نشست علمی پردازش هوشمند زبان عربی

با موضوع بررسی تحلیل‌گر صرفی نور

به کوشش: هیئت تحریریه فصلنامه ره‌آورد نور

اشاره

مرکز تحقیقات کامپیوتری علوم اسلامی، در حوزه متن‌کاوی فعالیت‌ها و دستاوردهای متعدد و ارزنده‌ای داشته است. متن‌کاوی در سه مرحله: آماده‌سازی، پردازش و تحلیل متون، در روند تولید محصولات نور جلوه دارد. در این میان، پردازش متون عربی، به سبب دشواری‌ها و پیچیدگی‌های خاصی که این زبان برخوردار است، همواره با چالش‌های جدی مواجه است. ازاین‌رو، تحلیل‌گرهای صرفی در دنیا، با مشکلات و کاستی‌هایی جدی روبه‌رو بوده‌اند. این امر، مرکز نور را بر آن داشت تا از سال ۱۳۸۹ اقدام به تولید یک تحلیل‌گر صرفی بومی نماید.

در این راستا، هشتمین نشست از سلسله نشست‌های علمی علوم اسلامی و انسانی دیجیتال با موضوع «پردازش هوشمند زبان عربی (جلسه اوّل) با محوریت تحلیل‌گر صرفی نور»، در ۱۹ خرداد ۱۴۰۱ش در سالن اجتماعات مرکز تحقیقات کامپیوتری علوم اسلامی نور برگزار گردید. دکتر حبیب سریانی (پژوهشگر گروه علمی قرآن و لغت نور) و حجت‌الاسلام سید محمد دانش (توسعه‌دهنده متن‌کاوی نور) به عنوان کارشناس، و جناب دکتر محمود شکراللهی (عضو هیئت علمی دانشگاه تبریز) و حجت‌الاسلام والمسلمین محمدرضا مدرسی (پژوهشگر حوزه علمیه) به عنوان ناقد در این نشست علمی حضور داشتند. در ذیل، خلاصه‌ای از این نشست علمی، به همراه دسته‌بندی و تیترگذاری جدید، از نظر خوانندگان عزیز می‌گذرد. گزارش تفصیلی این نشست، در وبگاه نورسافت انعکاس یافته است.

کارشناس: دکتر حبیب سریانی

اهمیت و کارکردهای متن‌کاوی

«مرکز کامپیوتری علوم اسلامی نور، دارای دو ماهیت است: یکی «تحقیقات کامپیوتری» و دیگری عنوان «علوم اسلامی». این مرکز، به ازای ماهیت «تحقیقات کامپیوتری»، تلاش کرده در تمام این سال‌ها مطابق با دانسته‌ها و متدهای جدیدِ فناوری هوشمند که در دنیا عرضه شده، حرکت کند و حتی پیشرو باشد. ازاین‌رو، یکی از موضوعاتی که ما را ملزم به تحقیق و مطالعه می‌کند، بحث «متن‌کاوی» است. متن‌کاوی یا کشف دانش از متن، به طور ساده یعنی «به‌کارگیری تکنیک‌های هوشمند پردازش متن جهت کشف یک‌سری اطلاعات نهفته از درون متن». متن‌کاوی، غالبا در داده‌هایی مطرح می‌شود که این داده‌ها به شکل جدولی، فرم‌دار یا ساخت‌یافته نیستند، ساختار مشخصی ندارند و یا اینکه نیمه‌ساخت‌یافته بوده و ساختار محدودی داشته باشند.

در داده‌های رقومی بنا بر آماری که ارائه شده، بیش از هشتاد درصد داده‌های رایانه‌ای بدون ساختار هستند. در مرکز نور نیز اگر با متن خام «علوم اسلامی» مواجه شویم، مانند: فقه، قرآن، حدیث یا کتاب‌های تاریخی، خواهیم دید به‌نوعی ساختار مشخص ندارند. البته ما در مرکز نور، در احادیث یک‌سری نشانه‌ها (فرمت‌ها) گذاشته‌ایم که به طور مثال، متن حدیث را از سند جدا کرده است و آن متن خام بدون ساخت، به شکل «نیمه‌ساخت‌یافته» در آمده است. اینک از دل این داده‌هایی که خیلی ساختار مشخصی ندارند، وظیفه دانش متن‌کاوی استخراج دانش‌های متعدد است؛ آن هم با توجه به علوم اسلامی که عموماً یک داده ثابت هستند و در قرون متقدم وجود داشته‌اند؛ همچون: قرآن، احادیث و تاریخ.

اگر سه مرحله برای متن‌کاوی متصور باشیم، یعنی: آماده‌سازی متون، پردازش متون و تحلیل آنها، در تمامی این مراحل، مرکز نور توانسته ورود مؤثری داشته باشد. در ورود اطلاعات و همچنین جست‌وجو در مقالات و کتاب‌هایی که تنها تصویر آنها در دسترس است، OCR در دسترس است. متن‌کاوی، در بخش تصحیح نقش داشته است؛ مانند برنامه «پاک‌نویس» (ویراستیار سابق) که توسط بخش تهران انجام پذیرفت یا با همین تکنیک‌های متن‌کاوی، در داده‌های عربی مرکز نور، بیش از دویست و هشتاد هزار غلط متنی یا غلط فونتی کشف شده است.

متن‌کاوی، در پردازش متن هم کارکرد داشته است؛ همچون: برچسب‌گذاری نوع متن، آیات، احادیث، ریشه‌گذاری، پیراسته کردن و برچسب‌گذاری موضوعی. نهایت، در تحلیل متن هم توانسته‌ایم موفقیت‌هایی با استفاده از فناوری‌های متن‌کاوی داشته باشیم؛ مانند: مشابهت‌یابی لفظی و معنایی، پیشنهاد مقالات مرتبط در نور مگز، سمیم نور که سامانه تقلب‌یاب است یا نمایه‌زنی ماشینی که در حوزه‌نت استفاده می‌شود.»

تصویر شماره ۱

چالش‌های هوشمندسازی واژگان عربی

«پردازش لفظی متن، به‌خصوص در متن‌کاوی زبان عربی، جایگاه ویژه‌ای داراست؛ از این جهت که مهم‌ترین دیتای علوم اسلامی، به زبان عربی کلاسیک است؛ زبانی که از شاخه‌های زبان سامی بوده و به علت ریشه‌محور بودن و پیچیدگی‌های خاص زبان عربی در لغت، صرف

«مرکز کامپیوتری علوم اسلامی نور، دارای دو ماهیت است: یکی «تحقیقات کامپیوتری» و دیگری عنوان «علوم اسلامی». این مرکز، به ازای ماهیت «تحقیقات کامپیوتری»، تلاش کرده در تمام این سال‌ها مطابق با دانسته‌ها و متدهای جدیدِ فناوری هوشمند که در دنیا عرضه شده، حرکت کند و حتی پیشرو باشد. ازاین‌رو، یکی از موضوعاتی که ما را ملزم به تحقیق و مطالعه می‌کند، بحث «متن‌کاوی» است

و نحو، ضرورت پردازش‌های لفظی بیشتر مشخص می‌شود؛ حتی پردازش لفظی متن، تأثیر مستقیمی بر پردازشِ نحوی و محتوایی دارد؛ مثلاً در بحث ادبیاتی مثل نقش «حال» را داریم که غالباً مشتق است؛ یعنی اگر شما بتوانید در یک عبارتی فعل و فاعل را پیدا کنید، آن کلمه مشتقِ منصوب را غالباً به عنوان حال در نظر می‌گیرند. مثال دیگر در بحث متن، مفعولَ مطلق است که مصدری از ریشه فعل پیش از خود است؛ یعنی شما به محض اینکه یک مصدری پیدا کنید که منصوب باشد و هم‌ریشه با فعل پیش از خود باشد، غالباً مفعول مطلق را تشخیص داده‌اید.

