شباهت‌سنجی لغوی در وب‌گاه‌های تخصصی فارسی به کمک سیستم نروفازی

حمید آهنگربهان[۱] ، غلامعلی منتظر[۲]

[۱] دانشجوی دکتری مهندسی فنّاوری اطلاعات،دانشگاه تربیت مدرّس، h.ahangarbahan@modares.ac.ir
[۲] دانشیار مهندسی فنّاوری اطلاعات،دانشگاه تربیت مدرّس، montazer@modares.ac.ir

چکیده

تاکنون روش‌های مختلفی برای تشخیص شباهت دو متن ارائه شده که کارایی آنها به محتوای متن و منابع مورد استفاده برای مقایسه بین واژه‌های آنها محدود بوده و هیچگونه تطبیقی با داده مورد بررسی نداشته‌اند به طوری که نیاز به آموزش سنجه با توجه به متن مورد بررسی به خصوص در متون تخصصی فارسی احساس می‌گردد. در این مقاله روشی ارائه شده است که با توجه به کیفی و ناکامل بودن عوامل اثرگذار بر سنجش شباهت لغوی بین دو متن تخصصی و همچنین نیازمندی به آموزش سنجه، از استنتاج نروفازی که قابلیت یادگیری از داده آموزشی را دارد، استفاده می‌کند. این روش، جمله‌های موجود را به دو بخش عمومی و تخصصی تقسیم کرده و سپس با استفاده از سیستم استنتاج نروفازی و پس از یادگیری از طریق دادهٔ آموزش، میزان شباهت بین جمله‌ها را در دو بخش محاسبه و سپس با هم ترکیب کرده و در نهایت شباهت بین دو جمله ارزیابی می‌شود. روش مذکور بر روی بخش آزمون پیکره مقاله‌های حوزه یادگیری الکترونیکی (پیکرهٔ همیافت) مورد ارزیابی قرار گرفته که با دقت بیش از ۸۲٪ امکان شناسایی زوج‌جمله‌ها مشابه را داراست.

کلمات کلیدی

شباهت‌سنجی لغوی، روش نروفازی، سنجه شباهت‌سنجی یادگیر، متون تخصصی فارسی، پیکرهٔ همیافت، انفیس

۱- مقدمه

در حال حاضر اطلاعات فراوانی بر روی اینترنت در دسترس است و افراد به راحتی می‌توانند با استفاده از این اطلاعات سند جدیدی به نام خود تدوین کنند. یکی از مسائل اصلی در این زمینه، یافتن اسناد مشابه و تکراری در وب‌گاه‌های تخصصی تا محققان مدت زمان کمتری را صرف یافتن اطلاع اصیل کنند که علاقه‌مندی‌های فراوانی را به خود جلب کرده است. یافتن متون مشابه در حوزه‌های مختلف علمی همانند ترجمهٔ ماشینی، خوشه‌بندی اطلاعات، پیش‌بینی روند حرکتی اطلاعات و تحلیل داده کاربرد دارد. از آنجا که پردازش زبان طبیعی[۱] فرایندی سخت و مبهم است در نتیجه یافتن شباهت بین واژه‌ها نیز عملی دشوار است. تاکنون مطالعات فراوانی در این زمینه انجام شده است؛ در دو مقالهٔ الظهرانی و عثمان [۴,۵] مرور کاملی بر روش‌های پیشنهادی برای شباهت‌سنجی بین متون ارائه شده است.

در این مقاله هدف اصلی تشخیص میزان شباهت لغوی بین دو متن تخصصی است که بتواند به وب‌گاه‌های که در حوزه‌ای تخصصی فعالیت دارند کمک کند تا اسناد اطلاعاتی مشابه را شناسایی کنند و به کاربران نتایج بهتری را نمایش دهد. با توجه به بیان مسئله، فرایند اصلی تشخیص شباهت، مقایسهٔ دو تکه متن است. باید توجه داشت داشت در مقایسه دو تکه‌متن، از یک‌سو،

خبرگان مختلف به علت آنکه شباهت بین متون تخصصی موضوعی مفهومی و مبهم بوده ممکن است امتیازهای متفاوتی به میزان شباهت بین دو متن دهند و از سوی دیگر، منابع مورد استفاده برای شباهت‌سنجی همانند واژه‌نامه[۲] و یا هستان‌نگار[۳] دارای کیفیت‌های متفاوتی بوده که منجر به مقایسه‌ای نه چندان مطمئن بین دو متن خواهد شد. در نتیجه شباهت‌سنجی(مقایسه دو تکه‌متن) بین متون عملی غیرقطعی[۴] بوده و نیازمند یادگیری از نظرات خبرگان و محتوای متن است که تاکنون تعداد محدودی روش تطبیقی با داده و یادگیر در حوزه شناسایی متون مشابه ارائه شده است[۶,۷].

