عنوان زیرپروژه:

امکان سنجی تولید سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری

فهرست مطالب

عنوان **شماره صفحه**

1. مقدمه

در متون تفسیری فارسی که گاه ترجمه‌ای فارسی از متن عربی‌اند و گاه از ابتدا به فارسی نگارش شده‌اند بطور معمول هدف تفسیر آیات قرآن کریم به ترتیب ذیل دنبال می‌شود. ابتدا آیه مورد نظر مطرح می شود؛ سپس آیه بطور یکجا و یا قطعه قطعه به استناد آیات قرآنی دیگر و یا احادیثی از ائمه اطهار (س) و روایات دیگر تفسیر می‌شوند، که بطور معمول متن عربی حدیث نیز توضیح فارسی آنرا همراهی می‌کند. بنابراین متون تفسیری فارسی در داخل متون فارسی، که بخش عمده تفسیر را تشکیل می‌دهند، در کنار آیات قرآن کریم، حاوی متون حدیثی و روایتی است که آنها نیز مانند قرآن کریم به زبان عربی‌اند. بنابراین مسئله پژوهش حاضر را می‌توان به دو زیر مسئله تبدیل کرد: تشخیص بخش‌های عربی در متون فارسی تفسیری، و سپس تشخیص آیات قرآن کریم از سایر متون عربی. بدین ترتیب مسئله اول به عنوان یک مسئله تشخیص زبان[۱] و مسئله دوم به عنوان یک مسئله تشخیص منبع[۲] از دیدگاه فنون رده بندی متن[۳] که از زیر شاخه‌های متن کاوی[۴] قابل بررسی‌اند. از آنجا که مسئله تشخیص زبان در پژوهشی دیگر تحت عنوان *مطالعه و بررسی روش‌های تشخیص هوشمند جمله فارسی از عربی در پیکره‌های مخلوط فارسی و عربی* توسط پژوهشگر حاضر تحت بررسی است، با فرض موفقیت سامانه فوق در تفکیک بخش‌های عربی از فارسی، پژوهش حاضر در مسئله دوم متمرکز شده و در آن تلاش بر تفکیک آیات قرآن کریم از سایر متون عربی خواهد بود.

برای دستیابی به راه‌کاری مناسب برای حل این مسئله طرح پژوهشی حاضر حصول به پاسخ سؤالات ذیل را سر لوحه پژوهش امکان سنجی خود قرار داده است تا روش‌های تشخیص هوشمند آیات قرآن کریم از سایر متون عربی در متون تفسیری فارسی تبیین شده و زمینه را برای تولید سامانه‌های هوشمند مربوطه فراهم کند.

1. کیفیت و کمیت پیکره‌های متنی فارسی تفسیری قرآن کریم برای تولید سامانه هوشمند تشخیص آیات قرآن در این متون چگونه باید باشد؟

---

[۱] Language Detection

[۲] Source Verification

[۳] Text Classification

[۴] Text Mining

2. چه نرم افزارهای برای پیش پردازش پیکره‌های متنی فارسی تفسیری قرآن کریم مورد نیاز است؟

3. آیا پایگاه دانش سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری مبتنی بر قواعد پیاده سازی شده انسانی خواهد بود یا آموخته‌های ماشین از پیکره‌های متنی فارسی تفسیری قرآن کریم؟

4. آیا سامانه‌ای مشابه سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری تا کنون بررسی یا تولید شده است؟

5. چه سامانه‌هایی مکمل سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری است؟

6. سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری چه معماری باید داشته باشد؟

7. اهمیت و کاربرد سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری چیست؟

8. تولید سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری با چه چالشهایی مواجه است؟

سؤالات فوق پژوهش حاضر را از بررسی‌های ذیل به پاسخهای لازم سوق داده است.

1. بررسی سامانه‌هایی که می‌توانند در داخل متنی که تمام آن با کاراکترهای یکسان نگارش شده، رشته‌ای را که به زبانی به غیر از زبان کلی متن است تشخیص دهند،

2. بررسی معماری سامانه‌های طبقه بندی متن،

3. بررسی کیفیت و کمیت پیکره‌های متنی فارسی تفسیری قرآن کریم برای پردازش،

4. بررسی نحوه جدا سازی اولیهٔ بخش‌های عربی از متون فارسی تفسیری و سپس نحوه تفکیک آیات قرآن کریم از دیگر بخش‌های عربی،

5. بررسی پژوهش‌ها و سامانه‌هایی که با هدف پردازش و طبقه بندی متون عربی وجود دارند،

6. بررسی پژوهش‌ها و سامانه‌هایی که با هدف پردازش و طبقه بندی آیات قرآن کریم وجود دارند،

7. بررسی پژوهش‌ها و سامانه‌هایی که با هدف پردازش و طبقه بندی متون فارسی وجود دارند،

8. بررسی پژوهش‌ا و سامانه‌هایی که با هدف پردازش و طبقه بندی متون فارسی تفسیری وجود دارند.

