پیشگفتار

زبان فارسی یکی از مهم‌ترین ارکان فرهنگ و هویت ملی ما ایرانیان است. برای بقا و بالندگی، این زبان باید بتواند با شرایط دنیای رقومی و بسترهای نوظهور مجازی همزیستی داشته باشد. در جهان کنونی، هرروزه حجم زیادی از اطلاعات در قالب محتوای زبانی و مستندات متنی بر بستر فضای مجازی عرضه می‌شود و ماشین‌های هوشمند، ضمن پردازش یا تولید محتوا و استخراج اطلاعات و دانش از آن‌ها، اطلاعات تولیدشده را براساس نیازهای کاربر یا تمایلات تولیدکننده، در اختیار کاربران قرار می‌دهند. در چنین فضایی، لازمهٔ ادامهٔ حیات زبان فارسی رفع موانع فنی و ارتقای دانش و فناوری‌های مرتبط با پردازش این زبان است. در غیر این صورت، این زبان در آینده‌ای نه‌چندان دور، کنار زده خواهد شد، و لاجرم، در دنیای هزارهٔ سوم، که دنیای حضور و رقابت در فضای مجازی است، دیگر سخن گفتن از هویت و فرهنگ ایرانی بی‌معنی و ممتنع خواهد بود.

در سال‌های اخیر، شاهد رشد بسیار سریع تحقیقات هوش مصنوعی و زیرحوزه‌های آن چه در گسترش مبانی نظری و مرزهای دانش و چه در تولید فناوری‌های کاربردی بوده‌ایم. در اسناد سیاست‌گذاری و برنامه‌ریزی کلان بسیاری از کشورهای جهان، به هوش مصنوعی به‌عنوان یکی از علومی که تصویر آیندهٔ جهان را شکل خواهد داد، اشاره شده و سرمایه‌گذاری‌های هنگفتی جهت آموزش و تربیت متخصصان در این حوزه و تولید دانش و فناوری‌های مرتبط انجام شده است. زبان‌شناسی رایانشی و پردازش زبان‌های طبیعی با موضوع آموزش (و ارزیابی) مهارت‌های زبانی انسان به رایانه، ازجمله شاخه‌های قدیمی و مطرح در هوش مصنوعی است. در این شاخه، تلفیق شناخت داده‌های زبانی توسط زبان‌شناسان و پردازش داده‌های زبانی و مدل‌سازی فرایندهای زبانی توسط مهندسانِ زبان، به تولید دانش و توسعهٔ محصولات هوشمند زبانی در حوزهٔ علم و فناوری منجر شده است. از جملهٔ این محصولات می‌توان به دستیارهای هوشمند، مترجم‌های ماشینی، منشی‌های تلفنی، واسط‌های هوشمند انسان ـ ماشین، موتورهای جستجو، سامانه‌های پرسش و پاسخ، هویت‌سنجی براساس نوشته یا گفته و ماشین‌های نویسنده یا سخنگو اشاره کرد.

سیر تکامل و تحول پژوهش‌های علمی و صنعتی در حوزهٔ فناوری‌های زبانی در سه دههٔ اخیر، حرکت از سامانه‌های قاعده‌بنیان به پردازش‌های آماری و سپس به یادگیری ماشین خصوصاً یادگیری عمیق را نشان می‌دهد. سامانه‌های قاعده‌بنیان که عمدتاً برای کاربردهای کوچک یا محدود، با قوانین قابل استخراج و برشماری به کار می‌رفته‌اند، سعی در استخراج و به‌کارگیری قوانین زبان و مدل‌سازی دانش و مهارت‌های زبانی انسان در قالب این قوانین داشتند. این سامانه‌ها گرچه در مواردی از خود کارایی خوبی نشان می‌دادند، ولی با بزرگ‌تر شدن حوزه و واقعی شدن مسائل مورد توجه، امکان تهیهٔ قواعد و پوشش همهٔ پدیده‌های زبانی سخت و گاه غیرممکن می‌شد. لذا حرکت به سمت روش‌های آماری آغاز گردید. در این روش‌ها فرض بر این است که با داشتن نمونه‌های زبانی زیاد می‌توان قواعد و روابط حاکم بر پدیده‌های زبانی را با روش‌های آماری استخراج نمود. این روش‌ها گرچه در سطح روساخت قادر به استخراج اطلاعاتی در مورد زبان بودند و در سطح گسترده‌تری نسبت به روش‌های قاعده‌بنیان کار می‌کردند، ولی همچنان در تبیین جزئیات ساختاری زبان طبیعی خصوصاً در سطح معنا و بالاتر ضعف داشتند. در سال‌های اخیر جمع‌آوری حجم زیادی از داده‌های زبانی به صورت خام و برچسب‌خورده و افزایش توان پردازشی و محاسباتی رایانه‌ها این امکان را فراهم کرده است تا با حرکت به سمت مدل‌های نورونی، شبکه‌های عصبی و یادگیری عمیق به سطح بالاتری از کارایی نسبت به روش‌های قبل دست یابیم. شبکه‌های عمیق سعی در مدل‌سازی نحوهٔ آموزش و پردازش زبان در مغز انسان دارند

