زدایی مفهوم کلمات

در این بخش، چهار روش یادگیری در ابهام زدایی مفهوم کلمات مورد بررسی قرار می گیرد که شامل: روش با ناظر[۷]، روش بدون ناظر[۸]، روش نیمه ناظر[۹]، روش مبتنی بر پایگاه دانش و فرهنگ لغت[۱۰] است.

٦-١- روش با ناظر

در روش‌های باناظر از یک مجموعه داده آموزشی برای آموزش استفاده شده و سپس براساس این داده آموزشی تصمیم‌گیری های مناسب صورت می‌پذیرد. روش‌های باناظر برروی مسئله ابهام زدایی مفهوم کلمات، دارای کارایی بالایی است و نتایج بسیار خوبی را نیز بدست آورده است ولی دارای مشکلاتی است. مانند: تهیه کردن یک مجموعه داده آموزشی که بتواند کلیه یا اکثریت مفاهیم کلمه‌ها را تحت پوشش خودش داشته باشد. این روش، به طور عمده زمینه ای از واژه‌ها را برای ابهام زدایی انتخاب می کند. روش با ناظر شامل مراحل تحت نظارت و آموزش و مرحله ی آزمایش است. روش‌های باناظر گران هستند و به دلیل کمبود اطلاعات، شکننده و انعطاف ناپذیرند و برای حاشیه نویسی و شرح اطلاعات دشوار است. ابهام زدایی مفهوم کلمات مبتنی بر یادگیری روش باناظر بهترین عملکرد را در SenseEval و کارگاه‌های SemEval دارند (Zhong et al, ۲۰۰۸). الگوریتم‌های یادگیری باناظر شامل: روش کلاسه بندی ماشین بردار پشتیبانی، درخت تصمیم گیری، شبکه‌های عصبی، لیست تصمیم گیری، بیزین، k نزدیک ترین همسایه و بوستینگ است.

٦-٢- روش بدون ناظر

روش‌های مبتنی بر بدون ناظر با روش‌های باناظر و مبتنی بر پایگاه دانش بسیار متفاوت هستند. به عبارتی دیگر در روش‌های بدون ناظر هدف تعیین کردن دسته هر کدام از مفاهیم است. ولی روش‌های دیگر برای برچسب گذاری هر کدام از کلمه‌ها استفاده می‌شود. به هر حال هر دو مورد تفکیک کردن مفاهیم و برچسب گذاری از چالش‌های موجود در ابهام زدایی مفاهیم است. در سال‌های اخیر، روش بدون ناظر، در استفاده از منابع دانش گسترده با دامنه‌ی خاصی از اطلاعات، موفق آمیز بوده

---

[۷] Superviser
[۸] Unsupervised
[۹] Semi-supervised
[۱۰] Dictionary and knowledge based methods

است که می‌توان به فعالیت چن[۱۱] و همکارانش در سال ۲۰۰۹ اشاره داشت (Raviv and Markovitch, ۲۰۱۲). در سال ۲۰۰۱، بریل[۱۲] و همکارانش روش‌های بدون ناظر، نیمه ناظر و یا یادگیری فعال را با استفاده از مجموعه داده‌های بزرگ، هنگامی که برچسب‌ها گران قیمت بود پیشنهاد کردند (Yuret and Yatbaz, ۲۰۱۰). روش بدون ناظر بسیار قوی و قابل حمل[۱۳] و در واقع به طور دستی است و به منابعی مانند مفاهیم سلسله مراتبی[۱۴] و واژه نامه‌ها و منابع دانش دستی نیاز ندارد (Tomar et al, ۲۰۱۳). الگوریتم‌های ابهام زدایی مفهوم کلمات بدون ناظر به دو دسته‌ی کلی تقسیم می‌شوند:

- ابهام زدایی مفهوم کلمات مبتنی بر Token[۱۵]: کسانی که عمل ابهام زدایی مبتنی بر Token را به وسیله کاوش شباهت‌ها و یا ارتباط بین یک کلمه‌ی مبهم و متن آن را می‌یابند.
- ابهام زدایی مفهوم کلمات مبتنی بر نوع[۱۶]: آنهایی که عمل ابهام زدایی مفهوم کلمات را بر اساس نوع، به سادگی با اختصاص تمام موارد و مثال‌های رایج برای یک کلمه‌ی مبهم انجام می‌دهند (Brody et al, ۲۰۰۶).