در گزارش پژوهشی حاضر ابتدا میزان و چگونگی استفاده از آیات قرآن کریم در متون تفسیری فارسی بررسی شده، آنگاه با تبیین مشکل تشخیص هوشمند آیات فوق در این متون از لحاظ علمی، شاخه علمی که برای حل این مسئله از آن مدد جسته شده است، یعنی رده‌بندی متون، توضیح داده می‌شود. در این راستا به الگوریتم Benedetteo, Caglioti & Loreto اشاره می‌شود که می‌تواند با تکیه بر شیوه استفاده الگوریتم‌های فشرده سازی از مفهوم آنتروپی در نظریه اطلاعات رشته مورد نظر را حتی در طول بسیار کوچک رده‌بندی نماید. در پایان شیوه‌ای خاص که توسط شکرالهی و همکارانش برای تشخیص آیات قرآنی از یکدیگر مورد استفاده قرار می‌گیرد تشریح خواهد شد.

2. مروری بر کیفیت و کمیت آیات قرآن کریم در تفاسیر فارسی

در بررسی تفاسیر فارسی در خصوص کیفیت و کمیت استفاده از آیات قرآن در آنها، بخشی از ترجمه فارسی تفسیر مجمع البیان در ذیل ارائه می‌شود.

«تِلْکَ آیاتُ الْکِتابِ الْحَکِیمِ» یعنی آیه‌هایی که ذکر آن در پیش شده یا آیه‌هایی که بر محمد نازل گردیده آنها آیات کریم قرآنی است که از هر باطلی محفوظ و از هر فساد و تباهی ممنوع است. اختلاف و دروغی در آن نیست.

و برخی گفته‌اند: «تلک» اشاره بسوره‌های قرآن است. یعنی این سوره‌ها آیه‌های کتاب حکیم یعنی لوح محفوظ است، و اینکه آن را محکم نامیده چون گویای بحکمت است، و یا برای آنکه علوم و حکمت را جمع کرده. و قولی است که علت توصیف کتاب به «حکیم» آن است که کتاب دلیل و راهنمای بر حق است مانند کسی که گویای بحق باشد، و هم برای آنکه بانسان معرفت و بینشی می‌دهد که بدانوسیله راه هلاکت را از طریق نجات و رستگاری تمیز می‌دهد.

«أَ کانَ لِلنَّاسِ عَجَباً أَنْ أَوْحَیْنا إِلی رَجُلٍ مِنْهُمْ أَنْ أَنْذِرِ النَّاسَ» جمله بصورت استفهام است اما منظور انکار است (و باصطلاح استفهام انکاری است) و چنانچه بعضی گفته‌اند:

منظور از «ناس»- مردم- نیز مردم مکه هستند، یعنی ما در شگفتیم که آیا خدای سبحان جز یتیم ابی طالب کسی را پیدا نکرد که وی را بعنوان پیغمبری برای مردم بفرستد. و روی این ترتیب معنای آیه این می‌شود که: آیا اینکه ما به یکی از این مردم وحی کردیم تا مردم را بترساند موجب تعجب و شگفت است! یعنی برای چه تعجب می‌کنند و چرا باید تعجب کنند که ما بمردی از خودشان وحی کنیم! با

اینکه اینجا جای تعجب نیست، و پیش نظر هر عاقلی چنین برنامه‌ای لازم و واجب بود، زیرا وقتی خدای- تعالی عقل بندگانش را کامل فرمود، و معرفت خویش و ادای شکرش را بر آنان واجب ساخت، و این هم معلوم بود که آنها باین راه نیایند و اصلاح نگردند جز بوسیله شخصی که از نزدیک بیاید و آنها را بخدا دعوت کند، و کسی باشد که آنها را بیاد خدا انداخته و متنبهشان سازد، در این صورت از روی حکمت انجام چنین کاری بر خداوند لازم بود.

سپس بدنبال این مطلب خداوند آن وجهی را که بخاطر آن پیغمبر را فرستاد و آنچه را بدو وحی کرده ذکر فرموده با این جمله که گوید: ما بدو وحی کردیم که مردم را بعذاب آگاه کن، و بدان بیمشان ده.

«وَ بَشِّرِ الَّذِينَ آمَنُوا أَنَّ لَهُمْ قَدَمَ صِدْقٍ عِنْدَ رَبِّهِمْ» و وسیله شرافت و خلود در نعمت- های بهشتی را برای آنها معرفی کن، تا ضمناً ارزش و ارج اعمال صالح و کارهای نیک هم معلوم گردد. و برخی گفته‌اند: منظور از «قدم صدق» پاداش نیک و مقام ارجمندی است که در نتیجه پیش فرستادن اعمال نیک بدان داده شود. و یا منظور سعادتی است که قبلا برای آنها سبق ذکر یافته. و این هر دو وجه از ابن عباس نقل شده. و مؤید این قول دوم است آیه شریفه: «إِنَّ الَّذِينَ سَبَقَتْ لَهُمْ مِنَّا الْحُسْنى ...» «1».

