عنوان زیرپروژه:

مطالعه و بررسی روش‌های تشخیص هوشمند جمله فارسی از عربی در پیکره‌های مخلوط فارسی و عربی

فهرست مطالب

عنوان **شماره صفحه**

1. مقدمه

به جهت امتزاج فرهنگ فارسی با فرهنگ اسلامی، متون نظم و نثر فارسی مملو از جملات و عبارت‌های عربی است. بخش عمده این قبیل متون متعلق به حوزه علوم اسلامی شامل علوم قرآن و حدیث و نیز فقه است.

قدر مسلم پردازش رایانه‌ای چنین متونی مستلزم سامانه‌های جداگانه برای فارسی و عربی است که ایجاب می‌کند سامانه‌ای مستقل ابتدا بخش‌های فارسی را از عربی تفکیک نماید. برای دست‌یابی به راه-کاری مناسب برای این مهم طرح پژوهشی حاضر حصول به پاسخ سؤالات ذیل را سرلوحه پژوهش امکان سنجی خود قرار داده است تا روش‌های تشخیص هوشمند جمله فارسی از عربی در پیکره‌های مخلوط فارسی و عربی تبیین شده و در پژوهش‌های آتی به سامانه‌های هوشمند تبدیل شوند.

1. روش‌های کلی تشخیص هوشمند جمله‌های یک زبان در پیکره‌های مخلوط این زبان با زبان‌های مشابه چیست؟
2. کیفیت و کمیت پیکره‌های متنی مخلوط فارسی و عربی برای تشخیص هوشمند جمله فارسی از عربی در این پیکره‌ها چگونه باید باشد؟
3. چه نرم افزارهایی برای پردازش پیکره‌های متنی مخلوط فارسی و عربی برای تشخیص هوشمند جمله فارسی از عربی در این پیکره‌ها مورد نیاز است؟
4. در صورت نیاز به پایگاه دانش برای تشخیص هوشمند جمله فارسی از عربی در پیکره‌های متنی مخلوط فارسی و عربی، آیا این پایگاه دانش مبتنی بر قواعد پیاده‌سازی شده انسانی خواهد بود یا آموخته‌های ماشین از پیکره‌های متنی مخلوط فارسی و عربی؟
5. در پیکره‌های متنی مخلوط فارسی و عربی چه ویژگی‌هایی وجه ممیزه بین جمله‌های فارسی و عربی است و بهترین شیوه گزینش این ویژگی‌ها چیست ؟

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

1. روش‌های کلی رده بندی متون،

2. سامانه‌هایی در دنیا که توانسته‌اند 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،
3. کیفیت و کمیت پیکره‌های متنی مخلوط فارسی و عربی،
4. کیفیت و کمیت بخش‌های عربی، از قبیل آیات قرآن کریم، احادیث و اشعار عربی، در داخل پیکره‌های متنی مخلوط فارسی و عربی،
5. نرم افزارهای پردازش دستوری پیکره‌های متنی فارسی و عربی،
6. پژوهش‌ها و سامانه‌هایی که با هدف پردازش و طبقه بندی متون عربی وجود دارند،
7. پژوهش‌ها و سامانه‌هایی که با هدف پردازش و طبقه بندی متون فارسی وجود دارند،
8. پژوهش‌ها و سامانه‌هایی که با هدف پردازش و طبقه بندی پیکره‌های متنی مخلوط فارسی، عربی و زبان‌های مشابه دیگر وجود دارند.

در گزارش پژوهشی حاضر ابتدا کیفیت پیکره‌های مخلوط فارسی و عربی مورد بررسی قرار می‌گیرد. آنگاه مشکل تشخیص هوشمند جمله فارسی از عربی در پیکره‌های مخلوط فارسی و عربی از لحاظ علمی تبیین شده و شاخه علمی که برای حل این مسئله از آن مدد جسته خواهد شد، یعنی رده بندی متون، توضیح داده می‌شود. به عنوان نمونه کاربرد این نگرش علمی، به سامانه‌ای که بطور اولیه به این منظور با همکاری نگارنده این گزارش پژوهشی و گروهی از دانشجویان در دانشگاه نبی اکرم (ص)، تبریز، پیاده سازی شده است اشاره خواهد گردید. سپس در همین راستا، به عنوان نمونه بهینه که در طی دهه گذشته جامع ترین الگوریتم عرضه عمومی شده است، الگوریتم[1] Ludovik and Zacharski بطور کامل تشریح می‌شود. در پایان برخی وب‌گاه‌های عمومی که در همین راستا خدمات on-line ارائه می کنند معرفی خواهند شد.

[1] Ludovik and Zacharski

2.مروری بر کیفیت پیکره‌های مخلوط فارسی و عربی

در بررسی کیفیت و کمیت امتزاج متون فارسی و عربی در یک دیگر آنچه در پژوهش حاضر بیشتر مورد توجه قرار گرفته است متونی است که عمده آنها به زبان فارسی‌اند و تنها بخش‌هایی از آنها به زبان عربی است. اینگونه متون بطور غالب متون اسلامی یا ادبی اند که در آنها گونه عربی استفاده شده در بخش‌های عربی شامل آیاتی از قرآن کریم، احادیث و روایات، و نیز اشعار عربی است. در این میان، تفاسیر فارسی و یا ترجمه شده از عربی برای قرآن کریم بهترین نمونه‌ای هستند که، در عین حجم بسیار بالا، از تمامی گونه‌های عربی فوق در آنها یافت می شود. بنابراین در پژوهش حاضر این متون به عنوان نمونه اصلی مورد بررسی قرار گرفته‌اند.

