دسته‌بندی پرسش‌ها با استفاده از ترکیب دسته‌بندها

هادی قائمی* و محسن کاهانی

آزمایشگاه فناوری وب معنایی دانشگاه فردوسی مشهد، مشهد، ایران

چکیده

هدف از تولید و گسترش سامانه‌های پرسش و پاسخ، ایجاد پاسخ دقیق برای پرسش داده‌شده به زبان طبیعی است. یکی ازمهم‌ترین بخش‌های سامانه‌های پرسش و پاسخ، دسته‌بندی پرسش است. عمل دسته‌بندی پرسش، پیش‌بینی نوع پاسخ مورد نیاز برای پرسش داده‌شده به زبان طبیعی است. کارهای انجام‌شده در این زمینه را می‌توان در دو دستهٔ مبتنی بر قانون و مبتنی بر یادگیری تقسیم کرد. در این مقاله یک معماری جدید برای دسته‌بندی ترکیبی پرسش‌ها ارائه شده است، نتایج هریک از دسته‌بندها توسط پنج روش رأی‌گیری وزن‌دار، فضای دانش رفتار، بیز ساده، کلیشهٔ تصمیم و دمپستر شفر ترکیب شده و خروجی نهایی را شکل می‌دهد. این روش ترکیبی متشکل از دو دسته‌بند مبتنی بر یادگیری ماشین (ماشین بردار پشتیبان و نمایش پراکنده) و یک دسته‌بند مبتنی بر قانون استفاده شده است. عملیات دسته‌بندی مبتنی بر یادگیری با توجه به مجموعه ویژگی‌های استخراج شده از پرسش‌ها انجام می‌پذیرد. این ویژگی‌ها براساس ساختار لغوی و نحوی پرسش‌ها استخراج شده‌اند. در پایان نتایج حاصل از دسته‌بندها با روش‌های معمول در ترکیب دسته‌بندهای تک‌طبقه ترکیب شده‌اند و نتایج حاصل بیان‌کنندهٔ بهبود عملیات دسته‌بندی نسبت به روش‌های موجود است.

واژگان کلیدی:دسته‌بندی پرسش‌ها، مبتنی برقانون، مبتنی بر یادگیری، نمایش پراکنده، ماشین بردار پشتیبان، پرسش‌وپاسخ.

Question Classification using Ensemble Classifiers

Hadi Ghaemi* & Mohsen Kahani

Web Technology Lab. Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

Question answering systems are produced and developed to provide exact answers to the question posted in natural language. One of the most important parts of question answering systems is question classification. The purpose of question classification is predicting the kind of answer needed for the question in natural language. The literature works can be categorized as rule-based and learning-based methods. This paper proposes a novel architecture for hybrid classification of questions. The results of the classifiers were combined by five methods of Weighted Voting, Behavior Knowledge space, Naive Bayes, Decision Template and Dempster-Shafer. The method uses a combination of two classifiers based on machine learning (Support Vector Machine and Sparse Representation) and one rule-based classifier. The learning-based classification uses the set of features extracted from the questions. The features are extracted on the basis of the lexical and syntactic structure of the questions. The results from the classifiers were combined by the methods that are common in the combination of one-class classifiers and the Obtained results indicate the improvement of the classification operations in comparison with the present methods.

Keywords: Question classification, Rule-based, Learning-based, Sparse Representation, Support Vector Machine, Question answering.

* نویسندهٔ عهده‌دار مکاتبات

۱- مقدّمه

با افزایش و توسعهٔ وب، نیاز است تا موتورهای جستجو هوشمندتر از قبل رفتار کنند. در بیشتر موارد کاربران به جای فهرستی از اسناد به بخش کوچکی از اطلاعات نیاز دارند تا به جای مطالعه کل اسناد، بخش کوچکی از سند را مطالعه کنند. با توجه به نیاز کاربران، نسل بعدی موتورهای جستجو، سامانه‌های پرسش و پاسخ است، سامانه‌هایی که به کاربر اجازه می‌دهد؛ سؤال را در ساختار زبان طبیعی به سامانه بدهد و پاسخ را به‌صورت دقیق دریافت کند. هدف تولید و گسترش سامانه‌های پرسش و پاسخ، دادن پاسخ دقیق و کوتاه به پرسش داده شده به زبان طبیعی است. دسته‌بندی پرسش جزء مهم‌ترین بخش‌های سامانه‌های پرسش و پاسخ است.

دسته‌بندی پرسش را می‌توان، به عمل تعیین یک مقدار منطقی به دوتایی $\ll q_j, c_i \gg \in \mathrm{Q} \times \mathrm{C}$ تعریف کرد. که دراین جا Q مجموعهٔ پرسش‌ها و $C = \{c_1, \ldots, c_{|C|}\}$ مجموعهٔ از پیش تعیین شده برای دسته‌های پرسش است. مقداردهی $\ll q_j, c_i \gg$ با T بیان‌کنندهٔ تعلق پرسش q_j به c_i و مقدار F بیان کنندهٔ عدم تعلق این پرسش به دستهٔ c_i می‌باشد.

سامانه‌های پرسش و پاسخ مختلف، معماری‌های متفاوت دارند؛ اما در همهٔ آنها بخشی برای دسته‌بندی پرسش وجود دارد که نقشی مهم را ایفا می‌کنند [11]. حدود ۳۶/۴٪ از خطاهای یک سامانه پرسش و پاسخ، به‌علت دسته‌بندی نادرست سئوال است[14]. تأثیر سامانه‌های دسته‌بندی پرسش‌ها در افزایش دقت سامانه‌های پرسش و پاسخ غیر قابل انکار و در [14] این تأثیر نشان داده شده است. به‌عنوان مثال برای پرسش ۱ برچسب تعلق‌گرفته Location است، که مشخص می‌کند پاسخ مورد نیاز به پرسش داده شده از نوع مکان است. درواقع نوع پاسخ مربوط به پرسش را پیش‌بینی می‌کند، به همین دلیل به دسته‌بندی پرسش پیش‌بینی‌کنندهٔ نوع پاسخ گفته می‌شود.

Q1: *What country's capital is Tirana?*

دسته‌های متفاوتی برای پرسش‌ها مشخص شده است؛ اما بیشترین کاربرد را دسته‌های معرفی‌شده در [10] دارد، که در آن برای پرسش‌ها دو دسته‌بندی ریز[1] با ۶ دسته و درشت[2] با ۵۰ دسته ارائه شده است. این دسته‌بندی‌ها در جدول۱ آورده شده است. در این مقاله تمرکز ما بر روی دستهٔ ریز با ۶ دسته است.

به طور کلی برای دسته‌بندی پرسش‌ها سه انگیزهٔ اصلی کاهش فضای جستجو، مشخص‌کردن پاسخ و انتخاب راهبرد جستجو را می‌توان درنظر گرفت. در ادامه هر یک از دلایل توضیح داده خواهند شد.

(جدول- ۱): دسته‌های پرسش [11]
(Table-1): Question classes [11]

Coarse	Fine
ABBR	اختصار، گسترش
DESC	تعریف، توضیح، چگونگی، دلیل
ENTY	حیوان، انسان، رنگ، موجود زنده، ارز، مریضی، رویداد، خوراک، ابزارآلات، زبان، نامه، زمین، کالا، مذهب، ورزش، اجسام، سمبل‌ها، تکنیک، اصطلاح، رسانه، واژه
HUM	شخص، تفسیر، گروه، عنوان
LOC	شهر، کشور، کوه، ایالت، استان
NUM	کد، عدد، زمان، فاصله، پول، رتبه، درصد، مدت، سرعت، دمای هوا، سایز، وزن

کاهش فضای جستجو: با توجه به نوع طبقه می‌توان فضای جستجو را کاهش داد. به عنوان مثال با دانستن برچسب پرسشی که از نوع مکان است، سامانه بدنبال متن‌ها و پاراگراف‌هایی خواهد گشت که اطلاعاتی از این نوع را داشته‌باشد.

مشخص‌کردن پاسخ: دانستن نوع طبقه علاوه‌بر این‌که فضای جستجو را کاهش می‌دهد، در یافتن پاسخ صحیح از مجموعه پاسخ‌های نامزد تأثیر گذار است؛ به‌عنوان مثال برای پرسش۱، برچسب تعلق‌گرفته از نوع مکانی است، درنتیجه از بین پاسخ‌های نامزد انتخاب‌شده، پاسخ نهایی باید از نوع طبقه پرسش باشد.