و قول دیگر آن است که معنای «قَدَمَ صِدْقٍ» پیش انداختن خداوند آنها را در روز قیامت میباشد، چنانچه در آن حدیث نیز فرموده است: «نحن الآخرون السابقون یوم القیامة»- مائیم آن امتی که از نظر زمان آخرین امت هستیم ولی در روز قیامت بر دیگران سبقت جوییم- و برخی گویند: «قَدَمَ صِدْقٍ» یعنی شفاعت محمد- صلی اللّه علیه و آله- در روز قیامت. و این وجهی است که ابو سعید خدری گفته و از امام صادق علیه السلام نیز روایت شده.

«قالَ الْكافِرُونَ إِنَّ هذا لَساحِرٌ مُبِينٌ» کافران گویند: این پیغمبر ساحری است که سحرش آشکار است. و این سخن دلیل بر عجز و ناتوانی آنها از معارضه با قرآن کریم بوده که به این گفتار متوسل شده‌اند، و آن را بسحر و جادو منسوب داشته‌اند.

همانطور که در این نمونه دیده می شود درصد کمی از متن به آیات قرآنی اختصاص دارد و بدنه اصلی متن فارسی است، که این مطلب در مورد متون تفسیری قرآن کریم صادق است. نکته دیگر اینکه آیات قرآن بکار رفته در متن دارای حرکه هستند که این امر براحتی آنها را از بخشهای اصلی متن جدا می-کند؛ ولی این امر همیشه صادق نیست. در همین خصوص نمونه متنی که از همان منبع در پی می‌آید قابل توجه است.

این بود خلاصه آنچه مفسرین در این باره گفته‌اند، لیکن سیاق آیات و نیز روال قرآن در بیاناتش با این گفتار سازگار نیست. توضیح اینکه اگر کلمه" ارحام" را صله‌ای مستقل برای موصول" الذی" بگیریم تقدیر کلام چنین می‌شود:" و اتقوا اللّه الذی تسائلون بالارحام- بترسید از خدایی که یکدیگر را به رحمها سوگند می‌دهید" و معلوم است که این عبارت ناتمام است، چون صله نامبرده خالی از ضمیر است و این جایز نیست.

(بله اگر از این صله ضمیری به موصول بر می گشت مثلا می گفتیم:" و اتقوا اللّه الذی جعل بینکم و بین ارحامکم مودة فتساءلون بهم" «1» آن وقت می‌توانستیم کلمه" و الارحام" را صله مستقلی بگیریم" مترجم").

برخلاف نمونه متن تفسیری قبلی، در این نمونه متن آیات قرآنی بکار رفته دارای حرکه نیستند. اینگونه آیات بدون حرکه شباهت بسیاری به احادیث ائمه معصومین (ع) بکار رفته در چنین تفاسیری دارند که امر تشخیص آیات قرآن کریم را بسیار مشکل‌تر می‌کند. متن تفسیری که در پی می‌آید از این گونه است.

امام صادق علیه السلام فرمود:

«اصبروا علی الفرائض» در برابر واجبات صبر کنید.

«صابروا علی المصائب» در برابر مشکلات صبر کنید.

«و رابطوا علی الأئمه» از پیشوایان خود دفاع کنید. «1»

رسول خدا صلی اللّه علیه و آله فرمودند: «اصبروا علی الصلوات الخمس و صابروا علی قتال عدوکم بالسیف و رابطوا فی سبیل الله لعلکم تفلحون» بر نمازهای شبانه روزی پایداری کنید و در جهاد با دشمن، فعّال و در راه خدا با یکدیگر هماهنگ باشید تا رستگار شوید.

از همین قبیل‌اند اشعار عربی که در چنین متونی بکار رفته‌اند، مانند متن تفسیری ذیل:

آیة: نشانه‌ای است که از جهت مخصوصی بمقطع سخن آگاهی دهد.

حکیم: در اینجا بمعنای محکم است. چنانچه اعشی در شعر خود گوید:

و غریبة تأتی الملوک حکیمة قد قلتها لیقال من ذا قالها «1»

و برخی گفته‌اند: حکیم بمعنای حاکم (و داور) است بدلیل آنکه خداوند در جای دیگر درباره قرآن فرمود: «لِيَحْكُمَ بَيْنَ النَّاسِ فِيمَا اخْتَلَفُوا فِيهِ».

قدم: بگفته ازهری چیزی است که انسان آن را پیش از خود می‌فرستد که برای وی ذخیره‌ای باشد. و ابن اعرابی گفته: قدم بمعنای متقدم در شرف است. و ابوعبیده و کسایی گفته‌اند: هر کس در کار خیر و یا کار شری پیشقدم شود عرب او را «قدم» گویند، چنانچه گویند: فلانی را در اسلام قدمی است. یعنی تقدم دارد.

