امکان‌سنجی برای طرح مدل سازی زبان فارسی

اثر: دکتر محمود بی جن خان
استادیار دانشکدهٔ ادبیات و علوم انسانی دانشگاه تهران
(از ص ۸۱ تا ۹۶)

چکیده:

تقطیع و برچسب دهی نحوی - معنایی داده‌های نوشتاری یکی از فعالیتهای اصلی در طراحی و ساخت هر دادگان زبانی برای استخراج مدل زبانی است. در این مقاله مشکلاتی که نگارنده در انجام این فعالیت برای طرح "امکان سنجی برای طرح مدل سازی زبان فارسی" داشته، توضیح داده شده، همچنین برای حل مشکلات از معیارهای زبان‌شناختی و مهندسی استفاده شده‌است. در نهایت برای استخراج مدل زبان فارسی یک بسته نرم‌افزاری نوشته شده، که در چارچوب فرآیند مارکف صفر تا سه مرحله‌ای، توزیع احتمال مشروط کلمات فارسی را در چهار حالت به طور مستقل ازو وابسته به مقولهٔ نحوی - معنایی به دست می‌دهد.

واژه‌های کلیدی: تقطیع و برچسب‌دهی، مدل سازی زبانی، فرآیند مارکف.

مقدمه :

مدل زبانی در پردازش زبان و گفتار به یک نوع بازنمایی زبان طبیعی در رایانه اطلاق می‌شود که از رهگذر آن می‌توان یک دنبالهٔ بهینه از کلمات زبان را بر اساس محدودیت‌های نحوی-معنایی پیش بینی کرد. اهمیت مدل زبانی در طرح بازشناسی رایانه‌ای گفتار، ناشی از آن است که می‌توان به کمک یک مدل زبانی بهترین گزینهٔ دنبالهٔ کلمات را از بین چند دنباله انتخاب کرد. در بسیاری از این طرح‌ها گزارش شده‌است که با استفاده از مدل زبانی درصد قابل ملاحظه‌ای به کارآیی سیستمهای بازشناسی گفتار اضافه شده‌است (رن - یوان، ۱۹۹۵؛ زو، ۱۹۹۰ و جواو، ۱۹۹۸). ویکتورزو (۱۹۹۰) در اجرای طرح سامیت (SUMMIT) تا ۴۰٪ افزایش را گزارش داده‌است.

بر این اساس گروه پردازش گفتار در پژوهشکدهٔ پردازش هوشمند علائم، بر آن شد تا برای بهبود عملکرد سیستم بازشناسی گفتار پیوستهٔ فارسی (شنوا) اقدام به طراحی و ساخت یک مدل زبانی مناسب کند. تجربهٔ اولیهٔ این گروه در طراحی و تهیه یک مدل زبانی از نوع "کلمات جفت" (Word Pairs) موفق نبود. زیرا این مدل از نوع مدل‌های قطعی و غیر احتمالی (deterministic) بود؛ علاوه بر آن آرایش کلمات در ساخت جمله‌های فارسی از محدودیت بسیار کمتری در مقایسه با آرایش کلمات در ساخت جمله‌های سایر زبانها (بخصوص زبان انگلیسی) برخوردار است. واضح است که ساخت یک مدل زبانی مناسب برای زبانهایی چون فارسی که آرایش کلمات در ساخت جمله تقریبا" آزاد است، مستلزم استفاده از یک مدل زبانی احتمالی است که احتمال وقوع یک کلمه، مشروط به یک بافت زبانی بیش از یک کلمه باشد.

به این ترتیب گروه پردازش گفتار به طور جدی برای طراحی و ساخت یک مدل آماری زبان فارسی اقدام نمود. در این راستا یک طرح مشترک پژوهشی بین

پژوهشکده و معاونت پژوهشی دانشگاه تهران با عنوان "امکان سنجی برای طرح مدلسازی زبان فارسی" تعریف شد و به تصویب رسید. در این گزارش مشکلات مربوط به تقطیع و برچسب‌دهی داده‌های طرح، که متون روزنامه‌ای هستند، توضیح داده می‌شوند.