برای غلبه بر این مشکل‌ها، در این مقاله که توسعه‌یافته تحقیق نویسندگان در مقاله [۱] است روش جدیدی ارائه شده است که ویژگی آموزش‌پذیری و تطبیق با دادهٔ مورد بررسی، را دارست که به بهبود میزان تشخیص متون مشابه منتهی می‌شود. در مقاله [۱] نویسندگان ازنظریهٔ فازی که قابلیت برخورد با ابهام در داده و همچنین بازنمایی اطلاعات خبرگان به صورت قواعد فازی را دارد؛ استفاده‌کرده و روشی برای برخورد ابهام موجود در مسئله شباهت‌سنجی ارائه کردند. از مشکلات روش [۱] عدم تطبیق‌پذیری قواعد فازی با داده مورد بررسی بوده که این مشکل در مقاله حاضر به کمک استنتاج نروفازی[۵] که دارای قابلیت همزمان ابهام‌زدای و رفع عدم‌قطعیت و همچنین آموزش و تطبیق با داده بوده، مرتفع گردیده است. به منظور ارزیابی عملکرد روش پیشنهادی از پیکرهٔ مجموعه داده‌های پایان‌نامه‌های حوزهٔ یادگیری الکترونیکی فارسی (همیافت) استفاده شده است.

ادامهٔ این مقاله بدین صورت تنظیم شده است: در بخش ۲ مروری بر پژوهش‌های مرتبط در این حوزه ارائه خواهد شد. بخش ۳ به بیان روش نروفازی مورد استفاده اختصاص یافته است و در بخش ۴ روش پیشنهادی برای شباهت‌سنجی ارائه شده و در نهایت در بخش‌های ۵ و ۶ به ترتیب نتایج عددی حاصل از پیاده‌سازی روش و همچنین نتیجه‌گیری بیان شده است.

۲- پژوهش‌های مرتبط

تشخیص شباهت عبارت است از قضاوت در مورد اینکه «آیا دو عبارت متنی دارای معنای یکسانی هستند یا خیر؟»[۵]. روش‌های متفاوتی در سطوح مختلف متن همانند لغوی[۶]، معنایی[۷]، نحوی[۸] و یا ترکیبی از آنها برای تشخیص شباهت در زبان انگلیسی ارائه شده است. در زبان فارسی نیز تحقیقاتی در زمینهٔ بازیابی اطلاعات انجام شده و تحقیقات کمی برای تشخیص شباهت بین دو متن انجام گرفته است. در این بخش مروری بر تحقیقات انجام شده در زمینهٔ تشخیص شباهت لغوی در زبان‌های مختلف ارائه خواهد شد.

استفاده از سنجه‌های فضای برداری همانند سنجه کسینوسی و n-گرم از عمده‌ترین روش‌هایی است که در زمینه شباهت‌سنجی لغوی متون کاربرد دارد. در این راستا می‌توان از تحقیق بارون‌سدانو و همکاران نام برد که برای مقایسه سندهای اصل با مشکوک از روش n-گرم استفاده کردند و با nهای مختلف آزمایش‌های خود را بر پیکره METER انجام دادند و در نهایت به این نتیجه رسیدند که بهترین n برای n-گرم سطح واژه، عدد ۲ و۳ است که به ترتیب معیار F برای آنها در پیاده‌سازی ۶۸٪و ۶۶٪ بوده است[۹]. کومار و تریپثی نیز از مفهوم ۳-گرم پیوسته برای تشخیص دستبرد ادبی به کمک پیکره استفاده کردند. آنها در این روش بلندترین رشته از واژه‌های جمله را برای تشخیص متن دستبردی انتخاب و سپس بر مبنای آن متون مشابه را استخراج می‌کنند[۱۰]. از دیگر روش‌های پرکاربرد و ساده برای تشخیص شباهت استفاده از سطح ظاهر متون بدون استفاده از دیگر سطوح نمایش متن همانند نحوی و معنایی است[۱۱-۱۳].

کارهای فراوانی نیز در سطح نحوی متن انجام شده است[۱۲,۱۳] در این سطح پس از پیش‌پردازش، پارسرهای وابستگی گرامر[۹] برای تشخیص شباهت استفاده می‌گردد. خروجی این پارسرها گراف یا درخت است. پس از تشکیل گراف یا درخت از روش‌های شباهت‌سنجی مبتنی بر گراف همانند شمارش تعداد گره‌ها یا یال‌ها برای تشخیص میزان شباهت دو متن استفاده می‌شود [۱۱-۱۳].

در زبان فارسی نیز، برای نمونه، ارومچیان و همکاران از روش‌های موجود برای شباهت‌سنجی استفاده کرده و در نهایت نتیجه گرفتند که رویکرد فضای‌برداری با وزن‌دهی بهترین کارایی را در بازیابی اطلاعات دارد[۱۴]. نیری و ارومچیان روش فازی را برای بازیابی اطلاعات فارسی ارائه دادند که از کمّی‌سازی‌های فازی برای تشخیص شباهت بین واژه‌ها بهره برده است. نویسندگان روش پیشنهادی را بر روی پیکرهٔ همشهری پیاده‌سازی کرده و نشان دادند که این روش بهتر از رویکرد فضای برداری عمل می‌کند[۱۵]. زمانی‌فر و همکاران روشی برای شناسایی متون مشابه در فارسی ارائه دادند. از آنجا که هر واژه دارای چندین معناست روش پیشنهادی آنها از ویژگی هم‌رخدادی و همچنین ترتیب واژه‌ها در متن برای جلوگیری از تشخیص اشتباه متون به عنوان متونی مشابه استفاده کرده است. از آنجا که هیچ مجموعه داده استانداردی در زبان فارسی وجود ندارد. نویسندگان ۱۸۰ سند را از وب انتخاب کرده و پیکره‌ای را برای تحقیق ساخته‌اند و روش آنها را بر روی این پیکره پیاده‌سازی کرده با روش‌های فضای برداری و شاخص‌گذاری معنایی پنهان مقایسه کردند و در نهایت دقت ۶۶٪ را به دست آورده‌اند[۱۶]. در کارهای [۲,۳] نویسندگان برای تعیین میزان شباهت اسناد، روش‌های کلونی مورچگان و اتوماتای یادگیر توزیع‌شده را به کار گرفتند و از اطلاعات چگونگی استفاده کاربران از وب استفاده کرده‌اند. روش آنها، بر مبنای این ایده بوده است که اگر تعدادی از کاربران تعدادی از صفحات وب را به صورت متوالی درخواست کنند، احتمالا این صفحات به نیازهای اطلاعاتی یکسانی پاسخ داده‌اند و در این صورت با همدیگر شباهت دارند. نتایج پیاده‌سازی در هر دو کار نشان‌دهنده آن است که در مقایسه با روش هب و تنها روش گزارش شده مبتنی بر اتوماتای توزیع‌شده در تشخیص شباهت صفحات از کارایی بالاتری برخوردار است[۲,۳]. این ایده در شناسایی متون مشابه تخصصی به علت رفتار کاربران مختلف می‌تواند متفاوت باشد چندان دقیق نیست.