بنابراین اگر برخی نشانه گزاری‌هایی که به صورت دستی در نمونه متن‌های فوق بکار رفته است در نظر گرفته نشوند، آیات قرآن کریم در متون تفسیری فارسی در عمده موارد هیچ مشخصه ظاهری تمییز دهنده ندارند.

از لحاظ طول آیات قرآنی بکار رفته نیز در نمونه ذیل از تفسیر نور نکات قابل توجهی وجود دارد.

«قَدَمَ صِدْقٍ» چند معنی دارد:

1 سابقه‌ی خوب. مثل «قدمٌ فی الحرب، قدم فی الاسلام»، یعنی در مبارزه و در اسلام سابقه دارد.

2 مقام و منزلتِ صدق و نیکو.

3 رهبر و پیشوای صدق. در روایات شیعه و سنّی «قَدَمَ صِدْقٍ» را رسول اللّه صلی اللّه علیه و آله و علی علیه السلام دانسته‌اند. «1»

4 شفاعت. در روایتی مراد از «قَدَمَ صِدْقٍ» مقام شفاعت معرّفی شده است. «2»

کافران در ردّ رسالت پیامبر هیچ گونه دلیل و منطقی نداشتند، بلکه با بعید دانستن آن، از پذیرش آن شانه خالی می‌کردند. چنان که اصول اعتقادی را نیز فقط با تعجّب انکار می‌کردند:

در توحید: «أَ جَعَلَ الْآلِهَةَ إِلهاً واحِداً» «3» در نبوّت: «أَ هذَا الَّذِی بَعَثَ اللَّهُ رَسُولًا» «4»

در معاد: «مَنْ یُحْیِ الْعِظامَ وَ هِیَ رَمِیمٌ» «5» در امامت: «أَنَّی یَکُونُ لَهُ الْمُلْکُ عَلَیْنا وَ نَحْنُ أَحَقُّ بِالْمُلْکِ مِنْهُ وَ لَمْ یُؤْتَ سَعَةً مِنَ الْمالِ» «6»

در این نمونه آیه مورد نظر هم بطور کامل بکار رفته است، و هم در حد عبارت و یا حتی کلمه تنها شکسته شده است. بنابراین در متون فارسی تفسیری، اول اینکه، نه تنها تمامی بخش‌های فارسی، بلکه بسیاری از آیات قرآنی مخلوط شده نیز فاقد حرکه می‌باشند؛ دوم اینکه، این آیات در کنار بخش‌های عربی دیگر نظیر احادیث و اشعار قرار گرفته‌اند؛ و مطلب آخر اینکه، این آیات در طول‌های مختلف، از جمله طولانی گرفته تا یک کلمه تنها، در داخل متون فارسی مخلوط شده‌اند.

3. رده‌بندی و تشخیص منبع متن

مسئلهٔ تشخیص منبع یک متن در کنار دیگر متونی که از منابعی دیگرند ولی هم متن مورد نظر و هم دیگر متون به زبانی یکسان نگاشته شده‌اند حوزه‌ای است که از منظر رده‌بندی متن نگریسته می‌شود. بطور کلی رده‌بندی متن، که انتساب اسناد متنی بر اساس محتوی به یک یا چند طبقه از قبل تعیین شده است، یکی از مهمترین مسایل در متن کاوی است. مرتب کردن بلادرنگ نامه‌های الکترونیکی یا فایل‌ها در سلسله مراتبی از پوشه‌ها، تشخیص موضوع متن، جستجوی ساخت یافته و یا پیدا کردن اسنادی که در راستای علایق کاربر می‌باشد، از جمله کاربردهای مبحث رده بندی (طبقه‌بندی، دسته-بندی یا کلاس‌بندی) متن است که در آن سه مرحلهٔ پیش-پردازش[۱]، آموزش رده بند[۲] و رده‌بندی[۳] وجود دارد.

در مرحله پیش-پردازش, دانش موجود در هر متن باید بازنمایی[٤] شود تا قابل استفاده نرم افزارهای رده بندی گردد. این بازنمایی به شکل مدل برداری[٥] از ویژگی‌های[٦] متن که برگرفته از عناصر موجود در آن است انجام می‌گیرد. رایجترین این ویژگی‌ها کلمه[۷] است که برخی مواقع بوسیلهٔ نرم افزارهای پردازش زبان طبیعی[۸] با برچسب‌هایی[۹] همراه می شوند که حاوی اطلاعات صرفی- نحوی[۱۰] آنها است. در برخی موارد این ویژگی‌ها به صورت‌های پیچیده‌تری نیز تبدیل می‌شوند که رایج‌ترین شیوه استفاده از مدل سازی چند-گرمی[۱۱] است. برای بازنمایی متون با استفاده از ویژگی‌های مناسب استخراج یا ساخته شده

---

[۱] Pre- Processing
[۲] Training Classifier
[۳] Classification
[٤] Knowledge Representation
[٥] Vector Model
[٦] Features
[۷] Word
[۸] Natural language processing
[۹] Tags
[۱۰] Morpho syntactic
[۱۱] N- gram

از همان متون هر یک از این ویژگی‌ها به ارزشی عددی نگاشت داده می‌شود که به عنوان وزن[۱] آن ویژگی در متن مورد نظر محاسبه می‌شود. برای این منظور ماتریس $m \times n$ ایجاد می‌شود که در آن m کل متن‌های موجود در رده‌ها، n کل ویژگی‌های ایجاد شده، و A_{ij} تعداد تکرار ویژگی i یا به عبارتی وزن آن در متن j است.

