ابهام‌زدایی از معنای کلمه با الگوریتم لسک ساده و گسترش‌یافته

علی فارسی نژاد
دانشجوی کارشناسی ارشد زبان‌شناسی رایانشی
دانشگاه صنعتی شریف
farsinejad@mehr.sharif.edu

پروانه خسروی زاده
استادیار
دانشگاه صنعتی شریف
khosravizadeh@sharif.ir

چکیده:
رفع ابهام معنایی از یک واژه؛ فرآیند خودکار تشخیص معنای آن واژه در بافت است. این وظیفه از پردازش زبان طبیعی کاربردهای زیادی در ترجمه ماشینی، استخراج اطلاعات و غیره دارد. در این مقاله برای رفع ابهام معنای کلمه، از الگوریتم لسک[1] ساده و گسترش‌یافته برای محاسبه اشتراک هر یک از معانی کلمات هدف و جمله بافت استفاده شده است.

کلیدواژه‌ها: زبان‌شناسی رایانشی، رفع ابهام معنایی، الگوریتم لسک، فارس‌نت.

۱. مقدمه

ابهام‌زدایی از معنای کلمه[2] یا تشخیص اینکه واژه هدف در کدام یک از معانی ممکنش در یک بافت خاص به‌کار رفته است، از جمله پردازش‌های معنایی زبان طبیعی است. انسان‌ها به طور طبیعی و با استفاده از دانش خود در مورد دنیای اطراف، ابهام‌زدایی از معنای کلمه را تقریباً به صورت ناخودآگاه انجام می‌دهند، اما ابهام‌زدایی از معنای کلمه توسط کامپیوتر، یکی از مسائلی است که هنوز در پردازش رایانشی زبان و هوش مصنوعی حل نشده است. به‌طور کلی روش‌های رفع ابهام معنایی را به سه دسته تقسیم می‌کنند:

۱. الگوریتم‌های مبتنی بر فرهنگ لغات[3] که نیاز به یک فرهنگ لغت برای تعاریف معانی کلمه دارند؛

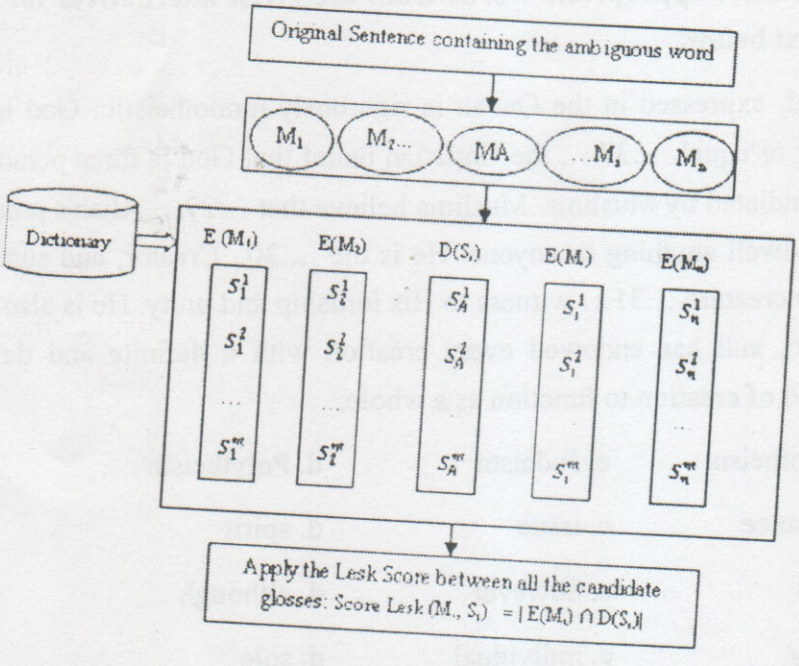
---

[1] Lesk
[2] Word Sense Disambiguation
[3] Dictionary based WSD

۲. الگوریتم‌های رفع ابهام بانظارت[۴] که نیاز به پیکره برچسب‌خورده معنایی دارند؛

۳. الگوریتم‌های بدون نظارت رفع ابهام[۵].

الگوریتم لسک که یکی از قدیمی‌ترین روش‌های رفع ابهام معنایی دسته اول است، بر این فرض استوار است که وقتی دو کلمه در یک جمله کنار هم به کار می‌روند، به موضوعی واحد اشاره می‌کنند و تعاریف فرهنگ لغت این کلمات مجاور با هم اشتراکاتی خواهند داشت. شکل ۱، دیاگرام الگوریتم اولیه لسک را نشان می‌دهد. برای معرفی الگوریتم اصلی لسک[۶]، مثالی که لسک در مقاله (۱۹۸۶) ارائه کرده است، یعنی عبارت pine cone ، در زیر مطرح می‌شود.