روش‌های ارائه‌شده در منابع تحقیق از ویژگی‌های متفاوت همانند لغوی، نحوی و معنایی استفاده کرده‌اند همانطور که اشاره شد روش‌های که ویژگی لغوی استفاده می‌کند توانایی تشخیص واژه‌های هم‌معنا منطبق با دادهٔ مورد نظر را ندارند. از طرفی در روش‌های مورد استفاده نیز، می‌توان به کندی روش‌های مبتنی بر فضای برداری، محاسبات زیاد در یادگیری ماشینی، فقدان دانش پیشین برای تدوین قواعد فازی، عدم‌تطبیق قواعد با داده و تشکیک در کیفیت منابع مورد استفاده اشاره کرد. در نتیجه ارائه روش یا سنجه‌ای که بتواند این مسائل را برطرف کند می‌تواند در حوزه زبان بسیار کارگشا باشد.

۳- گذاری اجمالی بر روش نروفازی

روش‌های نروفازی ترکیبی از سیستم استنتاج فازی و شبکه‌های عصبی مصنوعی است که روش‌های مکمل یکدیگر در طراحی سیستم‌های هوشمند تطبیقی هستند. شبکه‌های عصبی زمانی‌که: الف) هیچ گونه دانش پیشین درباره مسأله وجود نداشته باشد. ب) نمونه‌های کافی برای آموزش در اختیار باشد. ج) هیچ‌گونه روش مستقیمی برای استخراج قوانین از ساختار شبکه وجود نداشته باشد؛ قابل به‌کارگیری بوده و نتایج خوبی تولید خواهند کرد. در مقابل، سیستم فازی به جای نمونه‌های یادگیری برای تولید دانش پیشین، نیاز به مجموعه‌ای از قواعد زبانی دارد. متغیرهای ورودی و خروجی نیز باید به صورت زبانی توصیف شوند. اگر دانش کامل نباشد و یا در تناقض باشد سیستم فازی باید تنظیم شود و از آنجا که هیچ رویکرد فرموله‌ای برای این امر وجود ندارد، تنظیم با روش‌های ابتکاری انجام می‌شود که خود می‌تواند منبع خطا باشد. بنابراین در یک سیستم فازی داشتن رویکردی تطبیقی خودکار مطلوب خواهد بود که این نیاز را می‌توان با شبکه عصبی پاسخ داد[۱۷]. شبکه‌های عصبی فازی (نروفازی) را می‌توان به سه نوع مشارکتی[۱۰]، همزمان[۱۱] و یکپارچه[۱۲] تقسیم کرد[۱۸]. که در این تحقیق شبکه عصبی فازی یکپارچه به‌کار گرفته می‌شود. در مدل شبکه عصبی فازی یکپارچه، الگوریتم‌های یادگیری شبکه برای تشخیص پارامترهای سیستم‌های استنتاج فازی به کار گرفته می‌شوند. سیستم‌های نروفازی یکپارچه از استنتاج فازی تاکاگی- سوگنو[۱۳] استفاده می‌کند که در آن از ترکیب پس انتشار برای یادگیری توابع عضویت و تخمین کمترین خطا مربع برای تشخیص ضرایب قواعد در تجمیع آنها استفاده می‌کند. در این سیستم‌ها فرایند یادگیری دو بخش دارد: در بخش اول در حالی‌که پارامترهای مقدم (توابع عضویت) برای چرخهٔ آموزش ثابت در نظر گرفته می‌شود پارامترهای ورودی منتشر شده و تخمین بهینه‌ای به کمک فرایند کمترین خطا انجام می‌گیرد. در بخش دوم در حالی‌که پارامترهای خروجی ثابت باقی مانده‌اند دوباره الگوهای آموزش منتشرشده و در این چرخه، پارامترهای مقدم به کمک خروجی پس‌انتشار ویرایش می‌شوند. این فرایند تا به دست آوردن نتیجه دلخواه تکرار می‌گردد. جزییات عملکرد هر لایه در این معماری (شکل (۱)) به صورت زیر است [۱۷]:

الف- لایهٔ ورودی: هیچ محاسبه‌ای در این لایه انجام نمی‌شود. هر گره در این لایه مرتبط با یک متغیر ورودی است و تنها مقادیر ورودی را به لایهٔ بعدی انتقال می‌دهد. وزن هر ارتباط در لایهٔ اول برابر یک است.