انتخاب راهبرد جستجو: با توجه به نوع طبقه می‌توان راهبرد جستجو را تعیین کرد. به‌عنوان مثال در پرسش۲ با توجه به اینکه نوع پرسش definition می‌باشد، می‌توان راهبرد یافتن پاسخ را به‌صورت از پیش تعریف‌شدهٔ "... ethology is a" در نظر گرفت.

Q2: *What is ethology?*

در ادامهٔ این مقاله ابتدا در بخش ۲ کارهای انجام‌شده در زمینهٔ دسته‌بندی پرسش‌ها بیان می‌شود، در بخش ۲-۱ دسته‌بندهای مورد استفاده در این روش پیشنهادی معرفی می‌شود. در بخش ۳ روش پیشنهادی بیان شده‌است و در بخش ۴ پیاده‌سازی و ارزیابی و درنهایت در بخش ۵ نتیجه‌گیری بیان می‌شود.

[1] Coarse

[2] Fine

۲- مروری بر کارهای گذشته

به‌نظر می‌رسد ساده‌ترین راه برای دسته‌بندی پرسش‌ها استفاده از قواعد دست‌نویس از پیش تعریف‌شده است. قوانین مورد استفاده می‌تواند ساده و پیچیده باشد. ساده مثل "اگر پرسشی که با who یا whom شروع شود به کلاس شخص دسته‌بندی می‌شود" و "اگر پرسشی که با where شروع شود با برچسب مکانی برچسب زده می‌شود". همچنین این قوانین می‌توانند پیچیده باشند و برای بررسی، نیاز باشد توسط ابزارهای پردازش زبان طبیعی، مورد پردازش واقع شوند از این سامانه‌ها می‌توان به [18] و [16] اشاره کرد.

در اکثر موارد استخراج قوانین مورد استفاده به‌صورت دستی انجام می‌پذیرد [4]، [16] و [13]. اما در بعضی موارد این قوانین به‌صورت خودکار استخراج می‌شوند[6]. در [6] قواعد در قالب ماشین‌های حالات متناهی، بیان شده است، این قواعد از روشی خودکار و با استفاده از مثال‌های آماده، و پس از حذف ایست‌واژه‌ها، و با استفاده از کلیدواژه‌ها استخراج شده‌اند و براساس فراوانی رخداد کلیدواژه‌ها در ترتیب پرسش مورد بررسی واقع می‌شوند؛ اما در کل روش‌های مبتنی بر یادگیری مزایا و معایب خاص خود را دارد. برخی مزایا و معایب این روش‌ها عبارتنداز:

- محاسبات سریع و ارزان دارند.
- این قوانین، حداکثر انعطاف‌پذیری را دارند.
- قوانین فقط برای مجموعه دادهٔ که ایجاد شده‌اند، دقت خوبی دارند.
- فرآیند استخراج قوانین از پرسش‌ها کاری خسته‌کننده و وقت گیر و نیاز به کار و صرف وقت زیادی است.
- تعدا قوانین استخراجی ممکن است بسیار زیاد باشد، به‌نحوی که مدیریت آن مشکل شود.

روش‌های مبتنی بر یادگیری، برای بخش یادگیری، ابتدا نیازمند داده‌های برچسب خورده هستند؛ سپس مدل‌های یادگیری بر روی این داده‌ها، آموزش داده می‌شوند. در این روش‌ها برای دسته‌بندی، پرسش‌ها در قالب مجموعه‌ای از ویژگی‌ها در نظر گرفته می‌شوند؛ و سعی می‌شود با توجه به ویژگی‌های استخراج‌شده از داده‌های برچسب‌خورده، الگوها را به‌صورت خودکار استخراج کنند. بنابراین در این روش‌ها انتخاب ویژگی‌ها و دسته‌بند اهمیت زیادی دارد. ویژگی‌های مورد استفاده می‌تواند خیلی ساده یا پیچیده باشد؛ ساده در حد ویژگی‌های کلامی یا پیچیده مثل ویژگی‌ها نحوی و معنایی.

علاوه‌براین، روش‌های ترکیبی نیز وجود دارد که از هر دو روش مبتنی بر یادگیری و مبتنی بر قانون استفاده می‌کنند [18]، [17]. روش ارائه‌شده در این مقاله نیز در این دسته قرار دارد.

۲-۱- دسته‌بندها

از الگوریتم‌های دسته‌بندی مورد استفاده برای دسته‌بندی پرسش‌ها می‌توان به ماشین بردار پشتیبان [11]، شبکه‌های بیزی[1][19]، درخت‌های تصمیم[2][19]، SNoW [3] [10] و BPNN [11] اشاره کرد. در این میان، از مجموع کارهای انجام‌شده ماشین بردار پشتیبان بهترین کارآیی را داشته است.

در این مقاله از دو دسته‌بند مبتنی بر ماشین بردار پشتیبان و مبتنی بر نمایش پراکنده استفاده شده است. همچنین ویژگی‌های مورد استفاده برای هر دسته‌بند، متفاوت است، که انتخاب براساس نتایج دسته‌بندها صورت پذیرفته شده است.

۲-۱-۱- ماشین بردار پشتیبان

ماشین بردار پشتیبان، با این فرض که دسته‌ها به‌صورت خطی جداپذیرند، ابرصفحه‌هایی با حداکثر حاشیه[4] را به‌دست می‌آورد که دسته‌ها را جدا کند. در داده‌های با ابعاد زیاد این دسته‌بند جزء سریع‌ترین دسته‌بندها است. این دسته‌بند نوع کرنلی محسوب می‌شود؛ چرا که در این روش برای جداکردن دسته‌ها، داده‌ها را توسط تابعی تحت عنوان تابع کرنل به ابعادی بیشتر نگاشت می‌کند. در دسته‌بندی پرسش‌ها به‌طورمعمول از کرنل خطی استفاده می‌شود. در این مقاله نیز از کرنل خطی استفاده شده است.

۲-۱-۲- نمایش پراکنده

در سامانه معادلات خطی $A \in \mathbb{R}^{m \times n}$، اگر $m < n$ باشد، آن سامانه را فرومعین می‌گویند. در این سامانه تعداد مجهول‌ها بیشتر از تعداد معادلات است؛ درنتیجه بی‌نهایت جواب برای آن وجود دارد. به‌منظور محدودکردن تعداد پاسخ‌ها و ارزیابی مطلوبیت هر پاسخ ممکن x، از یک تابع $J(x)$ استفاده شده

[1] Naïve Bayes
[2] Decision Trees
[3] Sparse Network of Winnows
[4] Maximum margin

اســـت و پاســـخی با کمترین مقدار تابع ارزیابی، به‌عنوان مطلوب‌ترین پاسخ است. در این صورت مسئله به یک مسئله بهینه‌سازی تبدیل می‌شود. مسئله بهینه‌سازی کلی به‌صورت رابطهٔ ۱ می‌شود.

$$(P_J):\ min_x J(x) \quad Subject\ to\ b = AX \qquad (۱)$$

شـناخته‌شـده‌ترین انتخاب برای تابع $J(x)$، مربع نرم اقلیدسی، یعنی $J(x) = \|X\|_2^2$ است.

۲-۲- دسته‌بندی مبتنی بر نمایش پراکنده

در [21] بـا اســتفاده از نمایش پراکنده، الگوریتمی برای دسـته‌بندی ارائه شـده اسـت که در آن برای هر طبقه iام، $\delta_i: \mathbb{R}^n \to \mathbb{R}^n$ را به‌عنوان یک تابع مشـخصه در نظر می‌گیرد تا ضـرایب مرتبط با نمونه‌های طبقه iام را از بردار ضـرایب پراکنده، انتخاب کند. برای $x \in \mathbb{R}^n$ حاصـل از مسـئلهٔ بهینه‌سازی، $\delta_i(x) \in \mathbb{R}^n$ یک بردار جدید همانند x می‌شود، با این تفاوت که در آن فقط درایه‌های مرتبط با کلاس iام مقدار دارند و مابقـی، درایه‌های صـفر هستند. با استفاده از ضـرایب مرتبط با طبقه iام انتخاب‌شـده توسط $\delta_i(x)$، نمونه آزمون y تقریب زده می‌شـود، به‌طوری که $\hat{y}_i = A\delta_i(x)$؛ سـپس مبتنی بر کمینه‌ترین باقیمانده بین y و $\hat{y}_i$ در میان تمام طبقه‌ها، برچسـب نمونه آزمون را تعیین می‌کند؛ یعنی همانند رابطه ۲:

