اولین کنفرانس ملی محاسبات نرم و فناوری اطلاعات
1st National Conference on Soft Computing and Information Technology
دانشگاه آزاد اسلامی واحد ماهشهر
۱۲و۱۳ اسفند ماه ۸۹

یک روش بیزی برای رفع ابهام معنایی کلمات در زبان فارسی با تأکید بر ویژگیهای محلی کلمه

بابک مسعودی (۱) – سعید راحتی قوچانی (۲) – اعظم استاجی (۳)
(۱) گروه کامپیوتر - دانشگاه آزاد اسلامی واحد مشهد
Babakmasoudi282@yahoo.com
(۲) گروه برق – دانشگاه آزاد اسلامی واحد مشهد
Rahati@Mshdiau.ac.ir
(۳) گروه زبانشناسی – دانشگاه فردوسی مشهد
Estaji@um.ac.ir

خلاصه: در هر زبان کلماتی مبهم وجود دارندکه دارای چند معنی متفاوت هستند و یا بدلیل دارا بودن ساختار نوشتاری یکسان و تلفظ متفاوت (هم‌نویسه) مبهم به شمار می‌آیند. انتساب اتوماتیک معنی صحیح یک کلمه، از مسایل جاری در حوزه پردازش زبانهای طبیعی محسوب می‌شود. در این مقاله با در نظر گرفتن خصوصیات نوشتاری زبان فارسی، یک روش مبتنی بر پیکره جهت یافتن معنی صحیح کلمات پیشنهاد شده است. بر این اساس ویژگی کلمه و نشانه‌هایی که بلافاصله قبل و بعد از کلمه مبهم بکار رفته‌اند، علاوه بر ویژگی سبد لغات همراه هر معنی کلمه، به منظور رفع ابهام، مورد استفاده قرار گرفته است. نتایج دسته‌بندی با استفاده از یک روش بیزی (Naïve Bayes) تغییریافته برای تعدادی از کلمات مبهم زبان فارسی که از پیکره پژوهشکده پردازش هوشمند علایم استخراج شد، نشان می‌دهد استفاده از ویژگیهای مذکور در مقایسه با روشهایی که تنها از سبد لغات استفاده می‌کنند، باعث بهبود دقت بازشناسی می‌شود.

کلمات کلیدی: بیزین، چندمعنی، رفع ابهام معنایی، هم‌نویسه.

۱. مقدمه

یکی از مشکلاتی که سیستمهای پردازش زبان طبیعی با آن مواجه هستند، مسأله رفع ابهام معنایی کلمات است. رفع ابهام معنایی[1](WSD)، عبارتست از انتساب معنای صحیح - که قابل تشخیص از دیگر معانی بالقوه است - به یک واژه خاص در یک متن یا گفتمان [۱]. در واقع، رفع ابهام از معنای یک کلمه بر می‌گردد به اینکه آن کلمه در چه جمله‌ای به کار رفته و با چه کلماتی همنشین شده است.

ابهام‌زدایی از معنی یک کلمه شاید به تنهایی کاربردی نباشد، اما بعنوان عملیات واسطه در کاربردهایی نظیر ترجمه ماشینی انجام آن ضروری است. بعنوان مثال مترجم زبان فارسی به انگلیسی برای ترجمه کلمه "شیر" با سه معنی "Lion"، "Valve"و"Milk" مواجه است که باید با توجه به بافت متن مناسبترین واژه را انتخاب کند. برخی از کاربردهای رفع ابهام معنایی کلمات عبارتند از:

- ترجمه ماشینی
- بازیابی اطلاعات
- استخراج اطلاعات
- سیستمهای تبدیل متن به گفتار

به دلیل کاربردهای گسترده، روشهای متعددی جهت ابهام‌زدایی اتوماتیک از کلمات پیشنهاد شده است. این روشها عمدتاً در دو دسته قرار می‌گیرند: روشهایی که بر مبنای منابع[2] لغوی نظیر لغتنامه قرار دارند و روشهایی که بر پایه پیکره[3] انجام می‌شوند.

اگر اطلاعاتی در مورد طبقه‌بندی معنایی یک کلمه وجود نداشته باشد، در این صورت می‌توان از مشخصات عمومی معنای یک کلمه در فرهنگ لغت استفاده نمود. در سال ۱۹۸۶، Lesk [۲] روشی را پیشنهاد کرد که بر اساس آن همه معانی کلمه از فرهنگ لغت

1 Word Sense Disambiguation

2 Knowledge-Based Approaches
3 Corpus-Based Approaches

استخراج شده و هر کدام از معانی با تعاریف فرهنگ لغت از سایر کلمات بافت مقایسه می‌گردد، واژه‌ای که بیشترین همپوشانی را داشته باشد بعنوان معنی صحیح انتخاب می‌شود. کار Lesk اگرچه برای نمونه‌های کوتاه متن دقتی برابر ۷۰-۵۰ درصد داشت اما بعنوان یک کار پایه‌ای در روش بر مبنای منابع لغوی به حساب می‌آید.