ب- لایهٔ فازی‌گر: هر گره در این لایه مرتبط با یک برچسب زبانی (همانند کم، متوسط و زیاد) برای هر یک از متغیرهای لایه ورودی است. به زبان دیگر ارتباط خروجی در این لایه نشان‌دهندهٔ مقدار عضویت است که میزان تعلق هر ورودی به مجموعهٔ فازی را مشخص می‌کند. تصمیم دربارهٔ مقدار اولیه و نوع توابع عضویت هر متغیر ورودی با استفاده از الگوریتم خوشه‌بندی انجام می‌شود. همچنین شکل نهایی توابع عضویت در طول یادگیری شبکه تنظیم خواهد شد.

ج- لایهٔ مقدم قواعد: هر گره در این لایه نشان‌دهندهٔ مقدم قواعد است. معمولاً عملگر نُرم t در این گره استفاده می‌شود. خروجی گره در این لایه نشان‌دهنده قوت آتش (شدت فعال شدن قاعده) برای قاعدهٔ فازی مرتبط است.

د- لایهٔ هنجارسازی قوت قواعد: هر گره در این لایه نسبت قوت آتش i- مُین قاعده به مجموع قوت آتش تمام قواعد را محاسبه می‌کند.

$$\overline{w}_i = \frac{w_1}{w_1+w_2} \quad i = 1,2 \ldots \qquad (۱)$$

ه- لایهٔ تالی قواعد: عملکرد هر گره i در این لایه به صورت تابع زیر است:

$$\overline{w}_i f_i = \overline{w}_i \, (p_i x_i + q_i x_i + r_i) \qquad (۲)$$

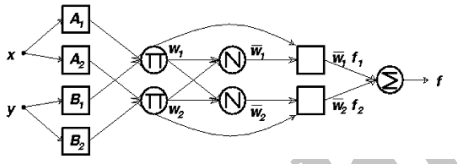
که $\overline{w}_i$ خروجی لایهٔ هنجارسازی قوت قواعد است و $\{p_i, q_i, r_i\}$ مجموعه پارامتر است. یکی از روش‌های مناسب برای تشخیص پارمترهای تالی استفاده از الگوریتم کمترین مربع خطا است.

ز- لایهٔ استنتاج قواعد: هر گره در این لایه مجموع تمام سیگنال‌های ورودی را به صورت زیر محاسبه می‌کند.

$$\text{خروجی نهایی} = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \qquad (۳)$$

۳-۱- شبکهٔ تطبیقی مبتنی بر استنتاج فازی[۱۴](انفیس)

انفیس ساختار شبیه استنتاج تاکاگی سوگنو دارد. نسخه ویرایش شده‌ای از آن در شکل (۱) نشان داده شده است که توانایی پیاده‌سازی سیستم استنتاج فازی سوکاموتو[۱۵] را دارد. در استنتاج سوکاموتو، خروجی نهایی میانگین وزنی خروجی قطعی هر قاعده است که از طریق قوت آتش قاعده (آستانه فعال شدن قاعده) استنتاج و خروجی توابع عضویت می‌شود. در انفیس، فرایند تطبیق (یادگیری) تنها در سطح پارامترها با ساختار ثابت انجام می‌شود. در این شبکه ترکیبی از پس انتشار و فرایند حداقل مربعات خطا برای آموزش سیستم استفاده می‌شود[۱۷].



شکل (۱): معماری انفیس مبتنی بر استنتاج فازی سوکاموتو

٤- چارچوب پیشنهادی سنجه یادگیر

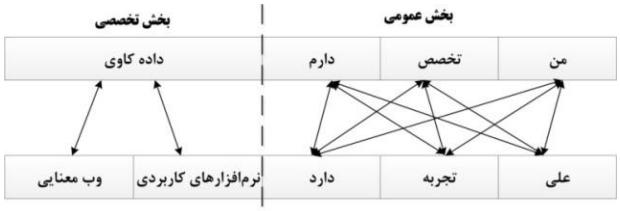
در این بخش، سنجه یادگیر پیشنهادی نروفازی(انفیس) برای ارزیابی شباهت لغوی بین دو متن تخصصی فارسی ارائه می‌شود. تشخیص شباهت لغوی در متون تخصصی دارای مشکلاتی از قبیل وجود ابهام در معنای دقیق واژه و همچنین عدم‌اطمینان در غنای منابع برای سنجش است (که آیا ارتباط بین تمام واژه‌ها را پوشش می‌دهند) که منجر خواهد شد که با داده و اطلاعات ناقص روبرو باشیم. از طرفی برای تشخیص دقیق شباهت در متون تخصصی نیازمند استفاده از سنجه‌های مختلف رویکردهای شباهت‌سنجی «لغوی» و همچنین آموزش برخط سیستم برای تطبیق قواعد با دادهٔ مورد بررسی هستیم. روش پیشنهادی در این مقاله از استنتاج نروفازی برای حل این مشکلات استفاده کرده است. شکل (۲) ساختار سنجه پیشنهادی برای حل مسئله سنجش شباهت در متون تخصصی را نشان می‌دهد. جزییات هر بخش در ادامه آورده خواهد شد.

