

فصلنامه علمی - پژوهشی علوم انسانی دانشگاه الزهرا (س)

سال شانزدهم، شماره ۵۸، تابستان ۱۳۸۵

ارایه یک مدل جدید آماری برای شناسایی و تفکیک هوشمند متون مختلف از یکدیگر

دکتر محمد علی صنیعی منفرد^۱

چکیده

کمتر کسی در مورد تفاوت های موجود بین زبان های مختلف تردید دارد، اما آیا می توان این تفاوت ها را مدل سازی نموده و الگویی را استخراج کرد که به کمک آن بتوان یک زبان مورد نظر را از میان سایر زبان ها و بصورتی اتوماتیک شناسایی کرد؟ منظور از مدل سازی، بررسی ساختاری زبان های مختلف در یک چارچوب یکسان و در ارتباط با فیزیک و کالبد کلمات است.

در این مقاله، طول کلمه و توزیع آماری آن را تحلیل کرده و نشان خواهیم داد که شناسایی پنج زبان زنده دنیا با مطالعه طول کلمات آن ها کاملاً امکان پذیر است. این مدل سازی امکانات و تسهیلات جدیدی را در فضای فناوری اطلاعات و بهینه سازی فرآیند های داده کاوی فراهم خواهد آورد.

۱. استادیار دانشگاه الزهرا (س). پست الکترونیک: mas_monfared@alzahra.ac.ir

واژه‌های کلیدی: مدل‌سازی آماری، شناسایی زبان، تشخیص الگو،

هوش مصنوعی، داده کاوی^۳.

مقدمه

مطالعه و تحقیق روی زبان همیشه مورد علاقه محققین رشته‌های مختلف قرار داشته است. به این علاقه در چند دهه اخیر، ابعاد جدیدی اضافه شده است که از جمله شناسایی و ترجمه اتوماتیک از یک زبان به زبان دیگر است. این مطالعات جدید، کاربردهای گسترده‌ای را نیز به وجود آورده است. به عنوان مثال ما می‌خواهیم در فضای اینترنتی بتوانیم متون مورد نظر خود در میان انبوه اطلاعات ارسالی جدا کرده و مورد استفاده قرار دهیم. این کار بدون شناسایی اتوماتیک زبان امکان‌پذیر نمی‌باشد. از طرف دیگر، ترجمه اتوماتیک از یک زبان به زبان دیگر استفاده از مطالب متنوع موجود در فضای اینترنت را بسیار گسترش خواهد.

از طرف دیگر، ما هم‌چنین علاقمندیم امکان گفتگوی بین دو فرد با دو زبان مختلف را به گونه‌ای فراهم آوریم که هر کدام درحالی که به زبان خود سخن می‌گویند بتوانند یکدیگر را درک کنند و این کار مستلزم ترجمه شنیداری اتوماتیک است. با این حال ترجمه شنیداری از ترجمه نوشتاری پیچیده‌تر است، چراکه لهجه‌های متفاوت گویندگان، عادت‌های گفتاری مختلف، بیان احساسی کلمات و جملات، تلفظ‌های اشتباه و نیز تفاوت‌های که به خاطر وسیله انتقال یا کانال ارتباطی ایجاد می‌کند همه موجب پیچیدگی بیشتر می‌شوند. بیان نوشتاری فاقد این همه پیچیدگی است، چراکه معمولاً نوشته‌ها با فرمت صحیح (قابل تفکیک به حروف) در اختیار ما قرار می‌گیرند. البته اگر متنی به صورت دست نوشته باشد مجدداً پیچیدگی زیادی پیدا خواهد کرد. در مجموع، بررسی ادبیات موضوعی نشان می‌دهد که بیشترین حجم تحقیقات بر روی ترجمه شنیداری متمرکز است.

1. Language identification
2. Pattern recognition
3. Data mining

ترجمه از یک زبان به زبان دیگر چه به صورت نوشتاری و چه به صورت شنیداری آن در مقایسه با موضوع شناسایی اتوماتیک زبان‌ها حد اعلای پیشرفت و توسعه در این رشته را تعقیب می‌نماید. این درحالی است که شناسایی اتوماتیک زبان‌ها به مراتب ساده‌تر از ترجمه زبان‌هاست. با این وجود، علی‌رغم بیش از چهار دهه مطالعه و تحقیق هنوز راه طولانی در پیش است و تحقیقات زیادی در حوزه شناسایی زبان‌ها باید به‌انجام برسد.

پیشینه

برای شناسایی اتوماتیک زبان نیاز داریم نوشته مورد نظر به صورت حروف قابل تفکیک در اختیار ما قرار گیرد. دو حالت متصور است یا این که متن مورد نظر ما برای شناسایی به صورت کامپیوتری موجود است که در این صورت چنین نوشته‌ای خودبه‌خود فرمت مورد نظر ما را دارد، یعنی قابل تفکیک به حروف می‌باشد. اما اگر متن مورد نظر در کتاب یا روزنامه قرار دارد و ما فقط می‌توانیم تصویری از آن را در اختیار بگیریم، آنگاه باید روی این تصویر تبدیلی انجام گیرد. این تبدیل در واقع تبدیل نوشته تصویری به نوشته‌ی قابل تفکیک به حروف است. چنین تبدیلی به کمک فناوری شناسایی نوری حروف انجام می‌پذیرد. با این حساب، در روشی که ما در این مقاله ارائه می‌نماییم اطلاعات خام مورد نیاز آن نوشته‌های قابل تفکیک به حروف هستند که به هر یک از دو طریقه بالا به‌دست آمده باشد قابل قبول خواهد بود.

اولین تحقیق انجام گرفته در این حوزه به کار راثو در پایان نامه کارشناسی ارشد خود برمی‌گردد (راثو، ۱۹۷۴). راثو ابتدا توزیع احتمالی را برای هر حرف به‌دست آورد و سپس برای کلمات که در واقع از ترکیب حروف به‌دست می‌آیند توزیع مشترک را محاسبه کرد. راثو روش خود را برای شناسایی زبان انگلیسی و اسپانیایی آزمایش کرد و نتایج آزمایشات خود را در پایان‌نامه خود ارائه نمود.