نمونه متنی که در پی می‌آید بخشی کوچک از کتاب تفسیر نور در باره سوره یونس است.

> دهمین سوره‌ی قرآن کریم که در اوایل بعثت در مکّه نازل شده است، «یونس» نام دارد. این سوره یکصد و نه آیه دارد و عمده‌ی مطالب آن، پیرامون توحید و حقّانیّت قرآن، پاسخ به منکران وحی، بیم دادن مشرکان، بیان عظمت آفرینش و آفریدگار، ناپایداری دنیا و توجّه دادن به آخرت است.
>
> ...
>
> أَ كانَ لِلنَّاسِ عَجَباً أَنْ أَوْحَيْنا إِلى رَجُلٍ مِنْهُمْ أَنْ أَنْذِرِ النَّاسَ وَ بَشِّرِ الَّذِينَ آمَنُوا أَنَّ لَهُمْ قَدَمَ صِدْقٍ عِنْدَ رَبِّهِمْ قالَ الْكافِرُونَ إِنَّ هذا لَساحِرٌ مُبِينٌ (2)
>
> آیا برای مردم شگفت‌آور است که به مردی از خود آنان وحی کردیم که مردم را بیم و هشدار بده و به مؤمنان بشارت بده که برای آنان نزد پروردگارشان جایگاه نیکویی است؟ کافران گفتند: همانا این مرد جادوگری آشکار است!
>
> سوره‌ی قبل، یعنی توبه، رفتار منافقان و کیفر آنها را بیان می‌کرد این سوره به بیان رفتار مشرکان پرداخته است.
>
> ...

پیام ها:

1 ریشه‌ی کفر، غالباً استبعاد و تعجّب از وحی است. «عَجَباً أَنْ أَوْحَيْنا» 2 سرچشمه‌ی دعوت انبیا، وحی الهی است. «أَوْحَيْنا إِلى رَجُلٍ» 3 لیاقت‌های معنوی افراد، در ظاهر دیده نمی‌شود. اگر کسی مورد لطف خاص خدا قرار گرفت، او را تحمّل کنیم. «رَجُلٍ مِنْهُمْ» 4 پیامبران به خاطر الگو بودنشان باید از مردم و درد آشنا باشند. «مِنْهُمْ» 5 وظیفه‌ی پیامبران، بشارت و انذار است. «أَنْذِرِ»، «بَشِّرِ» 6 هر آنچه نمی‌فهمیم، آن را رد نکنیم. زیرا که انکار و تهمت زدن، شیوه‌ی کفّار است. «قالَ الْكافِرُونَ إِنَّ هذا لَساحِرٌ مُبِينٌ» 7 ایمان، زمینه‌ی قدم صدق و جایگاه ویژه نزد خداوند است. «بِأَنَّ لَهُمْ الْجَنَّةَ»

همان طور که مشاهده می‌شود در نیمه اول این متن آیه مورد نظر به طور مجزا در بین دو بند توضیح فارسی آمده است؛ ولی در ادامه متن همان آیه شکسته شده و به صورت عبارتها و یا کلماتی در دل توضیحات فارسی ظاهر شده است. این به آن معناست که، در پردازش این متن و جدا کردن بخش‌های عربی از فارسی، واحد پردازش می‌تواند جمله کامل، عبارت و یا حتی کلمه تنها باشد.

البته باید توجه داشت در این متن بخش‌های عربی استفاده شده نه تنها دارای حرکه می‌باشند، بلکه به هنگام آماده سازی متن به طور دستی از بخش‌های فارسی جدا شده اند. این شرایط همیشه برقرار نیست. به عنوان مثال در کتاب تفسیر نمونه از قرآن کریم در خصوص همان سوره که بخشی از آن در پی می‌آید در ابتدا و انتهای متن از عبارتها و جمله‌های عربی استفاده شده است که مشخصه‌های فوق را ندارند.

سوره یونس (ع) این سوره در مکه نازل شده و 109 آیه است

بسم اللّه الرحمن الرحیم

محتوی و فضیلت این سوره

این سوره که از سوره‌های مکی است، و به گفته بعضی از مفسران بعد از سوره اسراء و قبل از سوره هود نازل شده است همانند بسیاری از سوره‌های مکی روی

چند مساله اصولی و زیر بنائی تکیه می‌کند، که از همه مهمتر مساله" مبدء" و" معاد" است.

منتها نخست از مساله وحی و مقام پیامبر ص سخن می‌گوید، سپس به نشانه‌هایی از عظمت آفرینش که نشانه عظمت خدا است می‌پردازد، بعد، مردم را به ناپایداری زندگی مادی دنیا و لزوم توجه به سرای آخرت و آمادگی برای آن از طریق ایمان و عمل صالح متوجه می‌سازد.

و به تناسب همین مسائل قسمت‌های مختلفی از زندگی پیامبران بزرگ از جمله نوح و موسی و یونس (ع) را بازگو می‌کند و به همین مناسبت نام سوره یونس بر آن گذارده شده است.

