سومین همایش ملی
زبان شناسی رایانشی
۲۸ و ۲۹ آبان ۱۳۹۳
دانشگاه صنعتی شریف

دانشگاه صنعتی شریف
انجمن زبان‌شناسی ایران

الگوریتم PMI (اطلاعات متقابل نقطه‌ای)

معیاری جهت تشخیص شباهت معنایی کلمات

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

سیما رضایی پور
دانشجوی کارشناسی ارشد زبان‌شناسی رایانشی
دانشگاه صنعتی شریف
rezaiepour_s@mehr.sharif.ir

سومین همایش ملی
زبان شناسی رایانشی
۲۸ و ۲۹ آبان ۱۳۹۳
دانشگاه صنعتی شریف

دانشگاه صنعتی شریف
انجمن زبان‌شناسی ایران

الگوریتم PMI (اطلاعات متقابل نقطه‌ای)

معیاری جهت تشخیص شباهت معنایی کلمات

چکیده

در فرایند استخراج و بازیابی اطلاعات، تشابه‌یابی داده‌های موجود در فضای مجازی از اهمیت خاصی برخوردار است. در پژوهش حاضر، از الگوریتم اطلاعات متقابل نقطه‌ای (PMI) جهت تشخیص شباهت معنایی جفت واژه‌ها استفاده شده است. داده‌های آماری این پژوهش از جستارهای موتور جستجوی وب بدست آمده است. جهت سنجش درصد درستی الگوریتم از نمونه سوالات چهار گزینه‌ای کلمات مترادف تافل استفاده شده و مقادیر بدست آمده از ترجمه فارسی سوالات، با نمونه مشابه انگلیسی آن مقایسه شده است. دقت عملکرد این الگوریتم در تشخیص شباهت معنایی جفت واژه‌ها در نمونه مشابه انگلیسی 74 درصد و در پژوهش حاضر 60 درصد بوده است.

کلید واژه‌ها: الگوریتم اطلاعات متقابل نقطه‌ای، اطلاعات مشترک، استخراج اطلاعات، احتمال شرطی، هم‌معنایی

مقدمه

در زبان طبیعی و نظام معنایی آن، روابط مفهومی پیچیده‌ای در سطح واژگان وجود دارد. یکی از مهمترین و پرکاربردترین روابط مفهومی در سطح واژگان زبان رابطه ترادف یا هم‌معنایی است. از دیدگاه نظری، در هیچ زبانی هم‌معنایی مطلق وجود ندارد و بر اساس اصل اقتصاد زبانی، از دو واژه هم‌معنای مطلق در زبان یکی محکوم به فنااست. این بدان معنی است که هیچ دو واژه‌ای را نمی‌توان در نظام یک زبان یافت که بتوانند در تمامی جملات آن زبان به جای یکدیگر به کار روند و تغییری در معنی آن زنجیره پدید نیاورند (صفوی، 1379، ص106). محیط کاربردی واژه‌های هم‌معنی، هم‌نشینی آنها با واژه‌های دیگر بر روی زنجیره کلام و واژه‌های مرکب زبان شواهدی در اثبات این دیدگاه نظری هستند. برای مثال، جانشینی سه واژه هم‌معنای "جهان"، "گیتی" و "دنیا" (همان، ص107) در ترکیب‌هایی نظیر "جهان‌خوار"، "گیتی‌نورد" و "دنیاپرست" یا براساس شم زبانی گویشوران فارسی زبان امکان‌پذیر نیست و یا تغییری در معنای کلی آن واژه مرکب به وجود می‌آورد. گویشوران هر زبان بر اساس تجربه و شم زبانی خود قادر به تشخیص این امر هستند و در انتخاب واژه‌های هم‌معنا در تعاملات زبانی خود خطا نمی‌کنند درحالیکه تشخیص و انتخاب واژه صحیح در پایگاه دانش و منابع اطلاعاتی معضلی در پردازش زبان طبیعی است. کاربران سامانه‌های اطلاعاتی معمولا جستجوی خود را با به کارگیری چند کلیدواژه انجام می‌دهند. اندازه‌گیری میزان ارتباط هر