به دلیل پیچیدگی‌های زبان عربی، تحلیل‌گرهای صرفی در دنیا، همواره با چالش‌ها و نقایص جدی روبه‌رو بوده‌اند. این امر، ما را بر آن داشت تا از سال ۱۳۸۹ اقدام به تولید یک موتور صرف بومی نماییم؛ مثلاً در ساخت کلمات عربی اساساً اِعرابی مثل: فتحه، کسره، ضمه و سکون، دارای نقش بیشتری نسبت به مصوت‌هاست؛ برای مثال، کلمه «بکر» را به گونه‌های مختلف می‌توان خواند و یا مانند «حسن»، به دو گونه «حُسن» و «حَسَن» قابل خوانش است. از سوی دیگر، اکثر داده‌های علوم اسلامی بدون حرکت هستند؛ البته در یک پروسه طولانی و با زحمت فراوان، برخی کتاب‌ها را حرکت و اِعراب زده‌ایم؛ اما در بسیاری از کتاب‌های دیگر فقهی و حدیثی و تاریخی، نمی‌توان این فرآیند طولانی را به شکل دستی ادامه داد. پس، لازم است با فناوری ماشینی، اقدام به تکمیل حرکت و اِعراب کنیم.

در زبان عربی، برخی ویژگی‌های دیگر مثل: قواعد اعلال، قواعد تخفیف و قواعد ادغام هم بر پیچیدگی صرفی آن افزوده است؛ مثلا کلمه «قِ» می‌تواند فعل امر «وَقَیَ یقی» باشد. اگر همین کلمه مشکل را در تحلیل‌گر صَرفی مرکز نور جست‌وجو کنیم، اعلام می‌کند که کلمه «قِ»، فعل مذکر مخاطب معلوم است، لازم یا متعدی است، امر است و صیغه هفتم و از چه ریشه‌ای. مثال دیگر، «یُمَدُّونَ» است که اگر کسی به زبان عربی مسلط نباشد، ممکن است برای آن ریشه «مدو/ مدی» هم در نظر بگیرد؛ درحالی‌که این ریشه‌ها، ثلاثی مجرد نداشته و در نتیجه، ما هم در پاسخ‌ها اعلام نکرده و تنها پاسخ‌های واقعی و استعمالی را اعلام می‌نماییم.

از دیگر چالش‌های عربی، بحث سماعیات است که یکی از اشکال‌های اساسی موتورهای تحلیل‌گر دنیاست؛ اما در تحلیل‌گر صرفی نور، توانسته‌ایم بسیاری از کلمات سماعی همچون صفت مشبهه‌ها و ساختارهای بی‌قاعده مانند جمع مکسرها را تشخیص دهیم؛ به عنوان مثال، بیش از ۱۸ هزار جمع مکسر را از کتاب‌ها شناسایی کرده و در موتور صرف نور در نظر گرفته‌ایم.»

طراحی تحلیل‌گر صرفی نور

«در مورد طراحی اولیه موتور صرف باید عرض کنم در ابتدای امر واقعاً این بحث برای ما خیلی غریب بود؛ یعنی احساس کردیم که اگر حقیقتا بخواهیم به یک ماشین قواعد سنگین اعلال را بفهمانیم، بسیار سخت خواهد بود؛ لذا در ابتدا تلاش کردیم با استفاده از وزن‌ها (Pa– tern)، کلمات را شناسایی کنیم. با یک محاسبه، نسبت به ریشه‌ها، بیش از یک میلیون و پانصد هزار وزن ساخته می‌شد؛ درحالی‌که نسبت به افعال پیچیده معتل، مضاعف یا مهموز، کار بسیار سخت شده و در نظر گرفتن تمامی وزن‌ها در تمامی حالات (به‌ویژه به علت استثنائات)، تقریباً ناممکن می‌شد. ازاین‌رو، به جای استفاده از اوزان پیش‌ساخته، به سمت استفاده از قواعد مربوط به حروف زاید «سألتمونیها» رفتیم؛ یعنی اینکه در این کلمه، کدام حروف می‌تواند اصلی نباشد و بر

اساس این حروف زاید و اصلی، کلمه دارای چه ریشه و چه ساختاری است.

این روشِ قاعده‌محور (rule-based)، کارایی بیشتری نسبت به طرح قبلی داشت؛ زیرا با تشخیص جایگاه تک‌تک این حروف زاید در زبان عربی، درک بهتری از ساختار کلمه پیدا می‌کردیم. اینکه نون، الف، یاء و سین در زبان عربی، کجای یک کلمه، زایده یا اصلی خواهد بود، ما را در تشخیص ساختارهای پیچیده هم یاری می‌کرد. البته برای تشخیص کلمات بی‌ساختار هم یک‌سری بانک‌های داده مرکز نور تولید شد که در تشخیص کلمات سماعی، بسیار راهگشا بود. یکی از علل اصلی در اینکه شش سال طراحی تحلیل‌گر صرفی نور طول کشید، تهیه همین بانک‌های داده از میان معاجم و کتاب‌های معتبر در موضوع ادبیات عرب بود که از آن جمله می‌توان به داده‌های مربوط به: حروف (حدود ۱۳۰ کلمه)، جوامد غیرمصدری (حدود ۱۵۰۰۰ کلمه)، جمع‌های مکسر (بیش از ۱۸۰۰۰ هزار کلمه)، صفات مشبهه و صیغه مبالغه (بیش از ۷۵۰۰ کلمه) و مصادر ثلاثی مجرد (بیش از ۱۰۰۰۰ کلمه) اشاره نمود.

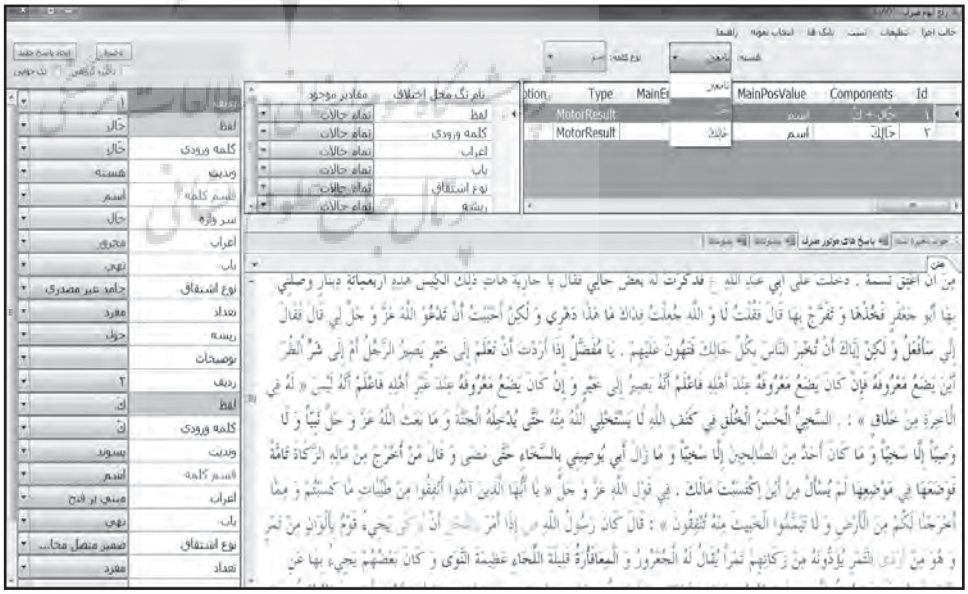
در ادامه نیز ابزاری برای رفع ابهام صرفی و تهیه دیتای استاندارد Gold آماده شد که این اجازه را به ما می‌داد برای بیش از ۲۰ ویژگی صرفی کلمه و از میان ده‌ها مقادیر مختلف، برچسب‌های صحیح را با توجه به جایگاه کلمه در متن انتخاب کنیم. ما این کار را ابتدا با قرآن آغاز کردیم. بیش از هشتاد درصد کلمات قرآن که برخی از آن کلمات هم دارای نوشتار خاص قرآنی هستند، توسط تحلیل‌گر صرفی نور شناسایی و بازبینی نهایی شد. با این پروژه، تمام قرآن با دقت کامل تجزیه صرفی شد. (تصویر شماره ۲) در حال حاضر، خروجی این پروژه، در پایگاه جامع قرآنی در اختیار عموم قرار گرفته و به‌عنوان یکی از پُربازدیدترین بخش‌های پایگاه به شمار می‌رود... .»