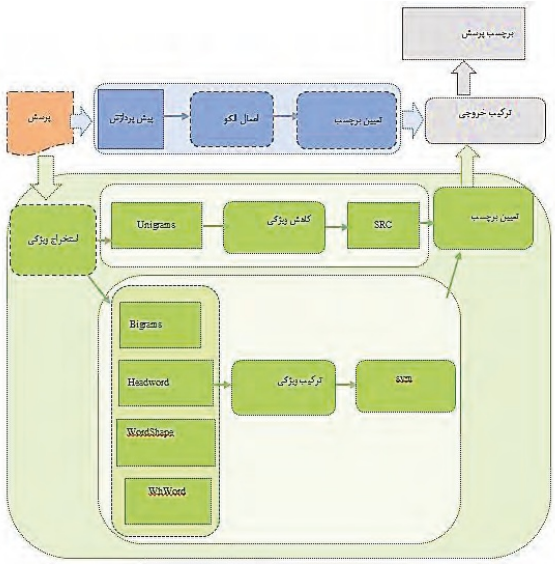
$$\min_i\ r_i(y) \triangleq \|y - A\delta_i(x_1)\|_2 \qquad (۲)$$

۳- روش پیشنهادی

همان‌طور که بیان شـد در روش ارائه‌شـده از یک دسـته‌بند مبتنی بر قاعده و دو دسـته‌بند مبتنی بر یادگیری اسـتفاده شده است. معماری روش پیشنهادی در شکل (۱) آورده شده است.

مجموعه قوانین اسـتخراج‌شـده در بخش مبتنی بر قانون از داده‌های یادگیری اسـتخراج می‌شود و توسط آنها دسته‌های پرسـش‌های مرحلهٔ آزمون مشـخص می‌شود. برای روش‌های مبتنی بر یـادگیری، ابتدا مجموعه ویژگی‌ها از پرسـش‌ها اسـتخراج می‌شـود و در ادامه دسـته‌بندی داده‌ها توسط این مجموعـه ویژگی طی دو مرحلـهٔ یـادگیری و آزمون انجـام می‌پذیرد.

بـه‌طورکلی از نوآوری‌های روش ارائه‌شـده می‌توان به‌ایدهٔ اسـتفاده از دسـته‌بند مبتنی بر نمایش پراکنده برای پردازش متن برای نخسـتین بار و همچنین ارائهٔ معماری با استفاده از ترکیب دسته‌بندها اشاره کرد که باعث بهبود قابل ملاحظه در نتایج نهایی شده است.



(شکل- ۱): معماری سیستم دسته‌بند ترکیبی
(Figure-1): Architecture of Ensemble Classifier System

۳-۱- دسته‌بند مبتنی بر قانون

برای دسته‌بندی مبتنی بر قانون ابتدا طی یک مرحلهٔ پیش‌پردازش درخت تجزیهٔ پرسـش ایجاد و اسـامی خاص پرسـش توسـط ابزار شـناسـایی اسـامی خاص و Gazetteer اسـامی خاص پرسـش شناسایی می‌شوند. البته Gazetteer را می‌توان شناسایی‌کنندهٔ اسامی خاص در نظر گرفت که در آن اسـامی در گروه‌های جدا به‌صـورت فهرسـتی از قبل آماده، گردآوری شـده اسـت. با توجه به این‌که اسـامی در گروه‌های جدا بیان می‌شـوند، می‌توان در موارد خاص از آنها اسـتفاده کرد. برای اکثر شـناسـایی‌کننده‌های اسـامی خاص، اسـامی شـهرها و کشـورها تفاوتی ندارد و برای هر دو برچسب مکان تعلق می‌گیرد؛ اما در Gazetteer این اسـامی به صـورت جدا فهرسـت شـده است. همچنین شناسایی‌کنندهٔ اسامی خاص، اسـامی مشـاغل را شـناسـایی نمی‌کنند. در ادامه برای ایجاد قوانین برای دسـته‌بندی پرسـش از درخت تجزیه و اسـامی شناسایی‌شده استفاده می‌شود. قوانین ایجادشده در دسته‌بند مبتنی بر قانون در قالب عبارات منظم می‌باشـند. این قوانین با توجه به ساختار لغوی و نحوی پرسش‌ها ایجاد شده است.

(جدول- ۲): برخی از قواعد مورد استفاده برای دسته‌بند مبتنی بر قاعده

(Table-2): Some of the rules used in the rule-based classifier

	Label	Pattern
1	*DESC*	پرسش‌هایی که با یک *what is/are* and شروع می‌شوند و با همراه شده *a, an* or *the* در ادامه با یک یا چند کلمه همراه هستند.
2	*DESC*	پرسش‌هایی که با what do/does شروع شده و با mean خاتمه می‌یابند.
3	*DESC*	پرسش‌هایی که با what causes/cause آغاز می‌شوند.
4	*DESC*	پرسش‌هایی که با what is/are آغاز شده و با known for خاتمه می‌یابند.
5	*DESC*	پرسش‌هایی که با why آغاز می‌شوند.
6	*DESC*	پرسش‌هایی که با what is/are آغاز می‌شوند و با فعل همراه نیستند.
7	*DESC*	پرسش‌هایی که با how is/are/did/does/do/ can آغاز می‌شوند.
8	*HUM*	پرسش‌هایی که با what/which همراه با اسامی شغل همراه هستند.
9	*ENTY*	پرسش‌هایی که با How do you say آغاز می‌شوند.
10	*ABBR*	پرسش‌هایی که با what does/do آغاز شده و با stand for همراه هستند.

در جدول (۲) بخشی از قوانین مورد استفاده آورده شده است. ترتیب اعمال این قوانین در نتیجهٔ دسته‌بند تأثیرگذار است و باید به آن توجه داشت، چون بعضی از الگوها خاص و بعضی عام هستند و قوانین خاص باید قبل از قوانین عام قرار گیرند. به‌عنوان مثال در قانون ۷ بیان شده است، "درصورت وجود how is/are/did/does/do/can در ابتدای پرسش، نوع پرسش DESC خواهد بود"، همچنین در قانون ۹ بیان شده‌است "درصورت وجود عبارت How do you say در ابتدای پرسش، نوع پرسش ENTY خواهد بود". قانون ۷ قانونی کلی و قانون ۹، قانونی خاص است. درصورت عدم توجه به ترتیب این قوانین، موارد مربوط به قانون ۹ به‌اشتباه توسط قانون هفت دسته‌بندی می‌شوند.

به‌دلیل بروز حالات استثنای زیاد و برای افزایش دقت قوانین استخراج‌شده از ابزارهای پردازش زبان طبیعی مثل برچسب‌زن اجزای سخن و شناسایی‌کنندهٔ اسامی خاص و Gazetteer استفاده شده است.

به‌عنوان مثال، قانون شش در حالتی رخ می‌دهد که "بعد از *what* یکی از افعال *is/are* رخ دهد؛ و در ادامه دیگر هیچ فعلی رخ ندهد؛ در این صورت پرسش از نوع *DESC* خواهد بود". برای اعمال این قانون نیاز به دانستن برچسب هر یک از کلمات است که برای این منظور نیاز به استفاده از برچسب‌گذار اجزای سخن است. به عنوان مثال برای این دسته می‌توان به پرسش سه اشاره کرد:

Q3: *What is the Olympic motto?*

خروجی حاصل از برچسب‌گذار اجزای کلام مربوط به این پرسش به‌صورت زیر است.

WP/What VBZ/is DT/the NNP/Olympic NN/motto.?

$$x=(x_1, x_2, \dots, x_N) \quad (۳)$$

با توجه به خروجی برچسب‌گذار اجزای سخن، این پرسش شرایط اعمال این قانون را دارد.

به‌عنوان مثالی دیگر پرسش هشت بیان می‌کند، "درصورتی‌که پرسش با *what/which* شروع و در ادامه اسمی از نوع شغل انسانی آورده شود، این پرسش از نوع *HUM* خواهد بود". برای تشخیص اسامی شغل انسانی از ابزار gazetteer استفاده می‌شود. به‌عنوان مثال در پرسش چهار کلمه *actor* اسم شغل انسانی می‌باشد و بعد از کلمهٔ پرسش what آمده است درنتیجه شرط اعمال قانون را دارا دارد.

Q4: *What actor first portrayed James Bond?*

باید توجه داشت ترتیب اعمال قواعد اهمیت زیادی دارد و همچنین سعی شده است تا قواعد به جهت افزایش در دقت به صورت ترکیبی از شروط باشد؛ به‌عنوان مثال در پرسش پنج به جهت وجود شغل انسانی ممکن است، پرسش به *HUM* برچسب‌گذاری شود؛ اما با توجه به قاعده هفت از جدول ۲ نوع پرسش DESC برچسب‌گذاری می‌شود. با توجه به اینکه قواعد جدول ۲ از قواعد با اولویت هستند و در ابتدا اعمال می‌شوند سامانه به‌درستی فرآیند برچسب‌گذاری را انجام می‌دهد.