شکل ۱ - دیاگرام الگوریتم اولیه لسک

۴ Supervised WSD
۵ Unsupervised WSD
۶ Original Lesk

در فرهنگ لغت آکسفورد کلمه "pine" با دو معنی مختلف ذکر شده است:

sense #1: kind of evergreen tree with needle-shaped leaves

sense #2: waste away through sorrow or illness

کلمه "cone" نیز دارای سه معنی در همین فرهنگ است:

sense #1: solid body which narrows to a point

sense #2: something of this shape whether solid or hollow

sense #3: fruit of a certain evergreen tree

حال هر کدام از معانی بالا را با هم مقایسه می‌کنیم تا ببینیم کدام دو معنی تعداد کلمات مشترک بیشتری دارند.

Pine #1 ∩ cone #1 = 0
Pine #2 ∩ cone #1 = 0
Pine #1 ∩ cone #2 = 0
Pine #2 ∩ cone #2 = 0
Pine #1 ∩ cone #3 = 2
Pine #2 ∩ cone #3 = 0

هیچکدام از این تعاریف به جز تعریف معنی اول "pine" و معنی سوم "cone" با هم اشتراکی ندارند. بنابراین می‌توان دریافت که "pine" در معنی اولش یعنی درخت کاج و "cone" در معنی سومش یعنی میوه درختان سوزنی به کار رفته است.

(کیلگاریف[7]، ۲۰۰۰) معتقد است، الگوریتم ساده‌شده لسک که در این پژوهش مورد استفاده قرار گرفته است، معنی واژه‌های مجاور را درنظر نمی‌گیرد و برای هر واژه اشتراک معانی آن را با بافت محاسبه می‌کند. شبه‌کد این الگوریتم در زیر ارائه شده است.

Function SIMPLIFIED LESK (*word, sentence*) **returns** best sense of *word*
    max-overlap ← 0
    context ← set of non-stop words in *sentence*
    for each sense in senses of word do
     signature ← set of words in the gloss and examples of sense
     overlap ← number of common words between signature and context
     if overlap > max overlap then
     max-overlap ← overlap
     best-sense ← sense
end

[7] A. Kilgarriff

Return best-sense

اشکال الگوریتم لسک اولیه و ساده شده در این نکته است که تعاریف فرهنگ لغت اغلب کوتاه هستند و اطلاعات کافی را برای رفع ابهام از معنی کلمه ندارند و همچنین در بیشتر موارد کلمه مشترکی با بافت ندارند.

۱.۱. لسک بهبودیافته

برای بهبود الگوریتم لسک، گسترش‌های بسیاری پیشنهاد شده است. ایده اصلی آن است که کلمات شاخص معنی هر واژه افزایش یابد تا بتواند اطلاعات کاملی در مورد معنی آن به دست دهد. در الگوریتم لسک گسترش‌یافته از یک ساختار سلسله‌مراتبی مانند وردنت برای غنی‌کردن تعاریف استفاده می‌شود (پترسن[8]، ۲۰۰۲: ۱۳۶-۴۵). در این روش تعاریف مربوط به واژه‌های زیرشمول یا/و شامل معنایی کلمه نیز به تعریف هر معنی افزوده می‌شوند. صراف‌زاده (۲۰۱۱) در یک رویکرد دوزبانه، به رفع ابهام معنایی از الگوریتم لسک در زبان فارسی پرداخته است. فرمول زیر نحوه محاسبه این تعریف گسترش‌یافته را نشان می‌دهد.

$$ExtendedGloss(A) = gloss(A) + gloss(hypo(A)) + gloss\ (hyper(A))$$

۲.آزمایش

۱.۲. داده و پیش‌پردازش

داده‌های این پژوهش شامل دو بخش است:۱. مخزن معانی کلمه برای ارجاع و استفاده از تعاریفش ۲. پیکره آزمایشی برای بافت و سنجیدن دقت الگوریتم. در این پژوهش از فارس‌نت به عنوان انبار تعاریف معانی کلمه[9] و از پیکره همشهری به عنوان داده آزمایشی[10] استفاده شده است.

فارس‌نت (شمس‌فرد،۲۰۱۰) اولین هستان‌شناسی زبان فارسی است. نخستین نسخه از این هستان‌شناسی ۱۰،۰۰۰مجموعه هم معنا و ۱۸،۰۰۰ واژه فارسی را دربر می‌گیرد. کلمات تحت پوشش این محصول از ۳ مقوله نحوی (اسم، فعل و صفت) و از بین پرخدادترین واژه‌های زبان فارسی انتخاب شده‌اند. ساختار این هستان‌شناسی بر اساس مجموعه‌های هم‌معنی[11] است و هر واژه با یکی از معانی ممکنش در این مجموعه‌ها شراکت دارند. هر مجموعه هم‌معنی با یک تعریف و یک مثال مشخص می‌شود. روابط is-a و has-parts of شمول و زیرشمول معنایی را مشخص می‌کنند.