در این هنگام بطور معمول مشکل گستره و پراکندگی زیاد وزن این ویژگی‌ها پیش می‌آید که برای حل آن به شیوه‌های گزینش ویژگی‌ها[۲] رجوع می‌شود تا مقادیری از ویژگی‌ها که بیش از دیگران قابلیت تمییز دهندگی[۳] متون را دارند گلچین شوند.

در مرحله آموزش رده‌بند، برای ایجاد سامانه‌های رده بندی از سامانه های یادگیری ماشین با ناظر[۴] بهره‌گیری می‌شود. این سامانه‌ها بوسیله ویژگی-وزن‌های به دست آمده در مرحله پیش-پردازش از متونی که تحت عنوان متون آموزشی[۵] از قبل رده بندی شده‌اند آموزش داده شده و به یک رده‌بند متون[۶] تبدیل می شوند.

در مرحله رده‌بندی، متون مورد نظر برای رده‌بندی، پس از گذر از مرحله پیش-پردازش و تبدیل به بردارهای مقدار-وزن در قالب ویژگی‌های انتخاب شده در مرحله آموزش، به رده بند داده می‌شوند تا، مطابق با رده‌های از قبل آموزش داده شده به رده‌بند، در یکی از رده‌های آن رده‌بندی شوند. میزان موفقیت رده‌بند در این رده‌بندی معادل ارزیابی انجام شده بر روی آن در مرحله آموزش فرض می شود.

در اجرای مراحل فوق برای ایجاد یک رده‌بند، آنچه رده‌بندهای مختلف را از یکدیگر متمایز می‌کند بیشتر نوع و چگونگی استفاده از ویژگیا در مرحله پیش-پردازش است بعلاوه نوع سامانه یادگیری ماشینی که انتخاب شده و برای منظور خاص تنظیم می‌شود. در دو بخش بعدی به دو نمونه از فنون موفق که برای استفاده از ویژگی‌های متون به منظور تأیید منبع بکار رفته‌اند اشاره می‌شود.

---

[۱] Weight
[۲] Feature selection
[۳] Discrimination
[۴] supervised machine learning
[۵] Train
[۶] Text Classifier

4. الگوریتم Benedetteo, Caglioti & Loreto

یکی از شیوه های غالب کنونی در ایجاد، استخراج و گزینش ویژگیهای متن روش مورد استفاده توسط Darrio Benedetteo, Emanuele Caglioti and Vittorio Loreto است که در آن برای تشخیص منبع رشته فرضی از ویژگی توالی کلمات استفاده شده و برای وزن دهی مقادیر ویژگی‌ها از فنون نظریه اطلاعات بهره‌گیری شده است.

به نظر گروه فوق، بسیاری از پدیده‌های طبیعی، همچون زبان و توالی DNA و پروتئین، اغلب به صورت رشته‌هایی[۱] از کاراکترها[۲] بازنمایی می‌شوند که با تجزیه آنها می‌توان اطلاعات‌شان را استخراج کرد. بطور مثال در مورد توالی DNA و پروتئین می‌توان زیر-توالی‌هایی[۳] که نشانگر ژن‌ها و کارکردشان است تشخیص داد. و در متون مکتوب می‌توان زبان، نگارنده و موضوع آنرا مشخص ساخت.

4-1. الگوریتم Benedetteo, Caglioti & Loreto و تئوری اطلاعات

چنین اهدافی از تجزیه رشته‌ها را می‌توان از منظر نظریه اطلاعات[۴] واصل شد که در آن اطلاعات کلمه معنای خاصی می‌یابد: آنتروپی[۵] رشته، و آن میزان[۶]غریبی[۷] است که منشأ ساطع کنندهٔ رشته می‌تواند برای ما محفوظ نگاه دارد.