Q۵: How do doctors diagnose bone cancer?

۳-۲- دسته‌بند مبتنی بر یادگیری

برای دسته‌بند مبتنی بر یادگیری از دسته‌بند مبتنی بر نمایش پراکنده و ماشین بردار پشتیبان استفاده شده است. ویژگی‌های مورد استفاده برای دسته‌بندها، مستقل از یکدیگرند و این انتخاب با توجه به عملکرد آنها بر روی ویژگی‌ها است. همان‌طورکه در بخش بعد بیان خواهد شد با توجه به نتایج حاصل‌شده، ویژگی‌های مورد استفاده برای دسته‌بند مبتنی بر نمایش پراکنده، کاهش دادهٔ Unigrams و برای ماشین بردار پشتیبان Bigrams، Headword، WordShape و WhWord است.

در دسته‌بندی مبتنی بر یادگیری از ویژگی‌های متفاوتی استفاده شده است که می‌توان آنها را در دو دسته تقسیم کرد. ویژگی‌های لغوی، نحوی دسته‌بندی نمود.

۳-۳- ویژگی‌های لغوی

ویژگی‌هایی لغوی مربوط به پرسش‌ها به‌طورمعمول براساس کلمات درون پرسش استخراج می‌شوند. در دسته‌بندی پرسش‌ها همانند نمایش اسناد، پرسش‌ها به‌صورت بردار نمایش داده می‌شود. به‌عنوان مثال پرسش X به‌صورت رابطهٔ ۳ نمایش داده می‌شود.

در این رابطه i بیان‌کنندهٔ فرکانس تعداد ویژگی xi ام در پرسش x و N تعدا کل ویژگی‌ها است. ویژگی‌های لغوی مورد استفاده به‌صورت زیر است:

Unigram: درواقع همان فهرست کلمات درون پرسش است. رخداد تک‌تک کلمات درون پرسش است. به‌عنوان مثال برای پرسش ۵ این مجموعه ویژگی به صورت زیر می‌باشد:

Q5: *How many Grammys did Michael Jackson win in 1983?*

x = {(How, 1), (many, 1), (Grammys, 1), (did, 1), (Michael, 1), (Jackson, 1), (win, 1), (in, 1), (1983, 1), (?, 1)}

همان‌طور که در ادامه بیان می‌شود، ابعاد این ویژگی خیلی بیشتر از تعداد داده‌ها است؛ به همین جهت ابعاد این ویژگی توسط SVD کاهش داده و از کاهش‌یافته آن استفاده می‌شود. طبق [11] این ویژگی با ابعاد ۴۰۰ بهترین نتیجه را به همراه دارد به همین جهت، ابعاد این ویژگی را تا ۴۰۰ کاهش داده و به جای کل ویژگی‌ها استفاده می‌شود.

برای کاهش ویژگی با استفاده از SVD ابتدا ماتریس Q ایجاد می‌شود که ردیف‌های آن بیان‌کنندهٔ ویژگی‌ها و ستون‌های آن بیان‌کنندهٔ پرسش‌ها است. ابعاد این ماتریس $d \times n$ است d تعداد ویژگی‌ها و n تعداد پرسش‌ها می‌باشد. هر عضو همانند $Q_{i,j}$ بیان‌کنندهٔ مقدار ویژگی jام برای پرسش i ام است. SVD ماتریس Q با ابعاد $m \times n$ را به سه ماتریس $Q = USV^T$ تجزیه می‌کند که $U = \left[u_{ij}\right]$ ماتریس متعامد[1] ستونی $m \times m$ بوده و ستون‌های آن بردارهای یکه سمت چپ نامیده می‌شوند؛ $S = diagonal\left(\sigma_1, \sigma_2, \ldots, \sigma_n\right)$ ماتریس قطری $m \times n$ است که عناصر قطر اصلی آن مقادیر یکه غیرمنفی هستند که به‌صورت نزولی مرتبط شده‌اند و $V = \left[v_{ij}\right]$ هم ماتریس متعامد سطری بوده و ستون‌های آن بردارهای یکه سمت راست نامیده می‌شوند.

برای کاهش ویژگی تا ابعاد k، از ماتریس U، زیر ماتریس U_k با ابعاد $d \times k$ ساخته می‌شود که شامل k ستون ابتدایی ماتریس U است ماتریس با ابعاد کاهش‌یافته به‌صورت رابطهٔ ۴ ایجاد می‌شود.

$$R = Q^T U_K \qquad (۴)$$

ابعاد این ماتریس $n \times k$ خواهد بود که سطرها بیان‌کنندهٔ پرسش‌ها و ستون‌ها ویژگی پرسش‌ها است.

Bigram: این ویژگی بخشی از n تایی‌ها است. به جهت زیادبودن ابعاد این ویژگی، فقط از دوتایی ابتدایی هر پرسش استفاده می‌شود. به‌عنوان مثال برای پرسش پنج این ویژگی *How-many* است.

Wh-words: این ویژگی کلمات پرسشی درون پرسش است. با توجه به اینکه انواع این کلمات هشت مورد بیشتر نیستند (*what, which, when, where, who, how, why* و *rest*) ابعاد این ویژگی هشت است. به عنوان مثال برای پرسش شش این ویژگی what است.

Q6: *What is the longest river in the world?*

Word shapes: این ویژگی مربوط به ظاهر کلمات درون پرسش است. پنج نوع ظاهری برای کلمات مشخص شده است *All digits, lower case, upper case, mixed* و *other*. است. به‌عنوان مثال برای پرسش دو این ویژگی‌ها به‌صورت زیر است.

Word-shapes = {(lowercase, 4) (mixed, 4) (digit, 1) (other, 1)}

۳-۴- ویژگی‌های نحوی

ویژگی نحوی مورد استفاده برای دسته‌بندی کلمهٔ اصلی است که توسط درخت تجزیه از پرسش استخراج می‌شود.

کلمهٔ اصلی[2]: کلمهٔ اصلی پرسش، به‌طورمعمول به‌عنوان مفیدترین کلمهٔ یک پرسش از نظر اطلاعاتی تعریف می‌شود. درواقع هدف پرسش را بیان می‌کند، مشخص می‌کند پرسش انتظار چه نوع پاسخی را دارد[7]. تشخیص صحیح کلمهٔ اصلی در پرسش می‌تواند باعث افزایش دقت در تشخیص دستهٔ پرسش شود. به‌عنوان مثال در پرسش سه کلمهٔ city، کلمهٔ اصلی پرسش است. در این پرسش وجود کلمهٔ city کمک زیادی در تشخیص دستهٔ پرسش که LOC است، می‌کند.

[1] Orthonormal

[2] Head Word

استخراج کلمهٔ اصلی از پرسش یکی از چالش‌های مهم می‌باشد، استخراج کلمهٔ اصلی با استفاده از ساختار نحوی پرسش استخراج می‌شود، برای این منظور از درخت تجزیه پرسش استفاده می‌شود.

استخراج کلمهٔ اصلی از پرسش یکی از چالش‌های مهم می‌باشد، استخراج کلمهٔ اصلی با استفاده از ساختار نحوی پرسش استخراج می‌شود، برای این منظور از درخت تجزیه پرسش استفاده می‌شود.

نخستین‌بار ایدهٔ استخراج کلمهٔ اصلی از ساختار درخت تجزیه توسط Collins در [2] ارائه شد. در [2] تعدادی قانون برای استخراج کلید اصلی ارائه شده است که به قانون‌های Collins شناخته می‌شوند.