تقطیع و برچسب‌دهی داده‌ها

داده‌ها عبارتند از ۳۱۰ پروندهٔ متنی از دادگان آموزش که شامل ۵۰۷۵۳۰ کلمه از متون روزنامه همشهری در ماه‌های دی و بهمن ۷۶ است. تقطیع و برچسب‌دهی داده‌ها در این فعالیت عبارت است از تعیین مرزکلمات در داده‌ها و اختصاص یک مقولهٔ نحوی معنایی به هر کلمه با توجه به بافت نحوی کلمه در متن. علاوه برآن، چون در مرحلهٔ اول طرح پس از ساخت واژگان، تقطیع و برچسب دهی اولیه داده‌ها مستقل از بافت نحوی برای هر کلمه انجام شده بود، بنابراین تقطیع و برچسب دهی داده‌ها در این مرحله شامل دو فعالیت مهم بود:

۱- تأیید یا اصلاح مرزکلمات.

۲- تأیید یا اصلاح بر چسب نحوی- معنایی برای هر کلمه با توجه به بافت نحوی.

مجری طرح برای انجام این دو فعالیت با مسائل متعددی روبرو شد. حال به طرح این مسائل و ارائه مثال از داده‌ها برای هر کدام می‌پردازیم.

مسألهٔ اوّل - کسرهٔ اضافه:

کسرهٔ اضافه واکهٔ /ِ/ است که به آخر بعضی اسم‌ها و صفات اضافه می‌شود.

مسأله این است که اگر اسم و صفتی که به آخر آنها کسره اضافه می‌شود، به همخوان یا واکهٔ /ا/ ختم شوند، کسرهٔ اضافه بدون نشانه خطی است و در غیر این

صورت کسرهٔ اضافه با نشانه "ی" در خط ظاهر می‌شود. به عنوان مثال در عبارت "کتابهای جالب خواندنی زیادی" سه کلمهٔ اوّل دارای کسرهٔ اضافه هستند ولی کسرهٔ اضافه فقط درکلمه اوّل نمود خطی دارد و به صورت "ی" به آخر "کتابها" چسبیده است در حالی که چون "جالب" به همخوان /h/ و "خواندنی" به واکهٔ /i/ ختم می‌شود، کسرهٔ اضافه نمود خطی در پایان آنها ندارد.

البته گاهی کسرهٔ اضافه اختیاری است. به عنوان مثال کلمهٔ "رفتن" را در عبارت "رفتن به آنجا" می‌توان با کسرهٔ اضافه و بدون کسرهٔ اضافه تولید کرد. در چنین مواردی مجری طرح کلمه را بدون کسرهٔ اضافه برچسب داده‌است.

به هر حال اختصاص "کسرهٔ اضافه" به برچسب کلمات مورد نظر یکی از فعالیتهای زمان بر در اجرای این طرح بود.

مسأله دوم، سوم و چهارم ناشی از هم‌نویسه‌ها (Homograph) و شبه هم‌نویسه‌های موجود در خط فارسی هستند. این مسائل را با استفاده از اصطلاح ادبی جناس (تجنیس) توضیح می‌دهیم (شفیعی کدکنی، ۱۳۷۹، ص ۳۰۳).

مسأله دوم - جناس تام

جناس تام به کلماتی اطلاق می‌شود که صورت نوشتاری‌شان یکسان و البته هم‌معنی نیستند. برای جناس تام بر حسب صورت واجی کلمات دو حالت وجود دارد:

حالت اوّل - صورت واجی کلمات نیز یکسان است.

مثال ۱- "من" دو معنی دارد: معنی اول آن "ضمیر اول شخص مفرد" و معنی دوم آن "سه‌کیلوگرم" است.

مثال ۲- "در" دو معنی دارد: معنی اوّل آن به مفهوم حرف اضافه و معنی دوم آن به معنی مفهوم کلمه door انگلیسی است.

مثال ۳- "کجا" دو معنی دارد: معنی اول آن به مفهوم قید پرسشی کلی در جمله‌ای چون "کجا می‌روی"؟ و معنی دوم آن به مفهوم اسمی "جا" در جمله‌ای چون "هر کجا که بروی..." است.

مثال ۴- "داشت" سه معنی دارد: معنی اول آن به مفهوم فعل اصلی جمله در جمله‌ای چون "چند ماشین داشت". معنی دوم آن به مفهوم اسمی "پروراندن" در جمله‌ای چون "مرحلهٔ داشت غلات شروع شد" و معنی سوم آن به مفهوم فعل کمکی در جمله‌ای چون "داشت می‌رفت که ..." است.