٤-۱- پیش‌پردازش و قطعه‌بندی

هر جمله شامل یک یا چند واژه است که مفهوم خاصی را بیان می‌کند؛ این مفهوم معمولاً در محتوای تخصصی بیشتر از طریق واژه‌های تخصصی به

دست می‌آید؛ بنابراین برای سنجش دقیق‌تر در حوزه‌های تخصصی، نیاز است که بخش‌های مختلف هر متن را با منابع ویژه‌ای که برای آن بخش طراحی‌شده، سنجید. به همین منظور در روش پیشنهادی هر متن به دو بخش «تخصصی» و «عمومی» تقسیم‌بندی شده است. در بخش عمومی، واژه‌های عمومی همانند واژهٔ «کتاب» که در اکثر محتواها استفاده شده و در بخش تخصصی واژه‌های تخصصی یک حوزهٔ خاص همانند واژهٔ «یادگیری الکترونیکی » قرار می‌گیرد. البته واژه‌هایی همانند «پردازش» را می‌توان در هر دو بخش قرار داد ولی از آنجا که این واژه‌ها بیشتر در حوزهٔ تخصصی قرار می‌گیرند در این مقاله در بخش تخصصی دسته‌بندی می‌شوند. برای نمونه دو جمله «من در داده‌کاوی تخصص دارم» و «علی در نرم‌افزارهای کاربردی و وب‌معنایی تجربه دارد» در نظر بگیرید. این جمله‌ها با استفاده از هستان‌نگار تخصصی مطابق با شکل (۳) به دو بخش عمومی و تخصصی تقسیم می‌شوند.

شکل (۲): روش نروفازی برای شباهت‌سنجی متون



شکل (۳): نحوهٔ مقایسهٔ زوج‌واژه‌ها در زوج‌جمله‌ها

٤-۲- استخراج ویژگی‌ها و سنجه‌های شباهت‌سنجی

پس از بخش‌بندی هر متن، بخش‌های مرتبط با هم مقایسه می‌شود بدین ترتیب که بخش تخصصی متن اول با بخش تخصصی متن دوم و همین‌طور برای بخش عمومی هر دو متن نیز مقایسه انجام می‌گیرد. در روش پیشنهادی برای شباهت‌سنجی از رویکرد شباهت لغوی برای دو متن استفاده شده است. در این رویکرد سعی بر آن است تا ویژگی‌ها و سنجه‌هایی که در سطح ظاهر شباهت بین دو متن را می‌سنجد، به کار گرفته شود.

هم‌پوشانی واژه‌ها: این معیار تعداد واژه‌های مشترک بین دو متن را از طریق تولید n-گرم‌ها در سطح واژه یا نویسه به ترتیب بر طول متن اول و دوم تقسیم کرده و سپس از این دو نسبت میانگین هندسی گرفته می‌شود. مقایسهٔ متن اول با متن دوم به صورت زیر انجام می‌دهد[۱۹].

$$S(d_1,d_2)=\frac{|d_1\cap d_2|}{|d_1|} \quad (۴)$$

که $|d_1 \cap d_2|$ بیانگر تعداد واژه‌های مشترک برای دو متن و $|d_1|$ تعداد کل واژه‌ها برای متن اوّل است.

پرش-گرم: این سنجه همانند n-گرم است با این تفاوت که می‌تواند برای تولید گرم‌ها در سطح واژه پرشی به اندازهٔ n داشته باشد. طبق تحقیقات انجام گرفته تعداد کارای n، سه و چهار است. این سنجه به نوعی ویژگی ترتیب و همچنین وجود عبارت‌های چندواژه‌ای مشترک در هر بخش را در نظر می‌گیرد[۲۰].

نسبت واژه‌های تخصصی: این سنجه، نسبت تعداد واژه‌های تخصصی به تعداد واژه‌های عمومی در اجتماع دو متن را مشخص می‌کند. این نسبت برای تشخیص میزان اهمیت واژه‌ها در متن توسط نویسندگان تعریف شده است.

٤-۳- شباهت‌سنجی متون مبتنی بر انفیس

پس از تعریف سنجهٔ مناسب با هر رویکرد شباهت‌سنجی، از آنجا که سنجش شباهت بین دو متن تخصصی، بر اساس دانش افراد و خبرگان متفاوت و مبهم است و برای درک درست از این عمل بهتر است آن را به صورت فازی مدل کرد تا قادر به برخورد با ابهام و عدم‌قطعیت سنجش باشیم. از طرفی استفادهٔ صرف از سیستم فازی امکان یادگیری را به سیستم و انطباق با شرایط داده را نمی‌دهد در نتیجه بهتر است از شبکه‌های عصبی مصنوعی که قابلیت یادگیری دارند به همراه سیستم استنتاج فازی استفاده کنیم. پس از تعریف متغیر شباهت به صورت فازی، می‌توان با استفاده از رویکرد لغوی گفته‌شده برای شباهت‌سنجی و همین‌طور سنجه‌ها و ویژگی‌ها استخراج شده از دو بخش تخصصی و عمومی سیستم استنتاج نروفازی لغوی برای سنجش به صورت نروفازی طراحی کرد. از آنجا که ارزیابی شباهت بین دو متن فازی تعریف شده، سنجه‌های که در ورودی سیستم‌های استنتاج محاسبه می‌شوند نیز به صورت فازی مدل شده‌اند. چون بازهٔ عددی و جنس برخی متغیرهای به کار گرفته‌شده در سیستم استنتاج نروفازی یکی بوده؛ برای راحتی فهم در جدول (۱) انواع متغیرهای فازی و همچنین بازهٔ عددی و متغیر زبانی آورده شده و در جدول (۲) متغیرهای استفاده شده در هر سیستم به همراه جنس آن ذکر شده است.