با توجه به این که فراوانی سه حرفی‌ها در زبان‌های مختلف متفاوت هستند اشمیت از این سه حرفی‌ها برای تمایز و تشخیص زبان‌های مختلف استفاده کرد. بدین صورت که از متن مورد بررسی سه حرفی‌های آن استخراج نموده و فراوانی نسبی سه حرفی‌ها را برای مقایسه زبان‌ها به کار

برد. با این که اشمیت روش خود را به ثبت رساند اما نتایج به کارگیری این روش در منابع علمی گزارش نگردیده است (اشمیت، ۱۹۹۱). رساله‌های کارشناسی ارشد و دکتری زیادی در حوزه زبان‌شناسی زبان‌ها انجام گرفته و می‌گیرد از جمله گیزت تفکیک جمله‌ها در زبان‌های مختلف را با استفاده از اطلاعات زبان‌شناسی و دانش آماری بررسی کرد و نتایج تحقیقات خود را در کنفرانس‌هایی از جمله در پراگ چکسلواکی ارایه نمود. (گیزت، ۱۹۹۵) کوان و هیروس در ۱۹۹۷ روشی برای شناسایی زبان‌های غریبه پیشنهاد نمودند که به کمک آن متون غریبه فیلتر شوند. (کوان و هیروس، ۱۹۷۷) آنها از روش شبکه‌های عصبی فیدبک‌دار برای پیاده‌سازی استفاده نمودند. (صنعی منفرد، ۱۳۷۳) هاریک و اوهرلر در ۱۹۹۹ زوی شناسایی زبان با استفاده از گراماژ یا طول کلمات مطالعاتی را انجام داد (هاریک و اوهرلر، ۱۹۹۹).

روش دیگری که برای مدل کردن زبان به کار گرفته شده است استفاده از مدل‌های احتمالی یا استوکاستیکی به خصوص زنجیره‌های مارکوفی است. (صنعی منفرد، ۱۳۸۰) در این روش احتمالات انتقال از یک حرف به حرف دیگر محاسبه می‌شود و در صورتی که این احتمالات مستقل از پیشینه حروف باشد مارکوفی خواهد بود. در زبان انگلیسی احتمال این که حرف S با حرف t دنبال شود بستگی به این دارد که قبل از S چه حرفی آمده است. به عنوان نمونه دو حرف es خیلی زیاد قبل از t می‌آید در حالی که دو حرف ds به ندرت قبل از t مشاهده می‌شود. حال اگر توالی حروف در کلمات انگلیسی از قانون زنجیره‌های مارکوفی تبعیت نماید باید آمدن S قبل از t در هر دو حالت es و ds با احتمال برابری اتفاق می‌افتاد که البته این طور نیست. (اشبی، ۱۹۵۶) از جمله می‌توان از کار ماتروف و همکارانش در ۱۹۹۸ (ماتروف و همکاران، ۱۹۹۸) و کریشهف و همکارانش در ۲۰۰۲ نامبرد (کریشهف و همکاران، ۲۰۰۲).

علی‌رغم انجام این تحقیقات هنوز کار زیادی روی شناسایی نوشتاری زبان‌ها باید انجام گیرد، به خصوص با توجه به این که بیشتر مطالعات موجود روی زبان‌های لاتینی انجام شده است، مانند کار داماشک در ۱۹۹۵ (داماشک، ۱۹۹۵) و یا کار هافمن در سال ۲۰۰۰ (هافمن، ۲۰۰۰) زیر عنوان الگوریتم آشنایی که قادر است زبان‌هایی مانند هلندی، انگلیسی، استونیایی، ایرلندی، لهستانی، پرتغالی و باسکی را تفکیک نماید. هم‌چنین روی زبان‌های غیر لاتینی مانند ژاپنی

مطالعه‌ای توسط ماتسورا (ماتسورا، ۲۰۰۰) در سال ۲۰۰۰ انجام گرفته است. با این حال روی زبان‌هایی که نوشتار آن‌ها با استفاده از سیستم الفبایی نیست مانند نوشتارهای چینی مطالعه‌ای انجام نگرفته است. ما در این مقاله برای اولین بار دو گروه زبان را با استفاده از یک مدل آماری مورد مطالعه قرار می‌دهیم. گروه اول، زبان‌هایی با الفبای لاتین شامل انگلیسی، فرانسوی و ترکی استانبولی هستند و مطالعه این‌ها دارای سابقه در ادبیات موضوعی است. از طرف دیگر، گروه دوم زبان‌های مورد بررسی در این تحقیق زبان‌هایی با الفبای متفاوت فارسی و عربی هستند که نه بصورت جداگانه و نه در کنار زبان‌های لاتینی، بخصوص در چارچوب یک مدل، مطالعه نشده‌اند. سازمان این مقاله برتیب‌زیر خواهد بود. در بخش اول این مقاله روش کار برای بررسی کالبدی زبان‌های نوشتاری را معرفی و سپس نتایج آماری به دست آمده روی ۵ متن نمونه را در بخش دوم ارائه می‌کنیم. در بخش سوم الگوریتم شناسایی زبان را ارائه و بررسی خواهیم کرد و سپس به نتیجه‌گیری می‌پردازیم.

بخش اول: روش مطالعه

کلمه‌ای مانند «مادر» در زبان فارسی دارای چهار حرف «م»، «ا»، «د» و «ر» است. ما می‌گوییم مادر در زبان فارسی دارای طولی برابر ۴ است. همین کلمه در زبان انگلیسی (mother) ۶ حرف، در عربی (ام) ۲ حرف، در ترکی (آنا) ۳ حرف و در فرانسوی (mere) ۴ حرف دارد. بدین ترتیب کلمه مادر در پنج زبان بالا برتیب دارای طولی برابر پنج، شش، دو، سه و چهار حرف است. هم چنین، به جمله زیر توجه کنید:

بعد از پایان جنگ جهانی دوم شهری که در قلب آلمان شرقی قرار داشت به چهار قسمت تقسیم شد، مناطق تحت حفاظت فرانسه، انگلیس، آمریکا و شوروی.