و در بسیاری از زمینه‌ها ازجمله پردازش گفتار، ترجمهٔ ماشینی، تولید متن یا گفتار و دسته‌بندی متون توانسته‌اند جهشی در کارایی سامانه‌های پردازش زبان و گفتار ایجاد نمایند. ازجمله مشکلات مهم این شبکه‌ها نیاز به حجم زیاد داده، نیاز به توان پردازشی بسیار بالا و عدم توضیح‌پذیری است. دو مشکل اول برای زبان‌های با منابع محدود مانند زبان فارسی بسیار جدی‌تر و بازدارنده‌تر است و لازم است گویشوران و علاقه‌مندان به این زبان‌ها راه‌حل‌های مناسبی برای حل این چالش‌ها پیدا کنند تا در میانهٔ دنیای عظیم، پیچیده و سریع فناوری کارایی خود را از دست ندهند. به نظر می‌رسد ایجاد روش‌های جدید برای زبان‌های با منابع محدود و تلفیق رویکردهای مورد بحث در جای خود و به تناسب مسئلهٔ مورد بررسی می‌تواند بخشی از چالش‌ها را کم‌رنگ‌تر کند.

تاکنون کتاب‌های زیادی در حوزهٔ پردازش زبان شامل پردازش خط، متن و گفتار تألیف شده است. عمدهٔ این کتاب‌ها با تأکید بر زبان انگلیسی نوشته شده‌اند و حتی در تشریح روش‌های مستقل از زبان نیز، اعمال روش‌ها، گزارش نتایج و مثال‌های آن‌ها عمدتاً مربوط به زبان انگلیسی است. گرچه در تشریح مبانی نظری و کارهای انجام‌شده یا تطبیق‌یافته برای زبان‌های دیگر (مانند عربی) نیز کتاب‌هایی موجودند، ولی برای زبان فارسی کار برجسته و گسترده‌ای در این حوزه در قالب کتاب در سال‌های اخیر ارائه نشده است. تحقیقات این حوزه بیشتر در قالب مقالات (اعم از مقالات همایش یا نشریات) و یا در قالب پایان‌نامه‌های دانشگاهی عرضه گردیده‌اند و هیچ ارائهٔ جامعی در قالب کتابی از این دست، که تصویری کلان از مسیر پیموده‌شده و چشم‌اندازهای پیش رو به دست دهد، نداشته‌ایم. بدیهی است که تدوین چنین کتاب‌هایی برای ساماندهی و اطلاع‌رسانی در باب وضعیت پژوهش‌ها، خصوصاً در حوزهٔ زبان فارسی، ضروری است. به همین منظور و جهت آشنایی پژوهشگران با آخرین دستاوردهای پردازش زبان فارسی، نگارش، تصنیف و تدوین کتاب *پردازش متن و گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی*، به سفارش گروه زبان‌شناسی «سمت»، در قالب یک تألیف گروهی در سال ۱۳۹۷ هجری شمسی آغاز شد. هدف از تدوین این کتاب ارائهٔ گزارشی جامع و روزآمد از وضعیت تحقیقاتی پردازش متن و گفتار فارسی در زمینه‌های مختلف و زمینه‌سازی برای گسترش تحقیقات در این حوزه به منظور ارتقای دستاوردهای نظری و تولید محصولات کاربردیِ بهتر در جهت رفع نیازهای ملی و گسترش مرزهای دانش در سطح بین‌المللی بوده است.