جدول (۱): متغیرهای زبانی در ارزیابی شباهت بین دو متن

نوع متغیر	حرف	بازه عددی	متغیر زبانی
نسبت تعداد واژه‌های تخصصی بر تعداد واژه‌های عمومی	T1	(-∞ ,۰ ,۰/۱ ,۰/۳۵)	کم
		(۰/۱۵,۰/۳۵ ,۰/۵۵ ,۰/۸)	متوسط
		(۰/۵۰ ,۰/۷۵,۱,+∞)	زیاد
ارزیابی شباهت بین دو واژه/جمله	T2	(-∞ ,۰ ,۰/۱ ,۰/۳۵)	نامشابه
		(۰/۱,۰/۳۵ ,۰/۵۵ ,۰/۷۵)	مشابه متوسط
		(۰/۶۵ ,۰/۷۵,۱,+∞)	مشابه

جدول (۲): متغیرهای سیستم‌های نروفازی

بخش	نام متغیر	حرف اختصاری	جنس متغیر
عمومی	سنجه شباهت پرش-گرم عمومی	V1	T2
	سنجه شباهت هم‌پوشانی واژه عمومی	V2	T2
تخصصی	سنجه شباهت پرش-گرم تخصصی	V3	T2
	سنجه شباهت هم‌پوشانی واژه تخصصی	V4	T2
کل متن	نسبت تعداد واژه تخصصی به عمومی	V5	T1

٤-٤- قواعد سیستم فازی

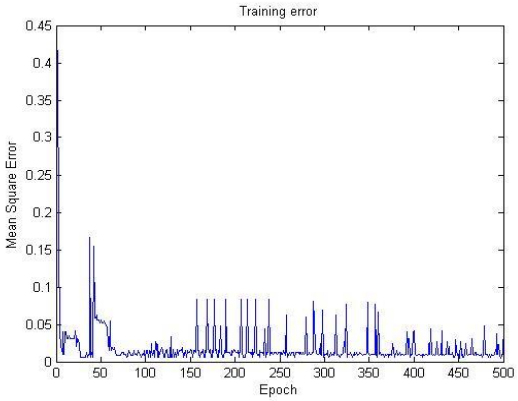
پس از تعریف متغیرهای فازی، قواعد با استفاده از مصاحبه با گروهی پنج نفره از خبرگان زبان‌شناسی و مختصصان حوزهٔ متن تخصصی (یادگیری الکترونیکی) استخراج شده است. برخی از این قواعد در جدول (۳) آورده شده است. برای واضح بودن متغیرهای شباهت‌سنجی ورودی سیستم‌ها به صورت «کم»، «متوسط» و «زیاد» بیان شده تا با خروجی سیستم تمایز داشته باشد.

جدول (۳): برخی قواعد برای سیستم نوروفازی شباهت‌سنجی لغوی

قاعده	نام متغیر					نتیجهٔ سیستم
	V1	V2	V3	V4	V5	
۱	–	زیاد	–	کم	کم	مشابه
۲	–	زیاد	کم	کم	متوسط	نیمه‌مشابه
۳	–	زیاد	متوسط	کم	متوسط	نیمه‌مشابه
۴	–	زیاد	کم	کم	زیاد	نامشابه
۵	–	زیاد	–	متوسط	کم	مشابه
۶	کم	زیاد	کم	متوسط	کم	نامشابه

٤-٥-سیستم نروفازی (انفیس)

پس از تعریف توابع عضویت و قواعد سیستم‌های فازی، برای به دست آوردن نتیجه بهتر در مرحله اول، بایستی سیستم نروفازی را با کمک داده آموزشی تطبیق داد. برای این کار بیش از ۶۵٪ داده‌های پیکرهٔ «همیافت» (۵۴۰ زوج‌جمله) برای آموزش و مابقی برای آزمون روش مورد استفاده قرار گرفته‌اند. پس از مرحلهٔ یادگیری که از طریق روش گرادیان نزولی انجام می پذیرد، توابع عضویت و قواعد فازی با توجه به داده منطبق خواهد شد. نمودار خطای آموزش پارامترهای شبکه عصبی برای داده آموزشی به ازای تعداد چرخه[۱۶] در شکل (۴) نمایش داده می‌شود. این نمودار نشان دهنده آن است که بعد از ۵۰۰ تکرار برای آموزش، سیستم نروفازی حدود ۵٪ خطا در تشخیص میزان شباهت زوج جمله های آموزشی دارد.