کارشناس: حجت‌الاسلام والمسلمین مهندس سید محمد دانش

شیوه پردازش در تحلیل‌گر صرفی نور

«در اینجا می‌خواهم بیشتر با تحلیل فنی و کامپیوتری، پروژه را با بیان نکاتی خدمت شما عزیزان عرض کنم. زبان عربی، یکی از پُرچالش‌ترین زبان‌های دنیاست و قرآن کریم هم به سبب پتانسیل‌هایی که زبان عربی داشته، به این زبان نازل شده است؛ اما اینکه قواعد آن را طوری پیاده‌سازی کنیم تا کامپیوتر بتواند آنها را تحلیل کند و بدون هیچ‌گونه ابهامی، مرحله به مرحله پیش رود و به جواب‌های مطلوب برسد، به عنوان یک پروژه، در دستور کار قرار گرفت.

در همین راستا، فعالیت‌های تخصصی در حوزه هوش مصنوعی با رویکردهای متفاوتی در تمام دنیا در حال گسترش است؛ مانند زبان‌شناسی رایانه‌ای یا پردازش زبان‌های طبیعی در زبان عربی. با بررسی پیشینه تحلیل‌گرهای صرفی جهت شناسایی نقاط قوت و ضعف‌ها، درمی‌یابیم تاکنون کارهای مشابه دیگری نیز در دنیا صورت پذیرفته است که از آن جمله می‌توان به تحلیل‌گر: بوک‌والتر، مبین، مادامیرا، الخلیل، کلمه و سفر اشاره کرد.



تصویر شماره ۲: نمونه ابزار استفاده‌شده جهت رفع ابهام صرفی متون اسلامی

شیوه پردازشی که در موتور صرفی یا تحلیل‌گر صرفی نور انجام می‌گیرد، این است که با توجه به ساختار یک کلمه در زبان عربی، ابتدا یک کلمه بماهوکلمه و فارغ از جایگاه آن در جمله، شناسایی و تحلیل صرفی می‌شود. در این مورد، دو رویکرد وجود دارد؛ برخی تحلیل‌گرها وابسته به جایگاه کلمه در جمله و کلمات قبل و بعد آن، پاسخ‌هایی برای کلمه ارائه می‌دهند و بعضی دیگر از تحلیل‌گرها برای کلمه بدون در نظر گرفتن جایگاه آن در جمله،

شیوه پردازشی که در موتور صرفی یا تحلیل‌گر صرفی نور انجام می‌گیرد، این است که با توجه به ساختار یک کلمه در زبان عربی، ابتدا یک کلمه بماهوکلمه و فارغ از جایگاه آن در جمله، شناسایی و تحلیل صرفی می‌شود. در این مورد، دو رویکرد وجود دارد؛ برخی تحلیل‌گرها وابسته به جایگاه کلمه در جمله و کلمات قبل و بعد آن، پاسخ‌هایی برای کلمه ارائه می‌دهند و بعضی دیگر از تحلیل‌گرها برای کلمه بدون در نظر گرفتن جایگاه آن در جمله، تمامی پاسخ‌های ممکن را اعلام می‌کنند و سپس، به سراغ رفع ابهام در جمله می‌روند

تمامی پاسخ‌های ممکن را اعلام می‌کنند و سپس، به سراغ رفع ابهام در جمله می‌روند. هدف ما با توجه به «تنوع متون علوم اسلامی» در مرکز نور، این بود که کلمه را به‌تنهایی بررسی کرده و تمام حالات صرفی معتبر را به دست آوریم و آنگاه به سراغ جداسازی پاسخ‌ها و رفع ابهام صرفی، با استفاده از اِعراب یا جایگاه کلمه در جمله می‌رویم. از این جهت است که این پروژه، بسیار حجیم شده است.

به طور کلی فرآیند تحلیل به دو دسته زیر تقسیم می‌شود: ۱. تولید حالات معتبر احتمالی؛ ۲. تطبیق و جداسازی حالات معتبر نهایی. تحلیل‌گر صرفی نور به جهت نگرش جامع آن، از چندین زیر موتور دیگر تشکیل شده است که به طور مثال، عبارت‌اند از: موتور شناسایی حروف، موتور شناسایی جمع‌های مکسر، موتور شناسایی اسامی جامد (جوامد غیرمصدری)، موتور شناسایی افعال غیرمتصرف، موتور شناسایی افعال متصرف یا موتور شناسایی اسامی مشتق و مصادر.

یکی از تقسیم‌بندی‌های اولیه در کار ما، این بود که کلمات در زبان عربی، اسم فعل یا حرف هستند و به دو دسته کلی سماعی یا قیاسی تقسیم‌بندی می‌شوند. سماعی‌ها، آن دسته از کلماتی هستند که در وزن و ساختار شکل‌گیری آنها قاعده معتبری وجود ندارد؛ مانند اسامی جمع مکسر، حروف و افعال غیرمتصرف. دسته دوم، قیاسی‌ها یا همان کلمات قاعده‌مند بودند؛ مانند اسمای مشتق یا افعال متصرف که طیف غالب کلمات زبان عربی، در این دسته قرار می‌گیرند. برای تشخیص سماعی‌ها، بانک‌های خیلی غنی‌ای آماده کردیم. در تشخیص قیاسی‌ها نیز به سراغ قواعد و کدنویسی رفتیم.

از جمله قابلیت‌های منحصربه‌فرد تحلیل‌گر صرفی نور، این است که می‌تواند کلمات را تماماً با حرکت، تماماً بی‌حرکت و یا تنها قسمتی با حرکت، مورد تحلیل قرار دهد. در واقع، حرکت داشتن یا نداشتن کلمه، هیچ خللی در تحلیل صرفی ما ایجاد نخواهد کرد؛ اما وقتی به مرحله تطبیق می‌رسیم، حروفی که دارای حرکت هستند، در کاهش پاسخ‌ها مؤثر واقع می‌شوند.

حال برای توضیح روند برنامه، تنها به سیستم شناسایی اسم‌های مشتق عربی اشاره می‌شود تا نمای کلی از فرآیندهای موجود در موتور تحلیل‌گر صرفی نور به دست آید:

* مرحله تولید حالات معتبر احتمالی: تشخیص پیشوند؛ تشخیص پسوندها؛ تشخیص ریشه‌های محتمل برای کلمه میانوند؛ بررسی احتمال وقوع اعلال در کلمه؛ بررسی باب‌های ممکن؛ بررسی کلمه از حیث مفرد، مثنی و جمع؛ بررسی احتمال اسم فاعل یا مفعول بودن کلمه؛ بررسی احتمال صیغه مبالغه بودن کلمه؛ بررسی احتمال صفت مشبهه بودن کلمه؛ بررسی احتمال اسم تفضیل بودن کلمه؛ بررسی احتمال اسم آلت بودن کلمه؛ بررسی احتمال اسم زمان و مکان بودن کلمه؛ بررسی احتمال مصدر میمی بودن کلمه؛ بررسی احتمال مصغر بودن کلمه.