امّا در روشهای مبتنی بر پیکره، سیستم به کمک نمونه‌های استخراج شده از پیکره زبانی بصورتهای با سرپرستی یا بدون سرپرستی آموزش دیده و از مدل ایجاد شده برای ابهام‌زدایی استفاده می‌کند.

معمول‌ترین روشهای یادگیری با سرپرستی که در زبانهای مختلف پیشنهاد شده، عبارتند از: دسته بندی بیزی، درختهای تصمیم، شبکه-های عصبی سیستمهای یادگیری منطقی و نزدیکترین همسایگی.

هرچند در زبان فارسی تعداد کلمات مبهم زیاد است و این موجب کاهش کارایی سیستمهایی نظیر ترجمه اتوماتیک متون فارسی شده است، متأسفانه در این زمینه فعالیتهای محدودی انجام گرفته که شاید یکی از علل آن نبود پیکره استاندارد دارای برچسب معنایی مناسب باشد.

از فعالیتهایی که در زمینه رفع ابهام کلمات فارسی صورت گرفته، می-توان به [۳] اشاره کرد. در این مقاله روشی مبتنی بر پیکره و یک فرهنگ لغت برای امتیازدهی به دسته تعلق مفهومی هر معنی کلمه مبهم پیشنهاد شده است. دقت میانگین این روش برای پانزده کلمه مبهم ٪۹۱,۴۶ گزارش شده است. همچنین یک روش مبتنی بر قاعده در [۴] پیشنهاد شد که با در نظر گرفتن این قواعد دقت آن برای یکی از پیکره‌های فارسی ٪۹۱ است.

در مقاله حاضر روشی مبتنی بر پیکره جهت رفع ابهام از کلمات مبهم پیشنهاد شده است. ما ابتدا ویژگیهای محلی کلمه مبهم، که عبارت از کلمات و نشانه‌هایی است که در پیکره به همراه کلمه هدف بکار برده شده‌اند و ویژگیهای عمومی که سبد لغات جمع‌آوری شده از جملات حاوی کلمه مبهم است، را از یک پیکره فارسی استخراج نموده و در ادامه با اعمال یک روش دسته‌بندی بیزی، یک مدل بازشناسی برای انتساب معنی صحیح هر کلمه مبهم متناسب با جمله‌ای که در آن بکار رفته است، را ایجاد می‌کنیم.

در ادامه، در بخشهای دوم و سوم روش دسته‌بندی مورد استفاده در مقاله مورد بررسی قرار خواهد گرفت. در بخش چهارم مجموعه ویژگیهای بکار رفته بیان خواهد شد و در بخش پنجم نتایج آزمایشات و بخش ششم نتیجه‌گیری خواهد بود.

۲. روش دسته‌بندی Naïve Bayes

الگوریتم یادگیری بیزی ابتدا در [۵] جهت ابهام‌زدایی از کلمات چند معنی پیشنهاد شد. ایده، توجه به کلمات همراه کلمه مبهم است که نشانگر بافت جمله هستند و هر کلمه حاوی مقداری اطلاعات مفید برای شناسایی بافت است.

دسته‌بندی کننده بیزی از قانون تصمیم‌گیری بیز بصورت زیر که در [۶] بیان شده استفاده می‌کند:

$$Decide\ s' \text{ if } p(s'|c) > p(s_k|c) \text{ for } s_k \neq s' \quad (۱)$$

اگر $s_1,\ldots,s_k,..s_K$ معانی مختلف کلمه مبهم w و $c_1,\ldots c_i,\ldots d_I$ بافتهای مختلف باشند. مقدار $P(s_k|c)$ بصورت زیر محاسبه می‌شود:

$$p(s_k|c) = \frac{p(c|s_k)}{p(c)}\,p(s_k) \quad (۲)$$

مقدار $P(c)$ ثابت است و اثری روی $P(s_k|c)$ ندارد، بنابراین معنی s' کلمه w خواهد بود:

$$s' = \arg\max s_k [\log p(s_k) + \log p(s_k)] \quad (۳)$$

همچنین Gale قانون تصمیم بیز را بصورت زیر تغییر داد:

$$s' = \arg\max s_k [\log p(s_k) + \sum_{v_j\ in\ c} \log p(v_j|s_k)] \quad (۴)$$