حالت دوّم - صورت واجی کلمات از نظر طرح تکیه‌ای متفاوت است.

مثال ۱- "دیگری" دو معنی دارد: اگر تکیهٔ آن روی "گر" باشد، به معنی "یک شیء دیگر" است، مانند "کتاب دیگری را دیدم"، که به معنی "یک کتاب دیگر را دیدم" است. اما اگر تکیهٔ آن روی "ری" باشد به معنی "یک نفر دیگر" یا ضمیر نکره است، مانند "کتاب دیگری را دیدم"، که به معنی "کتاب یک نفر دیگر را دیدم" است.

نکته - بسیاری از اسم‌ها و صفت‌ها، و فعل‌های فارسی بر حسب طرح تکیه‌ای از هم متمایز می‌شوند اگر چه صورت نوشتاری (و واجی) آنها یکسان است. مانند مثال‌های زیر در داده‌ها:

مثال ۲- "دریافت": اگر تکیه روی "در" باشد، به معنی فعلی است و اگر تکیه روی هجای آخر باشد، به معنی اسمی است.

مثال ۲- "نبود": اگر تکیه روی هجای اول باشد، به معنی فعلی است و اگر تکیه روی هجای آخر باشد، به معنی اسمی است.

مثال ۳- "دارد": اگر تکیه روی هجای اوّل یا دوّم باشد، به معنی اصلی است و اگر تکیه روی هجای آخر باشد، به معنی فعل کمکی است. به تفاوت معنایی "دارد"

در دو جملهٔ زیر توجه شود:

الف - "علی دارد می‌رود." "دارد" فعل کمکی است.

ب - علی کتاب دارد. "دارد" فعل اصلی است.

مثال ۵- "کتابی" دو معنی دارد. اگر تکیه آن روی "تا" باشد، به معنی "یک کتاب" است و اگر تکیه روی هجای آخر باشد، به معنی صفتی است.

در تقطیع و برچسب‌دهی متون، تعداد زیادی از کلمات تابع حالت دوّم بودند.

مسأله سوم: جناس ناقص

جناس ناقص به کلماتی اطلاق می‌شود که صورت نوشتاری‌شان یکسان ولی صورت واجی‌شان از نظر واکه‌ها متفاوت است. بعضی از مثال‌های جناس ناقص در داده‌ها عبارتند از:

مثال ۱- در /dar/ به معنی در ورودی و در /dor/ به معنی مروارید.

مثال ۲- جرم /Jorm/ به معنی گناه و جرم /Jerm/ به معنی وزن.

مثال ۳- مرد /mard/ به معنی انسان مذکر و مرد /mord/ به معنی فعل ماضی مردن.

مثال ۴- سر /sar/ به معنی کلّه، سر /ser/ به معنی بی حس و سر /sor/ به معنی لیز.

مثال ۵- سبک /sabok/ به معنی کم وزن و سبک /sabk/ به معنی روش.

مثال ۶- خرد /xord/ به معنی کوچک و خرد /xerad/ به معنی عقل.

مثال ۷- حکم /hokm/ به معنی فرمان، حکم /hakam/ به معنی قاضی و حکم /hekam/ به معنی حکمت‌ها.

مثال ۸- کر /kar/ به معنی ناشنوا و کر /kor/ به معنی آب شرعی.

مثال ۹- ببرد /beharad/ به معنی فعل التزامی بردن و ببرد /beborad/ به معنی فعل التزامی بریدن.

مثال ۱۰- نبرد /naharad/ به معنی فعل نفی بردن و برنده شدن، نبرد /nabord/ به معنی فعل نفی بردن و برنده شدن و نبرد /nabard/ به معنی اسمی جنگ.

در داده‌ها کلمات زیادی از نوع مثال ۹ و ۱۰ وجود داشت.

مثال ۱۱- نیل /nil/ به معنی رود نیل و نیل /ncyl/ به معنی نائل شدن.

مسأله چهارم: جناس زائد

جناس زائد به کلمات هم معنی اطلاق می‌شود که تفاوت صورت نوشتاری‌شان کمینه باشد.