و باز برای تایید مباحث فوق، سخن از لجاجت و سرسختی بت پرستان به میان می‌آورد، و حضور و شهود خدا را در همه جا برای آنها ترسیم می‌کند، و مخصوصا برای اثبات این مساله از اعماق فطرت آنان که به هنگام مشکلات آشکار می‌شود و به یاد خدای واحد یکتا می‌افتند، کمک می‌گیرد.

و بالآخره برای تکمیل بحثهای فوق در هر مورد مناسبی از بشارت و انذار، بشارت به نعمتهای بی پایان الهی برای صالحان و انذار و بیم دادن طاغیان و گردنکشان، استفاده می‌کند.

لذا در روایتی از امام صادق ع می‌خوانیم که فرمود:

من قرء سورهٔ یونس فی کل شهرین او ثلاثهٔ لم یخف علیه ان یکون من الجاهلین و کان یوم القیامهٔ من المقربین

نکته قابل توجه دیگر اینکه در نمونه متن‌های ارائه شده، از لحاظ کمی، بخش‌های عربی سهم کمی را به عهده دارند.

3. رده بندی و تشخیص زبان متن

مسئله تشخیص و تفکیک بخش‌های تک-زبانهٔ در داخل متون چند-زبانه بطور رایج به عنوان یک مسئله تشخیص زبان[۱] شناخته شده و از منظر رده بندی متن[۲] نگریسته می‌شود. بطور کلی رده بندی 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 (طبقه بندی، دسته بندی یا کلاس بندی) متن است. از این نظر بطور کلی در متون چند زبانه[۳] ، سامانه تشخیص زبان[۴] باید هر قطعه تک زبانه[۵] و نیز زبان هر قطعه را مشخص نماید. در ایجاد سامانه‌های ترجمه ماشینی[۶] اغلب پیکره‌های[۷] بسیار بزرگی مورد نیاز است و برای گردآوری خودکار چنین پیکره‌ای می‌توان یک سامانه تشخیص زبان را با یک عنکبوت وب[۸] ترکیب کرد. الگوریتم‌های تشخیص نه تنها زبان متن الکترونیکی بلکه encoding کاراکترهای زبان آن را تشخیص می‌دهند.

پس می‌توان مسئله تشخیص زبان قطعه‌ای از یک متن را به عنوان مسئله رده بندی متن دانست که در آن سه مرحله پیش-پردازش[۹] ، آموزش رده بند[۱۰] و رده بندی[۱۱] وجود دارد.

---

[۱] Language detection
[۲] Classification text
[۳] Multilingual documents
[۴] Language recognition system
[۵] Monolingual segment
[۶] Machine translation
[۷] Corpora
[۸] Web spider
[۹] Pre-processing
[۱۰] Training classifier
[۱۱] Classification

در مرحله پیش-پردازش، دانش موجود در هر متن باید بازنمایی[۱] شود تا قابل استفاده نرم افزارهای رده بندی گردد. این بازنمایی به شکل مدل برداری[۲] از ویژگی‌های[۳] متن که برگرفته از عناصر موجود در آن است انجام می‌گیرد. رایج‌ترین این ویژگی‌ها کلمه[۴] است که برخی مواقع به وسیله نرم افزارهای پردازش زبان طبیعی[۵] با برچسب‌هایی[۶] همراه می‌شوند که حاوی اطلاعات صرفی-نحوی[۷] آنها است. در برخی موارد این ویژگی‌ها به صورت‌های پیچیده‌تری نیز تبدیل می‌شوند که رایج‌ترین شیوه استفاده از مدل-سازی چند-گرمی[۸] است. برای بازنمایی متون با استفاده از ویژگی‌های مناسب استخراج یا ساخته شده از همان متون هر یک از این ویژگی‌ها به ارزشی عددی نگاشت داده می‌شود که به عنوان وزن[۹] آن ویژگی در متن مورد نظر محاسبه می‌شود. برای این منظور ماتریس $m \times n$ ایجاد می‌شود که در آن m کل متن-های 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 مقدار-وزن این ویژگی‌ها پیش می‌آید که برای حل آن به شیوه‌های گزینش ویژگی‌ها[۱۰] رجوع می‌شود تا مقادیری از ویژگی‌ها که بیش از دیگران قابلیت تمییز دهندگی[۱۱] متون را دارند گلچین شوند.

در مرحله آموزش رده بند، برای ایجاد سامانه‌های رده بندی از سامانه‌های یادگیری ماشین با ناظر[۱۲] بهره‌گیری می‌شود. این سامانه‌ها بوسیله مقدار-وزن‌های به دست آمده در مرحله پیش-پردازش از متونی

---

[۱] Knowledge representation
[۲] Model vector
[۳] Features
[۴] Word
[۵] Natural language processing
[۶] Tags
[۷] Morpho syntactic
[۸] N- gram
[۹] Weight
[۱۰] Selection feature
[۱۱] Discrimination
[۱۲] Supervised machine learning

که تحت عنوان متون آموزشی[1] از قبل رده بندی شده‌اند آموزش داده شده و به یک رده بند متون[2] تبدیل می‌شوند.

در مرحله رده بندی، متون مورد نظر برای رده بندی، 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 رده بند، در یکی از رده‌های آن رده بندی شوند. میزان موفقیت رده بند در این رده بندی معادل ارزیابی انجام شده بر روی آن در مرحله آموزش فرض می‌شود.