سند[۱] با کلیدواژه‌های کاربر برحسب الگوریتم‌هایی صورت می‌گیرد که برای تشابه‌یابی در متن نگاشته شده‌اند. در تشابه‌یابی متن که خود به دو شاخه تشابه‌یابی واژگانی[۲] و تشابه‌یابی معنایی[۳] تفکیک می‌گردد، برحسب نوع و حجم اطلاعات مورد بررسی و نیز هدف از انجام آن، روش‌ها و الگوریتم‌های مختلفی برای سنجش میزان شباهت متن مورد استفاده قرار می‌گیرند. آهنگربهان و منتظر (1391) مروری جامع بر روی روش‌های تشابه‌سنجی در متن انجام داده‌اند. برخی از روش‌های تشابه‌یابی معنایی متون که در اثر یادشده فهرست شده‌اند و در سنجه‌های مبتنی بر پیکره مورد استفاده قرار می‌گیرند عبارتند از؛ تحلیل معنایی پنهان[۴]، تحلیل معنایی پنهان احتمالی[۵]، تخصیص دیریکله پنهان[۶]، نمایه‌سازی تصادفی[۷]، مدل ابرفضای قیاس به زبان[۸] و اطلاعات متقابل نقطه‌ای[۹] (آهنگربهان و منتظر، 1391، ص230-233). در پژوهش حاضر از الگوریتم PMI یا اطلاعات متقابل نقطه‌ای استفاده شده است. از روش اطلاعات متقابل نقطه‌ای می‌توان به عنوان ابزاری برای ساخت پایگاه داده‌های واژگانی استفاده کرد. همچنین این روش می‌تواند در سیستم‌های بازیابی اطلاعات، همچون سیستم‌های بسط جستار[۱۰] کاربرد داشته باشد. به عنوان مثالی دیگر می‌توان به کاربرد آن در استخراج خودکار کلیدواژه‌ها اشاره کرد. در این کاربرد، بسامد وقوع کلمات، معیاری برای تشخیص کلیدواژه‌ها است. با توجه به این نکته که نویسندگان معمولا برای اجتناب از یکنواختی متن از کلمات مترادف و هم‌معنی استفاده می‌کنند، این الگوریتم می‌تواند در کاربردهای فوق بسیار راهگشا باشد. خلاصه‌سازی متن و طبقه‌بندی متن از کاربردهای دیگر روش اطلاعات متقابل نقطه‌ای است. روش اطلاعات متقابل نقطه‌ای نخستین بار در سال 1990 توسط چرچ[۱۱] و هنکس[۱۲] به عنوان روشی برای تخمین درجه ارتباط میان جفت واژه‌ها به‌کار گرفته شد. در این روش مقایسه احتمال مشاهده دو واژه x و y (احتمال وقوع مشترک)، با احتمال مشاهده هر یک از دو واژه x و y (احتمال وقوع مستقل) بر پایه معادله (1) محاسبه می‌شود:

---

[1] document
[2] lexical similarity
[3] semantic similarity
[4] latent semantic analysis
[5] probabilistic latent semantic analysis
[6] latent dirichlet allocation
[7] random indexing
[8] hyperspace analogues to language
[9] pointwise mutual information
[10] query expansion
[11] Church
[12] Hanks

$$1) \quad I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

مبانی نظری و روش پژوهش

بهره‌گیری از روش PMI در سنجش میزان شباهت واژگانی و با استفاده از سئوالات بخش واژگان آزمون تافل، نخستین بار توسط ترنی[۱۳] (2001) در آزمونی مشابه پژوهش حاضر به‌عنوان معیاری بدون نظارت جهت ارزیابی شباهت معنایی جفت واژه‌ها به کار گرفته شده است. پژوهش حاضر با الگوگیری از پژوهش ترنی، کاری مشابه بر روی سندهای زبان فارسی است. هدف اصلی از انجام این پژوهش، یافتن پاسخی برای پرسش‌های زیر بوده است؛ 1) آیا مفاهیم انتزاعی در ذهن بشر زیرساختی جهانی دارند؟ برای مثال در صورتیکه واژه "intertwine" در زبان انگلیسی مترادف معنایی واژه "twist" است آیا در زبان فارسی نیز معادل‌های این دو واژه یعنی "بافتن" و "تابیدن" دو واژه مترادف محسوب می‌شوند؟ و 2) با بهره‌گیری از روش PMI در سنجش میزان شباهت واژگانی و با استفاده از ترجمه سئوالات بخش واژگان آزمون تافل به چه درصدی از دقت در سندهای زبان فارسی می‌توان دست یافت؟ بدین منظور ابتدا تعدادی از سوالات تستی مربوط به کلمات مترادف آزمون تافل به فارسی ترجمه شده و واژه مورد سوال، با چهار گزینه موجود در تست مقایسه شده است.