(جدول-۳): قوانین استخراج کلید اصلی [11]

(Table-3): Primary key extraction rules [11]

Parent	Direction	Priority List
ROOT	L-b-C	S, SBARQ
S	L-b-C	VP, FRAG, SBAR, ADJP
SBARQ	L-b-C	SQ, S, SINV, SBARQ, FRAG
SQ	L-b-C	NP, VP, SQ
NP	R-b-P	NP, NN, NNP, NNPS, NNS, NX
PP	L-b-C	WHNP, NP, WHADVP, SBAR
WHNP	L-b-C	NP, NN, NNP, NNPS, NNS, NX
WHADVP	L-b-C	NP, NN, NNP, NNPS, NNS, NX
WHADJP	L-b-C	NP, NN, NNP, NNPS, NNS, NX
WHPP	R-b-C	WHNP, WHADVP, NP, SBAR
VP	R-b-C	NP, NN, NNP, NNPS, NNS, NX, SQ, PP
SINV	L-b-C	NP
NX	L-b-C	NP, NN, NNP, NNPS, NNS, NX, S

در یک قاعدهٔ گرامری آزاد از بافت به‌صورت $X \rightarrow Y_1 \ldots Y_2$ که X و Y_i در درخت تجزیه گرهٔ غیرپایانی هستند، قواعد اصلی مشخص می‌کنند که کدام یک از گره‌های غیرپایانی، قاعدهٔ اصلی X است. برای مثال برای دستور $SBARQ \rightarrow WHNP\ SQ$ قوانین Collins مشخص می‌کنند که کلمهٔ اصلی در گرهٔ غیرپایانی SQ قرار دارد. این روال به‌صورت تکرارشونده تا زمان رسیدن به گرهٔ پایانی ادامه می‌یابد.

در [11]، [18] و [7] مجموعه قوانین برای استخراج کلید اصلی از پرسش طراحی شده است، در این قواعد برای انتخاب کلید اصلی مربوط به پرسش، به اسامی اولویت بیشتری تعلق گرفته است. به‌عنوان مثال در جدول (۳) مجموعهٔ این قوانین ارائه شده است.

در این جدول ستون نخست گره‌های سمت چپ قوانین را مشخص می‌کند که گره‌های غیرپایانی است. ستون دوم جهت جستجوی سمت راست قوانین را مشخص می‌کند. که می‌تواند براساس دسته[1] یا موقعیت[2] صورت گیرد؛ و ستون سوم فهرست اولویت را بیان می‌کند

منظور از L-b-C، left by category است که بیان می‌کند از چپ‌ترین فرزند، به ترتیب عکس فهرست اولویت عمل جستجو انجام می‌شود و بروز هر تناظری به‌عنوان کلید اصلی برگردانده می‌شود؛ اما منظور از L-b-P، left by position است، الگوریتم ابتدا موارد درون فهرست اولویت را کنترل و تناظر این موارد در فرزندان را از چپ به راست بررسی می‌کند و نخستین رویداد به‌عنوان کلمه اصلی برگردانده می‌شود. R-b-C و R-b-C هم روالی همانند left دارند؛ اما عملیات در سمت عکس موارد قبلی اجرا می‌شود.

این قوانین، قوانین اصلاح‌شدهٔ Collins است که برای اعمال این قوانین نیاز است درخت تجزیهٔ پرسش توسط تجزیه‌گر ایجاد شود و این قوانین به‌صورت از بالا به پایین بر روی درخت تجزیه اعمال می‌شود.

اما قوانین ایجادشده برای همهٔ حالات نتیجهٔ صحیحی را به‌همراه نخواهد داشت. به‌عنوان مثال برای پرسش شش کلمهٔ اصلی city است، اما با اعمال این قوانینairport را برمی‌گرداند.

Q6: *What city is Logan Airport in?*

برای رفع این مشکل می‌توان قوانینی را به‌عنوان استثنا قبل از اعمال قوانین جدول ۳ در نظر گرفت. اگر *WHPP*، *WHNP*، *WHADVP* یا *WHADJP* بیش از یک فرزند داشته باشد، اسم یا گروه اسمی درون عبارت به‌عنوان کلمهٔ اصلی انتخاب شود. که در این حالت نتیجهٔ انتخاب کلمهٔ اصلی صحیح خواهد بود.

ابعاد هر یک از ویژگی‌ها به‌تفکیک نوع آن در جدول ۴ آورده شده است.

۳-۵- ترکیب دسته بندها

استفاده از نتایج چند دسته‌بند با عنوان یادگیری دسته‌جمعی، یک رویکرد مؤثر در یادگیری ماشینی است که در آن به‌منظور بهبود دقت یادگیری، نتایج دسته‌بندها با

[1] Category

[2] Position

یکدیگر ترکیب شده و یک سامانه مرکب شکل می‌گیرد. خروجی الگوریتم‌های مختلف دسته‌بندی در یکی از شکل‌های زیر بیان می‌شود[5].

الف- *خروجی تک طبقه*[1]: در این نوع، دسته‌بند فقط یک برچسب طبقه‌ای، برای الگوی ورودی ارائه می‌کند.

ب- *خروجی چند طبقهٔ مرتب‌شده*[2]: در این حالت، دسته‌بند تعلق الگوی ورودی به طبقه‌های مختلف را به‌صورت یک فهرست مرتب ارائه می‌کند. طبقه با بیش‌ترین تعلق در ابتدای فهرست و طبقه با کمترین تعلق در انتهای آن قرار دارد.

(جدول- ۴): تعداد ابعاد ویژگی‌های استخراجی
(Table-4): the number of extracted features dimensions

	Unigrams	W-S	W-W	L-B	H-W
ابعاد	9775	5	8	1010	1964
نوع	لغوی	لغوی	لغوی	نحوی	نحوی

ج- *خروجی چند طبقهٔ امتیازدار*[3]: در این شکل، دسته‌بند برای تعلق الگوی ورودی به هر کدام از طبقه‌ها یک مقدار عددی تخصیص می‌دهد. این مقدار عددی بیان‌گر میزان امتیازی است که با توجه به معیار دسته‌بندی، برای تعلق الگوی ورودی به آن طبقه محاسبه شده است.

متداول‌ترین تقسیم‌بندی برای روش‌های ترکیب خروجی دسته‌بندها، تقسیم‌بندی آنها بر حسب نوع خروجی الگوریتم دسته‌بندی است که بر اساس آن روش‌های ترکیب به سه دسته تقسیم می‌شوند. خروجی دسته‌بندهای مورد استفاده تک‌طبقه هستند به این جهت از روش‌های ترکیبی تک طبقه استفاده شده است، روش‌هایی چون رأی‌گیری[4]، فضای دانش رفتار[5]، بیز ساده[6]، کلیشه تصمیم[7] و دمپستر شفر[8] که نتایج هر یک در بخش بعد آورده شده است.

۴- پیاده‌سازی و ارزیابی

برای ابزارهای پردازش زبان طبیعی مورد استفاده (تجزیه‌گر، برچسب‌زن اجزای کلام و شناسایی‌کنندهٔ اسامی خاص) از ابزار دانشگاه استنفورد [3] ، [12] و [20]. و همچنین برای Gazetteer از ابزار موجود در مجموعهٔ Gate[9] استفاده شده است. برای ارزیابی عملیات دسته‌بندی پرسش‌ها از داده‌های استاندار UIUC استفاده شده است، این مجموعه شامل ۵۵۰۰ پرسش برچسب‌خورده به‌عنوان مجموعهٔ یادگیری و پانصد پرسش برچسب‌خورده به‌عنوان مجموعه آزمون است. با توجه به این‌که در این مجموعه داده از مجموعه پرسش‌های TREC استفاده شده است، به آن TREC نیز گفته می‌شود.

نحوهٔ توزیع سؤالات بخش یادگیری و آزمون در شکل (۲) آورده شده است. برای هر دسته‌بند ویژگی‌های متفاوتی استفاده شده است که این انتخاب براساس نتیجهٔ آنها صورت گرفته است.

(جدول- ۵): نتایج حاصل از دسته‌بندها برای ویژگی‌های متفاوت
(Table-5): Classifiers result on different features

	WH-H-WS-LB	H-WS-LB	H-WS	H	U
svm	92.2	90	72.6	68.3	82.2
src	88.3	87.5	68.6	66.4	91.2

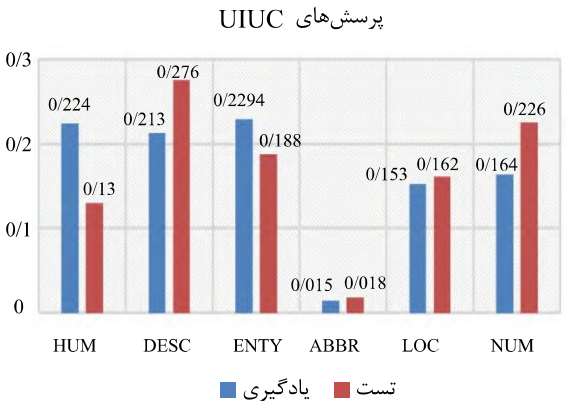
آزمایش بالا بر روی داده‌های آموزش با k-fold=10 اجرا شده است.