* مرحله تطبیق و جداسازی حالات معتبر نهایی: در این مرحله، پس از ساخت تمامی حالات معتبر ممکن برای یک کلمه، تلاش می‌کنیم بر اساس حرکت‌های آن کلمه و قواعد این‌چنینی، پاسخ‌ها را کمتر کرده و به مطلوب نزدیک‌تر شویم؛ مثلا «کتب» هم حالت فعلی دارد «کَتَبَ» و هم حالت اسمی جمع مکسر «کُتُب» که با لحاظ حرکت فتحه بر کاف، تنها حالت فعلی آن باقی می‌ماند؛ البته ابتدای کار برخی حرکات و اِعراب مورد بررسی قرار می‌گیرد تا وارد چرخه نشود؛ مثلاً اگر کلمه تنوین دارد، وارد سیستم تحلیل صرفی «فعل» نمی‌شود؛ زیرا در قواعد عربی، افعال

متصرف تنوین نمی‌گیرند. در این مرحله، فعالیت‌های زیر انجام می‌شود:

ـ هر کدام از وزن‌ها، یک ساختارند تا در این مرحله، اِعراب اصلی کلمه را مورد بررسی قرار دهد.

ـ امکان اِعلال قلبی بر کلمه بررسی می‌شود؛ مثلاً کلمه «قَالَ»، در اصل قَوَلَ بوده است.

ـ بررسی امکان اِعلال سکون.

ـ چون بعد از اِعلال سکون، امکان اِعلال قلب وجود دارد، مجدد اِعلال قلبی مورد بررسی قرار می‌گیرد.

ـ بررسی امکان اِعلال حذف؛ مانند حذف واو از «یقُل».

ـ بررسی امکان قاعده تخفیف در مهموزات.

ـ بررسی امکان قاعده ادغام در کلمات مضاعف.

کلمه در مرحله آخر، بر اساس آن اِعرابی که می‌تواند داشته باشد، بازسازی شده و در نهایت، با کلمه ورودی، حرف‌به‌حرف و حرکت‌به‌حرکت، مقایسه و تطبیق داده می‌شود تا پاسخ‌های اضافی حذف شود. تمامی این مراحل، تنها برای شناسایی یک اسم صورت گرفت؛ درحالی‌که ممکن است نزدیک به ۳۰ مرحله برای تشخیص فعل‌های زبان عربی داشته باشیم. گفتنی است که این مراحل، در خصوص سماعی‌ها مانند جمع مکسر، کمتر است.»

ناقد: دکتر محمود شکراللهی

ذکر چند نکته

«... بعد از حدود پانزده سال که بنده در خصوص موضوع متن‌کاوی با برخی مسئولان مرکز نور گفت‌وگو کرده بودم، اصلاً برایم قابل تصور نبود که امروز ببینیم در اینجا بحث متن‌کاوی با این زیبایی پیش رفته و کارهایی به این قشنگی انجام شده است؛ کارهایی که نه تنها در ایران، بلکه در دنیا قابل عرضه است و می‌تواند کنار کارهای مطرح دیگر قرار بگیرد و عرض اندام کند.

اما نکته اوّل، مربوط به تعریفی است که از متن‌کاوی ارائه شد. بنده می‌خواهم یک عبارتی به آن اضافه کنم. اگر متنی ساختارنیافته باشد، هیچ هوشی مصنوعی نمی‌تواند به آن ساختار بدهد. ببینید تعریف متن‌کاوی (Text Mining) این است که Unseen Structured را بتوانیم Seen کنیم؛ یعنی وجود دارند؛ ولی آن را نمی‌بینیم؛ مثلاً عبارت «پردازش هوشمند زبان عربی» یک Noun phrase است که شکل گرفته و ساختار آن وجود دارد؛ فقط هنوز سیستم نمی‌تواند آن را به عیان ببیند. هدف هم بیشتر، همین دیده شدن توسط ماشین است تا به ادامه کار بپردازد.

نکته دوم، درباره زبان‌های سامی است. یک ویژگی زبان سامی، این است که ریشه‌محور هستند و علاوه بر سه حرف ریشه، یک Pattern دیگر وجود دارد که به آن ریشه اضافه می‌شود و آن، «وزن» است. آن پترن، جدای از ریشه وجود دارد. آن پترن، یک هویت زبان‌شناختی دارد و شما بدون ریشه هم، با الگوریتم‌های پترنی و تشخیص الگو می‌توانید، آن را تشخیص بدهید. این، یکی از شاه‌کلیدهایی است که در «سامانه مبین» به کار گرفته شده است.

چیز دیگری که زبان‌های سامی را از تمام زبان‌های دیگر دنیا متمایز می‌کند، این است که ما همه زبان‌های دنیا را به غیر از زبان‌های سامی، می‌توانیم با الگوهای اصطلاحاً – Co catenative مدل‌سازی کنیم. تمام زبان‌های دنیا، ساختار Concatenative دارند؛ اما زبان‌های سامی، به‌خصوص عربی، نمی‌توان آنها را مدل‌سازی صوری و پیاده‌سازی کرد. به همین علت، در هیچ زبان دیگری، این همه مراحل پردازش نداریم.

نکته سوم اینکه در مورد سیستم‌هایی که اسم بردید، جای برخی سیستم‌ها در فهرست شما خیلی خالی بود. بعد از سیستم آمریکایی – Buc walter که اوّلین‌سیستم بسیار معروف دنیاست، سیستمKenneth R. Beesley است که سیستمی با عنوان – Finite-State Tec nology مدل‌سازی خیلی قشنگی از عربی ارائه می‌دهد.»

چند پرسش مهم

«در اینجا بنده سه سؤال مهم و کلان یادداشت کرده‌ام:

۱. جدول attribute value شما چیست؟ برای معرفی سیستم، چه برای کاربران و چه در سطح دنیا، اینها یک اصطلاحات استانداردی است که هر سیستمی بخواهد عرضه بشود، باید این tagset را ارائه بدهد.

۲. در برنامه خود، از چه approach فرمال و زبان‌شناختی استفاده کردید؟

سیستم‌های مبتنی بر هوش مصنوعی، دو دسته هستند: اوّل، سیستم‌هایی که knowledge base کار می‌شوند. دوم، سیستم‌هایی که data

اگرچه در مورد علم صرف مرسوم است که علم صرف، آنالیز و بررسی تحلیلِ کلمه در خارج از فضای جمله است، ولی شواهدی وجود دارد که این‌گونه نیست؛ مثلاً در مباحث مربوط به فعل، بحث «لازم و متعدی» مطرح است که باید ناظر به فضای جمله گزارش شود که این فعل، لازم است یا معتدی. نمونه دیگر، در بحث «معرفه و نکره» مطرح است که بعضی از اقسام معرفه، ناظر به فضای جمله است. یا بحث «معرب و مبنی» که در مورد بعضی از اقسام مبنی‌های عارضی، مثل «أیّ» پدید می‌آید، اینها ناظر به فضای جمله است

driven کار می‌شوند و مبتنی بر train data هستند. کار شما در مرکز نور، از نوع نخست است. سؤال بنده این است که در سیستم knowledge base از چه approach فرمالی استفاده کردید؟

۳. سؤال سوم بنده، دو قسمت دارد. در مقاله خودتان گفته بودید، این کار در آینده انجام خواهد شد. می‌خواهم ببینم الآن انجام شده است؟ قسمت دوم سؤالم، در مورد ارزیابی سیستم است:

یکی، ارزیابی داخلی سیستم (internal evaluation) است؛ مثلاً بدانیم به فرض مثال، precision سیستم چند بوده؟ یا به‌خصوص – ambig ity سیستم در ارزیابی چیست؟

نوع دوم ارزیابی سیستم، ارزیابی بیرونی یا external evaluation است؛ یعنی سیستم شما در مقایسه با سایر سیستم‌های مشابه چگونه است؟»

ناقد: حجت‌الاسلام والمسلمین محمدرضا مدرسی

نقد و بررسی پاسخ‌های تحلیل‌گر صرفی نور

«بسیار متشکرم از تیم فنی و پژوهشیِ موتور صرفی نور که انصافاً قدم قابل‌توجهی در ارائه محصولات علوم اسلامی برداشته‌اند. نکاتی را هم که در اینجا عرض می‌کنم، به معنای نادیده گرفتن تلاش‌های عزیزان نیست.