[1] Train
[2] Text classifier

4. تشخیص زبان در متون تک-زبانه

مطابق گزارش اخیر [3]، اصلی‌ترین تفاوت بین شیوه‌های متداول تشخیص زبان در متون تک-زبانه وابسته به مدل زبانی است که از متن مورد نظر می‌سازند (هرچند در مرحله آموزش رده‌بند در بکارگیری الگوریتم‌های مختلف یادگیری ماشین با هم تفاوت‌هایی دارند). بر این اساس این شیوه‌ها از سه نگرش کلی پیروی می‌کنند که وابسته به سه نوع ویژگی بکار رفته در مرحله پیش-پردازش‌اند:

1. کلمات کوتاه، که در آن مدل زبانی فقط از کلمات کوتاه با طول محدود و بدون توجه به تواتر این کلمات ساخته می‌شود. به طور نمونه، [4] از تمامی کلمات با طول حداکثر پنج کاراکتر که حداقل سه بار تکرار شده باشند بهره می‌برد و بدین ترتیب مدل زبانی‌اش از 980 تا 2750 چنین کلمه‌هایی استفاده می‌کند.
2. کلمات متواتر، که در آن مدل زبانی با استفاده از کلماتی ساخته می‌شود که بین کلیه کلمات متن بیشترین تواتر را دارند. به طور نمونه، [5] با استفاده از یک صد کلمه بابیشترین تواتر جداولی می‌سازد که در آنها هر کلمه ارزش تواتری خاصی دارد که خارج قسمت تقسیم تواتر نسبی هر کلمه در جدول بر تواتر نسبی کلمه اول آن جدول است.
3. چند-گرم کلمات، که در آن توالی کلیه کلمات متن در نظر گرفته شده و از آنها بر اساس چند-گرمی‌های با طول ثابت یا متغیر مدل زبانی ساخته می‌شود.

شیوه غالب کنونی در انتخاب ویژگی‌های تمییز دهنده برای شناسایی زبان استفاده از شیوه سوم است که به عنوان مثال در الگوریتم پیشنهادی [6] بکار رفته است و با ایجاد مدل زبانی با 300 چند-گرمی نتیجه 99.8% به دست آورده است. با این وجود و "در عین سادگی و شهودی بودن و نیز موفقیت در بسیاری موارد، این شیوه چندین نقطه ضعف دارد. اول اینکه، علیرغم کارکرد خوب در متون با اندازه متوسط، با کاهش طول متن عملکرد آن کاهش چشم‌گیری می‌یابد. دوم اینکه، به جهت اتکا بر واژه‌ها، این شیوه وابستگی بسیاری به توانایی سامانه در تقطیع متن به واژه ها دارد" [7]. ولی مهمترین کاستی تمامی این شیوه‌ها اختصاص آنها به متون تک-زبانه است و این در حالی است که پیکره مورد هدف این پژوهش بطور کلی پیکره‌ای دو زبانه حاوی فارسی و عربی‌است. بنابراین الگوریتم‌هایی در تشخیص زبان متون باید بیشتر مد نظر قرار گیرند که برای متون چند-زبانه طراحی شده‌اند.

5.تشخیص زبان در متون چند زبانه و الگـوریتم Ludovik and Zacharski

در بررسی الگوریتم‌های بکار رفته برای تشخیص زبان در متون چند-زبانه نکته اولی که جلب توجه می-کند نادر بودن چنین الگوریتم‌هایی است. نکته قابل توجه اصلی در این الگوریتم‌ها، از نمونـه‌هـای قبلـی مانند [8] گرفته تا الگوریتم جدید استفاده شده در [9] همگی برای متون حاوی مخلوطی از زبان‌هایی به غیر از فارسی و عربی طراحی شده‌اند. در تلاشی برای حل کاستی موجود و با هـدف تشـخیص زبـان در پیکره‌های مخلوط چند-زبانه که فارسی و عربی را نیز در بر می‌گیرد، به شیوه‌ای مشابه آنچه که در شیوه اول انتخاب ویژگی‌های متنی در تشخیص زبان متون تک-زبانه مورد استفاده است سامانه‌ای در دانشـگاه نبی اکرم (ص) در تبریز به صورت نمونه‌ی اولیه[۱] طراحی و پیاده سازی شده است[10]. مشابه موفقیت-های گزارش شده برای این شیوه در تشخیص زبان متون تک-زبانه، این سامانه در تمییز جملات فارسـی و عربی از یکدیگر توفیق زیادی داشته است. علیرغم این موفقیت نسبی دارای کاستی مهمی اسـت. ایـن سامانه فقط موقعی موفقیت دارد که متن بطور کامل در واحد جمله تقطیع شده باشـد؛ بنـابراین سـامانه فوق در تشخیص وا حدهای کوچکتر از جمله همانند عبارت یا کلمه هنوز توفیقی ندارد. و این در حـالی است که، مطابق بخش 2 این گزارش پژوهشی، در پیکره‌های مخلوط فارسی و عربی نه تنها از جمله‌های عربی بلکه از واحدهای کوچکتری چون عبارت و کلمه‌های عربی نیز استفاده شده است.