برای دسته‌بند مبتنی بر ماشین بردار پشتیبان بهترین ترکیب WH-H-WS-LB با ۹۲,۲ درصد و برای دسته‌بند مبتنی بر نمایش پراکنده ۹۱,۲ است، بنابراین از ترکیب دو دسته‌بند با ویژگی‌های یادشده عملیات دسته‌بندی را اعمال می‌کنیم.

نتایج حاصل از دسته‌بندها برای هریک از دسته‌ها در جدول (۶) آورده شده است.

با توجه اطلاعات جدول می‌توان نتیجه گرفت روش مبتنی بر نمایش پراکنده برای تمامی دسته‌ها عملکردی به‌طورتقریبی یکسان دارد؛ اما در روش مبتنی بر قانون اختلاف مابین صحت[10] و فراخوانی[11] بیش از روش مبتنی بر نمایش پراکنده است و این به‌علت نحوهٔ اعمال قوانین است، به عنوان مثال برای سؤالات با طبقه Enty، ایجاد قوانین برای این دسته به جهت نوع و ساختار پرسش‌ها به‌دلیل بروز استثناهای فراوان مشکل است. به همین جهت در ساختار و ترتیب قوانین مشخص‌کردن این دسته در انتها قرار گرفته است و در عدم تطابق پرسش با هیچ یک از قوانین، به این دسته تعلق می‌گیرد در نتیجه باید انتظار داشت فراخوانی مربوط به این دسته زیاد، اما صحت آن کم باشد که نتایج حاصله این موضوع را تأیید می‌کند.

[1] Abstract level
[2] Rank level
[3] Measurement level
[4] Voting
[5] Behavior knowledge space
[6] Naive Bayes
[7] Decision Template
[8] Dempster-Shafer Combination
[9] https://gate.ac.uk/
[10] precision
[11] recall

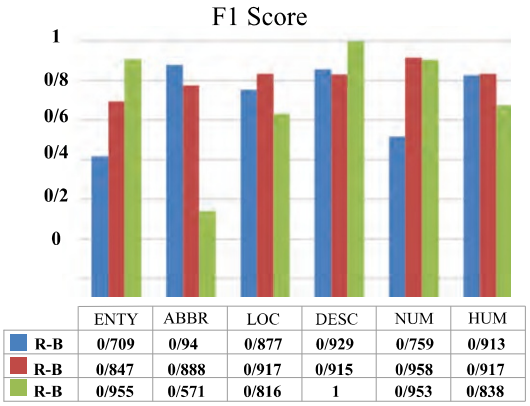


(شکل- ۲): نسبت برچسب پرسش‌های موجود در UIUC

(Figure-2): The UIUC questions tag ratio

همچنین دقت دسته‌بند مبتنی بر قانون برای بعضی از دسته‌ها بسیار خوب (مثل دستهٔ DESC) و برای بعضی دسته‌ها، نتیجهٔ قابل قبولی به همراه ندارد.(ABBR)

با توجه به نتایج دسته‌بندها برای دسته‌های متفاوت، می‌توان انتظار داشت، نتیجه حاصل از ترکیب این دسته‌بندها نتیجهٔ قابل قبول و بهتر از هر یک از دسته‌بندها به‌صورت جدا شود.



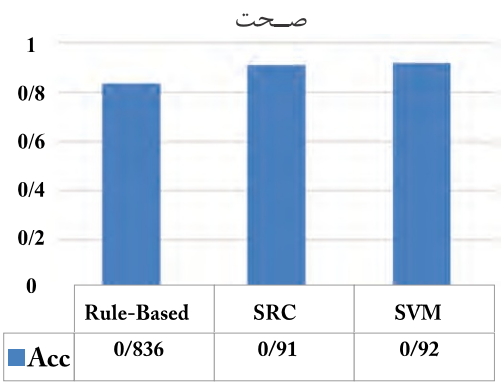
	ENTY	ABBR	LOC	DESC	NUM	HUM
R-B	0/709	0/94	0/877	0/929	0/759	0/913
R-B	0/847	0/888	0/917	0/915	0/958	0/917
R-B	0/955	0/571	0/816	1	0/953	0/838

(شکل- ۳): نتایج دسته‌بند مبتنی بر نمایش قانون، SVM و SRC

(Figure-3): The results from SRC, SVM and the Rule-based classifier

با توجه به نتایج شکل (۳)، امتیاز F1 مربوط به SR برای سه دستهٔ HUM، NUM، LOC بیشترین مقدار و دسته‌بند مبتنی بر قاعده‌ABBR و در انتها روش مبتنی بر SVM بهترین امتیاز را درDESC و ENTY دارد. در ساده‌ترین حالت، اگر دسته‌بندها به گونه‌ای ترکیب شوند که برای هر دسته از پرسش‌ها، با توجه به نتایج بیان شده، از دسته‌بندی استفاده شود که نتیجهٔ قابل قبول‌تری دارد، ترکیب سامانه‌ها ثمربخش خواهد بود.

با توجه به شکل (۴)، صحت[1] مربوط به هر یک از دسته‌ها در کل مجموعه داده‌ها آورده شده است، روش مبتنی بر ماشین بردار پشتیبان بالاترین امتیاز را دارد، اما همان‌طور که بیان شد در این دسته‌بند اختلاف مابین دسته‌های پرسش‌ها زیاد است و با اصلاح نتایج این دسته‌ها می‌توان به نتایج بهتری دست یافت.



	Rule-Based	SRC	SVM
Acc	0/836	0/91	0/92

(شکل- ۴): امتیاز صحت

(Figure-4): Accuracy Score

(جدول- ۶): نتایج حاصل از دسته‌بندها برای ویژگی‌های متفاوت

(Table-6): Classifiers results on different features

دسته	%ENTY		%ABBR		%LOC		%DESC		%NUM		
معیار	فراخوانی	دقت	فراخوانی	دقت	فراخوانی	دقت	فراخوانی	دقت	فراخوانی	دقت	دقت
Rule base	94	57	**100**	**90**	85	**90**	94	92	62	**98**	84
SRC	83	86	88	88	**94**	88	89	94	**100**	92	**93**
svm	**100**	**91**	50	66	80	82	**100**	**100**	**100**	91	92

[1] Accuracy

استفاده از ترکیب نتایج طبقه‌بند یکی از روش‌های افزایش کارایی سامانه‌های بازشناسی الگو است. برای مفید واقع‌شدن ترکیب طبقه‌بندها باید طبقه‌بندهای پایه ضمن برخورداری از کارایی قابل قبول، با یکدیگر متفاوت بوده و قاعدهٔ ترکیبی مناسبی برای تلفیق نتایج آنها به‌کار گرفته شود(Kabir, 1384). با توجه به نتایج جدول (۶)، می‌توان ادعا کرد هریک از دسته‌بندها به‌تنهایی نتیجهٔ قابل قبولی دارد، همچنین این دسته‌بندها در دسته‌های متفاوت، دقت‌های متفاوتی دارند. در نتیجه دو شرط را دارند و باید به دنبال قاعدهٔ ترکیبی مناسب بود.

در روش‌های دسته‌بندی، به‌الزام افزایش ویژگی‌ها با افزایش دقت همراه نیست و گاهی مواردی پیش می‌آید که با افزایش تعداد ویژگی‌ها، دقت به‌شدت کاهش می‌یابد. در دسته‌بند SRC با اضافه‌کردن بعضی ویژگی دقت به‌شدت کاهش می‌یابد. این مورد در ویژگی UNIGRAMS به‌وضوح مشخص است به جهت استفاده از ویژگی‌های مناسب، جهت دست‌یافتن به دقت بیشتر، هر مجموعه از ویژگی‌ها با دسته‌بندهای متفاوت مورد استفاده قرار گرفته است تا ضمن دست‌یابی به دقت‌های بالا در دسته‌های متفاوت بتوان با یک روش ترکیبی مناسب دقت کلی را افزایش داد.

همان‌طور که اشاره شد، برچسب نهایی هر پرسش از ترکیب خروجی سه دسته‌بند ایجاد می‌شود، برای این منظور با توجه به این که دسته‌بندهای مورد استفاده از نوع تک طبقه است. نتایج هر یک از دسته‌بندها را توسط روش‌های ترکیبی تک‌طبقه ترکیب می‌کنیم، روش‌های رأی‌گیری وزن‌دار، فضای دانش رفتار، بیز ساده، کلیشهٔ تصمیم و دمپستر شفر.

در روش رأی‌گیری وزن‌دار، نیاز است برای هر دسته وزنی تعلق گیرد، وزنی به‌عنوان ضریب اطمینان. در این جا برای هر دسته‌بند، وزن تعلق‌گرفته یک، بردار شش‌تایی است که همان صحت هر دسته‌بند برای شش دستهٔ موجود است.