مقادیر $P(v_j|s_k)$ و $P(s_k)$ بصورت زیر محاسبه می‌شوند:

$$p(v_j, s_k) = \frac{c(v_j, s_k)}{c(s_k)} \quad (۵)$$

$$p(s_k) = \frac{c(s_k)}{c(w)} \quad (۶)$$

بطوریکه $v_1,\ldots,v_j,\ldots,v_J$ کلمات همسایه، $C(v_j,s_k)$ تعداد رخدادهای واژه v_j در یک بافت با معنی خاص در مجموعه آموزش، $C(s_k)$ تعداد رخدادهای معنی s_k در مجموعه آموزشی و $C(w)$ تعداد رخدادهای کلمه چند معنی w است.

با توجه به موارد فوق الگوریتم بیز جهت تخصیص معنی مناسب یک کلمه چند معنی بصورت زیر است:

Comment: *Training*
For all senses s_k of w ***do***
For all words v_j in the vocabulary ***do***
$P(v_j|s_k)=C(v_j,s_k)/C(s_k)$
End
End
Comment: *disambiguation*
For all senses s_k of w do
Score (s_k)=log $P(s_k)$
For all words v_j in context window c do
Score (s_k)=score(s_k)+log $P(v_j|s_k)$
End
End
Choose s'=arg max score s_k (s_k)

۳. روش پیشنهادی

روش دسته‌بندی بکار رفته در این مقاله مطابق [۷] است. ما مجموعه ویژگی را F درنظر می‌گیریم، بطوریکه

$F=F_t \cup F_l - \{f_1,f_2,f_3,\ldots,f_m\}$

که F_t مجموعه ویژگی سبد لغات و F_l مجموعه ویژگیهای محلی کلمه است. برای هر کلمه مبهم w با n معنی، مجموعه معانی را بصورت زیر فرض می‌کنیم:

$S_w=\{w_{s1},w_{s2},\ldots,w_{sn}\}$

با بکار گیری الگوریتم بیزی داریم:

$$score_1(w_{si}) = \log p(w_{si}) + \sum_{f_j \in F_t} \log p(f_i|w_{si}) \quad (۷)$$

بطوریکه برای هر معنی کلمه مبهم داریم:

$$p(w_{si}) = \frac{freq(w_{si})}{freq(w)} \quad (۸)$$

$$p(f_i \mid w_{si}) = \frac{freq(f_j, w_{si})}{\sum_{f_t \in F_t} freq(f_t, w_{si})} \quad (۹)$$

برای در نظر گرفتن ویژگی کلمات و نشانه‌های اطراف کلمه مبهم ما از رابطه زیر استفاده می‌کنیم:

$$score(w_{si}) = score_1(w_{si}) + score_2(w_{si}) \quad (۱۰)$$

بطوریکه $score_1$ امتیازی است که به ویژگی اول استخراج شده از پیکره تعلق می‌گیرد و $score_2$ امتیاز تخصیص یافته به ویژگی دوم استخراج شده است.

برای محاسبه امتیاز ویژگی دوم رابطه ۱۱ را بصورت زیر تعریف می‌کنیم:

$$score_2(w_{si}) = \sum_{f_j \in F_l} \delta(f_j \mid w_{s\,j}) \quad (۱۱)$$

$$\begin{cases} if\ \mathrm{f_j} \in \mathrm{F_l}\ \delta(\mathrm{f_j} \mid \mathrm{w_{sj}}) = 1 \\ \text{else} \quad \delta(\mathrm{f_i} \mid \mathrm{w_{si}}) = 0 \end{cases}$$

سرانجام معنی صحیح کلمه بصورت زیر مشخص می‌شود:

$$s = Arg \max_{\mathrm{k}} \mathrm{s_k}\ \mathrm{score}(\mathrm{w_{sk}}) \quad (۱۲)$$

۴. مجموعه ویژگیها

ما برای مشخص کردن معنی صحیح یک کلمه چندمعنی استفاده از دو دسته ویژگی را پیشنهاد می‌کنیم.

از آنجا که لغات همراه کلمه هدف، نشانگر بافت جمله و عبارتی هستند که کلمه مبهم در آن استفاده شده است، ویژگی سبد لغات نقش مهمی را در یافتن معنی صحیح کلمه ایفا می‌کند. از این کلمات بدون درنظر گرفتن موقعیت و ترتیب بکار رفتن در جملات استفاده خواهیم کرد.