بر پایه این توضیحات، الگوریتمی مورد نیاز است که بتواند نه تنها در تشخیص طول رشته‌های بلند مانند جمله و کوتاه مانند عبارت موفق عمل کند بلکه در متون مخلوط حاوی فارسی و عربی نیز امتحان خـود را پس داده باشد. الگوریتم Ludovik and Zacharski در [7] چنین الگوریتمی اسـت و در واقـع از این جهت تنها الگوریتم گزارش شده و قابل دسترس است.Ludovik and Zacharski در الگـوریتم خود برای شناسایی زبان متون چند زبانه از ترکیب چند-گرم‌ها[۲] با مرتبه‌های مختلط[۳] ، زنجیره مارکف[۴] ، درستنمایی حداکثر[۵] و برنامه‌سازی پویا[۱] استفاده کرده‌اند که وابستگی این شیوه را به تشخیص کلمات

---

[۱] Prototype
[۲] N-gram
[۳] Mixed order
[۴] Markov chain
[۵] Maximum likelihood

متن می‌کاهد. این الگویتم دارای دو مرحله اصلی آموزش و رده بندی و نیز یک مرحله فرعی تقطیع[۲] است.

5-1. مرحله پیش پردازش و آموزش الگوریتم Ludovik and Zacharski

در مرحله پیش پردازش و آموزش، الگوریتمLudovik and Zacharski، برای ایجاد ویژگی‌ها و مقدار-وزن آنها، مجموعه‌ای خالی را مورد استفاده قرار می‌دهد که از چند-گرمی‌های با طول متغیر برگرفته شده از هر متن آموزشی پر خواهد شد (در پژوهش گزارش شده تا چهار-گرمی استفاده شده است). ابتدا فهرست‌های جداگانه‌ای برای تک-گرمی‌ها، جفت-گرمی‌ها و غیره ایجاد می‌شود. آنگاه بخشی از این چند-گرمی‌ها گزینش شده و به مجموعه اولیه اضافه می‌گردد – ابتدا تک-گرمی‌ها، سپس جفت-گرمی‌ها و به همین ترتیب. گزینش با محاسبه وزن آموزشی[۳] انجام می‌گیرد که بطور مثال برای تک-گرمی‌ها با استفاده از (1) است.

$$W(a_1) = -p(a_1) \log p(a_1) \qquad (1)$$

سپس برای محاسبه وزن آموزشی هر k-گرمی $(a_1a_2\ldots a_k)$ فرایند بازگشتی ذیل به ازای k = 2 … N طی می شود. اگر (k-1)-گرمی $(a_2\ldots a_k)$ هنوز در مجموعهٔ اولیه نیست وزن آموزشی با استفاده از (2) محاسبه می‌شود و در غیر این‌صورت از (3).

$$W(a_1 a_2 \ldots a_k) = -p(a_1 a_2 \ldots a_k) \log p(a_k \mid a_1 a_2 \ldots a_{k-1}) \qquad (2)$$

[۱] Dynamic programming
[۲] Segmentation
[۳] Training weight

$$W(a_1 a_2 \ldots a_k) = -p(a_1 a_2 \ldots a_k)(\log\ p(a_k \mid a_1 a_2 \ldots a_{k-1}) - \log p(a_k \mid a_2 a_3 \ldots a_{k-1})) \quad (3)$$

سپس با مرتب‌سازی نزولی تمامی چند-گرمی‌ها بر اساس وزن‌های آموزشی، بالاترین آنها انتخاب و در مجموعهٔ اولیه قرار می‌گیرند. در انتهای این فرایند، مجموعه اولیه شامل اجتماع چند-گرمی‌های گزینش شده برای همه L زبان‌ها در مجموعه آموزشی است.

آنگاه به ازای هر چند-گرمی (n = 1 … N)در مجموعه اولیه بردار L-بعدیِ وزن تشخیص اولیه[1] به ترتیب ذیل محاسبه می‌شود. هر i-امین عضو بردار، PRW، مطابق i-امین زبان بوده و شامل یک وزن تشخیص است. وزن تشخیص برای تک-گرمی‌ها به صورت (4) و برای تمامی دیگر چند-گرمی‌ها به صورت (5) خواهد بود.

$$PRW_i(a_1) = -\log\ p_i(a_1) \quad (4)$$

$$PRW_i(a_1 a_2 \ldots a_k) = -\log p_i(a_K \mid a_1 a_2 \ldots a_{k-1}) \quad (5)$$

در واقع (5) پاد-لگاریتم احتمال شرطیِ a_k است به شرط بافتار $a_1 a_2 \ldots a_{k-1}$ در مجموعه آموزشی مطابق با i-امین زبان. اگر یک چند-گرمی در مجموعه اولیه در دادهٔ آموزشی زبانی پیش نیاید، وزن تشخیص آن برای این زبان بعنوان ارزش حداکثری MAX تعریف می‌شود. در پایان محاسبه، بردارِ L-بعدیِ وزنِ تشخیصِ هر چند-گرمی بصورت (6) محاسبه می شود.

$$RW(a_1 a_2 \ldots a_N) = \begin{cases} PRW(a_1 a_2 \ldots a_N) & \mid a_1 a_2 \ldots a_N \in CNP \\ PRW(a_2 a_3 \ldots a_N) & \mid a_2 a_3 \ldots a_N \in CNP \\ \ldots & \mid \ \ldots \\ PRW(a_N) & \mid a_N \in CNP \end{cases} \quad (6)$$

[1] PRW: primary recognition weight

vector of MAX | *otherwise*

در پایان، میانگین وزن[1] و پراکندگی وزن[2] برای هر زبان محاسبه می‌شود که به ترتیب ذیل انجام می‌گیرد. هر متنِ آموزشیِ زبانِ فرضیِ i به K قطعه 500 بایتی ($x_1x_2\ldots x_{500}$) تقطیع می‌شود. آنگاه به ازای هر قطعه k (7) محاسبه می‌شود.