اوّلین‌نکته‌ای که عرض کنم، در مورد جایگاه موتور صرف و رابطه آن با علوم عربی است. اگرچه در مورد علم صرف مرسوم است که علم صرف، آنالیز و بررسی تحلیلِ کلمه در خارج از فضای جمله است، ولی شواهدی وجود دارد که این‌گونه نیست؛ مثلاً در مباحث مربوط به فعل، بحث «لازم و متعدی» مطرح است که باید ناظر به فضای جمله گزارش شود که این فعل، لازم است یا معتدی. نمونه دیگر، در بحث «معرفه و نکره» مطرح است که بعضی از اقسام معرفه، ناظر به فضای جمله است. یا بحث «معرب و مبنی» که در مورد بعضی از اقسام مبنی‌های عارضی، مثل «أیّ» پدید می‌آید، اینها ناظر به فضای جمله است. پس، این‌طوری نیست که یک ریلی داشته باشیم که مقصد و پایه اوّل آن، صرف باشد و بیاییم سراغ لغت و بعد، به نحو بپردازیم؛ زیرا بعضی وقت‌ها ممکن هست مجبور باشیم از صرف به لغت و یا از نحو به صرف برگردیم.

البته من احساس کردم، احیاناً این نکته رعایت شده؛ مثلاً در کلمه «اطهروا» که برای آن هیچ حرکتی هم گذاشته نشده، باب «افعال» در فهرست پاسخ‌ها نیست؛ یعنی نشان می‌دهد که ناظر به لغت بوده است؛ اما برخی جاها معلوم است نگاه‌هایی به لغت نبوده است؛ برای مثال، در تحلیل «ضربنا»، یکی از احتمالاتی که در پاسخ‌ها آمده، جمع مکسر صفت مشبهه و مصدر است؛ درحالی‌که اینها در لغت گزارش نشده است.

و یا در مورد کلمه «کُلوا» که فعل امر است و از «اَکِلَ یَأکَلُ» هم گزارش می‌دهد؛ درحالی‌که قاعده تخفیف همزه در این باب اجرا نمی‌شود. در فضای نحوی نیز باید یک‌سری قانون‌ها به موتور صرف

دیکته شود؛ مثلاً در تحلیل «کلوا»، یکی از پاسخ‌ها را «کاف» جاره بر سر «لام» موطئه قسم به همراه «وا»ی ندا در نظر می‌گیرد؛ درحالی‌که می‌دانیم چنین پاسخی با توجه به نحو، ناصحیح است؛ برای مثال، در مورد شرایط مؤکد بودن یک فعل، یکی از احتمالاتی که در مورد کلمه «لیَضربنانِّ» می‌گوید، این است که «لام» آن، حرف جر باشد. اگر لام حرف جر باشد که باید نون آن بیفتد! حالا می‌گوییم نونِ مؤکد فعل، دارای اعراب تقدیری است و «نون» نیفتاده. پس، این فعلِ قسم است که برای گرفتن نون تأکید، شرایطی دارد که در نحو ذکر شده است.

به نظر بنده، بعضی جایگاه‌های کلمات در تحلیل‌ها می‌تواند تأثیرگذار باشد؛ مثلاً فرض کنید «کُم» در مثل کلمه «فیکم و انّکم» که در پایگاه قرآن هم بارگذاری شده، در «فیکم» مجروری است و در «انّکم» منصوبی است؛ اما موتور صرف، فارغ از جایگاه کلمه، اعلام می‌کند: «منصوبی مجروری»!

مطلب دوم در مورد قواعد صرفی است که از مسلّمات صرف است؛ مثلاً در مورد کلمه «مؤمنون»، احتمالات مختلفی همچون: اسم فاعل، اسم مفعول، اسم مکان و زمان، مصدر میمی از باب‌های مختلف را اعلام می‌کند؛ درحالی‌که مستحضرید مصدر میمی، قابلیت جمع سالم مذکر را ندارد... . مثال دیگر، «کَوکَب» است که یک تناقضی هم در خود این کلمه وجود دارد. برای این کلمه، وزن «فَعلَب» را می‌گوید و «باء» را در آن، زایده گرفته و در وزن هم آورده تا بگوید زاید است.

نکته دیگر، یک‌سری بحث‌های صرفی است که تیم پژوهشگر شما باید در آنها به جمع‌بندی برسد؛ مثلاً آیا می‌توانیم یک رباعی داشته باشیم که یکی از حروف آن، حرف علّه باشد؟ بله؛ در مثل «وسوس» و در مضاعف‌های رباعی، می‌شود تکرار بشود؛ ولی در مثل «کوکب»، هم می‌شود مضاعف باشد و هم رباعی باشد. قدما می‌گویند نمی‌شود؛ مگر در محظوراتی که ابن‌جنی در خصائص گفته است. اگر بخواهیم در آنها، حرف علّه را حرف اصلی نگیریم، مشکلاتی پیش می‌آید؛ ولی در مثل «کوکب»، در واقع، باید حرف زاید باشد؛ درحالی‌که به عنوان «رباعی» مطرح شده است. می‌خواهم بگویم مسلّمات صرفی رعایت شود و در مورد بحث‌های اختلافی نیز به یک جمع‌بندی برسید.

نکته سوم، جایگاه اِعراب در مباحث صرفی است؛ البته یک خلطی هم بین «اِعراب» و «حرکت» در کلام دوستان شده بود و ظاهرا مقصود ایشان از اِعراب، همان حرکت است. اِعراب، تنها ناظر به آخر کلمه است. با این پیش‌فرض، ما در تحلیل صرفی، نیازی به اِعراب آخر کلمه نداریم؛ زیرا فرقی نمی‌کند که اعراب آخر کلمه چه باشد؛ درحالی‌که در پاسخ‌های موتور صرف، این پاسخ‌ها دیده شده و گاه با بیش از ۱۰۰ احتمال هم برای کلمه می‌دهد. از نظر ما، اِعراب باید در موتور تحلیل‌گر نحو باشد. باید برخی قواعد صرفی به موتور داده شود تا پاسخ‌های اضافی ندهد.

نکته چهارم، عملکرد ضعیف تحلیل‌گر در دسته‌ای از کلمات است؛ مانند فعل امر «اضرِب» که برای آن، پاسخ صحیحی ندارد. در ابواب غیرمشهور ثلاثی مزید نیز ضعیفَ است که البته خودتان هم متذکر شده بودید؛ ولی باید ابواب غیرمشهور هم مورد توجه قرار گیرد. دسته دیگر، کلمات مصغر است؛ مانند «عصیفر، جعیفر» یا کلمات منسوب، مانند «وفویّ، مرتضیّ» یا کلمات مثنی، مثل «اخوان، اخوین» که جوابی از این تحلیل‌گر برای آنها نمی‌بینیم.

پیشنهاد بنده، در نظر گرفتن یک الگوریتم جایگزین است. برداشتم این است که کلمات عربی در این موتور تحلیل‌گر، با نگاه پیشوندی پسوندی مثل «سألتمونیها» و به صورت تک‌حرف است؛ درحالی‌که ما با زبان عربی مواجه هستیم؛ برخلاف زبان فارسی که پسوندها و پیشوندها، باعث ترکیب و تعدد معانی می‌شود؛ ولی در زبان عربی، ما با ساختارها مواجه هستیم؛ یعنی این ساختارهاست که یک معنای جدید ایجاد می‌کند. خیلی معقول‌تر بود که بر اساس ساختارها حرکت کنید؛ مانند فعل‌ها که ساختار آن، خیلی محدود است. در واقع، بعد از حذف پیشوند و پسوندها در یک کلمه، به ساختاری می‌رسیم که اگر آن را با ساختارهای موجود بررسی کنیم، خیلی راحت‌تر می‌توان به نتیجه رسید و خیلی از باگ‌ها و اشکال‌ها برطرف خواهد شد.