این جمله را به کلمات سازنده آن خرد می‌کنیم. مثلاً «بعد»، «آز»، «پایان»، «جنگ»، «جهانی» پنج کلمه اول این جمله را می‌سازند. ما در روش خود فقط با طول این کلمات سروکار داریم و این که این کلمات از چه حروفی ساخته شده‌اند مورد علاقه ما نیستند. با این حساب کلمه «بعد» دارای

۳ حرف و کلمه «از» دارای ۲ حرف است و به همین ترتیب سه کلمه دیگر دارای ۵ حرف، ۳ حرف و ۵ حرف هستند. اینک می توانیم جمله بالا را بر حسب طول کلمات و فراوانی آنها مرتب می کنیم چنانکه در جدول ۱ نشان داده شده است.

جدول ۱

فراوانی	کلمات مورد نظر در جمله	طول کلمه به حرف
۱	و	یک
۵	از، که، در، به شد	دو
۵	بعد، جنگ، دوم، قلب، تحت	سه
۶	شهری، شرقی، چهار، قرار، داشت، قسمت	چهار
۷	پایان، جهانی، آلمان، تقسیم، مناطق، حفاظت، شوروی	پنج
۳	فرانسه، انگلیس، آمریکا	شش

منظور از فراوانی در جدول ۱ تواتر یا تعداد تکرار کلمات در جمله هستند. به همین ترتیب در جمله «I have been trying to cure it» ما هفت کلمه داریم که به ترتیب دارای طول ۱، ۴، ۴، ۶، ۲، ۴، ۲ حرف است. با این حساب اگر بخواهیم این جمله کوتاه را به صورت توزیع فراوانی نمایش دهیم مطابق با جدول ۲ داریم:

جدول ۲

فراوانی	نام کلمات	طول کلمه به حروف
۱	I	یک
۲	It, to	دو
۰	-	سه
۳	Have, been, cure	چهار
۰	-	پنج
۱	Trying	شش

با این حساب جمله انگلیسی بالا دارای توزیعی متفاوت با جمله فارسی نشان داده شده در جدول ۱ است. همین روش را برای سایر زبان‌های مورد مطالعه (ترکی استانبولی، فرانسوی و

عربی) به کار می‌گیریم. توجه داریم که فاصله فیزیکی بین کلمات در هر نوع متن نشان دهنده شروع و ختم کلمات هستند و نیازی به دانستن زبان برای پیش‌برد این تجربه و مطالعه آماری وجود ندارد.

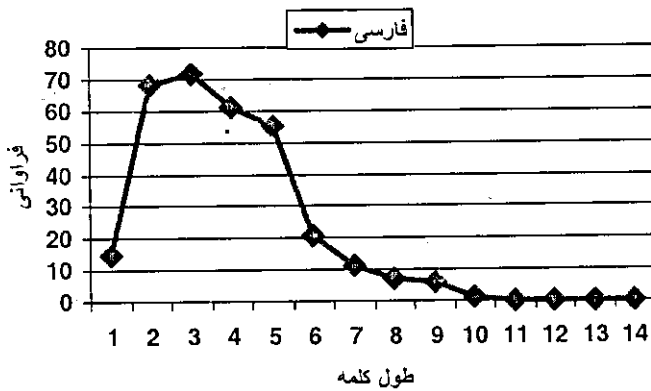
بخش دوم: نتایج آماری به‌دست آمده

در این قسمت نتایج بررسی‌ها انجام شده روی پنج متن انتخاب شده را نشان می‌دهیم. نتایج را در جدول شماره ۳ مشاهده می‌نمایید. به‌عنوان نمونه در متن عربی که ۵۰۲ کلمه وجود دارد ۵۴ کلمه یک حرفی و ۸۵ کلمه دو حرفی وجود داشت. به‌همین ترتیب برای بقیه کلمات، به شرحی که در جدول ۳ مشاهده می‌گردند.

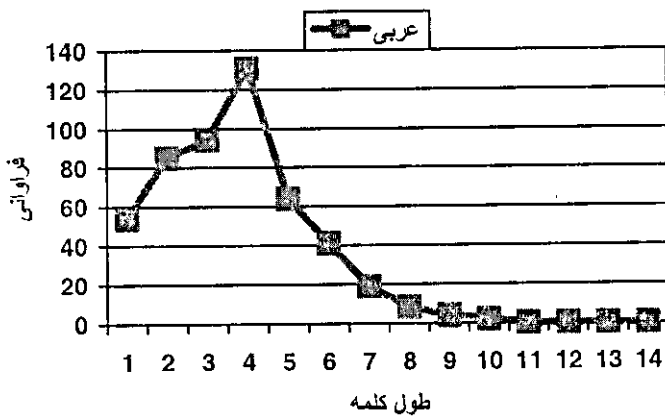
جدول ۳

فراوانی کلمه در متن مورد بررسی $f(x)$					طول کلمه
ترکی	فرانسه	انگلیسی	عربی	فارسی	X
۰	۲۳	۱۵	۵۴	۱۵	۱
۲۷	۱۱۶	۹۰	۸۵	۶۸	۲
۳۰	۶۴	۱۴۴	۹۴	۷۲	۳
۴۶	۴۶	۷۸	۱۳۱	۶۱	۴
۴۱	۳۹	۶۵	۶۴	۵۵	۵
۵۰	۵۰	۴۵	۴۱	۲۱	۶
۶۵	۴۰	۳۰	۱۹	۱۱	۷
۵۰	۲۳	۲۲	۸	۷	۸
۳۱	۱۶	۵	۴	۶	۹
۲۲	۴	۰	۲	۱	۱۰
۲۶	۴	۱	۰	۰	۱۱
۷	۴	۱	۰	۰	۱۲
۳	۰	۰	۰	۰	۱۳
۲	۰	۰	۰	۰	۱۴
۴۰۰	۴۲۹	۴۹۶	۵۰۲	۳۱۷	مجموع کلمات

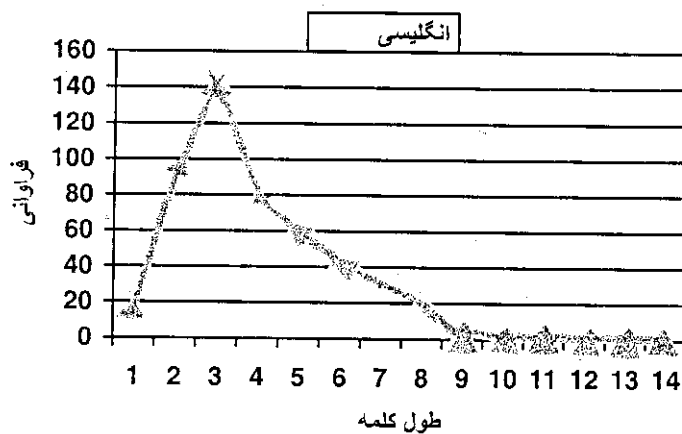
توزیع آماری مربوط به طول کلمات که نشان دهنده تفاوت های ساختار زبان های مورد مطالعه هستند را در اشکال ۱ الی ۵ نشان داده ایم.