$$d_{i,k} = \sum_{s=1}^{500} \frac{RW_i(x_{s-N+i}x_{s-N+2}\ldots x_s)}{500} \qquad (7)$$

میانگین وزن برای زبان i بر اساس (8) تعریف می‌شود.

$$WA_1 = \frac{\sum_{k=1}^{K} d_{i.k}}{K} \qquad (8)$$

و پراکندگی برای زبان i بر اساس (9) تعریف می‌شود.

$$\boldsymbol{D}_1^2 = \frac{\sum_{k=1}^{K} (d_{i.k} - WA_i)^2}{K} \qquad (9)$$

[1] Weight average
[2] Weight dispersion

2-5. مرحلــه رده بنــدی و تشــخیص الگــوریتم Ludovik and Zacharski

مرحله تشخیص شامل دو فرایند است: رده‌بندی و تأیید. در رده‌بندی متن تحت تشخیص به طـور اولیـه در رده یکی از زبان‌های تعیین شده در مرحله آموزش رده‌بندی می‌شود. آنگاه در فرایند تأیید اینکه متن مورد نظر با زبان پیشنهاد شده در فرایند اول چقدر مناسبت دارد تعیـین مـی‌شـود؛ اگـر تناسـب کـافی نباشد، زبان متن به عنوان ناشناس رده‌بندی می‌شود.

فرایند رده‌بندی: فرض می‌شود $x_0x_1...x_{s-1}$ رشته بایتی تحت رده‌بنـدی باشـد. یـک بـردار وزن نتیجـه تشخیص[1] به صورت (10) تعریف می‌شود. نتیجه تشخیص (زبان پیشنهادی برای متن) بـه صـورت (11) خواهد بود (زبان متن به عنوان زبان i^* رده‌بندی می‌شود که متناظر اسـت بـا آن عضـوی از بـردار وزن نتیجه تشخیص که کمترین مقدار را دارد).

$$RRW = \frac{\sum_{s=0}^{S-1} RW(x_{s-N+1}x_{s-N+2}...x_s)}{S} \qquad (10)$$

$$i^* = \arg_s \min RRW_i \qquad (11)$$

فرایند تأیید: اگر رابطه (12) برقرار باشد، آنگاه زبان متن به عنوان اینکه زبان آن i^* است رده‌بندی مـی‌شود. اما اگر این رابطه برقرار نباشد، آنگاه زبان آن به عنوان ناشناس رده‌بندی خواهدشد.

$$\frac{RRW_{i^*} - WA_{i^*}}{D_{i^*}} \le VER\_THR \qquad (12)$$

[1] Result Recognition Weight

اصلی‌ترین شرط موفقیت الگوریتم Ludovik and Zacharski در مرحله رده‌بندی و تشخیص این است که متن مورد نظر تک زبانه باشد. برآورده شدن این شرط به عهده مرحله تقطیع است.

3-5.مرحله تقطیع الگوریتم Ludovik and Zacharski

از آنجا که متون تحت بررسی چند زبانه هستند، این متون باید که ابتدا به قطعات تک زبانه تقطیع شده و سپس زبان هر قطعه تشخیص داده شود. این امر با افزودن یک الگوریتم برنامه‌نویسی پویا مبتنی بر مدل مارکوفی ساده از متون چند زبانه امکان پذیر است. این الگوریتم، که اولین بار توسط Vintsiuk برای تقطیع استفاده شد، گونه‌ای بسط یافته از الگوریتمی است که کارآیی محاسباتی آن توسط Ludovik افزایش یافته است و دارای دو مرحله است.

1. مرحله تشخیص، که الگوریتم برنامه‌نویسی پویا را بکار می‌گیرد تا با به حداکثر رسانی احتمال رشته کلی کاراکترهای متن تحت پردازش به تقطیع دست یابد،
2. مرحله تأیید، که تعیین می‌کند تا چه حدی هر قطعه با زبان تخصیص یافته مناسبت دارد.

در مرحله تشخیص، الگوریتم تقطیع بر مدل مارکوفی ذیل از متن چند زبانه بنا شده است. این مدل به ازای هر زبان یک حالت [1] دارد، $1 \leq i \leq L$ ، بعلاوه یک حالت 0 اضافی برای قطعاتی که در هیچ یک از زبان‌های آموزشی سامانه وجود ندارند، از جمله قطعاتی که به هیچ عنوان به زبانی متعلق نیست (مانند جدولی از اعداد). اگر سامانه در حالت i باشد، بسته به بافتار [2] و بر اساس مدل چند-گرمی، کاراکترهای زبان i را تولید خواهد کرد. تغییر وضعیت از زبان i_1 به زبان i_2 بر اساس احتمالات انتقال $p(i_2|i_1)$ و توزیع احتمال طول قطعه r $p(r)$، $r_{min} \leq r \leq r_{max}$ ، با گستره تغییر بین طول قطعه حداقل r_{min} تا حداکثر r_{max} تعریف می شود. الگوریتمی که در ذیل ارائه می‌شود با توجه به مدل مارکوفی نسبت به حداکثر درستنمایی متن مورد مشاهده حصول اطمینان کرده، قطعه‌ها و زبان‌های قطعه‌ها را می‌یابد.