شکل (٤): نمودار خطای بر روی پیکره در ۵۰۰ تکرار الگوریتم آموزش

٥- نتایج عددی و ارزیابی

در این بخش نتایج عددی روش پیشنهادی با استفاده از داده‌های واقعی بیان می‌شود. روش پیشنهادی در محیطهای نرم‌افزاری Visual Studio .Net و با استفاده از پایگاه داده Microsoft SQL و نرم‌افزار متلب برای بخش سیستم نروفازی پیاده‌سازی شده است. معماری روش به صورت شئ‌گرا انجام گرفته تا توانایی سنجش هر دو سند را بر اساس سنجه‌های متفاوت و همچنین پایگاه دانش‌های مختلف داشته باشد. از آنجا که هدف در این مقاله بررسی شباهت در متون تخصصی فارسی است و در حال حاضر هیچگونه پیکره یا مجموعه داده‌ای با این ویژگی در زبان فارسی وجود ندارد، از مجموعهٔ داده‌های پایان‌نامه‌های حوزهٔ «یادگیری الکترونیکی» برای تولید پیکره استفاده شده است که «همیافت» نام‌گذاری شده است. این مجموعه داده شامل ۸۱۰ زوج‌جمله بوده که توسط خبرگان به سه کلاس «مشابه»، «نیمه مشابه» و «نامشابه» دسته‌بندی شده است. جملهٔ اول این زوج‌جمله‌ها از پایان‌نامه‌های یادگیری الکترونیکی دانشجویان کارشناسی ارشد و دکتری ایران اتخاذ شده است و جملهٔ دوم توسط خبرگان به صورت تصادفی در یکی از این سه دسته ساخته و با آن زوج شده است و در نهایت توسط گروهی از خبرگان دیگر (مستقل از خبرگان تولیدکنندهٔ زوج‌جمله) این زوج‌جمله‌ها نیز ارزیابی و میانگین امتیاز این افراد به عنوان دسته نهایی این زوج‌جمله‌ها در نظر گرفته‌شده است. جمله‌ها در این مجموعه داده بین ۱۲ تا ۳۲ واژه و به طور میانگین ۱۸ واژه دارند. تعداد زوج واژه‌ها در هر دسته یکسان بوده به طوری که هر کدام یک سوم از مجموعه داده را دربر می‌گیرند.

برای ارزیابی عملکرد کارایی روش‌پیشنهادی از شاخص‌های «دقت»، «بازخوانی» و «سنجهٔ F» استفاده شده است. جدول (۴) نتایج ماتریس درهم‌ریختگی روش پیشنهادی بر روی بخش دادهٔ آزمون (۲۷۰زوج‌جمله) پیکره را نشان می‌دهد. برای محاسبه معیارهای کارایی از آنجا که هدف در این مقاله تشخیص شباهت لغوی، دو کلاس «مشابه» و «نیمه‌مشابه» به عنوان کلاس مثبت و کلاس «نامشابه» در نظر گرفته شده است.

جدول (٤): ماتریس درهم‌ریختگی نتایج سنجه یادگیر پیشنهادی

		کلاس پیش بینی شده	
		مشابه	نامشابه
کلاس واقعی	مشابه	۱۶۸	۱۲
	نامشابه	۱۳	۷۷

از آنجا که روش پیشنهادی بر روی متون تخصصی فارسی پیاده‌سازی شده است برای مقایسهٔ بهتر روش پیشنهادی، سه روش طرح شده در تحقیقات قبلی و بر روی پیکرهٔ فارسی پیاده‌سازی شده است. ضریب جاکارد به عنوان خط پایه آزمایش درنظر گرفته شده است. نتایج پیاده‌سازی این روش‌ها در جدول (۵) نشان داده شده است.

جدول (٥): نتایج کارایی روش پیشنهادی

روش	سنجهٔ F	دقت	بازخوانی
سنجه یادگیر نروفازی	٪۸۴	٪۸۲	٪۸۲
گوپتا و همکاران(نظریهٔ فازی) [۲۱]	٪۶۵	٪۶۴	٪۶۵
میهالیشا و همکاران (مقایسه جفت‌واژه) [۲۲]	٪۶۱	٪۶۳	٪۵۹
خط پایه: ضریب جاکارد[۲۳]	٪۵۵	٪۵۸	٪۵۳

٦- نتیجه‌گیری

در این مقاله روش جدیدی مبتنی بر سیستم استنتاج نروفازی برای ابهام‌زادیی و قابلیت تطبیق قواعد با داده ارائه‌شده که از نظریه‌های ریاضی که توانایی

حل این‌گونه عدم‌قطعیت داده (همانند نظریه‌های فازی و همچنین شبکه‌های عصبی مصنوعی) را دارند استفاده‌کرده است. نتایج روش پیشنهادی نشان‌دهنده آن است که کارایی و دقت این روش نسبت به حالت قطعی و می‌تواند به عنوان سنجه‌ای که توانایی انطباق‌پذیری با محتوا را دارد به کار گرفته شود. از طرفی روش پیشنهادی همان اشکال روش‌های ترکیبی یعنی حجم محاسبات بالا است را داراست. با توجه به پیچیدگی محاسباتی از مرتبه $O(n^2)$، n بُعد بردارهای ورودی، این روش‌ها برای مسائل با ابعاد بالا در حوزهٔ متنی قابل استفاده نیستند. یکی از روش‌های بهبود مدت زمان انجام الگوریتم استفاده از الگوریتم‌های موازی است. از طرفی نبود یک هستان‌نگار که تمام واژه‌های زبان فارسی را پوشش دهد از محدودیت‌های مسئله است.این روش برای شباهت‌سنجی بین تعدادی از پایان‌نامه دانشجویان رشته‌ی یادگیری الکترونیکی به کارگرفته شده است. نتایج نشان می‌دهد روش پیشنهادی در بیش از ۸۲٪ موارد، تحلیل درستی از شباهت زوج‌جمله ارائه می‌دهد و می‌توان از این روش برای کشف شباهت لغوی و در نهایت دستبرد ادبی استفاده کرد.