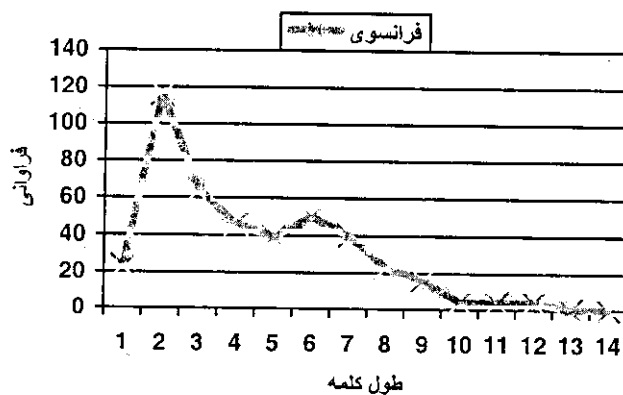
شکل ۱. تابع توزیع طول کلمات فارسی



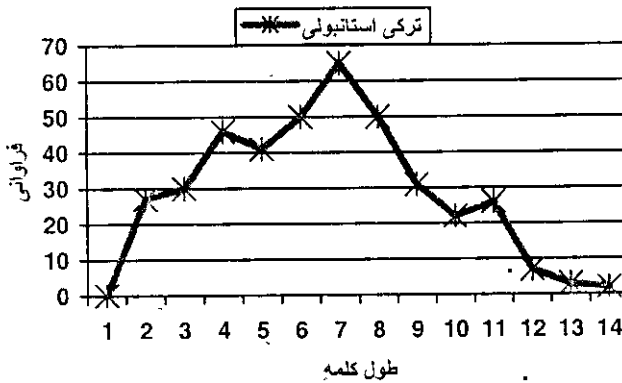
شکل ۲. تابع توزیع طول کلمات عربی



شکل ۳. تابع توزیع طول کلمات انگلیسی



شکل ۴. تابع توزیع طول کلمات فرانسه



شکل ۵. تابع توزیع طول کلمات ترکی استانبولی

تفاوت‌ها و تمایزات مشاهده شده در بین این توزیعات فراوانی امیدوارکننده هستند، به این معنا که می‌توان توزیع طول کلمات در یک زبان را به‌عنوان یکی از ویژگی‌های منحصر به فرد آن زبان در نظر گرفت. چگونگی بهره‌برداری از این ویژگی در شناسایی نوشتاری یک زبان موضوع بخش بعدی این مطالعه خواهد بود.

بخش سوم: تحلیل آماری و طراحی الگوریتم شناسایی زبان

مشاهده اشکال ۱ الی ۵ به‌خوبی تفاوت‌های رفتاری بین توابع توزیع طول کلمات در زبان‌های مختلف را نشان می‌دهد. با این حال، چنین قضاوتی در حال حاضر یک کیفی است در حالی که شناسایی اتوماتیک زبان‌ها نیاز به کمسازی و عددی کردن این قضاوت دارد. با این حساب، لازم است خواص کمی توابع توزیع طول کلمات را مورد مطالعه بیشتر قرار دهیم.

در این رابطه، ابتدا تمام توابع توزیع فراوانی طول کلمات را به‌صورت توابع توزیع احتمالی در می‌آوریم. در یک تابع توزیع احتمالی، جمع احتمالات روی حالات مختلف برابر با یک خواهد شد و در این صورت خواهیم توانست برخی از خواص کمی را برای آن‌ها محاسبه کنیم. از جمله این خواص کمی محاسبه میانگین، نما، میانه و انحراف معیار برای طول کلمات است. مجموعه این

محاسبات ما را قادر به ساختن پایگاه دانشی^۱ برای پنج زبان مورد نظر می نماید که بکسک این پایگاه دانش امکان ساخت الگوریتم شناسایی اتوماتیک زبان ها به وجود می آید. این پایگاه دانش را در جدول ۴ مشاهده نمایید.

این جدول در واقع هسته اصلی سازنده الگو یا مدل ما برای ارزیابی و تشخیص زبانهای مختلف خواهد بود. با این حساب، اگر متنی بصورت مجهول در اختیار ما قرار داده شود ما برای این متن محاسباتی را انجام می دهیم و نتیجه محاسبات را با نتایج معلوم موجود در جدول ۴ تطبیق می دهیم و مشخص خواهیم ساخت که متن مجهول ما به کدام یک از ۵ زبان موجود نزدیک تر است. توجه دارید که نمی توانیم با اطمینان کامل و به صورت یقینی ادعای شناسایی یک زبان را داشته باشیم زیرا مساله مورد مطالعه ما اساساً مساله ای غیردقیق و از مقولات احتمالی است. با این حال، انتظار داریم مدل ما بتواند نزدیک ترین زبان را شناسایی کرده و مثلاً بگوید: این متن مجهول انگلیسی است. انجام این کار نیازمند طراحی یک الگوریتم برای شناسایی خودکار زبانها است. اینک فرض کنید متن مجهول ما به صورت زیر است:

Tüm kullanma suyu tesisati düz olarak ve kullanma yerine doguk borular alta gelecek sekilde dösen melidir.

ما البته می دانیم این متن ترکی است اما می خواهیم بدانیم با توجه به الگوی موجود در جدول ۴ چگونه می توانیم آن را شناسایی نماییم. قاعدتاً در قدم اول ما باید توزیع احتمالی و خواص آماری این عبارت را محاسبه کنیم، چنانچه در جدول ۵ مشاهده می نمایید.