[1] State
[2] Context

اگر kامین قطعه در s_{k-1} شروع شده و در s_k-1 تمام شود، آنگاه با استفاده از پادلگاریتم، معیار حداقل سازی بر اساس (13)است که در آن TSW یک وزن قطعه است مطابق (14) به شرط صفر نبودن i_k ، وگرنه مطابق (15) (ضریب ثابت JUNK_THR[1]، به این منظور به کار می‌رود تا قطعه‌های متنی را که در هیچ یک از زبان‌های آموزشی نیستند حذف نماید):

$$Q(\{x_s, 0 \le s \le S\}, \{s_k, i_k, 0 \le k \le K\}) = \sum_{k=1}^{K} TSW\left(i_{k-1}, i_k, s_{k-1}, s_k\right) \qquad (13)$$

$$TSW\left(i_{k-1}, i_k, s_{k-1}, s_k\right) = \sum_{s=s_{k-1}}^{s_k-1} RW_{i_k}(x_{s-N+1} x_{s-N+2} \dots x_s) - \log p(i_k|i_{k-1}) - \log p(s_k - s_{k-1}) \qquad (14)$$

$$TSW\left(i_{k-1}, i_k, s_{k-1}, s_k\right) = \text{JUNK\_THR}^*(s_k - s_{k-1}) - \min_{i:1\le i\le L}\left[\sum_{s=s_{k-1}}^{s_k-1} RW_i(x_{s-N+1} \dots x_s) - \log p(i_k|i_{k-1})\right] - \log p(s_k - s_{k-1}) \qquad (15)$$

الگوریتم برنامه‌نویسی پویا که مقادیر بهینه$\{0<k<K, i_k, s_k\}$ را می‌یابد شامل گام تکراری (17) برای (16) با مقادیر اولیه (18) می شود.

[1] Junk threshold

$$Ind\left(i_{\mathrm{k}},s_{\mathrm{k}}\right)=\left(i_{\mathrm{k}-1}*,s_{\mathrm{k}-1}*\right)=\underset{\substack{\left(i_{\mathrm{k}-1},s_{\mathrm{k}-1}\right);\\ i_{\mathrm{k}-1};\ 0\le i_{\mathrm{k}-1}\le L;\\ s_{\mathrm{k}-1};\ r_{\min}\le(s_{\mathrm{k}}-s_{\mathrm{k}-1})\le r_{\max};\ s_{\mathrm{k}-1}\ge 0}}{\arg\min}\left(F(s_{\mathrm{k}-1},\ i_{\mathrm{k}-1})\right.$$

$$+\ TSW\left(i_{\mathrm{k}-1},i_{\mathrm{k}},\ s_{\mathrm{k}-1},s_{\mathrm{k}}\right) \qquad (16)$$

$$F\left(i_{\mathrm{k}},s_{\mathrm{k}}\right)=F\left(i_{\mathrm{k}-1}*,s_{\mathrm{k}-1}*\right)+\ TSW\left(i_{\mathrm{k}-1}*,i_{\mathrm{k}},\ s_{\mathrm{k}-1}*,s_{\mathrm{k}}\right) \qquad (17)$$

$$F(i,0)=0,\ 0<=\mathrm{i}<=L,\ F(i,s)=\mathrm{INFINTY},\ s:\ 0<\mathrm{s}<r_{\min},\ 0<=i<=L \qquad (18)$$

بعد از اینکه الگوریتم گام‌های بازگشتی را تمام می‌کند، زبان آخرین قطعه در مجموعه بهینه قطعات (19) یافت می‌شود که در آن S طول کل مربوط به متن است.

$$i*=\underset{i:0\le i\le 1=L}{\arg\min}\ F(i,S) \qquad (19)$$

مقدار $i*$ به همراه اطلاعات ذخیره‌شده در آرایه $Ind(i,s)$ اجازه خواهدداد تا تمامی قطعه ها با برچسب‌های زبان مربوطه از مجموعه بهینه دریافت شوند. در طی این فرایند، تعداد بهینه قطعه‌ها به طور خودکار تعیین می‌شود.

فرایند تأیید در مرحله تقطیع همانند فرایند تأیید در مرحله تشخیص و رده بندی متون تک زبانه است.

4-5. ارزیابی الگوریتم Ludovik and Zacharski

برای اندازه‌گیری نرخ خطای این الگوریتم در تقطیع متون چند-زبانه، الگوریتم روی 6 متن که مخلوطی از 1000 قطعه با طول‌های به طور نسبی مساوی (بین 20 تا 1000 بایت) از زبان‌های مختلف از جمله عربی و فارسی بودند توسط طراحانش ارزیابی شده است. نرخ خطای این الگوریتم 0.47% برای قطعه متن‌های 1000 بایتی، 0.69% برای قطعه متن‌های 540 بایتی، 1.4% برای قطعه متن‌های 202 بایتی، 2.08% برای قطعه متنهای 101 بایتی، 4.7 % برای قطعه متنهای 49 بایتی، و 12.88% برای قطعه متن-های 20 بایتی (حدود سه کلمه) گزارش شده است و این نرخ آخری بدین معناست که الگوریتم فقط در تشخیص بایت‌های ابتدایی و انتهایی کلمه خطا می‌کند، که ممکن است فاصله و یا علامت‌های نقطه-گزاری باشند.