مثال‌های جناس زائد در داده‌های ادبی و نوشتار قدیمی فارسی مشاهده شدند. بعضی از آنها عبارتند از:

مثال ۱- "ز" به معنی "از" در شعر فارسی.

مثال ۲- "همی" به معنی پیشوند استمرار در نثر قدیم فارسی و به معنی "می" در فعل‌های مضارع فارسی امروز.

مثال ۳- "ور" به معنی "واگر" در شعر فارسی.

مسأله پنجم: تجزیه یا ترکیب

یکی از مسائل بسیار مهم و وقت‌گیر در تقطیع داده‌ها به کلمات این بود که آیا چند کلمه مجزا از هم تایپ شده را باید ترکیب کرد و به یک کلمهٔ مستقل تبدیل نمود، یا بالعکس یک کلمه را به چند کلمهٔ سازنده‌اش تجزیه کرد و آن را به تعدادی کلمهٔ مستقل از هم تبدل نمود. این مسأله را تحت عنوان "تجزیه یا ترکیب" مطرح می‌کنیم.

به عنوان مثال "غیر قابل قبول" یک کلمه است یا دو کلمه یا سه کلمه؟ "ثبت نام" یک کلمه است یا دو کلمه؟ "بیست و پنج کیلومتری" یک کلمه است، یا دو کلمه یا چهار کلمه؟ "رئیس جمهور" یک کلمه است یا دو کلمه؟ "ریاست جمهوری" یک کلمه است یا دو کلمه؟ "قابل پیش بینی" یا "قابل x" یک کلمه است یا دو کلمه یا به

طور کلی n کلمه ؟ "این گونه" یا "هر وقت" یک کلمه هستند یا دو کلمه؟ "از دست داد" یک کلمه است یا دو کلمه یا سه کلمه؟ "بدین ترتیب" یک کلمه است یا دو کلمه یا سه کلمه؟ "از آن پس" یا "پس از آن" یک کلمه است یا دو کلمه یا سه کلمه؟ "دو ساعته" "سه ساعته"، "بیست و چهار ساعته"، "دو روزه"، "سه روزه"، "سه شبه" یک کلمه هستند یا دو کلمه یا سه کلمه ؟ "در حالی که" یک کلمه است یا دو کلمه یا سه کلمه؟ "در این حال" یا "در عین حال" یک کلمه است یا دو کلمه یا سه کلمه؟ "غیر قابل استاندارد بودن" یک کلمه است یا دو کلمه یا سه کلمه؟ "به اجرا گذاشتن" یک کلمه است یا دو کلمه یا سه کلمه؟ "پس از این که"، پس از آن که"، پیش از این که"، "پیش از آن که" یک کلمه هستند یا دو کلمه یا سه کلمه یا چهار کلمه؟ و بسیاری مثال‌های دیگر.

نکته مهم این بود که تایپیست‌های داده‌ها نیز در این موضوع اتفاق نظر نداشته‌اند.

به عنوان مثال عباراتی چون "قابل x" یا "غیر قابل x" را بعضی تایپیست‌ها به صورت یک کلمه و بعضی به صورت دو یا سه کلمه تایپ کرده‌اند. البته باید توجه داشت که تایپیست‌ها متعلق به یک مؤسسه روزنامهٔ همشهری بوده‌اند. اگر تنوع تایپیست‌ها بیشتر شود، این مسأله پیچیده‌تر خواهد شد. اما اگر حتی داده‌ها به طور یکنواخت با رعایت یک معیار دلخواه ویراستاری تایپ می‌شدند، مسأله پنجم را می‌توانستیم حداقل در مرحلهٔ امکان سنجی در نظر نگیریم.

به هر حال برای حل این مسأله دو ملاحظه زبانشناختی و مهندسی را در نظر گرفتیم:

ملاحظات زبانشناختی

ملاحظات زبانشناختی شامل معیارهای واجی، صرفی، نحوی و معنایی است.

۱- معیار واجی

یک گروه از کلمات داده شده را می‌توان ترکیب کرد و به یک کلمه مستقل تبدیل نمود، اگر تکیهٔ اصلی روی دورترین وابسته‌های پیشین و پسین هستهٔ گروه نباشد. (اسلامی، ۱۳۷۹).