این کتاب ابتدا در سه قسمت پردازش متن، پردازش گفتار و پردازش خط فارسی طراحی شد که در ویراست اول تنها دو قسمت از سه قسمت تدوین شده است. بدین ترتیب کتاب حاضر دارای دو قسمت پردازش متن و پردازش گفتار فارسی است و امید می‌رود قسمت پردازش خط و سایر موضوعاتی که به دلیل موانع موجود، در این ویراست، فصلی را به خود اختصاص نداده‌اند، در ویراست‌های بعدی، فصول جدید و مستقلی بیابند و به نسخهٔ به‌روزشدهٔ فصول کنونی اضافه شوند.

در کتاب حاضر هر قسمت، با بخش زیرساخت‌های داده‌ای آغاز می‌گردد و پس از مرور کارهای انجام‌شده در توسعهٔ ابزارها و پردازش‌های پایه و میانی، به بخش کاربردهای سطح بالاتر ختم می‌شود. هر بخش شامل چند فصل است، و در هر فصل، سعی بر آن بوده تا ضمن بررسی وضعیت موجود و تحلیلی از آن، افق‌های پیش رو و مسائل باز حوزهٔ مربوط ترسیم و تبیین گردد. به بیان دیگر هر فصل که به شکل یک مقالهٔ مروری دربارهٔ عنوان آن فصل، به قلم یکی از خبرگان و متخصصان موضوع نوشته شده، با مروری بر گذشته و تحلیلی بر حال، دریچه‌ای به آینده می‌گشاید.

بخش اول کتاب که به منابع زبانی و دادگان‌های متن‌محور و مدخل‌محور در پردازش متن اختصاص یافته، شامل ۵ فصل است. فصل اول نگاهی کلی به منابع زبانی و پیکره‌های متنی و واژی دارد و جایی که به پیکره‌های مورد توجه در فصول دیگر می‌رسد صرفاً به آن‌ها ارجاع می‌دهد تا حتی‌الامکان از تکرار مطالب پرهیز شود. در واقع فصل اول علاوه بر معرفی حوزهٔ کار، خود پیونددهندهٔ فصول دیگر و درآمدی برای ورود به فصول ۲ تا ۵ نیز هست. پنج فصلِ نخست به معرفی دادگان‌ها و منابع زبانی‌ای اختصاص داده شده‌اند که در ادامهٔ کتاب در فصول مختلف مورد استفاده قرار می‌گیرند. معرفی و تشریح پیکره‌های متنی و واژی زبان فارسی که با نشانه‌های صرفی، نحویِ سازه‌ای و وابستگی، معنایی و گفتمانی غنی شده‌اند، در این فصول صورت گرفته است.

در بخش دوم، ابزارهـا و پردازش‌های پایه و میانیِ متن معرفی می‌شـوند. پیش‌پردازش‌های پایه که عمدتاً بـرای هر کاربردی از متن اسـتفاده می‌شـوند، مانند واحدسازی، هنجارسـازی، تحلیل ساخت‌واژی، بن‌واژه‌یابی و ریشه‌یابی، در فصل ۶ مورد مطالعه قرار می‌گیرند. این فصل ضمن بیان مبانی نظری در این حوزه به بررسی و مقایسـهٔ ابزارهای پیش‌پردازش موجود برای زبان فارسی خواهد پرداخت. فصول ۷ تا ۱۲ به میان‌ابزارها یا (پیش)پردازش‌های میانی اختصاص دارند. منظور از (پیش)پردازش‌های میانی پردازش‌هایی هستند که معمولاً نه به‌عنوان یک کاربرد مستقل، بلکه به‌عنوان یک زیروظیفـه از وظایف یک برنامهٔ کاربـردی و در مراحل ابتدایی آن مورد اسـتفاده قرار می‌گیرنـد، اما به انـدازهٔ پیش‌پردازش‌های پایه، ابتدایی و سـطح پایین نیستند و الزاماً در مراحل اولیهٔ همهٔ کاربردها، نیازی به آن‌ها نیست. بازشناسی موجودیت‌های نامدار جهت شناسایی اسامی افراد، سازمان‌ها، مکان‌ها، زمان‌ها و مانند آن در متن، بازشناسی هم‌مرجع‌ها به منظور تشـخیص مرجع ضمایر و همچنین شناسـایی کلمات مختلفی که در متن به یک موجودیت واحد اشـاره دارند، شناسـایی اصطلاحات چندکلمه‌ای برای شناسایی افعال و کلمات مرکـب و چندواحدی و اصطلاحات و عبـارات زبانی، و در آخر رده‌بندی متون در این دسـته قرار می‌گیرنـد. همچنین فرایندهایی مانند تعبیهٔ کلمات و ساخت بردار جاسازی آن‌ها در فضای برداری معنایی و اسـتخراج مدل زبانی و همایندها نیز وظایفی هسـتند که در این بخش به آن‌ها توجه می‌شـود و برای بسـیاری از کاربردها ضروری‌اند.