در نتیجه آزمایش‌های مختلف Ludovik and Zacharski، عملکرد این الگوریتم برای تشخیص زبان متون تک-زبانه شامل زبان‌های عربی و فارسی وقتی خوب است که اندازه متن تحت بررسی بین 500 تا 1000 بایت باشد. نرخ خطای آن 0.27% برای متون 1000 بایتی، 0.52% برای متون 500 بایتی، 2.02% برای متون 100 بایتی، 4.01 % برای متون 50 بایتی و 11.92% برای متون 20 بایتی گزارش شده است. باید توجه داشت که این الگوریتم تنها یک ششم نرخ خطای الگوریتم [6] را دارد.

6.نتیجه‌گیری

در پژوهش حاضر مشکل تشخیص هوشمند جمله فارسی را در پیکره‌های مخلوط فارسی و عربی به عنوان یک مسئله تشخیص زبان در حوزه رده بندی متن که از فنون یادگیری ماشین استفاده می‌کنند نگریسته شد. آنگاه با مروری بر نمونه‌ای از این پیکره‌ها، یعنی برخی متون تفسیر قرآن کریم، مشخص شد با وجود اینکه چنین پیکره‌هایی حجم عظیمی را به خود اختصاص داده‌اند و در داخل بخش‌هایی که بطور عمده فارسی هستند تنها بخش‌های کوچکی با متون عربی مخلوط شده‌اند، اما این بخش‌های عربی همیشه به صورت جملات مستقل ظاهر نشده بلکه در قالب عبارت و یا حتی کلمه تنها نیز به کار رفته-اند. این امر الگوریتم‌های قبلی را که با تکیه بر خود واژه‌ها و آن هم در قالب جملات از قبل از هم تفکیک شده قادر به کار بودند را به چالش کشید. در نهایت الگوریتم تشخیص زبان Ludovik and Zacharski که برای متون چند زبانه طراحی شده است به عنوان چاره کار معرفی شد. این الگوریتم، با تکیه بر گزینشی خاص از ترکیبی خاص از واژه ها و با استفاده از ترکیبی خاص از فنون یادگیری ماشین با ناظر، پس از تقطیع متون چند زبانه به متون تک زبانه می‌تواند زبان آنها را در بخش‌هایی با طول تنها دو کلمه متوسط الی سه کلمه کوتاه تشخیص دهد. الگوریتم فوق که روی متون مخلوط فارسی و عربی نیز امتحان شده است هم در تقطیع متون چند-زبانه به متون تک-زبانه و هم در تشخیص زبان این قطعه‌ها بسیار موفق تر از الگوریتم‌های قبلی گزارش شده است. آنچه حتی در این الگوریتم هنوز حل نشده است تشخیص زبان بخش‌های تک-کلمه‌ای است که با حجمی قابل توجه در پیکره‌های مخلوط فارسی و عربی به کار رفته‌اند. امید است در پژوهش‌های آتی بتوان الگوریتمی طراحی کرد که بتواند الگوریتم Ludovik and Zacharski را تکمیل نماید.

7. پایگاه های on-line برای تشخیص زبان متن

در صورتی که استفاده از وب‌گاه‌های اینترنتی برای کاربر مشکل ساز نباشد و بتواند به نرخ خطای متوسطی بسنده نماید، نشانی‌های اینترنتی ذیل وب‌گاه‌هایی هستند که وظیفه تشخیص زبان متن را، به طور عمده در متونی که تک-زبانه شده‌اند، انجام می‌دهند.

http://rali.iro.umontreal.ca/Silc/index.jsp?lang=fr

http://www.eidetica.com/services/guesser

http://www.fuzzums.nl/~joost/talenknobbel/

http://odur.let.rug.nl/~vannoord/TextCat/Demo/textcat.html

http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser.en.html

http://labs.translated.net/language-identifier/

مراجع

[1] نرم افزار "جامع التفاسیر". مرکز تحقیقات کامپیوتری علوم اسلامی، نور. 1385.

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

[3] Lena Grothe, Ernesto William De Luca and Andreas Nürnberger. A Comparative Study on Language Identification Methods. In *Proceedings of LREC'08*. 2008.

[4] G. Grefenstette. Comparing Two Language Identification Schemes. In *Proceedings of 3rd International Conference on Statistical Analysis of Textual Data*. 1994.

[5] M.J. Martino and R.C. Paulsen. Natural Language Detemination Using Partial Words. In *U.S. Patent No. 6216102 B1*.

[6] W.B. Cavner and J.M. Trenkle. N-gram Based Text Categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*. 1994.

[7] Yevgeny Ludovik and Ron Zacharski. Multilingual Document Language Recognition for Creating Corpora. In *Proceedings of MT Summit VII*. 1999.

[8] Chew Lim Tan, Peck Yoke Leong and Shoujie He. Language Identification in Multilingual Documents. In *Proceedings of International Symposium on Intelligent Multimedia*. 1999.

[9] Yo Ehara and Kumiko Tanaka-Ishii. Multilingual Text Entry Using Automatic Language Detection. In *Proceedings of IJCNLP*. 2008.

[10] محمود شکرالهی فر، مهدی عنایتی، اردلان حسن زاده، بهاره نظرزاده. سامانه تشخیص زبان متون فارسی و عربی. پژوهش منتشر نشده در دانشگاه نبی اکرم (ص).1388.