٦-۳- روش نیمه ناظر

یادگیری نیمه ناظر بین یادگیری بدون ناظر و یادگیری باناظر قرار گرفته و در اینجا بسیار مناسب است. زیرا طبقه بندی بهتری را با استفاده از تعداد زیادی از داده‌های بدون برچسب و تعداد محدودی داده‌ی برچسب دار ایجاد می‌کند. در واقع یادگیری نیمه ناظر، از هر دوی داده‌های برچسب دار و بدون برچسب برای آموزش استفاده می‌کند. در حقیقت تعداد داده‌های بدون برچسب در این نوع یادگیری نسبت به داده‌های برچسب دار بسیار بیشتر است. تحقیقات در زمینه‌ی یادگیری ماشین وجود دارد که نشان داده اند زمانی که داده‌های بدون برچسب به همراه تعداد اندک داده برچسب دار به کار می‌روند، می‌توانند بهبود چشمگیری در دقت یادگیری ایجاد نمایند (عارفیان و افتخاری، ۱۳۹۲). بسیاری از الگوریتم‌های نیمه ناظر مختلف برای وظایف پردازش زبان طبیعی استفاده شده‌اند (Søgaard, ۲۰۱۱). نام یادگیری نیمه ناظر، حاکی از این واقعیت است که، داده بین یادگیری باناظر (با داده‌های آموزشی برچسب دار) و یادگیری بدون ناظر (بدون هیچ گونه داده‌های آموزشی با برچسب) مورد استفاده قرار می‌گیرد. بسیاری از محققان یادگیری ماشین یافتند که داده‌های بدون برچسب، زمانی که در ارتباط با مقدار کمی از داده‌های برچسب دار استفاده می‌شود، می‌تواند بهبود قابل توجهی در دقت یادگیری ایجاد کند. کسب داده‌های برچسب

[۱۱] chen
[۱۲] Brill
[۱۳] protable
[۱۴] hierarchies concept
[۱۵] Token_based WSD
[۱۶] Type_based WSD

دار برای مشکل یادگیری، اغلب به یک عامل انسانی ماهر و فرآیند گران قیمت و زمانبر نیاز دارد. در حالی که دستیابی به داده‌های بدون برچسب نسبتا ارزان است و وقت کمتری صرف می‌کند (Sati, ۲۰۱۳).

٦-٤- روش مبتنی بر پایگاه دانش و فرهنگ لغت

این روش با استفاده از دانش واژگانی[۱۷] مانند واژه نامه‌ها و اصطلاحنامه و فرض اینکه دانش را می‌توان از تعاریفی از کلمات استخراج کرد. در واقع در این روش، اگر اطلاعی درباره‌ی طبقه بندی معنایی یک کلمه وجود نداشته باشد، می‌توان از مشخصات عمومی معنای یک کلمه در فرهنگ لغت استفاده کرد. به عنوان مثال الگوریتم لسک که در این روش اول مفاهیم کلمه‌های مورد نظر را مقایسه می‌کند و برای محاسبه‌ی همپوشانی مفاهیم، کلمات مورد نظر استفاده می‌شود و با بیشترین همپوشانی[۱۸] کلمه در تعریف فرهنگ لغت خود همراه است. روش‌های مبتنی بر دانش، به ابهام زدایی مفهوم کلمات با مقایسه‌ی متن خود با اطلاعات از یک منبع واژگانی از پیش تعریف شده مانند وردنت می‌پردازد (Cabrera et al, ۲۰۰۹).