پژوهش‌های مرتبط

استفاده از سئوالات بخش واژگان آزمون تافل در بررسی میزان دقت سنجه‌های تشابه‌یابی از سابقه‌ای نسبتا غنی برخوردار است. نخستین بار لندوور[۱۴] و دومی[۱۵] (1997) با استفاده از سئوالات آزمون تافل، روش تحلیل معنایی پنهان (LSA) را در سنجش میزان شباهت واژگانی به کار بسته‌اند. این روش در تشخیص هم‌معنایی واژه‌ها تا 64٪ موفق عمل کرده است. لندوور و دومی در گزارش خود از این تحقیق اعلام داشته‌اند هیچ ابزار خودکار دیگری در یادگیری و بازنمایی دانش سراغ ندارند که بدون اتکا به دانشی که توسط انسان در اختیار ماشین قرار می‌گیرد و تنها بر اساس نوع تجربه‌ای که انسان در یادگیری زبانی خود بر آن متکی است، بتواند در آزمونی با مقیاس کاملا مشابه با آنچه برای بزرگسالان به کار می‌رود عملکرد خوبی را ارائه کند (Landauer&Dumais,1997: 220). ترنی (2001) در تحقیقی مشابه از روش اطلاعات متقابل نقطه‌ای برای شناسایی هم‌معنایی واژه‌ها استفاده کرده است. ترنی با استفاده از این روش به مقدار

[13] Turney
[14] Landauer
[15] Dumais

قابل ملاحظه‌ای نتیجه را بهبود بخشیده و به دقتی معادل 74٪ دست یافته است (Turney, 2001). ترا[16] و کلارک[17] (2003) نیز با استفاده از سئوالات آزمون تافل، ضمن تایید نظر ترنی (2001) در برتری روش PMI، موفق شدند با خزش در وب و جمع‌آوری داده‌ای به حجم 1000 گیگابایت نتیجه را تا 81/25٪ بهبود بخشند (Terra&Clarke, 2003:16). بولیناریا[18] و لوی[19] (2006) با استفاده از سئوالات همین آزمون، مقایسه‌ای میان استفاده از دو روش تحلیل معنایی پنهان و اطلاعات متقابل نقطه‌ای انجام داده‌اند. نتیجه این تحقیق نشان می‌دهد درصورتیکه حجم و کیفیت پیکره یکسان باشد، مقایسه بردارهای مقادیر PMI با استفاده از فاصله کسینوسی متریک نشان می‌دهد که حتی در پیکره‌هایی با حجم کم نیز روش PMI نتایج بهتری را نسبت به روش LSA در بر دارد (Bullinaria&Levy, 2006).

آزمون و ارزیابی

این تحقیق با الگوگیری از پژوهش‌های مشابه یادشده صورت گرفته است. بدین منظور ابتدا تعدادی از سوالات تستی مربوط به کلمات مترادف آزمون تافل به فارسی ترجمه شده و واژه مورد سوال، با چهار گزینه موجود در تست مقایسه شده است. الگوریتم با قراردادن این کلمات در جستارهای موتور جستجو و تحلیل پاسخ‌های بدست آمده، به هر کلمه عددی را به‌عنوان درصد شباهت اختصاص داده است. در این پژوهش، از بین چهار نوع جستار مطرح شده توسط ترنی(2001) ، جستار نوع اول یعنی AND به کار رفته است. PMI برای دو واژه w_1 و w_2 به صورت زیر محاسبه می‌شود:

1) $$PMI - IR(w_1, w_2) = log_2 \frac{P(w_1 \& w_2)}{P(w_1) * P(w_2)}$$

برای مقایسه شباهت این چهار گزینه {Choice1,Choice2, Choice3, Choice4} با واژه اصلی مورد آزمون (Problem Word)، در هر مرحله از پردازش، دو کلمه مسئله (واژه اصلی مورد آزمون) و گزینه مورد نظر را در فرمول بدست آمده از الگوریتم قرار داده و هربار به درصدی رسیده‌ایم، در نهایت گزینه‌ای که بیشترین درصد را کسب کرده است به عنوان پاسخ سوال و به عنوان شبیه‌ترین کلمه از نظر معنایی شناخته شده است.