---

۱ Strings
۲ Characters
۳ Subsequences
۴ Information theory
۵ Entropy
۶ Measure
۷ Surprise

بدین ترتیب می‌توان بر اساس محتوای اطلاعاتی نسبی[۱] زوج رشته ها مفهومی عام از دوری یا نزدیکی بین آنها تعریف کرد که در آن این فاصله[۲] با شیوه‌ای منبطق بر فنون فشرده‌سازی داده[۳] اندازه‌گیری می شود و هدف از آن استفاده از این فاصله اطلاعاتی[۴] بین زوج رشته‌ها بعنوان تفاوت معنایی[۵] واقعی بین آنها است. مفهوم آنتروپی ارتباط بسیار نزدیکی با مشکل دیرینه کدگزاری بهینه[۶] متن دارد؛ Shannon کشف کرد که حدی[۷] برای امکان کدگزاری رشته فرضی وجود دارد و آن آنتروپی رشته است. بهترین تعریف آنتروپی توسط Chaitin-Kolmogorov ارائه شده است: آنتروپی رشته‌ای از کاراکترها طول کوچکترین برنامه‌ای است (به واحد بیت) که رشته را به عنوان خروجی تولید می‌کند. این حدی است که الگوریتم‌های فشرده‌سازی سعی دارند به آن برسند؛ الگورتم‌هایی که تلاش می‌کنند فایل ورودی خود را به کوتاه‌ترین فایل ممکن تبدیل نمایند. بطور مثال الگوریتم LZ77 رشته‌های با نسخه دوم[۸]را در داده ورودی به ترتیب ذیل پیدا می‌کند. به دنبال طولانی‌ترین نظیر بافر پیش بینی[۹] گشته و بوسیله دو عدد نشانگری را برای آن نظیر تولید می‌کند: فاصله، که نمایانگر بعد مسافتی است که نظیر آغاز می‌شود، و طول، که نمایانگر تعداد کاراکترهای متناظر است. بعنوان مثال، نظیر توالی $\sigma_1 \ldots \sigma_n$ بوسیله نشانگر (d, n) ارائه می‌شود که در آن d فاصله از جایی است که نظیر آغاز می‌شود. سپس توالی متناظر با عددی کدگزاری می‌شود که برابر است با (1): یا به عبارتی، تعداد بیت‌هایی که برای کدگزاری d و n ضروری است.

$$\text{Log}_2 (d) + \log_2 (n) \qquad (1)$$

بطور خلاصه، میانگین فاصله بین دو $\sigma_1 \ldots \sigma_n$ پی در پی از مرتبه[۱۰] معکوس احتمال وقوع آن است. بر این اساس، برنامه فشرده‌ساز توالی‌های با تواتر بیشتر را با بایت‌های کمی کدگزاری می‌کند و بایت‌های

---

۱ relative informational content
۲ distance
۳ data-compression
۴ informational distance
۵ semantic
۶ optimal coding
۷ limit
۸ duplicate
۹ lookahead buffer
۱۰ order

بیشتر را برای توالی‌های نادر نگاه می‌دارد. اگر الگوریتم LZ77 یک توالی با طول L را که از منبعی همه سویی[1] ساتع شده و آنتروپی آن به ازای هر کاراکتر s است کدگزاری کند، آنگاه طول فایل فشرده شده تقسیم بر طول فایل اولیه به سمت s میل می‌کند در حالیکه طول متن به سمت ∞ میل می‌کند. به کلام دیگر، این الگوریتم فایل را به بهترین شکل کدگزاری نمی‌کند، اما با افزایش طول فایل عملکرد آن بهتر و بهتر می‌شود. بنابراین الگوریتم‌های فشرده‌سازی ابزاری قوی برای اندازه گیری آنتروپی هستند.

مفهوم آنتروپی نسبی در الگوریتم‌های فشرده‌سازی با مثال ذیل آسانتر قابل توصیف است. فرض می کنیم دو منبع همه سویی A و B رشته هایی از 0 و 1 را ساتع می‌کنند: A یک 0 با احتمال p و یک 1 را با احتمال $1-p$ ساتع می‌کند، در حالیکه B نیز یک 0 با احتمال q و یک 1 را با احتمال $1-q$ ساتع می‌نماید. الگوریتمی که رشته‌ای ساتع از A را فشرده‌سازی کند، آنرا بطور نسبی به صورت بهینه کدگزاری می‌کند: به عبارت دیگر، کدگزاری یک 0 با $-\log_2 p$ بیت و کدگزاری یک 1 با $-\log_2 (1-p)$ بیت. این کدگزاری بهینه برای رشته‌های ساتع از B بهینه نخواهد بود. آنتروپی به ازای هر کاراکتر رشته ساتع از B در کدگزاری بهینه برای A به صورت (2) است، در حالیکه آنتروپی به ازای هر کاراکتر رشته ساتع از B در کدگزاری بهینه آن (3) خواهد بود.

$$-q \log_2 p - (1-q) \log_2 (1-p) \qquad (2)$$

$$-q \log_2 q - (1-q) \log_2 (1-q) \qquad (3)$$

تعداد بیت‌های تلف شده برای کدگزاری رشته ساتع از B در کدگزاری بهینه برای A آنتروپی نسبی A و B است:

$$-q \log_2 \frac{p}{q} - (1-q) \log_2 \frac{1-p}{1-q} \qquad (4)$$

بر پایه این شیوه، اندازه‌گیری آنتروپی نسبی بین دو منبع A و B مطابق الگوریتم ذیل است: یک رشته طولانی A از منبع A و یک رشته طولانی B بهمراه یک رشته کوتاه b از منبع B استخراج می‌شود. سپس رشته جدید $A + b$ با الحاق b به بعد از A ایجاد می‌شود. آنگاه رشته $A + b$ با یک الگوریتم