مراجع

[۱] آهنگربهان، حمید، و غلامعلی منتظر، ۱۳۹۴، مدل سازی عدم قطعیت در سنجش شباهت لغوی محتوای منابع وب فارسی، اولین کنفرانس بین المللی وب پژوهی، تهران،فروردین ۱۳۹۴، دانشگاه علم و فرهنگ.

[۲] برادران هاشمی، علی، محمدرضا میبدی، و سعید شیری قیداری، ۱۳۸۶، کاوش استفاده از وب با استفاده از کلونی مورچه ها، پانزدهیمن کنفرانس مهندسی برق ایران، تهران، مرکز تحقیقات مخابرات ایران

[۳] برادران هاشمی، علی، و محمدرضا میبدی، ۱۳۸۵، داده کاوی استفاده از وب با استفاده از اتوماتای یادگیر توزیع شده، دوازدهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، دانشگاه شهید بهشتی

[4] Osman, Ahmed Hamza, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda. "An improved plagiarism detection scheme based on semantic role labeling." Applied Soft Computing 12, no. 5 (2012): 1493-1502.

[5] S. Alzahrani, N. Salim, A. Abraham, Understanding plagiarism linguistic pat-terns, textual features, and detection methods, IEEE Trans. Syst. Man Cybernet. C: Appl. Rev. 42 (2012) 133–149.

[6] El-Alfy, El-Sayed M., Radwan E. Abdel-Aal, Wasfi G. Al-Khatib, and Faisal Alvi. "Boosting paraphrase detection through textual similarity metrics with abductive networks." Applied Soft Computing (2014).

[7] Bilenko Mikhail and Mooney Raymond J. Adaptive Duplicate Detection Using Learnable String Similarity Measures. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), Washington DC, pp.39-48, August, 2003.

[8] Androutsopoulos, Ion, and Prodromos Malakasiotis. "A survey of paraphrasing and textual entailment methods." Journal of Artificial Intelligence Research (2010): 135-187.

[9] Barrón-Cedeño, Alberto, and Paolo Rosso. "On automatic plagiarism detection based on n-grams comparison." In Advances in Information Retrieval, pp. 696-700. Springer Berlin Heidelberg, 2009.

[10] A. Rajkumar and A. Chitra, "Paraphrase recognition using neural network classification," International Journal of Computer Applications,vol. 1, no. 29, pp. 42–47, February 2010.

[11] Finch, A., and H, Y.S., and Sumita, E. (2005). Using machine translation evaluation techniques to determine sentence-level semantic equivalence, "Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)", Jeju Island, South Korea, pp. 17-24.

[12] Wan, S., Dras, M., Dale, R., and Paris, C. (2006). Using dependency-based features to take the "para-farce" out of paraphrase, *Proceedings of the Australasian Language Technology Workshop (ALTW 2006)*, pp. 131-138.

[13] Zanzotto, F. M., & Dell' Arciprete, L. (2009). Efficient kernels for sentence pair classification. In Proc. of the Conf. onEMNLP, pp. 91–100, Singapore.

[14] Oroumchian, Farhad, and Firooz M. Garamalek. An evaluation of retrieval performance using farsi text. na, 2002.

[15] Nayyeri, Amir, and Farhad Oroumchian. "Fufair: a fuzzy farsi information retrieval system." University of Wollongong in Dubai-Papers (2006): 11.

[16] Zamanifar, Kamran, et al. "A new approach for semantic web matching." Security-Enriched Urban Computing and Smart Grid. Springer Berlin Heidelberg, 2010. 77-85.

[17] Abraham, Ajith. "Adaptation of fuzzy inference system using neural learning." Fuzzy systems engineering. Springer Berlin Heidelberg, 2005. 53-83.

[18] Nauck, Detlef, Frank Klawonn, and Rudolf Kruse. Foundations of neuro-fuzzy systems. John Wiley & Sons, Inc., 1997.

[19] Metzler, D. , Bernstein, Y. , Croft, W. B. , Moffat, A. , and Zobel, J. Similarity measures for tracking information flow. In Proceedings of CIKM '05, pp. 517-524. 2005.

[20] Šarić, Frane, et al. "Takelab: Systems for measuring semantic text similarity." Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2012.

[21] Rohit Gupta, Hanna Béchara, Ismail El Maarouf, and Constantin Orasan. 2014. UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. SemEval 2014, 785.

[22] Rada,Mihalcea, Courtney Corley, and Carlo Strapparava. . 2006 "Corpus-based and knowledge-based measures of text semantic similarity." In AAAI, vol. 6, pp. 775-780.

[23] Anna. Huang, 2008. "Similarity measures for text document clustering." In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49-56.

زیرنویس‌ها

[1] Natural Language Processing
[2] Dictionary
[3] Ontology
[4] Uncertain
[5] Neuro-Fuzzy Inference
[6] Lexical
[7] Semantic
[8] Syntactic
[9] Dependency Grammar Parsers
[10] Cooperative
[11] Concurrent

[12] Integrated
[13] Takagi Sugeno
[14] Adaptive Network Based Fuzzy Inference System (ANFIS)
[15] Tsukamoto
[16] Epoch