مثال - "به هر حال" یک گروه سه کلمه‌ای است، که هستهٔ آن "به" است به طوری که "هر حال" به عنوان یک گروه اسمی وابستهٔ پسین آن محسوب می‌شود. چون تکیه اصلی این گروه روی دورترین وابستهٔ پسین یعنی "هر" قرار ندارد، بنابراین می‌توان "به هر حال" را یک کلمهٔ مستقل در نظر گرفت، و در واژگان یک مدخل مجزا برای آن تعریف کرد.

۲- معیار صرفی

یک گروه از کلمات داده شده را می‌توان ترکیب کرد و به یک کلمه مستقل تبدیل نمود، اگر بتوان به ابتدا یا انتهای گروه، پیشوندها یا پسوندهای تصریفی را اضافه کرد، به طوری که گروه کلمات متعلق به یک طبقهٔ نحوی پایه، مانند اسم، صفت، فعل، ...، باشد.

مثال ۱- "قابل پیش بینی" یک گروه دو کلمه‌ای است. چون می‌توان به آخر آن پسوند جمع و کسرهٔ اضافه "های" را اضافه کرد و نتیجهٔ آن دو کلمهٔ مجاز "قابل پیش‌بینی‌تر" و "قابل پیش‌بینی‌ترین" به عنوان صفت تفضیلی و عالی باشد، بنابراین "قابل پیش‌بینی" یا به طور کلی "قابل x" یک کلمهٔ مستقل است و باید در واژگان یک مدخل مجزا داشته باشد.

مثال ۲- "به اجرا گذاشتن" یک گروه سه کلمه‌ای است. چون می‌توان به آخر آن پسوند جمع و کسرهٔ اضافه "های" را اضافه کرد و نتیجهٔ آن کلمهٔ مجاز "به اجرا گذاشتن‌های" به عنوان اسم مصدر جمع با کسرهٔ اضافه باشد، بنابراین "به اجرا

گذاشتن" یک کلمهٔ مستقل است و باید در واژگان یک مدخل مجزا داشته باشد.

۳- معیار نحوی

یک گروه از کلمات داده شده را می‌توان ترکیب کرد و به یک کلمه مستقل تبدیل نمود، اگر نتوان هر کدام از کلمات سازنده‌اش را با استفاده از وندها یا کلمات دیگر توسعه داد و یک گروه کلمات مجاز به دست آورد.

مثال ۱- "به هر حال" یک گروه سه کلمه‌ای است. "حال" اسم است و می‌توان آن را با وابستهٔ پیشین "گونه" توسعه داد. اما "به هر گونه حال" یک گروه کلمات مجاز نیست. چون نمی‌توان این گروه سه کلمه‌ای را توسعهٔ مجاز داد، بنابراین کلمهٔ مستقل است و باید در واژگان یک مدخل مجزا داشته باشد.

مثال ۲- "ریاست جمهوری" یک گروه دو کلمه‌ای است. "ریاست" اسم است و می‌توان آن را با صفت توسعه داد و گروه مجاز "ریاست محترم جمهوری" را به دست آورد. بنابراین "ریاست جمهوری" یک کلمهٔ مستقل نیست و نباید در واژگان مدخل مجزا داشته باشد.

مثال ۳- "نماز جمعه" یک گروه دو کلمه‌ای است. "نماز" اسم است و می‌توان آن را با صفت توسعه داد و گروه مجاز "نماز دشمن شکن جمعه" را به دست آورد. بنابراین "نماز جمعه" یک کلمهٔ مستقل نیست و نباید در واژگان مدخل مجزا داشته باشد.

مثال ۴- "مدیرکل" و "مدیرعامل" گروه‌های دو کلمه‌ای هستند. "مدیر" اسم است و می‌توان آن را با صفت توسعه داد اما گروه‌های حاصل مانند "مدیر محترم کل" و "مدیر محترم عامل" مجاز نیستند. بنابراین این دو گروه یک کلمه مستقل هستند و باید در واژگان مدخل مجزا داشته باشند.

۴- معیار معنایی

یک گروه از کلمات داده شده را می‌توان ترکیب کرد و به یک کلمه مستقل تبدیل نمود اگر معنی گروه از مجموع معانی کلمات سازنده گروه به دست نیاید.

مثال ۱- "به هر حال" یک گروه سه کلمه‌ای است که معنی آن از مجموع معانی کلمات "به" و "هر" و "حال" به دست نمی‌آید. بنابراین یک کلمهٔ مستقل است و باید یک مدخل مجزا در واژگان داشته باشد.