برای مثال، سوال زیر را که یکی از سوالات ترجمه شده این آزمون است، در نظر بگیرید :

To braid hair one must **twist** three strands together

[16] Terra
[17] Clarke
[18] Bullinaria
[19] Levy

برای بافتن مو، سه رشته را باید به یکدیگر **تاب‌داد (بافت)**

پیچاندن : Curl

بافتن : Intertwine

بستن : Fasten

نگه‌داشتن : Clip

در این سوال، واژه اصلی مورد آزمون که زیر آن خط کشیده شده‌است و هریک از چهار واژه موجود در گزینه‌ها {Choice1,Choice2, Choice3, Choice4} یا { پیچاندن ، بافتن ، بستن ، نگه‌داشتن} را بصورت جفت‌های دوتایی در نظر گرفته‌ایم.

الگوریتم PMI از اطلاعات متقابل نقطه‌ای بصورت معادله (2) استفاده می‌کند:

2) $score\ (choice_i) = log_2\ (P(problem, choice_i)/\ P(problem)\ *\ P(choice_i))$

در معادله بالا $P(problem, choicei)$ احتمال رخداد همزمان واژه مورد آزمون با گزینه i ام است. اگر $hits\ (x)$، تعداد اسناد بازیابی شده توسط موتور جستجوی وب باشد که شامل x می‌باشند، $hits\ (problem\ \text{AND}\ choice_i)$ نیز تعداد اسنادی است که شامل هر دو کلمه $problem$ و $choice_i$ هستند. بنابراین رابطه (1) بصورت زیر بازنویسی می‌گردد :

3) $score1(choice_i) = hits(problem\ \text{AND}\ choice_i)/hits(choice_i)$

با استفاده از موتور جستجوی وب تعداد سندهایی که شامل هر دو کلمه، یعنی واژه مورد آزمون و گزینه سوال باشند استخراج می‌شوند. نسبت این عدد به تعداد سندهایی که تنها شامل گزینه سئوال باشند، امتیاز کلمه گزینه را تخمین می‌زند.

از آنجا که باهم‌آیی‌ها لزوما به صورت دو واژه پشت‌سرهم اتفاق نمی‌افتند، جهت رسیدن به نتیجه بهتر می‌توان باهم‌آیی دو کلمه را در یک بازه‌ی 10 کلمه‌ای در نظر گرفت. بطوریکه چنانچه این دو واژه با فاصله 10 کلمه از یکدیگر ظاهر شوند، آنرا به عنوان یک باهم‌آیی تلقی کنیم و یا به عبارتی یک قاب[۲۰] به طول 10 کلمه در نظر گرفته و زمانی آن دو کلمه را باهم‌آیی تلقی کنیم که هر دو بصورت همزمان در این قاب قرار گیرند. در این حالت جستار نوع دوم یعنی NEAR را بکار برده‌ایم.

[20] frame

اگر واژه اصلی مورد آزمون و گزینه مورد نظر از نظر آماری مستقل از یکدیگر باشند، احتمال باهم‌آیی این دو واژه از حاصلضرب احتمال وقوع هر یک به تنهایی بدست می‌آید: $P(problem) * P(choice_i)$

در حالتی‌که این دو واژه مستقل از یکدیگر نباشند، آنگاه دو واژه گرایشی به باهم‌آیی داشته و بنابراین $P(problem, choice_i)$ بزرگتر از $P(problem) * P(choice_i)$ خواهد بود.

بنابراین نسبت $P(problem \text{ AND } choice_i)$ به $P(problem) * P(choice_i)$ معیاری است برای وابستگی آماری دو کلمه $problem$ و $choice_i$.