جدول ۴. پایگاه داده‌ها مربوط به زبان

تعداد حرف در کلمه	فارسی	عربی	انگلیسی	فوانسه	ترکی
۱	۰،۰۴۷	۰،۱۰۷	۰،۰۳۰	۰،۰۵۴	۰
۲	۰،۲۱۴	۰،۱۶۹	۰،۱۸۱	۰،۲۷۰	۰،۰۶۸
۳	۰،۲۲۷	۰،۱۸۷	۰،۲۹۰	۰،۱۴۹	۰،۰۷۵
۴	۰،۱۹۲	۰،۲۶۱	۰،۱۵۷	۰،۱۰۷	۰،۱۱۵
۵	۰،۱۷۳	۰،۱۲۷	۰،۱۳۱	۰،۰۹۱	۰،۱۰۳
۶	۰،۰۶۶	۰،۰۸۲	۰،۰۹۱	۰،۱۱۷	۰،۱۲۵
۷	۰،۳۵۰	۰،۰۳۸	۰،۰۶۰	۰،۰۹۳	۰،۱۶۳
۸	۰،۰۲۲	۰،۰۱۶	۰،۰۴۴	۰،۰۵۴	۰،۱۲۵
۹	۰،۰۱۹	۰،۰۰۸	۰،۰۱۰	۰،۰۲۷	۰،۰۷۸
۱۰	۰،۰۰۳	۰،۰۰۴	۰	۰،۰۰۹	۰،۰۵۵
۱۱	۰	۰	۰،۰۰۲	۰،۰۰۹	۰،۰۶۵
۱۲	۰	۰	۰،۰۰۲	۰،۰۰۹	۰،۰۱۸
۱۳	۰	۰	۰	۰	۰،۰۰۸
۱۴	۰	۰	۰	۰	۰،۰۰۵
میانگین	۳،۸۱۴	۳،۶۸۳	۴،۰۰۸	۴،۳۵۰	۶،۵۶۰
نما	۳	۴	۳	۳	۷
میانه	حدود ۳	حدود ۳	۳	۳	۶
انحراف معیار	۳،۱۶۷	۳،۱۲۵	۳،۵۱۶	۶،۱۶۷	۷،۶۸۶

جدول ۵. محاسبات یک متن مجهول

طول کلمه	فراوانی مشاهده شده	احتمال مشاهده شده
۱	۰	۰
۲	۱	$1 \div 16 = 0.063$
۳	۲	$2 \div 16 = 0.125$
۴	۲	$2 \div 16 = 0.125$
۵	۲	$2 \div 16 = 0.125$
۶	۲	$2 \div 16 = 0.125$
۷	۴	$4 \div 16 = 0.25$
۸	۳	$3 \div 16 = 0.188$
۹	۰	۰
۱۰	۰	۰
۱۱	۰	۰
۱۲	۰	۰
۱۳	۰	۰
۱۴	۰	۰
جمع	۱۶	۱
میانگین		۵/۶۲۵
نما		۷
میانه		۵.۵
انحراف معیار		۱/۹۶۲

اینک که متن مجهول را به صورت آماری ترجمه کرده ایم می توانیم بگویم که متن نوشتاری ما معادل یک بردار ۱۸ بعدی است. (۱۴ بعد مربوط به طول کلمه و ۴ بعد دیگر مربوط به میانگین،

نما، میانه و انحراف معیار است) این بردار ۱۸ بعدی را که کافی است با الگوهای موجود که آنها هم ۱۸ بعدی هستند (مطابق با جدول ۴) مطابقت داده، یا به زبان ساده تر تفریق کنیم تا بدین ترتیب بردار خطای ناشی از مقایسه ها را به دست آوریم و آنگاه بگوییم هر کدام خطای کمتری داشتند نزدیک ترین زبان به متن مجهول ما هستند. جدول ۶ این محاسبات را نشان می دهد.

جدول ۶. محاسبه خطای متن مجهول نسبت به پایگاه داده ها

خطا نسبت به عربی	خطا نسبت به فارسی	خطا نسبت به انگلیسی	خطا نسبت به فرانسه	خطا نسبت به ترکی	
۰/۸۵۲	۰/۸۷۷	۰/۷۳۳	۰/۷۰۰	۰/۴۶۶	روی تابع Ep توزیع
۸/۶۰۵	۹/۵۱۶	۹/۶۷۱	۱۱/۹۸۰	۷/۱۵۹	روی چهار مشخصه آماري Ec
۹/۴۵۷	۱۰/۳۹۳	۱۰/۴۰۴	۱۲/۶۸	۷/۶۲۵	مجموع خطا E

$$E_p = \sum_{i=1}^{14} |P_0^i - P_m^i|,$$

$$E_{\bar{x}} = |\bar{X}_0 - \bar{X}_m|,$$

$$E_{\hat{x}} = |\hat{X}_0 - \hat{X}_m|,$$

$$E_{\tilde{x}} = |\tilde{X}_0 - \tilde{X}_m|,$$

$$E_{\sigma} = |\sigma_0 - \sigma_m|,$$

$$E_c = E_{\bar{x}} + E_{\hat{x}} + E_{\tilde{x}} + E_{\sigma}$$

$$E = E_p + E_c$$

که در آن E_p مجموع خطای مطلق ناشی از تفاوت توزیع تجربی (یا مشاهده شده از روی نمونه) و توزیع الگو می‌باشد. E_C مجموع خطای روی میانگین، نما، میانه و انحراف معیار است. به این ترتیب p'_0 احتمال مشاهده شده و p'_m احتمال مطابق با الگو است. \bar{X}_0 میانگین مشاهده شده، \hat{X}_0 نمای مشاهده شده، \bar{X}_0 میانه مشاهده شده و σ_0 انحراف معیار مشاهده شده را نشان می‌دهد. به همین ترتیب \bar{X}_m میانه الگو و σ_m انحراف معیار الگو را نشان می‌دهد.

محاسبات انجام شده در جدول ۶ نشان می‌دهد که متن مجهول ما از نظر اندازه خطا روی تابع توزیع (E_p) روی زبان ترکی دارای کمترین مقدار است ($E_p = 0/466$). همچنین از نظر اندازه خطا روی چهار مشخصه آماری (E_C) هم با زبان ترکی کمترین خطا را نشان می‌دهد ($E = 7/625$). با این حساب نزدیک‌ترین زبان به متن مجهول زبان ترکی است.