در بخش سـوم، بـه تحلیل‌های لغوی، نحـوی و معنایی متون پرداخته شـده و فصولی در مـورد خطایابی متن، تجزیهٔ سـازه‌ای و وابسـتگی جملات و معناشناسـی رایانشـی را دربر گرفته است. فصل ۱۳ کـه به خطایابی و استانداردسـازی متون تخصیص یافته، به بررسـی نظری و کاربردی خطایابی در زبان فارسـی و معرفی و مقایسـهٔ سامانه‌های تهیه‌شده جهت ویرایش متون، استانداردسازی، خطایابـی و اصـلاح آن‌ها می‌پـردازد. فصول ۱۴ تـا ۱۶ به تجزیهٔ نحوی جملات فارسـی اختصاص دارند و به‌ترتیب تجزیهٔ سازه‌ای، تجزیهٔ سـطحی (چانکینگ) و تجزیهٔ وابستگی زبان را مورد مطالعه قرار می‌دهند. بررسـی مبانی نظری و الگوریتم‌های معرفی‌شده در سـطح جهانی و نحوهٔ انطبـاق آن‌ها برای زبان فارسـی از مباحث مورد بحث در این فصول هسـتند. در آخرین فصل از بخش سـوم به تحلیل‌های معنایی پرداخته می‌شـود و طیف وسـیعی از مباحث مطرح در معناشناسی رایانشی از بازنمایی معنایی و رفع ابهام معنایی کلمـات تا شباهت‌سـنجی معنایی کلمات و جمـلات، و بازنمایی معنای سـازه‌های بزرگ‌تـر از کلمه در دو حالـت ترکیب‌پذیر و غیرترکیب‌پذیر را شامل می‌شود. مباحثی مثل دگرنویسی، شناسایی نقش‌های موضوعی، اسـتلزامات متنی و شناسـایی استعاره در این فصل مورد بحث قرار می‌گیرند.

در نهایت در بخش چهارم، شـش مهـارت تخصصی زبان در چارچوب شش حوزهٔ کاربردی خلاصه‌سازی متن، مشابهت‌یابی و کشف تقلب، احساس‌کاوی، ترجمهٔ ماشینی، سامانه‌های پرسش و پاسخ و سامانه‌های جستجوگر معرفی می‌شوند.

بخش پنجم و ششـم به پردازش گفتار اختصـاص دارد. در این قسـمت نیز سـیر حرکت از دادگان‌ها و منابع زبانی آغاز می‌شود و سپس ابزارهای پایه و میانی پردازش گفتار فارسی معرفی می‌شوند و درنهایت کاربردها معرفی خواهند شد. بنابراین بخش پنجم به معرفی پیکره‌ها و ابزارهای پایهٔ پردازش گفتار فارسی اختصاص یافته است و سرانجام در بخش ششم فصول ۲۶ تا ۲۹ پردازش‌های میانی در حوزهٔ گفتار (که البته در اینجا خود می‌توانند کاربرد نهایی نیز باشند) مانند تبدیل گفتار به متن، تبدیل متن به گفتار، بازیابی اطلاعات گفتاری و بازشناسی زبان گفتاری معرفی می‌شوند و در فصل آخر (فصل ۳۰) به چند مهارت تخصصی گفتاری زبان فارسی در چارچوب برنامه‌های کاربردی پردازش گفتار پرداخته می‌شود.