در روش فضای دانش رفتار با ثبت نظرات دسته‌بندها در مورد الگوهایی که طبقه آنها معلوم است، رفتار جمعی دسته‌بندها را مدل کرده و بر اساس این مدل برچسب طبقه‌ای الگوی ناشناخته x را مشخص می‌کند. به‌عبارت دیگر در این روش، تصمیم‌گیری نهایی برای طبقه یک الگو با استفاده از رفتاری که دسته‌بندی‌های مختلف در هنگام یادگیری الگوها از خود نشان داده‌اند صورت می‌گیرد. در این روش باید تمامی حالات ممکن در نظر گرفته شود. در اینجا حالات ممکن ۶×۶×۶ می‌باشد و برای هر یک از حالات در صورت رخداد در مرحلهٔ یادگیری برچسبی در نظر گرفته می‌شود که این برچسب براساس برچسب واقعی پرسش‌ها می‌باشد[1].

در روش بیز ساده ابتدا با توجه به داده‌های آموزشی برای هر طبقه بند D_j، یک ماتریس سردرگمی، CM^j تشکیل می‌شود. عناصر این ماتریس به صورت $CM^j(k, s)$ هستند.

مقدار درایه $CM^j(k, s)$ بیان‌گر تعداد الگوهایی است که برچسب واقعی آنها k است، یعنی متعلق به طبقه ω_k هستند، ولی طبقه‌بند D_j آن‌ها را به طبقه ω_s نسبت داده است. با جمع‌کردن مقادیر عناصر ستون sام ماتریس سردرگمی طبقه‌بند D_j، می‌توان $CM^j(s)$ یعنی تعداد کل عناصری که این طبقه‌بند به طبقه ω_s نسبت داده است را مشخص کرد. در ادامه برچسب طبقه‌ای الگوی ناشناخته x را براساس احتمال رخداد هر یک از کلاس‌ها برای داده x حساب می‌شود. طبقه با بیشترین احتمال به‌عنوان برچسب نهایی انتخاب می‌شود[1].

در روش کلیشه تصمیم در ابتدا با استفاده از نمونه‌های آموزشی $\{x_1,\ldots,x_n\}$ که برچسب‌های طبقه‌ای آنها مشخص است، ماتریس $DP(x)$ ایجاد می‌شود که شامل خروجی طبقه‌بندهای $D_1(x)$، $D_2(x)$ و $D_3(x)$ است. در ادامه ماتریس‌هایی که مربوط به داده‌های با برچسب ω_k با k=1…6 را جدا کرده و با متوسط‌گیری از درایه‌های نظیر به نظیر ماتریس‌ها، ماتریس DT_k ایجاد می‌شود. در ادامه برای تعیین برچسب هر داده شباهت DP آنرا با تک‌تک DT ها محاسبه کرده و DT که بیشترین شباهت را داراست به‌عنوان برچسب نهایی انتخاب می‌شود [9].

در روش تئوری دمپستر شفر برای ترکیب طبقه‌بندها، ابتدا با توجه به رفتار طبقه‌بندها شواهدی از آنها استخراج می‌شود که به‌عنوان دانش اولیه در پیداکردن درجه عضویت الگو به هر یک از طبقه‌ها مورد استفاده قرار می‌گیرد. برای یک سامانه مرکب شامل L طبقه بند، L بردار مبنا برای هر طبقه به‌دست می‌آید که مجموعه این بردارها کلیشه تصمیم آن طبقه را می‌سازند. در مرحله آزمایش با توجه به شباهت ماتریس پروفایل تصمیم با کلیشه تصمیم هر طبقه ، شواهد مورد نیاز برای روش دمپستر شفر به‌دست آمده و میزان باور و سرانجام درجه عضویت هر طبقه محاسبه می‌شود. طبقه با بیشترین درجه عضویت، طبقه الگوی ورودی خواهد بود[1].

مواقعی که دسته‌بند مبتنی بر ماشین بردار پشتیبان نتیجهٔ قابل قبولی ندارد، سایر دسته‌بندها نتیجهٔ بهتری دارند.

روش‌های مورد استفاده در این مقاله برای ترکیب دسته‌بندها، تعیین برچسب نهایی براساس رفتار گذشتهٔ دسته‌بندها است. در واقع وضعیت کنونی داده در دسته‌بندها در نظر گرفته نمی‌شود. می‌توان با استفاده از دسته‌بندهای احتمالی مثل بیز و یا تبدیل دسته‌بند مبتنی بر ماشین بردار پشتیبان به‌صورت احتمالی بدین صورت که برای هر داده براساس امتیاز داده‌شده برای تعلق به هر طبقه وضعیت کنونی داده را در برچسب نهایی دخیل داد؛ علاوه بر این برای دسته‌بندهای مورد استفاده در این مقاله از ویژگی‌های لغوی و نحوی استفاده شده است که می‌توان با استفاده از ویژگی‌های معنایی مثل زیرشمول‌ها یا مترادفات مربوط به کلمه اصلی دقت عملیات دسته‌بندی را افزایش داد. همچنین در این مقاله برای دسته‌بندی مبتنی بر یادگیری از کرنل خطی استفاده شده است. با توجه به کارهای انجام‌شده، به‌ویژه برای دسته‌بندی مبتنی بر نمایش پراکنده و نتایج آنها، با استفاده از کرنل‌های دیگر مثل گوسی می‌تواند به نتایج بهتری دست یافت.

۶- مراجع 6-Reference

نبوی کریزی، سیدحسن.، کبیر، احسان اله.، "ترکیب طبقه بندها :ایجاد گوناگونی و قواعد ترکیب"، نشریه علمی پژوهشی انجمن کامپیوتر ایران مجلد ۳، شماره ۳ (الف)، صفحه ۹۵. پائیز ۱۳۸۴.

[1] S.H. Nabavi karizi, & E. Kabir, “Combining classifiers: Diversifying and rules of composition.” CSI Journal on Computer Science and Engineering, vol. 3, pp. 95. Autumn 1384

[2] M. Collins, “Head-driven statistical models for natural language parsing.” Computational linguistics, vol. 29, pp. 589-637, Dec 2003.

[3] J.R. Finkel, T. Grenager, and C. Manning, “Incorporating non-local information into information extraction systems by gibbs sampling,” In Association for Computational Linguistics: Proceedings of the 43rd annual meeting on association for computational linguistics, Michigan: ACL, 2005. pp. 363-370.

[4] Ulf. Hermjakob, “Parsing and question classification for question answering,” In Association for Computational Linguistics: Proceedings of the workshop on Open-domain question answering, Vol. 12, France: ACL, 2001, pp. 1-6.

[5] T.K. Ho, J.J. Hull, and S.N. Srihari, “Combination of decisions by multiple classifiers.” In Structured document image analysis, Springer Berlin Heidelberg 1992, pp. 188-202.

[6] Hoque, M. Moinul, T. Goncalves, and P. Quaresma, “Classifying Questions in Question Answering System Using Finite State Machines with a Simple Learning Approach.” In *PACLIC* Vol. *27*, Taiwan 2013 pp. 409-414.

[7] Z. Huang, M. Thint, and Z. Qin, “Question classification using head words and their hypernyms.” In Association for Computational Linguistics: Proceedings Conference on Empirical Methods in Natural Language Processing, Hawaii USA: ACL 2008, pp. 927-936.

[8] V. Krishnan, Das. Sujatha, and S. Chakrabarti, “Enhanced answer type inference from questions using sequential models.” In Association for Computational Linguistics: Proceedings Conference on Empirical Methods in Natural Language Processing, Vancouver Canada: ACL 2005, pp. 315-322.

[9] Kuncheva, I. Ludmila, Combining pattern classifiers: methods and algorithms. John Wiley & Sons, 2004.

[10] Li, Xin, and D. Roth, “Learning question classifiers.” In Association for Computational Linguistics: COLING '02 Proceedings of the 19th international conference on Computational linguistics, Taipei, Taiwan: COLING 2002, pp. 1-7.

[11] B. Loni, S. H. Khoshnevis, and P. Wiggers, “Latent semantic analysis for question classification with neural networks.” In Automatic Speech Recognition and Understanding, Hilton Waikoloa: ASRU, 2011, pp. 437-447.

[12] M. C. Marneffe, B. MacCartney, and D. Christopher, “Generating typed dependency parses from phrase structure parses.” In Conferences bring together a large number of people working and interested in HLT Proceedings: LREC, Genoa, Italy: May 2006, pp. 449-454.