دقت روی جدول ۴ نشان می‌دهد که زبان ترکی تنها زبانی است که بیشترین اندازه نما را دارد ($\hat{X} = 7$). با این حساب مطابق با جدول ۵ خیلی زود معلوم می‌شود که نمای زیاد در متن مجهول نشان‌دهنده ترکی بودن متن مورد نظر است. این تفاوت و تمایز بارز می‌تواند کار شناسایی زبان را برای ما آسانتر نماید چراکه ما قادر هستیم با سرعت بیشتری زبان متن مجهول را کشف کنیم. اما آیا این روش ساده همیشه جوابگو است؟ آیا پاسخ به این سوال بستگی به اندازه متن مجهول ما یا تعداد کلمات موجود در متن ندارد؟ این سوالی است که با انجام آزمایش قابل ارزیابی است.

نتایج انجام آزمایشات مختلف منجر به تدوین الگوریتمی گردید که در شکل ۶ نشان داده شده است. استفاده از این الگوریتم برای متن مجهول به سادگی نشان می‌دهد که متن مجهول متنی ترکی است زیرا اندازه نما برای متن مجهول $\hat{X} = 7.0$ است و مطابق با الگوریتم اگر نما بزرگتر از $5/5$ باشد (یا $\hat{X} > 5.5$) حتماً زبان متن ترکی است و بنابراین مطابق با الگوریتم پیشنهادی ما نیازی به بررسی بیشتر نخواهیم داشت.

اما آیا این الگوریتم همان‌طور که از عهده متن خیلی کوتاه ترکی برمی‌آید می‌تواند از عهده شناسایی سایر زبان‌ها هم برآید؟ این کار نیاز به بررسی و انجام آزمایشات بیشتر دارد. اینک برای

بررسی قدرت الگوریتم پیشنهاد شده چهار نمونه انگلیسی، عربی، فرانسوی و فارسی را مطابق با جدول ۷ مورد بررسی و آزمایش قرار می‌دهیم. هدف ما این است که نشان دهیم آیا الگوریتم ما قادر به شناسایی اتوماتیک این چهار نمونه هم هست یا خیر و اگر خیر دلیل آن چیست.

جدول ۷. نمونه‌های جدید

نمونه متن	زبان
Oh Mother, if I am ever half as good as you, I shall be satisfied.	انگلیسی
کیف تعرف الطيور وطنها و مقصدها و کیف لاتصل الطريق فی هذه المسافات الطويلة؟	عربی
Une pluie fine commence a tomber. La grisaille du ceil pese sur la ville pourtant toujours aussi active.	فرانسوی
البته باید مشخص شود که کمیته در چه سطحی قصد برخورد با این سؤالات را دارد و در این زمینه چگونه فعالیت‌های خود را شکل خواهد داد.	فارسی

برای استفاده از الگوریتم طراحی شده لازم است توابع توزیع فراوانی و مشخصه‌های آماری مورد نیاز را برای چهار متن جدید محاسبه کنیم. نتایج محاسبات مورد نیاز برای فعال کردن الگوریتم را در جدول ۸ نشان داده‌ایم.

حالا اگر الگوریتم خود را فعال کنیم مشاهده خواهیم کرد که به جز زبان فارسی الگوریتم ما در مورد زبان انگلیسی، عربی و فرانسوی نتایج نادرستی را بیار می‌آورد. زیرا براساس معیار E زبان فارسی کمترین خطا را دارد و بنابراین شناسایی به‌خوبی انجام گرفته است. اما در مورد زبان انگلیسی به اشتباه فارسی انتخاب شده و در مورد زبان عربی به اشتباه انگلیسی انتخاب شده و در مورد زبان فرانسوی به اشتباه فارسی انتخاب شده است.

جدول ۸. توزیع فراوانی مشاهده شده از نمونه‌های جدید

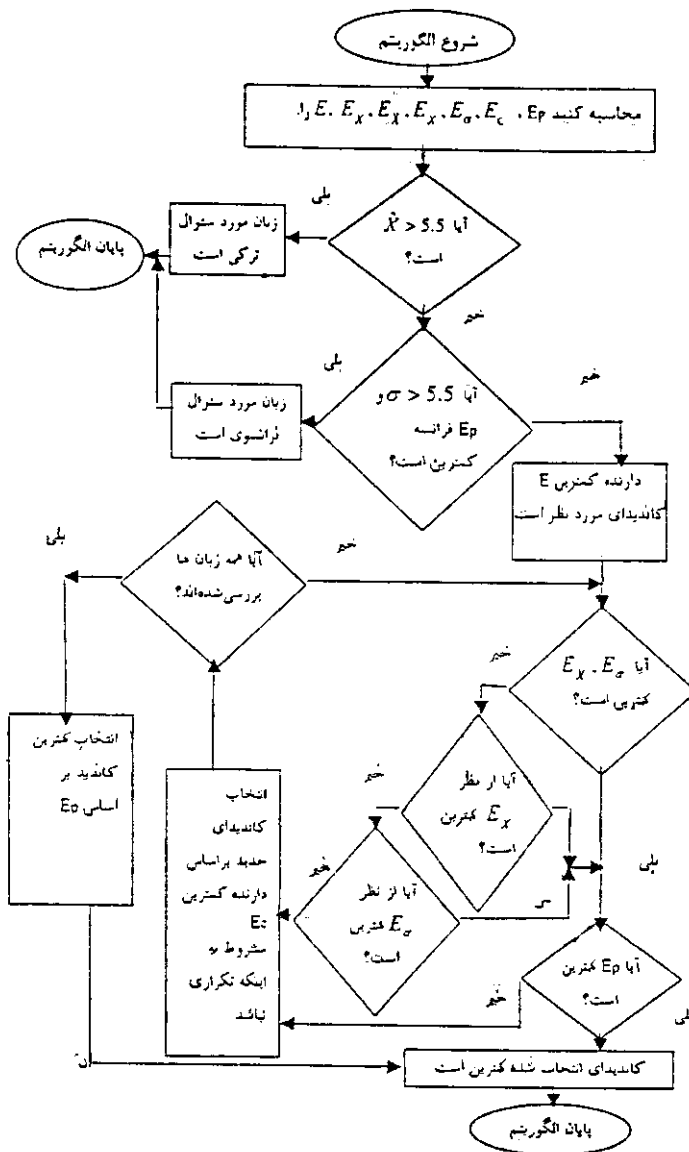
طول کلمه	نمونه انگلیسی	نمونه عربی	نمونه فرانسوی	نمونه فارسی
۱	۴	۹	۱	۱
۲	۱۳	۹	۱۳	۷
۳	۶	۹	۴	۷
۴	۱۴	۸	۶	۴
۵	۸	۹	۶	۵
۶	۴	۱۰	۹	۲
۷	۱	۵	۶	۰
۸	۰	۲	۴	۰
۹	۱	۰	۲	۱
۱۰	۰	۰	۳	۰
۱۱	۰	۰	۰	۰
۱۲	۰	۰	۰	۰
۱۳	۰	۰	۱	۰
جمع	۵۱	۶۱	۵۶	۲۷
میانگین	۳,۶۰۸	۳,۶۹۷	۵,۱۶۱	۳,۶۳۰
نما	۴	۶	۲	۲/۵
میانه	۴	۴	۵	۳
انحراف معیار	۱,۶۹۸	۲,۵۴۹	۲,۶۸۹	۱,۷۳۵