۷- نتیجه گیری

یکی از اولین مشکلاتی که در هر سیستم پردازش زبان طبیعی با آن برخورد می‌کنیم، مسئله ابهام معنایی و ساختاری کلمات است. یافتن مفهوم صحیح کلمات در یک جمله برای یک ماشین بسیار دشوار است. زیرا ماشین باید عبارت‌های انسان را که بدون ساختار هستند، به عبارت‌های با ساختار تبدیل نماید تا بتواند مفهوم صحیح کلمات را تشخیص دهد که به این تشخیص، ابهام زدایی مفهوم کلمات گفته می‌شود. روال اصلی هر سیستم ابهام زدایی دارای دو قسمت است: اول، یک پایگاه دانش و دوم، یک الگوریتمی که بتواند معنی صحیح را از پایگاه دانش استخراج نماید.

منابع به دو دسته منابع دارای ساختار و منابع بدون ساختار تقسیم بندی می‌شوند و برای کسب مفهوم صحیح کلمات باید بر روی آن پیش پردازش‌هایی انجام شود تا بهترین مفهوم بدست آید که در این مقاله بطور جامع به بررسی انواع منابع دارای ساختار و منابع بدون ساختار و پیش پردازش های ابهام زدایی مفهوم کلمات در پردازش زبان طبیعی پرداختیم و بر اساس بررسی های انجام شده، استفاده از وردنت پیشنهاد می شود که یک منبع ضروری برای ابهام زدایی مفهوم کلمات است و یک منبع دارای ساختار می باشد.

منابع

[۱۷] Lexical knowledge
[۱۸] Overlap

دکتر مهرنوش شمس فرد، دکتر احمد عبداله زاده بارفروش. " استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی". تازه‌های علوم شناختی، سال ۴، شماره‌ی۱، ۱۳۸۱.

مرتضی نوروزی، محسن طاهریان. "وب معنایی". مجری طرح: سازمان فناوری اطلاعات ایران، مدیر طرح: محمد رضا فروزنده دوست. چاپ اول ۱۳۹۰.

احسان حسامی، دکتر محمودی. " ابهام زدایی مفهوم کلمات". ارائه شده برای : سمینار کارشناسی ارشد مهندسی نرم افزار، دانشگاه آزاد اسلامی واحد قزوین، دانشکده برق، رایانه و فن آوری اطلاعات. سال تحصیلی ۹۰-۸۹.

فاطمه عارفیان، مهدی افتخاری. "روش جدید K نزدیکترین همسایه فازی و ناهموار برای طبقه بندی نیمه نظارتی". مهندسی کامپیوتر و توسعه پایدار با محوریت شبکه‌های کامپیوتری، مدل سازی و امنیت سیستم‌ها. برگزارکننده: موسسه آموزش عالی خاوران مشهد، ۲۸ آذرماه ۱۳۹۲.

E. Kilgarriff,"Introduction to the special issue on evaluating word sense disambiguation systems", Journal of Natural Language Engineering, ۸(۴): ۲۷۹-۲۹, ۲۰۰۲.

Ch.Ho, M.Murad, R.Kadir, Sh.Doraisamy. "Word Sense Disambiguation-based Sentence Similarity".Coling ۲۰۱۰: Poster Volume, pages ۴۱۸–۴۲۶, Beijing, August ۲۰۱۰.

T.Miller, N.Erbs, H.Zorn, T.Zesch, I.Gurevych. "DKPro WSD – A Generalized UIMA-based Framework forWord Sense Disambiguation". Proceedings of the ۵۱st Annual Meeting of the Association for Computational Linguistics, pages ۳۷–۴۲, Sofia, Bulgaria, August ۴-۹ ۲۰۱۳. ©۲۰۱۳ Association for Computational Linguistics.

T.Pedersen, R.Bruce. "Knowledge Lean Word-Sense Disambiguation". AAAI-۹۸ Proceedings. Copyright © ۱۹۹۸, AAAI (www.aaai.org). All rights reserved.

Z.Zhong, H.Tou, Y.Chan. "Word Sense Disambiguation Using OntoNotes: An Empirical Study". Proceedings of the ۲۰۰۸ Conference on Empirical Methods in Natural Language Processing, pages ۱۰۰۲-۱۰۱۰, Honolulu, October ۲۰۰۸. ©۲۰۰۸ Association for Computational Linguistics.

R.Navigli, "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. ۴۱, No. ۲, Article ۱۰, Publication date: February ۲۰۰۹.