نکته آخرین اینکه تا به حال، هرچه درباره کلام عرب گفتیم، با نگاه انفعالی بوده است. بهتر است در دو راهی قیاس و سماع، یک‌سری چالش‌هایی ایجاد کنیم؛ مثلاً چرا عرب یک جا «اختار» گفته و اعلال کرده؛ اما یک جا «ازدوج» گفته و اعلال نکرده است. امثال «ابن‌جنّی» در خصائص، مطالبی را ارائه داده‌اند که نقطه عطفی در این مباحث است. خیلی از واژگانی که راحت می‌گوییم سماعی است، تحلیل‌هایی قابل دفاع برای آن می‌آورند. ابن‌جنّی در کتاب خصائص، می‌گوید: علت‌هایی که نحویین حاذق می‌آورند، بیشتر شبیه کار متکلمان است؛ تا کار فقها. به نظرم جای طرح این دست مباحث نیز در این گونه جلسات هست و اینکه ببینیم اصلاً می‌شود این نظریه را به یک کار اجرایی تبدیل کرد یا خیر.»

کارشناس: دکتر حبیب سریانی

برخی دستاوردهای تحلیل‌گر صرفی نور

«در اینجا به بعضی از دستاوردهای فناورانه تحلیل‌گر صرفی نور به‌اختصار اشاره می‌کنم:

ـ استفاده از فناوری‌های هوشمند در یادگیری ماشینی جهت رفع ابهام صرفی؛

ـ بازبینی ریشه کلمات عربی در دیتای عظیم مرکز نور؛

ـ استفاده از پاسخ‌های رفع ابهام‌شده موتور صرف جهت هماهنگ‌سازی کلیدواژگان در معاجم موضوعی؛

ـ کاهش پاسخ‌های موتور صرف با استفاده از ریشه استعمالی؛

ـ استفاده از تحلیل‌گر صرفی نور در جست‌وجو بر اساس ریشه و پیراسته (میان‌وند کلمه)؛

ـ استفاده از موتور صرف در ساخت پیکره نحوی استاندارد قرآن؛

ـ بهبود و ارتقای عملکرد شناسایی و دسته‌بندی احادیث مشابه؛

ـ مدخل‌یابی هوشمند مشتقات لغوی؛

ـ جست‌وجوی هوشمند در پایگاه قاموس.»

پاسخ به اشکال‌های مطرح‌شده از سوی ناقدان

«در مورد اشکالاتی که مطرح شد، پاسخ برخی از آنها را به مهندس دانش می‌سپارم؛ اما در مورد برخی نقدها نکاتی را که یادداشت کرده‌ام که عرض می‌کنم.

ـ تگ‌ست‌ها و ارزشیابی آنها:

.N	.Att	Description	Type	Diversity Index
1	Stem	Stem	N, V	-
2	Root	Root	N, V	-
3	POS	Part Of Speech	N, V, P, R	4
4	Categ	Category	N, V	27
5	TON	Type of Name	N	56
6	TOP	Type of Particle	P	51
7	Num	Number	V	9
8	Tens	Tense	V	9
9	TOV	Type of Verb	V	16
10	Voic	Voice	V	5
11	Lemma	Lemma	N, V, P	-
12	Case	Case	N, V, P	14

جدول ۱: فهرست ویژگی‌های صرفی استفاده شده از آنالیزر صرفی نور در این تحقیقات. (Particle:P ،Name:N ،Verb:V)

نمونه استاندارد آنها به همراه ارزشیابی هریک در مقایسه با نمونه Trainشده در مقاله‌ای که ذکر شد، آمده است (noormags.ir/view/fa/articlepage/1795230). (جدول شماره ۱ و ۲)

■ رفع ابهام صرفی متون عربی کهن ■

Attribute	Total	Correct	Accuracy
Affix	48719	48719	1
Entry	48719	47736	0.98
Slice	48719	47703	0.98
Root	30236	29220	0.97
POS	48719	45695	0.94
Num	22031	20596	0.93
Categ	30139	27531	0.91
Time	8204	7351	0.90
DervT	22032	19321	0.88
Lemma	48719	43074	0.88
TOV	8204	7100	0.87
Voic	8203	7090	0.86
Case	41997	31062	0.74
TOP	11762	8525	0.72

جدول ۲: کیفیت روش پیشنهادی برروی دیتای ارزیابی پیکره صرفی نور.

البته ویژگی‌های صرفی که در تجزیه و رفع ابهام صرفی قرآن استفاده شد، بیش از این بود؛ ولی در مقایسه با موتورهای دیگر، مانند – MA AMIRA یا الخلیل، تنها برخی ویژگی‌ها مورد مقایسه قرار گرفت؛ چون تحلیل‌گرهای آنها به اندازه ما ویژگی‌های صرفی مختلف را اعلام نمی‌کردند. در مقایسه با این تحلیل‌گرها، حتی تحلیل‌گر SAFAR که در سال ۲۰۱۲ عرضه شد و پیشرفت‌های خوبی هم داشت، از نظر این تگ‌ستِ اعلامی (tagset)، بالاتر بودیم.

ـ پاسخ به اشکالات مربوط به تعدد پاسخ و عدم نگرش نسبت به جایگاه نحوی کلمه:

تحلیل‌گر صرفی نور، از یک پوسته «رفع ابهام

صرفی» بهره می‌برد که با استفاده از فناوری‌های هوشمند، کاملاً ناظر به جایگاه کلمه در جمله، اِعراب‌گذاری و رفع ابهام می‌کند. در مرحله تشخیص صرفی کلمه، به لحاظ تنوع زیاد کلمات در متون کهن عربی و به‌خصوص متون مذهبی، مجبور بودیم که یک نگاه عام داشته باشیم؛ اما برای کاهش پاسخ‌ها و توجه به جایگاه کلمه در جمله، از پوسته رفع ابهام صرفی استفاده کرده‌ایم.

البته در مورد ویژگیِ صرفی «اِعراب آخر کلمه» که فرمودید ناظر به نحو است و نباید اعلام شود، عرض شود که اگر قرار باشد از داده‌های مربوط به آخر کلمه در پردازش‌های نحوی استفاده کنیم، مثل اینکه چرا کلمه تنوین نگرفته یا اِعراب آن منصوب است، محلّ اِعلام و بررسی آن، همین صرف است. همچنین، در مورد ویژگی «لازم و متعدی» که کاملاً ناظر به نحو است، همین که در صرف اِعلام شود این فعل، لازم یا ناقصه است، موجب می‌شود در پردازش‌های نحوی به دنبال مفعول منصوب نباشیم. دقیقاً در همین مقامِ تحلیل صرفی است که باید تمامی این نوع اطلاعات جمع‌آوری شود.

ـ نگرش به بحث لغت: در موتور تحلیل‌گر صرفی به بحث‌های لغت نیز کاملاً توجه شده است؛ مثلاً در اشکال مربوط به کلمه «کلوا» که ذکر شد، آنچه از باب «اَکِلَ یَأکَلُ» گزارش می‌دهد، فعل ماضی از ریشه «کلو/ کلی» است و فعل امر از ریشه «أکل»، فقط از باب «فعَل/ یفعُل» گزارش می‌دهد؛ این مطلب، یعنی موتور صرف ناظر به استعمال لغت عمل کرده است.(تصویر شماره ۳)

اشکال دیگر در مورد «ضُرُب» که فرمودید در لغت استعمال نشده، ولی موتور صرف گزارش جمع مکسر می‌دهد؛ این مورد خاص را از کتاب «المعجم المفصل فی الجموع المکسرة» نوشته دکتر امیل بدیع یعقوب جمع‌آوری کرده‌ایم. ایشان آورده است که «ضُرُب» می‌تواند جمع مکسر «ضَرب» باشد. در کل، بیش از ۱۸ هزار جمع مکسر را از کتاب‌های لغت شناسایی کرده‌ایم و تلاش داشته‌ایم که همگی ناظر به لغت و باب‌های استعمالی باشد؛ البته ممکن است موردی هم در این میان، از قلم افتاده باشد.