[1] ergodic

فشرده سازی فشرده می‌شود. در نتیجه اندازه طول b در کدگزاری بهینه برای A از (5) محاسبه می شود، که در آن L_X به طول فایل فشرده شده X در واحد بیت اشاره دارد.

$$\Delta_{Ab} = L_{A+b} - L_A \qquad (5)$$

آنتروپی نسبی S_A به ازای هر کاراکتر بین A و B بر اساس (6) تخمین زده می‌شود، که در آن $|b|$ تعداد کاراکترهای رشته است.

$$S_{AB} = (\Delta_{Ab} - \Delta_{Bb}) / |b| \qquad (6)$$

تخمین آنتروپی منبع B نیز بر اساس (7) است.

$$\Delta_{Bb} / |b| = (L_{B+b} - L_B) / |b| \qquad (7)$$

بطور تجربی ثابت شده است که برای فایل A با مرتبه 32 تا 64 کیلو بایت، طول فایل b باید بین 1 تا 15 کیلو بایت باشد. بنابراین فایل ناشناخته X متعلق به منبع A_i است اگر به ازای رشته کوچک x از آن فایل مطابق رابطه (5) مقدار رابطه (8) کمینه باشد.

$$L_{Ai+x} - L_{Ai} \qquad (8)$$

4-2. ارزیابی الگوریتم Benedetteo, Caglioti & Loreto

ارزیابی گزارش شده توسط طراحان این الگوریتم در بکارگیری آن برای رده‌بندی متونی در زبان‌های رایج اروپایی حاکی از صحت 93.3% آن است. هرچند بکارگیری این الگوریتم برای رده‌بندی متون فارسی و عربی توسط طراحان آن گزارش نشده است، در آزمایش‌های اولیه در تشخیص آیات قرآن کریم و احادیث از یکدیگر بر اساس الگوریتم فوق توسط شکرالهی و همکارانش در دانشگاه نبی اکرم (ص)، تبریز،صحت مشابه حاصل شده است.

5. بهره‌گیری از سامانه خبره مبین

شیوه دیگری که توسط شکرالهی و همکارانش برای ایجاد و گزینش ویژگی‌های متنی در تشخیص آیات قرآن کریم و احادیث از یکدیگر استفاده شده است بکارگیری سامانه خبره مبین است که به منظور برچسب گزاری دستوری واژه‌ها در متون عربی کلاسیک، همانند قرآن کریم و احادیث، توسط شکرالهی و همکارانش طراحی و پیاده سازی شده است.

5-1. مرحله پیش-پردازش

برخلاف الگوریتم قبلی که در آن سنگ بنای ویژگی‌های ایجاد شده کلمه است، در این الگوریتم بجای کلمه از برچسب‌های دستوری کلمه‌ها بهره‌گیری می‌شود. به عنوان مثال رشته ذیل ترجمه یکی از آیه های قرآن کریم به تنها یک دسته از برچسب‌های دستوری کلمه‌های آن است.

PN PN PN PN @ PN PN V PN @ PN PN PN @ N N PN @ N V P N

در ایجاد ویژگی‌های متنی، از این رشته‌های ساده با استفاده از مدل‌سازی چند-گرمی[۱] رشته‌های ترکیبی ساخته می‌شود، که تا کنون آزمایشات بیشتر روی جفت-گرمی‌ها انجام یافته‌اند. گزینش بهینه در بین این ویژگی‌ها بر اساس تابع خی-دو[۲] انجام می گیرد. بر اساس گزارشات پژوهشی متعدد در ردەبندی متن، این تابع که در شماره (9) ارائه شده است بهتر از توابع مرسوم دیگر در داده‌کاوی[۳] نتیجه‌بخش بوده است.

$$X^2 = \frac{N \times (TP \times TN - FP \times FN)^2}{(TP+FN) \times (FP+TN) \times (TP+FP) \times (FN+TN)} \qquad (9)$$

[۱] N-gram
[۲] Chi-square
[۳] data mining

در این تابع که میزان وابستگی یک ویژگی مورد نظر و یک رده در مقایسه با سایر رده‌ها را با یک عدد بیان کرده و در صورت صفر بودن نشان‌دهنده عدم وابستگی معنی‌دار است، آرگومان‌های استفاده شده به شرح ذیل می‌باشند.