دسته دوم ویژگیها با در نظر گرفتن خصوصیات زبان فارسی انتخاب شده است. مبهم بودن بسیاری از کلمات در زبان فارسی از نبود اعراب‌گذاری در متون نوشتاری ناشی می‌شود. از اینرو کلمات و نشانه‌گذاریهایی که بلافاصله قبل و بعد از کلمه هدف قرار می‌گیرند، می‌تواند کمک موثری در رفع ابهام از اینگونه کلمات نماید. در مثال زیر با دقت در نشانه‌گذاری قبل و بعد از کلمه "شیر" معنی صحیح آنرا مشخص کرد:

شیر آب

شیر، آب

۵. آزمایشات

۵،۱. آماده سازی مجموعه داده

مجموعه داده مورد استفاده در این مقاله از متون پژوهشکده پردازش هوشمند علائم[4] که دارای برچسبهای POS است، جمع‌آوری گردیده است. حجم متون برچسب خورده این مجموعه داده نزدیک به ده میلیون کلمه است.

ابتدا از این مجموعه داده جملاتی را که حاوی چهار کلمه پر تکرار مبهم فارسی هستند، استخراج می‌کنیم.

جدول ۱ جزئیات کلمات مبهم مورد استفاده را نشان می‌دهد:

جدول ۱-کلمات مورد استفاده

کلمه	معانی	تعداد	تعداد جملات
شیر	Lion	۱۵۷	۶۸۰
	Milk	۴۴۸	
	Valve	۱۵۷	
سیر	Garlic	۱۰۷۲	۱۴۹۸
	Journey	۴۲۶	
مهر	Punch	۲۰۹	۷۲۹
	Love	۱۲۰	
	A Solar Month	۴۰۰	
جو	Atmosphere	۳۰۷	۴۱۷
	Barley	۱۱۰	

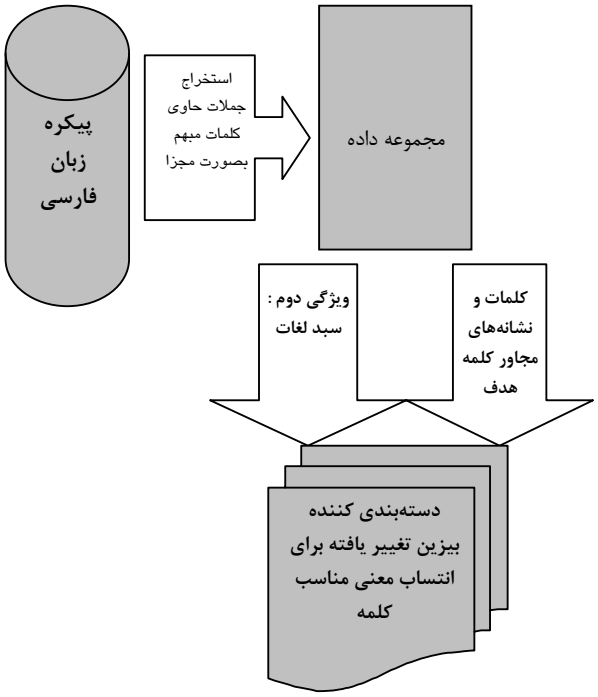
از این جملات دو دسته ویژگی شرح داده شده، بصورت زیر استخراج می‌گردند:

الف- سبد لغات: یک پنجره به مرکز لغت مورد نظر و عرض ۵± اطراف آن در نظر می‌گیریم. کلیه کلمات به استثنای حروف ربط، اضافه، اعداد و افعال را بعنوان کلماتی که بافت جمله مورد نظر را نشان می‌دهند، استخراج می‌کنیم.

ب- با در نظر گرفتن پنجره‌ای به مرکز لغت مورد نظر و عرض ۲± کلیه کلمات و نشانه‌گذاریهای قبل و بعد از آنرا بعنوان دسته ویژگی دوم استخراج می‌کنیم. بدیهی است درصورت قرار گرفتن کلمه در ابتدا یا انتهای جملات مقادیر این ویژگی فضای خالی خواهد بود که در نظر گرفته نمی‌شوند.

شکل ۱ دیاگرام روش پیشنهاد شده را نشان می‌دهد:

4 *www.rcisp.com*



شکل ۱-دیاگرام روش پیشنهاد شده

۵،۲. نتایج

برای هر آزمایش مجموعه داده بصورت تصادفی به دو مجموعه آموزش و تست تقسیم می‌شود. آزمایشات هر روش دسته‌بندی پنج بار تکرار و مقدار میانگین ثبت شده است.

در آزمایش اول روش دسته‌بندی پیشنهاد شده با استفاده از دو دسته ویژگی، آموزش دیده و میزان دقت دسته‌بندی محاسبه می‌شود و در آزمایش دوم با استفاده از روش بیزین و تنها با استفاده از ویژگی سبد لغات سیستم آموزش دیده و دقت دسته‌بندی محاسبه می‌گردد.