با توجه به اینکه در معادله (2) بدنبال یافتن بیشترین مقدار $\text{score}(choice_i)$ ، یعنی حداکثر امتیاز در بین گزینه‌ها هستیم بنابراین می‌توان از log_2 صرف نظر کرد زیرا این مقدار بطور یکنواخت در حال افزایش است. همچنین، از آنجا که مقدار $P(problem)$ برای همه گزینه‌ها مقدار ثابتی است آنرا نیز می‌توان حذف کرد. بنابراین معادله (2) را می‌توان بصورت زیر ساده کرد:

4) $score\ (choice_i) = (P(problem, choice_i) / P(choice_i))$

یکی از روش‌هایی که برای محاسبه احتمال $P(problem, choice_i)$ استفاده می‌شود، بدینصورت است که احتمال وقوع همزمان این دو واژه را در سندهای مختلف بدست می‌آوریم. صورت کسر، یعنی $P(problem, choice_i)$ تعداد سندهایی است که هر دو واژه در آن بکاررفته‌اند و مخرج کسر یا همان $P(choice_i)$ تعداد سندهایی است که تنها شامل واژه $choice_i$ می‌باشد.

همانطور که قبلا نیز اشاره شد، داده‌های ورودی استفاده شده در این بررسی، سوالات چهار گزینه‌ای آزمون تافل بوده که به زبان فارسی ترجمه شده‌اند. سپس هریک از گزینه‌ها بصورت دوبه‌دو با واژه اصلی مورد پرسش در معادله (2) قرار گرفته‌اند. ابتدا تعداد سندهای موجود در وب که شامل هر دو واژه هستند محاسبه شده است و پس از آن تعداد کل سندهایی که شامل گزینه مورد نظر بوده است. به هر گزینه امتیازی تعلق گرفته و در نهایت گزینه‌ای که بیشترین امتیاز را دارد به عنوان نزدیک‌ترین کلمه به واژه مورد پرسش انتخاب شده است. بعد از انتخاب تمام گزینه‌ها توسط الگوریتم و مقایسه آنها با پاسخ‌های صحیح به تخمین تقریبی 60٪ برای کلمات فارسی رسیده‌ایم که هرچند نسبت به تخمین بدست آمده در نمونه انگلیسی از دقت کمتری برخوردار است اما نتیجه قابل قبولی برای مطالعات آتی خواهد بود.

به عنوان نمونه‌ای از عملکرد الگوریتم، جزئیات محاسبه امتیاز در مورد مثال ذکر شده فوق، در ادامه بحث ارائه شده است:

Choice		Score
P (پیچاندن | تاب‌دادن)	2,110 / 127,000	= 0.01661
P (بافتن | تاب‌دادن)	4,000 / 493,000	= 0.08113
P (بستن | تاب‌دادن)	13,700 / 6,190,000	= 0.022132
P (نگه‌داشتن| تاب‌دادن)	14,200 / 694,000	= 0.02046

از آنجا که بیشترین امتیاز برای گزینه دوم بدست آمده است، این گزینه به عنوان گزینه صحیح انتخاب شده است. همین شیوه محاسبه بر روی تمام سوالات انجام شده و در نهایت پاسخ‌های بدست آمده با پاسخ‌های صحیح موجود در کلید سوالات مقایسه شده است. نتیجه این بررسی، دقتی حدود 60٪ را نشان داده است.

جمع‌بندی

در این پژوهش دو پرسش اساسی مدنظر بود. نخست آنکه آیا می‌توان ادعا کرد که مفاهیم انتزاعی در ذهن بشر زیرساختی جهانی دارند؟ در جهت دستیابی به پاسخی برای این پرسش، ویرزبیکا[21] (1996) نخستی‌های معنایی[22] را معرفی می‌کند. این مفاهیم، صورتهایی واژگانی هستند که خود در ساخت مفاهیم پیچیده‌تر شرکت می‌کنند. به اعتقاد ویرزبیکا، نخستی‌های معنایی مفاهیمی پایه هستند که در تمام زبان‌های طبیعی وجود دارند و نمی‌توان اجزای آنها را به عناصر کوچکتری فروکاست. از این رو، آنها را بازنمودی از مفاهیم معنی‌دار جهانی می‌دانند (Wierzbicka, 1996). در پژوهش حاضر، به سنجش شباهت واژگانی جفت واژه‌های فارسی و انگلیسی بر پایه ترجمه آزمونی با اعتبار استانده پرداخته‌ایم. این پژوهش با استفاده از الگوریتم PMI تشابه‌یابی واژگانی میان مفاهیم ترجمه شده را با دقت حدود 60٪ نشان می‌دهد که خود، در زمینه نظریه‌پردازی و انجام پژوهش‌های آتی نتیجه‌ای قابل ملاحظه است.