J.Chen. "Adaptive Word Sense Disambiguation Using Lexical Knowledge in a Machine-readable Dictionary".Computational Linguistics and Chinese Language Processing, Vol. ۵, No. ۲, August ۲۰۰۰, pages ۲-۵. © Computational Linguistics Society of R.O.C.

L.Tan. " Examining Crosslingual Word Sense Disambiguation". A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirement for the

degree of Masters of Arts, CHAPTER ۴. KNOWLEDGE SOURCES FOR WSD, pages ۲۸-۲۹. School of Humanities and Social Sciences. ۲۰۱۳.

S.Patwardhan, S.Banerjee, and T.Pedersen. "Using Measures of Semantic Relatedness for Word Sense Disambiguation". A. Gelbukh (Ed.): CICLing ۲۰۰۳, LNCS ۲۵۸۸, pp. ۲۴۱-۲۵۷, ۲۰۰۳. ©Springer-Verlag Berlin Heidelberg ۲۰۰۳.

T.Tanaka, F.Bond,T.Baldwin, S.Fujita, Ch.Hashimoto. "Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information". Proceedings of the ۲۰۰۷ Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. ۴۷۷-۴۸۵, Prague, June ۲۰۰۷. ©۲۰۰۷ Association for Computational Linguistics.

A.Philpot, E.Hovy, P.Pantel, "The Omega Ontology", In Proceedings of the IJCNLP Workshop on Ontologies and Lexical Resources (OntoLex, Jeju Island, South Korea), ۲۰۰۵.

S.Ponzetto , R.Navigli . "Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems". Proceedings of the ۴۸th Annual Meeting of the Association for Computational Linguistics, pages ۱۵۲۲-۱۵۳۱, Uppsala, Sweden, ۱۱-۱۶ July ۲۰۱۰.

D.Yuret, M.Yatbaz. " The Noisy ChannelModel for Unsupervised Word Sense Disambiguation". ©۲۰۱۰ Association for Computational Linguistics. Volume ۳۶, Number ۱. pages ۱۱۲.

A.Raviv, Sh.Markovitch. "Concept-Based Approach toWord-Sense Disambiguation". Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Copyright ©۲۰۱۲, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

G.Tomar, M.Singh, Sh.Rai, A.Kumar, R.Sanyal, S.Sanyal. "Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation". IJCSI International Journal of Computer Science Issues, Vol. ۱۰, Issue ۵, No ۲, September ۲۰۱۳ ISSN (Print): ۱۶۹۴-۰۸۱۴, ISSN (Online): ۱۶۹۴-۰۷۸۴ www.IJCSI.org. Copyright (c) ۲۰۱۳ International Journal of Computer Science Issues. All Rights Reserved.

S.Brody, R.Navigli, M.Lapata. "Ensemble Methods for Unsupervised WSD". Proceedings of the ۲۱st International Conference on Computational Linguistics and ۴۴th Annual

Meeting of the ACL, pages ۹۷-۱۰۴, Sydney, July ۲۰۰۶. ©۲۰۰۶ Association for Computational Linguistics.

Ms. Ankita Sati. "Review: Semi-Supervised Learning Methods for Word Sense Disambiguation". IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: ۲۲۷۸-۰۶۶۱, p- ISSN: ۲۲۷۸-۸۷۲۷Volume ۱۲, Issue ۴ (Jul. - Aug. ۲۰۱۳), PP ۶۳-۶۸. www.iosrjournals.org.

A.Søgaard. "Semisupervised condensed nearest neighbor for part-of-speech tagging". Proceedings of the ۴۹th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages ۴۸-۵۲, Portland, Oregon, June ۱۹-۲۴, ۲۰۱۱. ©۲۰۱۱ Association for Computational Linguistics.

R.Cabrera, P.Rosso, M.Gómez, L.Pineda, and D.Avendaño. "Semi-supervised Word Sense Disambiguation Using the Web as Corpus". A. Gelbukh (Ed.): CICLing ۲۰۰۹, LNCS ۵۴۴۹, pp. ۲۵۶–۲۶۵, ۲۰۰۹. © Springer-Verlag Berlin Heidelberg ۲۰۰۹.