[13] A. Merkel, and D. Klakow, “Improved methods for language model based question classification." In INTERSPEECH, Antwerp, Belgium :August 2007, pp. 322-325.

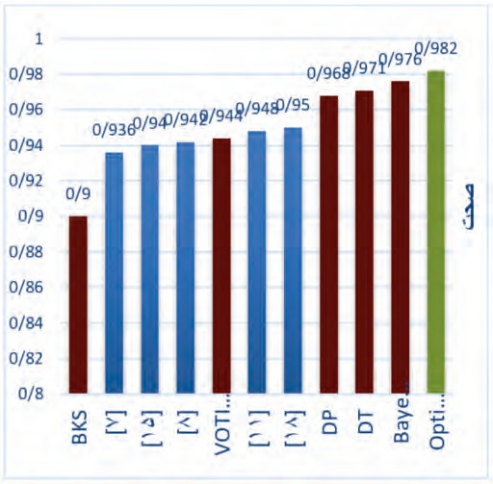
[14] D. Moldovan, M. Paşca, S. Harabagiu, & M. Surdeanu, “Performance issues and error analysis in an open-domain question answering system.” ACM Transactions on Information Systems, vol. 21, pp. 133-154, October 2003.

[15] Y. Pan, Y. Tang, L. Lin, & Y. Luo, “Question classification with semantic tree kernel.” In ACM SIGIR conference on Research and development in information retrieval: Proceedings of the 31st annual international, Singapore, Singapore, July 2008, pp. 837-838.

[16] D. Radev, W. Fan, H. Qi, H. Wu, & A.Grewal, “Probabilistic question answering on the

با توجه به نتایج حاصل از شکل ۵ نتایج حاصل از ترکیب دسته‌بندها به روش بیز ساده بهترین نتیجه را به همراه دارد. همچنین در طبقه‌بند فرضی بهینه دقت کسب‌شده ۰,۹۸۲ است که نشان‌دهندهٔ این است که در بهترین روش ترکیبی مورد استفاده که بیز ساده است، به اندازهٔ ۰,۰۰۶ خطا وجود دارد؛ چون‌که طبقه‌بند فرضی بهینه تنها زمانی اشتباه می‌کند که همه طبقه‌بندهای پایه اشتباه کنند. پس این اختلاف بیان‌کنندهٔ این نکته است که حداقل یکی از دسته‌بندها عملیات برچسب‌گذاری را به‌درستی انجام داده است؛ اما در ترکیب نتایج سامانه دچار اشکال شده است.

در روش ترکیبی فضای دانش رفتار علت پایین‌بودن نتیجه وجود حالت‌های بدون نتیجه است. حالت‌هایی که در مرحلهٔ آزمون رخ می‌دهند، اما در مرحلهٔ یادگیری همچنین حالتی رخ نمی‌دهد، رسیدن به نتیجه مطلوب با این روش نیازمند داده‌های آموزشی کافی یا وجود حالت‌های محدود است.



(شکل-۵):امتیاز نهایی سیستم و سیستم‌های موجود

(Figure-5): The final score from the proposed method and the other existing approaches

روش ترکیب وزن‌دار تنها به وضعیت همان داده وابسته است. چالش اصلی این روش ترکیبی در حالتی است که تعداد بیشتری از دسته‌بندها به‌اشتباه یک طبقه را انتخاب کنند که سبب اشتباه در برچسب‌گذاری نهایی می‌شود.

درستی روش‌های کلیشهٔ تصمیم و تئوری دمپستر شفر وابسته به وضعیت داده‌های مرحلهٔ آزمون است و برای طبقه‌های با تعداد دادهٔ کمتر احتمال بروز خطا بیشتر است. علاوه بر این در این دو روش در حالت‌هایی که هر یک از سه دسته‌بند، داده را با طبقه‌ای متفاوت برچسب‌گذاری کند میزان شباهت بردار داده برای هر یک از طبقه‌ها مشابه هم بوده و احتمال بروز خطا در انتخاب برچسب نهایی افزایش می‌یابد.

در روش ترکیبی ارائه‌شده، برچسب نهایی براساس رفتار گذشتهٔ دسته‌بندها مشخص می‌شود و وضعیت داده در دسته‌بندها در نظر گرفته نمی‌شود؛ این درحالی است که ممکن است برچسب‌دهی با اشتباه همراه باشد. به‌عنوان مثال برچسب تعلق گرفته برای پرسش ۶ توسط دسته‌بند مبتنی بر قانون و مبتنی بر نمایش پراکنده DESC و برچسب تعلق گرفته توسط دسته‌بند مبتنی بر ماشین بردار پشتیبان LOC است. در روش‌های ترکیبی برچسب تعلق گرفته برای این پرسش DESC است که این برچسب به اشتباه داده شده است. در SVM متلب برچسب داده‌شده توسط تابع نشانه[1] در آخرین مرحله مشخص می‌شود. درواقع برای مقادیر کم و زیاد به یک صورت برخورد می‌کند و یک برچسب تعیین می‌شود. این در صورتی است که برای داده‌های با مقدار زیاد می‌توان با اطمینان بیشتری برخورد کرد. در این مثال با توجه به اینکه مقدار داده‌شده توسط SVM مقداری زیادی است پس می‌توان به برچسب داده‌شده توسط SVM اطمینان بیشتری داشت.

۵- نتیجه‌گیری و کارهای آینده

هدف از سامانه‌های پرسش‌وپاسخ، پاسخ‌دهی خودکار به پرسش‌های زبان طبیعی است که توسط انسان پرسیده می‌شود. در این سامانه‌ها دسته‌بندی پرسش‌ها نقشی مهم در انتخاب پاسخ صحیح برای پرسش داده شده دارد. در این مقاله یک روش ترکیبی جدید برای دسته‌بندی پرسش ارائه شده است. در روش ترکیبی از یک دسته‌بند مبتنی بر قانون و دو دسته‌بند مبتنی بر یادگیری استفاده شده است. برای دسته‌بند مبتنی بر قانون از یک مجموعه قانون در قالب عبارات منظم استفاده شده است؛ و برای دسته‌بندهای مبتنی بر یادگیری عملیات دسته‌بندی براساس ویژگی‌های لغوی و نحوی پرسش‌ها انجام می‌گیرد. برای ترکیب نتایج دسته‌بندی از روش‌های رأی‌گیری، بیز ساده و فضای دانش رفتار استفاده شده است که با روش ترکیبی بیز ساده بهترین نتیجه حاصل شد. کسب نتیجهٔ خوب در روش ترکیبی، به‌دلیل رفتار متفاوت دسته‌بندها در روش‌های مختلف است. یعنی در

[1] sign

web," Journal of the American Society for Information Science and Technology, vol. 56, pp. 571-583, DEC 2005.

[17] S. K. Ray, S. Singh, & B. P. Joshi, "A semantic approach for question classification using WordNet and Wikipedia." Pattern Recognition Letters, vol. 31, pp. 1935-1943, October 2010.

[18] J. Silva, L.Coheur, A. C. Mendes, & A. Wichert, "From symbolic to sub-symbolic information in question classification." Artificial Intelligence Review, vol. 35, pp. 137-154, February 2011.

[19] H. Sundblad, "Question classification in question answering systems". Phd dissertation,. Institutionen för datavetenskap, 2007.

[20] K. Toutanova, D. Klein, C.D. Manning, & Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network". In Association for Computational Linguistics on Human Language Technology: Proceedings of the Conference of the North American Chapter, Washington, DC, USA, October 2003, pp. 173-180.

[21] J. Wright, et al. "Robust face recognition via sparse representation." IEEE transactions on pattern analysis and machine intelligence, vol 31, pp. 210-227, Feb 2009.

هادی قائمی مدرک کارشناسی ارشد خود را در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه فردوسی مشهد اخذ کرده است. زمینه پژوهشی مورد علاقه ایشان شامل پردازش زبان طبیعی و پردازش الگو است.

نشانی رایانامهٔ ایشان عبارت است از:

Hadi.qaemi@stu.um.ac.ir

محسن کاهانی استاد گروه مهندسی کامپیوتر دانشگاه فردوسی مشهد و مدیر آزمایشگاه فناوری وب است. ایشان دکترای خود را در رشته مهندسی کامپیوتر از دانشگاه ولونگونگ استرالیا در سال ۱۳۷۷ اخذ کرده است. زمینه پژوهشی مورد علاقه ایشان شامل وب معنایی، پردازش زبان طبیعی، سیستم‌های تصمیم یار و مهندسی نرم‌افزار است.

نشانی رایانامهٔ ایشان عبارت است از:

kahani@um.ac.ir