برای بررسی اشتباه در تشخیص لازم است تفصیل محاسبات را مورد توجه قرار دهیم. این کار را در جدول ۹ نشان داده‌ایم. در این جدول مشاهده می‌گردد که انتخاب نزدیک‌ترین زبان براساس این که چه معیاری (E، EC، یا Ep) مبنا قرار دارد نتایج متفاوتی ببار خواهد آورد. مثلاً اگر زبان متن مجهول ما واقعاً انگلیسی است و معیار ما Ep باشد آنگاه ما براساس این معیار به این نتیجه می‌رسیم که نزدیک‌ترین زبان عربی است در حالی که این اشتباه است. این نشان می‌دهد که معیار Ep به تنهایی برای تشخیص درست کافی نیست.

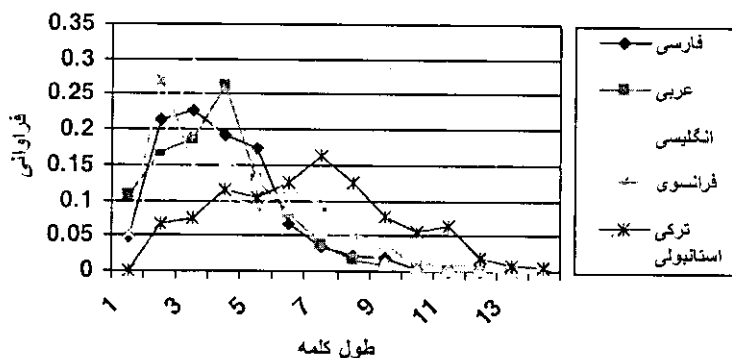
جدول ۹.

زبان متن مورد نظر	اگر انتخاب زبان متن مجهول بر مبنای کمترین خطای Ep باشد	اگر انتخاب زبان متن مجهول بر مبنای کمترین خطای EC باشد	اگر انتخاب زبان متن مجهول بر مبنای کمترین خطای E باشد
انگلیسی	۰،۶۳۱	۳،۵۶	۴،۲۱۵
	به اشتباه عربی انتخاب می‌شود	به اشتباه فارسی انتخاب می‌شود	به اشتباه فارسی انتخاب می‌شود
عربی	۰،۴۷۹۳	۱،۱۴۹	۱،۷۹۹
	انگلیسی	فارسی	انگلیسی
فرانسوی	۰،۴۶۷۷	۲،۶۷۹	۳،۱۴۷
	فارسی	فارسی	فارسی
فارسی	۰،۲۲۹۶	۲،۱۱۶	۲،۳۴۶
	فارسی	فارسی	فارسی

اما آیا اشتباه ناشی از فرض بنیادی ما در طراحی الگوریتم است که براساس آن تفاوت بین زبان‌های مختلف را در قالب تفاوت بین توابع توزیع فراوانی روی طول کلمات مدل کرده‌ایم؟ نگاهی مجدد که به این توابع مطابق با شکل ۷ نشان می‌دهد که تفاوت بین توابع توزیع بارز و قابل توجه است و این تفاوت می‌تواند مبنای تمایز و شناسایی قرار گیرد. اگر این قضاوت درست است آنگاه اشتباه در شناسایی فقط می‌تواند ناشی از کوتاه بودن اندازه متون مورد بررسی باشد.



شکل ۶. الگوریتم شناسایی زبان



شکل ۷. توزیع طول کلمات در زبان های مختلف بر اساس پایگاه دانش

برای آزمایش این وضعیت اندازه متون انگلیسی، عربی و فرانسوی را افزایش می دهیم. (نگاه کنید به پیوست مقاله) نتایج نهایی را در جدول ۱۰ آورده ایم و از آوردن محاسبات اولیه (توابع توزیع مشاهده شده) برای جلوگیری از طولانی شدن مقاله خودداری می کنیم.

جدول ۱۰.

زبان متن مورد نظر	اگر انتخاب زبان متن مجهول بر مبنای کمترین خطای E _p باشد	اگر انتخاب زبان متن مجهول بر مبنای کمترین خطای E _c باشد	اگر انتخاب زبان متن مجهول بر مبنای کمترین خطای E _e باشد
انگلیسی	۰،۷۰۴۸	۲،۵۰۲	۳،۲۰۶
عربی	فرانسه	عربی	عربی
	۰،۷۵۲۱	۳۸۶	۴،۶۱۴
فرانسوی	فارسی	عربی	عربی
	۰،۳۲۴	۴۸۲	۵،۴۸
فارسی	فرانسه	فارسی	فارسی
	۰،۲۲۹۶	۲،۱۱۶	۲،۳۴۶
	فارسی	فارسی	فارسی

در این آزمایش مشاهده می‌گردد که علاوه بر زبان فارسی زبان فرانسوی اگر E_p معیار باشد و زبان عربی اگر E معیار باشد به درستی شناسایی شده‌اند. با این حساب الگوریتم ما قادر به تشخیص سه زبان از میان چهار زبان مورد آزمایش است. این درحالی است که متن انگلیسی مجهول ما براساس معیار E_p فرانسوی و براساس معیار E با عربی اشتباه گرفته می‌شود. این نشان می‌دهد که اندازه متن مجهول ما هنوز باید افزایش داده شود. توجه دارید که اندازه متن‌های انتخابی ما کمتر از ۵۰ کلمه هستند درحالی که یک متن یک صفحه‌ای با ۲۴ سطر حداقل ۴۰۰ کلمه دارد و لزوماً استفاده از چنین متنی دقت شناسایی را بشدت افزایش خواهد داد. ارزیابی رابطه بین دقت و اندازه متن موضوعی است که نیاز به ادامه تحقیق و انجام آزمایشات بیشتر دارد.