جدول ۲ مقایسه میزان دقت دسته‌بندی را در دو روش پیشنهاد شده و بیزین نشان می‌دهد. همچنین از آنجا که مجموعه داده مورد استفاده در [۳] مشابه مجموعه داده استفاده شده در این مقاله است، میزان دقت دسته‌بندی ذکر شده برای هر کلمه نیز جهت مقایسه در ستون چهارم جدول قرار داده شده است.

جدول ۲- مقایسه نتایج دسته‌بندی

کلمه	روش پیشنهادی	بیزی	[۳]
شیر	۹۰،۵۶	۸۹،۴۴	-
سیر	۹۵،۲۷	۹۲،۹۱	۹۲،۳۳
مهر	۸۴،۳۵	۸۱،۶۶	۸۳،۹۷
جو	۹۵،۸۹	۹۴،۵۲	۹۵،۷۲

همانگونه که جدول شماره ۳ نشان می‌دهد استفاده از ویژگی دوم در مقایسه با روش بیزین و [۳] به بهبود دسته‌بندی در جملات فارسی کمک می‌کند.

همچنین برای نشان دادن میزان اثر ویژگی کلمات و نشانه‌های همراه کلمه مبهم در استخراج سبد لغات بجای پنجره‌ای به عرض ۵± از یک پنجره به عرض ۲± استفاده کرده‌ایم. جدول ۳ میزان دقت دسته‌بندی را در این حالت نشان می‌دهد.

جدول ۳- دقت دسته‌بندی با سبد لغات کمتر

کلمه	دقت
شیر	۸۷،۷۸
سیر	۹۲،۹۰
مهر	۸۲،۹۹
جو	۸۲،۱۹

مطابق نتایج بدست آمده مشاهده می‌شود علیرغم کاهش تعداد لغاتی که بیانگر بافت جمله هستند، دقت دسته بندی در حد مطلوب است. این مسأله نشان‌دهنده مؤثر بودن کلمات و حتی نشانه‌های بکار رفته در اطراف کلمات مبهم جهت رفع ابهام معنایی، می‌باشد.

۶. نتیجه‌گیری و کارهای آینده

در این مقاله روشی برای رفع ابهام معنایی (WSD) کلمات فارسی پیشنهاد شده است. نتایج آزمایشات نشان می‌دهد استفاده از ویژگی کلمات و نشانه‌هایی که در همسایگی کلمه مبهم بکار برده می‌شوند به همراه سبد لغات می‌تواند نقش مؤثری در بهبود سیستم رفع ابهام معنایی کلمات فارسی داشته باشد. برای انجام کار بعدی استخراج ویژگیهای بیشتر نظیر POS[5] کلمات و نیز استفاده از روشهای دسته-بندی نظیر ماکزیمم آنتروپی، درختهای تصمیم و نزدیکترین همسایگی که در رفع ابهام کلمات در زبانهای دیگر نتایج مطلوبی ارایه داده‌اند، پیشنهاد می‌گردد. همچنین درصورتی که پس از استخراج ویژگیها ریشه لغات (نظیر مفرد اسامی و یا افعال) قرار گیرد، نتیجه دسته‌بندی بهبود خواهد یافت.

مراجع

[۱] J. V. Nancy Ide, "Introduction to the special issue on word sense disambiguation: the state of the art," *Computational Linguistics,* vol. 24, 1998.

[۲] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," presented at the Proceedings of the 5th annual international conference on Systems documentation, Toronto, Ontario, Canada, 1986.

[۳] M. M. H. Raheleh Makki, "Word Sense Disambiguation of Farsi Homographs Using

[5] Part Of Speech

Thesaurus and Corpus," *Lecture Notes in Computer Science,* vol. 5221, pp. 315-323, 2008.

[۴] C. Saedi and M. Shamsfard, "Translating Persian documents into English using knowledge based WSD," in *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*, 2009, pp. 1-6.

[۵] W. Gale*, et al.*, "A method for disambiguating word senses in a large corpus," *Computers and the Humanities,* vol. 26, pp. 415-439, 1992.

[۶] R. O. Duda, *Pattern classification and scene analysis [by] Richard O. Duda [and] Peter E . Hart*. New York: Wiley, 1973.

[۷] Q. L. Wanyin Li, Wenjie Li, "Integrating Collocation Features in Chinese Word Sense Disambiguation " presented at the Proceedings of the Fourth Sighan Workshop on Chinese Language Processing 2005.