پرسش دوم در این پژوهش این بود که بهره‌گیری از روش PMI در سنجش میزان شباهت واژگانی و استفاده از ترجمه سئوالات بخش واژگان آزمون تافل چه درصدی از دقت در سندهای زبان فارسی را نشان

[21] Wierzbicka
[22] semantics primitives

می‌دهد؟ دستیابی به دقت حدود 60٪ نتیجه‌ای قابل قبول است که برای پژوهش‌های بعدی می‌تواند راهگشا باشد.

نتیجه‌گیری

در این مقاله از الگوریتم PMI برای تشخیص هم‌معنایی کلمات در زبان فارسی استفاده شده است. برای ارزیابی میزان دقت آن در تشخیص صحیح، از ترجمه فارسی سوالات چهارگزینه‌ای تافل بهره گرفته‌ایم و در نهایت به دقت 60٪ رسیدیم که در مقایسه با سوالات انگلیسی دقت کمتری را نشان می‌دهد. ترنی (2001) بر اساس یک نمونه‌گیری آماری از تعداد زیادی از متقاضیان ورود به دانشگاه‌های امریکای شمالی، متوسط نمره اخذ شده آزمون تافل را 64/5٪ اعلام کرده است. لندوور و دومی نیز در مقاله خود اظهار می‌کنند بر اساس آنچه به آنان اعلام شده این میانگین نمره برای اخذ پذیرش بسیاری از دانشگاه‌ها کافی است (Landauer & Dumais,1997: 220). از این رو، هرچند نتیجه حاصل از این پژوهش دقت کمتری را نسبت به نمونه مشابه انگلیسی آن نشان می‌دهد اما در مقایسه با دانش زبان طبیعی فردی که در آزمون زبان خارجی شرکت کرده است، به نظر می‌رسد نتیجه حاصل از استفاده از این روش و بهره‌گیری از متون ترجمه‌شده در یادگیری ماشین، بسیار به نتیجه قابل قبول نزدیک است.

نتایج حاصل از این پژوهش می‌تواند هم در بعد نظری و هم در مباحث کاربردی مفید باشد.

منابع

1- آهنگربهان، حمید و منتظر، غلامعلی (1391). مجموعه مقالات نخستین کنفرانس ملی مدیریت منابع اطلاعاتی وب: 219-248.

2- صفوی، کورش (1379) *درآمدی بر معنی‌شناسی*، پژوهشگاه فرهنگ و هنر اسلامی، تهران، ایران.

3. Aji, S . & Kaimal, R. (2012). Document Summarization Using Positive Pointwise Mutual Information, *International Journal of Computer Science & Information Technology (IJCSIT)*, 4(2).

4. Bullinaria, J. A. & Levy, J. P. (2006). Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study, Behavior Research Methods, 39: 510-526.
Retrieved on September 03, 2014 from;
www.researchgate.net/publication/221760580_Extracting_semantic_representations_from_word_co-occurrence_statistics_stop-lists_stemming_and_SVD/file/9c96052dd18f545590.pdf

5. Church, K. W. & Hanks, P. (1989). Word Association Norms, Mutual Information and Lexicography, In *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*: 76-83.

6. Landauer, T.K. & Dumais, S.T.(1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104: 211-240.

7. Terra, E. & Clarke, C. L. A. (2003). Frequency Estimates for Statistical Word Similarity Measures, In Proceedings of the 2003 Conference of the North American Chapter of HLT-NAACL, Edmonton, AL, Canada: HLT-NAACL: 165-172.

8. Turney, P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, In *Proceedings of theTwelfth European Conference on Machine Learning*, Berlin: Springer-Verlag.

9. Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press, Oxford. ISBN 0-19-870002-4.