[8] A. Kilgarriff
[9] Sense inventory
[10] Test data
[11] synsets

در این پژوهش از فارس‌نت به دو صورت استفاده شده است، نخست، به عنوان انبار معانی ممکن برای هر واژه به همراه شماره هر یک از معانی و دوم استفاده از تعاریف مجموعه‌های هم‌معنی و مثال‌های فارس‌نت به عنوان یک فرهنگ لغت. در مسیر انجام این پژوهش، داده‌های فارس‌نت که در اصل به صورت XML بود، با نوشتن یک برنامه پایتون[12] تجزیه شد و نتایج در یک پایگاه داده رابطه‌ای[13] اکسس مایکروسافت ذخیره شد. همچنین برای تلفیق این جداول، یک سری کوئری نوشته شد تا کلمات مختلف هر مجموعه هم‌معنی و روابط بین آنها نیز ذخیره شوند. این جداول، ورودی برنامه و الگوریتم لسک در پژوهش حاضر هستند.

برای آزمایش الگوریتم نیاز به یک پیکره آزمایشی بود تا موارد وقوع واژه هدف در آن پیکره به همراه بافتی که واژه هدف در آن به کار رفته است، مشخص شود. برای آزمایش برنامه از نسخه یک پیکره همشهری (آل احمد، ۲۰۰۹) استفاده شد که در گروه تحقیقاتی پایگاه داده دانشگاه تهران و بر اساس استاندارد TREC تهیه شده است. این پیکره بیش از ۱۵۰ هزار سند از متون روزنامه همشهری از سال ۱۹۹۶ تا ۲۰۰۲ را در قالب XML در بر دارد. هر دو نوع داده‌ها نیاز به پیش‌پردازش‌هایی مانند اصلاح کارکترهای فارسی داشتند که بر روی داده‌ها اعمال شد.

در این پژوهش واژه «شیر» به عنوان واژه هدف انتخاب شد. این کلمه ۲۸۰۹ بار در پیکره همشهری تکرار شده و دوازده معنی در فارس‌نت برای آن ذکر شده است. جدول ۱ ، دوازده معنی مختلف کلمه شیر را نشان می‌دهد.

جدول ۱ - معانی مختلف کلمه «شیر»

Sense ID	Word text	Gloss	Example
۰	شیر	مایعی خوراکی که از گاو و گوسفند می‌دوشند.	شیر را گرم کن و بعد بخور.
۱	شیر	قوطی یا بطری حاوی شیر گاو یا گوسفند.	شیر توی یخچال است.
۲	شیر	مایعی که از غده های پستانی انسان ماده ترشح می‌شود و به مصرف غذای کودک می‌رسد.	مادر بهتر است شیر خودش را به بچه‌اش بدهد.
۳	شیر	جانور پستاندار گوشتخوار بزرگ از گربه سانان، با پشم کوتاه زرد یا خرمایی که جنس نر آن در اطراف سر و گردن یال سیاه یا خرمایی دارد.	شیر به طرف گاوها حمله برد.
۴	شیر	هر وسیله ای که با بستن و باز کردن دهانه یا	شیر آب توی حیاط یخ زده است.

[12] Python
[13] relational

		دریچه‌ای برای تنظیم یا قطع و وصل جریان مایع یا گاز به‌کار می‌رود.	
۵	شیر	مایعی خوراکی که از پستان پستانداران ترشح می‌شود و نوزادان آنها برای تغذیه می‌خورند.	بچه های گرگ از شیر مادرشان برای تغذیه استفاده می‌نمایند.
۶	شیر	ماده‌ای خوراکی شامل شیری که آب آن را گرفته‌اند و به صورت گرد درآمده و برای تکمیل غذای شیرخوار مواد مغذی دیگری به آن افزوده می‌شود.	او برای پیدا کردن شیر همه ی داروخانه ها را زیر پا گذاشت.
۷	شیر	قوطی حاوی شیر خشک.	آن شیر را از روی قفسه بده تا برای بچه غذا درست کنم.
۸	شیر	مایعی سفید که در ساقه بعضی از گیاهان وجود دارد.	شیر درخت لباسش را کثیف کرد.
۹	شیر	شخص شجاع ، دلاور و پهلوان.	شیران دلاور عرصه‌های نبرد حق بر باطل.
۱۰	شیر	طرفی از سکه‌ها که روی آن عدد مربوط به ارزش سکه نوشته نشده است.	در اول مسابقه گفت که برای تعیین دروازه او شیر را انتخاب می‌کند.
۱۱	شیر	برج پنجم از برج‌های دوازده‌گانه، پس از سرطان و پیش از سنبله برابر با مرداد.	؟