در تدوین این کتاب تلاش شده است تا یکدستی و یکپارچگی فصول و هم‌راسـتایی آن‌هـا با هدف کتاب، با تهیهٔ شـیوه‌نامه‌های محتوایـی و قالبـی لازم بـرای مؤلفـان و همچنیـن ارزیابی دقیق هـر فصل به وسـیلهٔ نگارنـدگان این سـطور و مدیـر وقت گروه زبان‌شناسی «سمت»، جناب آقای دکتر احمدی، و بازبینی و اصلاح بازخوردهـای مؤلفـان در دو مرحلهٔ داوری، فراهم شـود. اگرچه در مقـام ویراسـتاری علمی، نگارندگان این سـطور کوشـیده‌اند محتـوای فصل‌ها در عین یکپارچگی و ارتباط و انسـجام کافی، از دوباره‌گویی و تکرار مطالب به‌دور باشـند، ولـی در بعضی موارد

به دلیل تفاوت نویسندگان فصول یا التزام به انسجام مطالب یک فصل، ممکن است برخی مطالب (خصوصاً در مورد ابزارها یا دادگان‌های مورد استفاده) در بیش از یک فصل تکرار شده باشد.

همچنین کار یکدست‌سازی معادل‌های فارسی اصطلاحات تخصصی، فعالیت زمان‌بَری بوده که طی فرایند ویرایش علمی، توسط نگارندگان این سطور، صورت گرفته و در ادارهٔ ویرایش «سمت» در کل کتاب اعمال گردیده است. نتیجهٔ این فعالیت علاوه بر هماهنگی نسبی متن کتاب از حیث به‌کارگیری معادل‌های فارسی، تهیهٔ واژه‌نامهٔ دوزبانهٔ ارزشمندی است که می‌تواند از سوی سایر محققان و دانشجویان علاقه‌مند به پژوهش در این حوزه مورد استفاده قرار گیرد و از تشتّت آراء در ترجمهٔ کلمات تخصصی انگلیسی به فارسی بکاهد. گفتنی است اصطلاحات تخصصی در متن کتاب با قلم ایتالیک مشخص شده‌اند تا خوانندگان، در صورت نیاز، بتوانند با مراجعه به واژه‌نامهٔ پایان کتاب، معادل انگلیسی این اصطلاحات را بیابند.

مخاطب کتاب عمدتاً دانشجویان تحصیلات تکمیلی رشته‌های مهندسی کامپیوتر و فناوری اطلاعات خصوصاً گرایش هوش مصنوعی، زبان‌شناسی رایانشی و زبان‌شناسی (و رشته‌های مرتبط) و پژوهشگران بالقوه در حوزهٔ پردازش زبان فارسی هستند که انتظار می‌رود با مبانی پردازش زبان طبیعی و پردازش گفتار آشنا باشند.

در پایان، شایسته است از افراد مؤثر در تدوین این کتاب که در زمان نسبتاً طولانی تألیف و تدوین کتاب (بیش از سه سال) با این مجموعه همکاری نموده‌اند تشکر و قدردانی شود:

ـ مؤلفان ۳۰ فصل کتاب، که نقش اصلی را در تدوین کتاب داشته‌اند، و طی مراحل مختلف ارزیابی و ویرایش، شکیبایی و مساعدت لازم را مبذول داشتند.

ـ آقای دکتر مهدی احمدی، مدیر وقت گروه زبان‌شناسی «سمت»، که پیشنهاد تدوین چنین کتابی را، ابتدا در سال ۱۳۹۵، با نگارندگان این سطور در میان گذاشتند و در تمامی مراحل تدوین و تدارک محتوای کتاب، همفکری و همکاری علمی و پشتیبانی و مدیریت اجرایی ایشان، بی‌تردید، بسیار مؤثر و کلیدی بوده است.

ـ کارشناسان معاونت پژوهشی «سمت»، خانم‌ها عیوضی و کشوری، که هماهنگی‌ها و پیگیری‌های گوناگون این کتاب برعهدهٔ ایشان بوده است.

ـ همکاران مدیریت تدوین «سمت» (واحد ویرایش)، خانم‌ها لشکری‌نژاد، ضرغامی، بیون، مشایخی، ربانی، جلال‌زاده که با تلاش فراوان کار بازخوانی و ویرایش زبانی و یکدست‌سازی و خانم‌ها علیزاده و فیض‌نژاد که کار صفحه‌آرایی و طراحی تصاویر کتاب را به سرانجام رساندند.

مهرنوش شمس‌فرد ـ محمود بی‌جن‌خان

۱۴۰۰/۱۲/۱۲