نتیجه‌گیری

در این مقاله نشان داده شد که خصوصیات آماری طول کلمه ما را قادر به تفکیک و شناسایی زبان‌های مختلف می‌نماید. پنج زبان فارسی، عربی، انگلیسی، فرانسوی و ترکی استانبولی در این مطالعه مورد توجه قرار گرفت و الگوریتمی طراحی شد که به کمک آن شناسایی اتوماتیک زبان امکان‌پذیر گردد. استحکام و دقت الگوریتم به این نکته بر می‌گردد که هیچ دو زبانی دارای تابع توزیع مشابهی نیستند، چنانچه در شکل ۷ به خوبی مشاهده گردید. این تفاوت و تمایز کاملاً در این مقاله مورد بهره برداری قرار گرفت و با انجام چندین آزمایش نشان داده شد که الگوریتم پیشنهادی به خوبی قادر به شناسایی زبان‌ها از یکدیگر است. البته هر چه تعداد کلماتی که در متن نمونه (مجهول) قرار دارد بیشتر شود دقت و سرعت کشف زبان هم بالاتر می‌رود.

نتایج این تحقیق را می‌توان از جهات مختلف مورد مطالعه بیشتر قرار داد:

۱. ارزیابی رابطه بین دقت و اندازه متن مجهول
۲. ارزیابی سرعت یا کارایی الگوریتم روی متون متنوع از منابع مختلف

۳. افزایش دقت الگوریتم با بکارگیری مشخصه‌های آماری پیچیده‌تری مانند ضریب چولگی^۱، ضریب کشیدگی^۲ و گشتاورهای^۳ درجات بالاتر و نیز استفاده از خواص تابع توزیع تجمعی^۴
۴. تحلیل خطای الگوریتم با تغییر متون مرجع مانند تغییر از متون ادبی کلاسیک به متون روزمره
۵. افزایش تعداد زبان‌ها

قدردانی

لازم می‌دانم از داوران محترم این مقاله که با طرح چالش‌ها و راهنمایی‌های ارزشمندشان به تبیین اهداف این تحقیق کمک بسزایی نمودند صمیمانه تشکر و قدردانی نمایم. هم‌چنین مایلیم از خانم فتحی، دانشجوی درس آمار مهندسی خود، که آزمایشات اولیه مربوط به پنج زبان را با گزینش متون مختلف به انجام رساندند و آقای کیانوش روایی که با انجام برنامه‌نویسی‌های لازم امکان انجام آزمایشات و محاسبات آماری مربوط به الگوریتم را فراهم آوردند تشکر و قدردانی نمایم.

منابع

- Beasley, K. R. (1988), "Language Identifier: A Computer program for automatic natural language identification on on-line text", *Proceedings of the 29th Annual Conference of American translator Association*, pp 47-54.
- Damashek, M. (1995), "Gauging Similarity with n-Grams: Language Independent Categorization of Text", *Science*, Vol. 267, pp 843-848.

1. Skewness
2. Kurtosis
3. Moments
4. Cumulative distribution function

- Giguet, E. (1995), "Categorization according to Language: A step toward combining Linguistic Knowledge and Statistic Learning", *Proceeding of the International Workshop of Parsing Technologies*, Prague: September.
- Harbech. S. and U. Ohler, (1999), "Milligrams for Language Identification", *Proceeding of the 6th European Conference on Speech Communication and Technology (Eurospeech 99)* Budapest: September.
- Huffinan, S.(2000), "The Acquaintance algorithm", GLBH@email.msn.com
- Kirchhoff, K. Parondekar, S. and J. Bilmes, (2002), "Mixed – memory markou models for Automatic Language Identification", *Proceeding of ICASSP 2002*, Orlando.
- Kwan, H. K. and K. Hirose, (1997), "Use of Recurrent Network for unknown language Rejection in Language Identification System", *Proceeding 5th European Conference on Speed Communication and Technology (Euro speech 97)*, Rhodes: September.
- Matrouf, D., Add a – Decker, M., Lamel L, L. and J. Gauvain (1998), "Language Identification Incorporating Lexical Information", *Proceeding of the International Conference on spoken Language Processing (ICSLP 98)*, Sydney: December.
- Matsuura, T. (2000), "Authorship Attribution in Japanese modern sentences via N-Gordon Distribution", *Mathematical Linguistic*, Vol. 22, No.6, pp. 223-238.
- Rau, M. D. (1974), *Language Identification by Statistical Analysis*, Master's thesis, Naval Postgraduate School.

Schmitt, J. C. (1991), Trigram-based method of Language Identification, *U. S. Patent Number 5062143*, October.

صنّعی منفرد، محمد علی، (۱۳۷۳)، استفاده از شبکه‌های عصبی، اولین کنفرانس ملی نگهداری و تعمیرات، دانشگاه صنعتی اصفهان.

صنّعی منفرد، محمد علی، (۱۳۸۰)، تحقیق در عملیات پیشرفته با نگرش کاربردی، انتشارات دانشگاه الزهرا (س).

پیوست مقاله

اضافه متن انگلیسی:

Cried Jo, much touched. I hope you will be a great deal better, dear, but you must keep watch over your bosom enemy, as father calls it, or it may sadden, if not spoil, your life.

اضافه متن عربی:

ان الطيور تعرف مقصدها من مواقع الشمس و القمر و النجوم فى السما . فتستعين بمواقع الاجرام السماويه على معرفه الزمن و الا تجاه الصحيح. لكن هنا سوال آخر و هو اذا تغيرت هذه المواقع بسبب حركه الارض و الشمس و القمر و النجوم. فكيف لا تضل الطيور طريقها؟

اضافه متن فرانسوي:

Un Camion de blanchisserie est arre'te' contre letrottoir, faitont e'cron'entre le bistrot et l'atelier de Gilory. Av dessus de ce dernier, un panneau me'tallique indique la raison sociale d'une enterprise quine travaille que pour un seul client.