ـ پذیرش برخی باگ‌ها و اشکالات:
برخی باگ‌ها و اشکالات، قابل پذیرش است؛ مانند نقصی که در مورد وزن بیان کردید، ما هم خودمان این اشکال را می‌پذیریم؛ چون وزن، واقعاً بحث مشکلی است. از سوی دیگر، تفهیم آن به ماشین، مشکل بود. در نتیجه، تصمیم گرفتیم به جای وزن، بر تقویت تشخیص ویژگی‌های صرفی دیگر تکیه کنیم.

ـ نقد مربوط به وزن‌ها و ساختارها:
در مورد تغییر روش ساختاری در الگوها هم که ناقدان محترم بیان کردند و تأکید داشتند که بر ساختارهای وزنی و پترن‌ها پیش برویم، در ابتدای بحث عرض شد که ما یک بار این مسیر را رفتیم و چون به نتیجه نرسیدیم و با تعدد بیش از حد ساختارها در حالت‌های مختلف اعلال و تخفیف و ادغام و بسیاری از استثنائات دیگر مواجه شدیم، از روش قاعده‌محور (rule-based) و دیتاهای آماده سماعی استفاده نمودیم. در واقع، وقتی شما پیشوند و پسوندها را حذف می‌کنید و به یک میانوند فعلی یا اسمی می‌رسید، می‌توانید با چند قاعده ساده در معمول افعال، نوع آن اسم یا فعل را تشخیص دهید.

البته در برخی موارد، می‌توان از ساختارهای وزنی نیز با روش‌های بهتری استفاده کرد؛ ولی در مجموع، روش پیاده‌شده را مؤثرتر می‌دانیم. ما


```
<Phrase Entry="كَلُوا">
  - <Ans number="1">
      <word Entry="كُل" Voic="مجهول" Trans="لازم" TOV="3" Time="ماضي" Temp="فُعِلُوا" RootT="ناقص واوي" Root="كلو" Prsn="غائب" Num="جمع"
        Genr="متصرف" Gend="مذكر" Decl="مبني" Categ="فَعَلَ يَفْعَلُ" Case="مبني بر ضم"
        BackTrackingLemma="كُلُوا/////يَكْلُي/////كُلِيَ/يَكْلَي/كُلَيَ//كَلَا.كِلَا///كلو" Lemma="كَلَا/كِلَا" Pos="فعل" Affix="هسته" Slice="كُل" Seq="1"/>
      <word Entry="وُ" Root="تهي" Num="جمع" Categ="تهي" Case="مبني بر سكون" Pos="اسم" Affix="پسوند" Slice="وُ" Seq="2" DervT="ضمير متصل
        مرفوعي"/>
      <word Entry="ا" Decl="مبني" Case="مبني بر سكون" Pos="حرف" Affix="پسوند" Slice="ا" Seq="3" Spc="فعل" Opr="تهي" Kol="الف فارقه"/>
    </Ans>
  - <Ans number="2">
      <word Entry="كُل" Voic="مجهول" Trans="متعدي" TOV="3" Time="ماضي" Temp="فُعِلُوا" RootT="ناقص يائي" Root="كلي" Prsn="غائب" Num="جمع"
        Genr="متصرف" Gend="مذكر" Decl="مبني" Categ="فَعَلَ يَفْعَلُ" Case="مبني بر ضم"
        BackTrackingLemma="كُلُوا/////يَكْلَي/////كُلِيَ/يَكْلَي/كُلَيَ//كَلَا///كلي" Lemma="كَلَي" Pos="فعل" Affix="هسته" Slice="كُل" Seq="1"/>
      <word Entry="وُ" Root="تهي" Num="جمع" Categ="تهي" Case="مبني بر سكون" Pos="اسم" Affix="پسوند" Slice="وُ" Seq="2" DervT="ضمير متصل
        مرفوعي"/>
      <word Entry="ا" Decl="مبني" Case="مبني بر سكون" Pos="حرف" Affix="پسوند" Slice="ا" Seq="3" Spc="فعل" Opr="تهي" Kol="الف فارقه"/>
```


تصویر شماره ۳

به جای اینکه سراغ ساختارها برویم و برای دسته‌های مختلف سماعی‌ها ساختار ایجاد کنیم، بانک‌های داده مربوط به سماعیات، مانند بانک داده: جمع مکسر، صفت مشبهه یا مصادر را غنی نموده‌ایم و بیشتر بر قواعد مربوط به وندیت تکیه کرده‌ایم. در حال حاضر، نتایج تحقیقات در دنیا نشان داده که تجزیه صرفی کلمات با استفاده از الگوی «استم» (Stem) و حذف پیشوند و پسوند، بیش از الگوهای دیگر همچون تشخیص ریشه (Root) و لِمّا (Lemma)، در فرآیندهای هوشمندسازی اثربخش بوده است. ازاین‌رو، یکی از کارهای اصلی موتور صرف نور، حذف دقیق پیشوند و پسوندها بر اساس قواعد ناظر به استعمال و واقع است.

ـ ضرورت نگرش مبتنی بر هوش مصنوعی:

به نظرم باید تحلیل‌گر صرفی نور، با نگرش هوش مصنوعی و عملیاتی به سامان برسد. ما در موتور صرف، از تحلیل‌های عجیب برخی نحویین، فاصله گرفته‌ایم و با نگرش هوش مصنوعی قواعد را به موتور صرف تفهیم کرده‌ایم؛ مثلاً در مثال «کوکب» که ذکر شد، درگیر بحث‌های رباعی یا ملحق به ثلاثی نشدیم؛ چون احکام یکسانی داشتند. یکی از موارد منحصربه‌فرد موتور صرف نور، این است که ما «ال» موصول نداشته، نظر اخفش و مازنی را در مورد «الِ» وارد بر مشتقات ترجیح داده‌ایم و آن را «ال» تعریف گرفته‌ایم. در میان متأخرین نیز دکتر سامرائی در کتاب «معانی النحو»، همین قول را پسندیده است. مقاله علمی این بحث هم نوشته شده و به‌زودی در پایگاه بارگذاری خواهد شد.

ـ ضرورت نگرش کلان به مسئله:

پاسخ نهایی به نقدهای شما عزیزان این است که مسائل مربوط به تحلیل‌گر صرفی نور و دستاوردهای آن، باید به طور کلان دیده بشود. با توجه به حجم عظیم داده‌ها در دانش صرفی کلام عرب، در چند درصد نتایج، دارای پاسخ درست و در چند درصد، دچار خطا و اشتباه هستیم. با این نگرش کلان، بهتر می‌توان تحلیل‌گر صرفی نور را بررسی نمود.»

نکته‌ها و پیشنهادها

ـ مهندس دانش: «ما در موتور صرف، هم از یک‌سری دیتابیس‌های آماده در سماعی‌ها استفاده کردیم و هم از روش قاعده‌محور rule-based در تشخیص قیاسی‌ها بهره بردیم؛ مثلاً ریشه‌های موجود در زبان عربی را به دست آوریم و گفتیم ریشه کلمه‌ای مثل «کتب» می‌تواند به باب افعال برود و بعد Pattern و ساختارها را خودمان تولید کردیم و «اَکتَبَ یُکتِبُ اِکتَاب» را ساختیم.»

ـ دکتر شکراللهی: «البته منظور این نبود.»

ـ دکتر سریانی: «در معماری موتور صرف، شالوده اصلی، همان Buckwalter بود؛ ولی تغییرات بسیاری ایجاد شد و ما قواعد زبان عربی را سعی کردیم به روش خودمان بومی‌سازی کنیم.»

ـ دکتر شکراللهی: «بله، همین درست است.»

ـ مهندس دانش: «ما از موتور Buckwalter الهام گرفتیم؛ ولی یکی از ضعف‌های آن موتور، این بود که آن جای‌گشت‌های پیشوندی و پسوندی را هم یک حالت در نظر می‌گرفت که موجب می‌شد تعداد حالات، بیش از اندازه تولید شود؛ اما با توجه به اینکه می‌توانستیم Tokenization داشته باشیم، حالت‌های جای‌گشتی را فقط بر میانوند و بیس کلمه بررسی کردیم... .