۲.۲ محاسبه اشتراک بافت و تعریف

برای هر واژه هدف ابتدا بافت مجاور به صورت پنجره‌ای حول آن واژه و با شعاع‌های مختلف ۵ کلمه و ۱۰ کلمه در نظر گرفته شد. اما بعد از آزمایشات مشخص شد بهترین اندازه بافت، همان جملهٔ حاوی واژه هدف است. برای محاسبه اشتراک بین بافت و هر یک از معانی کلمه، ابتدا کلمات «ایست» از بافت حذف شدند و هر جمله از بافت یک، به صورت سبد کلمات درآمد. ابتدا فقط از تعاریف و در مرحله بعد ترکیب تعریف و مثال برای محاسبه اشتراک استفاده شد. برای لسک گسترش‌یافته از بین معانی مختلف شیر فقط معانی شماره ۸،۷،۵،۳،۲ و ۱۱ شامل معنایی داشتند.

۲. نتایج

مقایسه جدول‌های زیر میزان افزایش دقت در رفع ابهام معنایی را با استفاده از روش یادشده نشان می‌دهد.

جدول۲ - درصد دقت رفع ابهام برای معانی مختلف شیر با استفاده ازلسک ساده

	تعداد یافت شده	درصد دقت
شیر وحشی	۱۱۷	%۱۶
شیر گاو و گوسفند	۱۴۲۰	%۵۴
شیر مادر	۵۰۰	%۶۸
شیر صنعتی	۱۲۰	%۳۳

جدول۲ - درصد دقت رفع ابهام برای معانی مختلف شیر با استفاده ازلسک گسترش یافته

	تعداد پیدا شده	درصد دقت
شیر وحشی	۳۶۵	%۳۷
شیر گاو و گوسفند	۱۷۲۱	%۶۶
شیر مادر	۶۲۰	%۶۵
شیر صنعتی	۱۲۰	%۳۳

۳. بحث

در لسک ساده در موارد زیادی هیچ اشتراکی بین تعریف معنی و بافت وجود نداشت. در برخی موارد واژه‌های مشترک بین دو تعریف معنی، موجب می‌شد که الگوریتم، چند معنی را برای یک بافت انتخاب کند. تعداد واژه‌های مشترک در لسک ساده حداکثر ۲ یا ۳ کلمه و در اغلب موارد یک کلمه بود. در لسک گسترش‌یافته همپوشانی بین بافت و تعریف معنی تا ۴ و ۵ کلمه هم افزایش یافت. به‌طور کلی، لسک گسترش‌یافته بهبود قابل قبولی را نسبت‌به لسک ساده نشان می‌دهد.

از آنجا که الگوریتم لسک بسیار به تعاریف معانی واژگان حساس است، یک دلیل دقت نسبتاً پایین آن، تعاریف فارس‌نت است. تعاریف اغلب واژه‌های کلیدی، شاخص هر معنی را در بر ندارند و مثال‌ها که می‌توانستند شامل کلمات همایند با هر معنی باشند، کمکی به افزایش همپوشانی با بافت نمی‌کنند. این احتمال وجود دارد که اشکال دیگر در ارتباط با ریزشدن بیش از حد بعضی از معانی باشد. به طور مثال،

دوازده معنی مختلفی که در فارس‌نت برای شیر ذکر شده است در این آزمایش به چهار معنی تقلیل داده شد. در مجموع، الگوریتم لسک و واریاسیون‌های مختلف آن می‌تواند سیستم پایه[14] مناسبی را برای رفع ابهام معنایی فراهم کند. این الگوریتم بسیار به کلمات به‌کار رفته در تعریف حساس است اما به میزان بافت انتخابی چندان حساس نیست.

منابع

M. Lesk (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. *In DeBuys V (eds.) Proceedings of SIGDOC- 86: 5th ACM international conference on systems documentation*, New York, US, pp. 24-26.

Kilgarriff, A. & Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers in the Humanities* 34(1–2), pp. 15–48.

Banerjee, S., and Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, pp. 136–45.

AlcAhmad, A., Amiri, H. , Darrudi, E. , Rahgozar, M. & Oroumchian, F. (2009). Hamshahri: *A standard Persian text collection, Journal of Knowledge-Based Systems*, 22(5), pp.382-387.

Shamsfard M., Hesabi A., Fadaei H., Mansoory N., Famian A., Bagherbeigi S., Fekri E., Monshizadeh M., & Assi M. (2010). Semi Automatic Development of FarsNet; *The Persian WordNet, 5th Global WordNet Conference* (GWA2010), Mumbai, India.

Sarrafzadeh. B., Yakovets, N., & Cercone, N. (2011). Cross Lingual Word Sense Disambiguation for Languages with Scarce Resources. In *Proceedings of the 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011. St. John's*, Canada.

[14] Baseline