- TP: تعداد تکرار ویژگی مورد نظر در رده مثبت
- FP: تعداد تکرار ویژگی مورد نظر در سایر رده‌ها
- FN: تعداد متون فاقد ویژگی مورد نظر در رده مثبت
- TN: تعداد متون فاقد ویژگی مورد نظر در سایر رده‌ها
- N: کل تعداد متن‌ها در تمامی رده‌ها

بر اساس حاصل این تابع به ازای هر ویژگی-رده، برای هر رده ویژگی‌هایی که دارای بالاترین مقدارند گزینش می‌شوند، که تعداد آنها بطور معمول بین 1% تا 10% تعداد کل ویژگی‌هاست. آنگاه نوبت وزن-دهی به این ویژگی‌هاست که تابع رایج آن بر اساس تابع $TF.IDF$ به ترتیب (10) محاسبه می‌شود.

$$TF.IDF = TF.\log \frac{N}{DF} \qquad (10)$$

در این تابع TF به تعداد تکرار ویژگی مورد نظر[۱] یا به اصطلاح وزن آن در متن مورد نظر، DF[۲] به تعداد متن‌های شامل آن ویژگی، IDF به معکوس[۳] DF، و N به تعداد کل متن‌ها اشاره دارند.

[۱] term frequency
[۲] document frequency
[۳] inverse

بدین ترتیب ماتریس بزرگ قبلی با سلول‌هایی که حاوی تعداد تکرار تمامی ویژگی‌های ایجاد شده در تمامی متن‌ها بود به ماتریسی بسیار کوچک تبدیل می‌شود که سلول‌های آن حاوی وزن‌های ویژگی‌های گزینش شده برای همان متن‌هاست.

5-2. مرحله آموزش رده‌بند

ماتریس نهایی ایجاد شده در مرحله پیش-پردازش به عنوان مجموعه داده آموزشی برای سامانه‌های یادگیری ماشین با ناظر استفاده می‌شود. تا کنون سامانه‌های یادگیری ماشین با ناظر[۱] SVM و نیز درختهای تصمیم گیری[۲] با ویژگی‌های فوق برای تشخیص دسته آیات قرآنی از یکدیگر آموزش داده شده اند، ولی در زمینه تشخیص آیات قرآنی از احادیث و متون مشابه آزمایشات هنوز ادامه دارند.

5-3. مرحله رده‌بندی و ارزیابی

سامانه‌های فوق موفقیت چشم‌گیری در تشخیص دسته آیات قرآنی از یکدیگر، مانند آیات مکی از آیات مدنی و نیز آیات جزءهای ابتدایی قرآن از جزءهای دیگر آن داشته‌اند. در هر دو مورد F-score حدود 95% مشاهده شده است.

[۱] support vector machine
[۲] decision trees

6. نتیجه‌گیری

بر اساس توضیحات عرضه شده در تشخیص هوشمند آیات قرآن در متون فارسی تفسیری، به جهت مخلوط بودن بخش‌های دیگر عربی مانند احادیث و اشعار عربی در این متون و شباهت آنها با آیات قرآنی، ابتدا باید سامانه‌ای دیگر کلیه بخشهای عربی را از داخل متون فارسی تفسیری جدا نماید و سپس در این بخش‌های عربی آیات قرآن تمییز داده شوند. از طرفی دیگر، این بخشهای عربی در قالب رشته-هایی با طول‌های زیاد به اندازه یک آیه کامل یا چند سطر تا طول‌های بسیار کوتاه در حد یک کلمه ظاهر شده‌اند.

با نگریستن به سامانه هوشمند تشخیص آیات قرآن در متون فارسی تفسیری به عنوان سامانه‌ای برای تأیید منبع متن، الگوریتم Benedetteo, Caglioti & Loreto، در کنار موفقیت قابل توجه در تشخیص منبع رشته‌های طولانی، توفیق شایانی در خصوص رشته‌های بسیار کوتاه نداشته است. سامانه-ای که برای حل این معضل با تلاش شکرالهی و همکارانش در حال ایجاد شدن است، با وجود نتیجه بخش بودن در تفکیک آیات قرآنی از یکدیگر، در تفکیک آنها از متون عربی دیگر از جمله احادیث هنوز به سرانجام نرسیده است. به نظر می‌رسد پژوهش بیشتری باید در راستای شناسایی وجه ممیزه زبان قرآن با سایر متون مشابه انجام پذیرد.

مراجع

[1] Dario Benedetteo, Emanuele Caglioti & Vittorio Loreto. Language Treesand Zipping. *PHYSICAL REVIEW LETTERS*. Volume 88, number 4, 2002.

[2] George Forman. Feature Selection for Text Classification. *Computational Methods of feature Selection*. CRC Press/Taylor and Francis Group, 2007.

[3] Mahmoud Shokrollahi-Far, Behrouz Minaei, Issa Barzegar, Hadi Hossein-Zadeh, Mojhdeh Ghasdi, Salman Hoseini. Bootstrapping Tagged Islamic Corpora. In *Proceedings of 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt. 2009.

[4] Mahmoud Shokrollahi-Far. Mobin Exper System: Grammatical Tagger for Classical Arabic. Submitted to *10th International Conference on the Statistical Analysis of Text Data*, Italy. 2010.

[5] Mahmoud Shokrollahi-Far. Source Verification of Quranic Texts. Submitted to *10th International Conference on the Statistical Analysis of Text Data*, Italy. 2010.