برای ارزیابی سیستم Evaluation ۲۰۵هزار کلمه را با تنوعی از کتب مختلف، مانند: قرآن، حدیث و فقه آماده‌سازی کردیم. ۵درصد کلمات، اصلاً جوابی نداشت و یا مانند کلمه «انبطال» بودند که ساختگی بودند و استعمالی در عرب نداشتند و یا مثل برخی اَعلام بودند و در موتور صرف نور، برنامه‌ای برای تشخیص اَعلام نداشتیم. برای ۱۷درصد کلمات، تنها یک پاسخ ارائه دادیم که با دقت صحت ۹۸درصد بود. برای ۳درصد کلمات، دو پاسخ ارائه شد؛ البته تحلیل، بر دو فرض انجام شد: ۱. یکی متنِ بدون اِعراب که موجب شد تعداد پاسخ‌ها زیاد شود؛ ۲. متن را با پوسته جانبی موتور صرف اِعراب‌دار کرده و بعد تحلیل صرفی کردیم که پاسخ‌ها محدودتر شد.»

ـ دکتر شکراللهی: «اینجا Precision اصلاً معنا ندارد. بنده چندین مورد از سیستم شما خروجی گرفتم و دیدم در برخی موارد، بیش از هشتاد مورد خروجی می‌آورد. درست است که الگوبرداری شما از موتور Buckwalter بوده؛ اما طبق اعدادی که خودشان اِعلام کردند، خیلی کمتر است.»

ـ دکتر سریانی: «البته در دیتای قرآن که اِعراب کامل دارد، پاسخ‌های موتور ما این‌طور زیاد نیست.»

ـ دکتر شکراللهی: «می‌خواهم بگویم اوّلاً، بنده اینجا عددی نمی‌بینم که بگوییم Ambiguity سیستم شما چند است؟ ثانیاً، جایی که سیستم Ambiguity دارد، به این معنا که به‌ازای هر ورودی چند خروجی دارد، دیگر هیچ‌کدام از این سیستم‌ها، به Precision نگاه

نمی‌کنند؛ زیرا دیگر نگاه یک‌به‌یک ندارید؛ بلکه نگاه یک‌به‌چند دارید. در اینجا فقط Recal را اِعلام می‌کنند که می‌بینم هست و با چند کلمه‌ای که دیدم، به نظر می‌رسد درست باشد. البته وقتی Precision معنا نداشته باشد، دیگر F-Measure هم معنا ندارد؛ چون F-Measure تابعی از Precision و R - cal است. چیزی که بسیار مهم بوده و جای آن خالی است و می‌توانید خود را با سیستم‌های بزرگ دنیا مثل Buckwalter مقایسه کنید، همین Ambig - ity سیستم است.

نکته‌ای که باید در مورد سیستم‌های معروف دنیا مثل حیفا یا MADA بگویم که به آن اشاره کردید و بر اساس ورژن دوم Buckwalter کار می‌کنند و بر اساس ورودی‌های بدون حرکت هم هستند، در اینها به یاد ندارم که بیشتر از ۴ مورد بدهند؛ درحالی‌که موتور شما، این تعداد خروجی مبهم می‌دهد. یکی از شگردهایی که همان Buckwalter استفاده می‌کند و متأسفانه از آن غفلت شده، این است که Buckwalter از چند دیتابیس استفاده می‌نماید. یک دیتابیس واژگانی است؛ شامل: اسامی، افعال، ریشه‌ها و غیره، و یکی هم سه یا چهار table دارد که مورد انتهایی آن، به ابهام‌زدایی مربوط است و مقایسه می‌کند که این وند، اصلاً در متون برای این ریشه استفاده شده است یا خیر. همان‌جا بسیاری از ابهامات را حذف می‌کند. این مرحله آخر، باید به موتور شما اضافه شود... .

ضروری است که سیستم‌های دنیا را در همه موارد مطالعه کنیم تا بفهمیم نقاط ضعف و قوت آنها چیست و کجای آنها به درد ما می‌خورد. اینجا زحمت بسیار زیادی کشیده شده و بر اساس Buckwalter هم پیش رفته است؛ ولی کاملاً هم منطبق بر آن نیست و به همین دلیل، پاسخ‌های بسیار هم می‌دهد. این سیستم، برای سال‌ها قبل بوده و در طیّ این مدت، سیستم‌های بسیار خوبی در دنیا درست شده است؛ مثل همین سیستم مبین. در حال حاضر، چهار مدل فرمال استاندارد در دنیا جا افتاده است که در بیس مدل‌سازی ریاضی از زبان عربی از این چهار شیوه استاندارد استفاده می‌شود... .

شما هم بسیار زحمت کشیده‌اید و یک تجربه فوق‌العاده گران‌بهایی به دست آورده‌اید و در این دیتای عظیمی که داشتید، مؤثر بوده و با آن مشکلات خود را حل کرده‌اید که خیلی هم خوب است؛ ولی حتماً این table آخر را مثل Buc - walter اضافه کنید تا یک الگوریتم و سامانه جدید ایجاد شود که اشکالات فعلی را نداشته باشد.»

ـ دکتر سریانی: «آنچه شما به عنوان مرحله آخر Buckwalter فرمودید، ما اینجا داریم و به‌عنوان یک پوسته نهایی رفع ابهام صرف استفاده می‌کنیم. علت اینکه موتور صرف را در حالت اوّلیه این قدر گسترده و با پاسخ‌های متعدد در نظر گرفتیم، این است که امثال Buckwalter در کلمات متون کهن عربی، بسیار ضعیف عمل می‌کنند؛ مانند چندوجهی بودن کلمه «فتاه» در متون مختلف تاریخی و فقهی و حدیثی. این ما را مجبور کرد که حالت پیش‌فرض موتور صرف را بسیار عام در نظر بگیریم و پس از جواب‌های متعدد، برای آن یک پوسته رفع ابهام صرفی در نظر بگیریم که به‌نوعی کار همان table آخر را می‌کند.»

ـ دکتر شکراللهی: «پیشنهاد من این است که دیگر از کلمه Adaptation استفاده نکنیم؛ چون در این بحث‌ها؛ چنین کلمه‌ای نداریم و به جای آن، از کلمه Disambiguation یا ابهام‌زدایی استفاده کنید... .

یک نکته را هم در مورد سیستم‌های صرفی دنیا از لحاظ approach کلام و هدف نهایی برای ساخت سیستم عرض کنم. ما دو نوع هدف داریم: یکی Generation و دیگری Ro - gnition. شما موتور را با هدف Gener - tion ساخته‌اید؛ درحالی‌که نیازتان با توجه به دیتای مرکز، Rocgnition بوده است و به همین دلیل، این همه پاسخ دارای ابهام هم دارد. Rocgnition بسیار سبک‌تر از Ge - eration است؛ به عنوان مثال، همین Buc - walter، با هدف Generation ساخته شده است... .

کلام آخر اینکه... نشست‌های خیلی خوبی را شروع کردید که بنای واقعاً بابرکتی است. خواهش می‌کنم هرازچندگاهی، این نشست‌ها را به شهر دیگر ببرید؛ مثلاً به تبریز یا مشهد یا اصفهان بروید. این شهرها، هم حوزه‌های علمیه قوی دارند و هم جامعه فناوری خوبی دارند که در این زمینه، فعالیت می‌کنند... ؛ حتی بنده می‌توانم برای برگزاری این نوع نشست‌ها تمهیداتی ایجاد کنم؛ مثلاً در هلند هم اجرا شود؛ یعنی بعد از اینکه در چند شهر دیگر این نشست‌ها را برگزار کردید، می‌توانید آنها را در خارج از کشور هم اجرا نمایید.»■