مثال ۲- "قابل پیش بینی" یک گروه دو کلمه‌ای است که معنی آن از مجموع معانی کلمات "قابل" و "پیش‌بینی" به دست می‌آید. بنابراین یک کلمهٔ مستقل نیست.

مثال ۳- "به دست آورد" یک گروه سه کلمه‌ای است، که معنی آن از مجموع معانی کلمات "به" و "دست" و "آورد" به دست نمی‌آید. بنابراین یک کلمهٔ مستقل است و باید یک مدخل مجزا در واژگان داشته باشد.

مثال ۴- "به اجرا گذاشتن" یک گروه سه کلمه‌ای است، که معنی آن از مجموع معانی کلمات "به" و "اجرا" و "گذاشتن" به دست نمی‌آید. بنابراین یک کلمهٔ مستقل است و باید یک مدخل مجزا در واژگان داشته باشد.

ملاحظات آماری

از نظر مهندسی گفتار، هر چه طول کلمه بیشتر باشد پردازش آن از جنبه‌های مختلف از جمله بازسازی (synthesis) و بازشناسی (recognition) گفتار ساده‌تر است. بنابراین بهتر است به ازای هر گروه از کلمات داده شده یک مدخل واژگانی تعریف شود. اما این کار باعث می‌شود از یک سو با افزایش حجم واژگان، زمان جستجو در واژگان افزایش یابد و از سوی دیگر چون بین کلمات موجود در یک گروه محدودیت نحوی وجود دارد، در صورتی که گروه کلمات به یک مدخل واژگانی

تبدیل شود، مدل زبانی که چیزی جز مدلسازی محدودیت کنار هم قرار گرفتن کلمات نیست، تضعیف می‌شود. به این ترتیب معیار مهندسی باید به گونه‌ای تعریف شود که به نوعی یک بده بستان بین سادگی در پردازش اکوستیکی گروه کلمات و پیچیدگی در افزایش حجم واژگان و تضعیف مدل زبانی را حل کند.

به این منظور معیار مهندسی را به صورت زیر تعریف می‌کنیم:

۵- معیار مهندسی

یک گروه از کلمات داده شده را می‌توان ترکیب کرد و به یک کلمهٔ مستقل تبدیل نمود اگر فراوانی آن گروه در زبان یا بخشی از زبان زیاد باشد.

مثال ۱- "پس از"، "قبل از" ، "در داخل" و بسیاری از حروف اضافهٔ مرکب از جمله گروه‌های پرکاربرد در داده‌هابودند. بنابراین می‌توان برای هر کدام یک مدخل واژگانی در نظر گرفت.

مثال ۲- "به هر حال"، "به طوری‌که" ، "از این پس" و بسیاری از حروف ربط و قیدها از جمله گروه‌های پرکاربرد در داده‌ها بودند. بنابراین می‌توان برای هر کدام یک مدخل واژگانی در نظر گرفت.

مثال ۳- "به اجرا گذاشتن"، "سر رسیدن" و بسیاری از اسم مصدرها از جمله گروه‌های کم کاربرد در داده‌ها بودند. بنابراین می‌توان آنها را به کلمات سازنده‌شان تجزیه کرد.

مثال ۴- "به اجرا گذاشت"، "سررسید" و بسیاری از افعال مرکب از جمله گروه‌های کم کاربرد در داده‌ها بودند. بنابراین می‌توان آنها را به کلمات سازنده‌شان تجزیه کرد.

نتیجه:

همانطور که در مثال‌های بالا مشاهده شد بعضی از گروه کلمات مانند "به هر حال" با تمام معیارها به یک نتیجه واحد می‌رسند، یعنی باید یک مدخل مجزا برای آن در واژگان تعریف کرد. امّا بسیاری از گروه کلمات بر اساس یک معیار باید به یک کلمهٔ مستقل تبدیل شوند، در حالی که بر اساس معیار یا معیارهای دیگر باید به کلمات سازنده‌شان تجزیه شوند. به عنوان مثال، گروه "به اجرا گذاشتن" بر اساس معیار صرفی و نحوی یک کلمهٔ مستقل، ولی بر اساس معیار مهندسی باید به کلمات سازنده‌اش تجزیه شود.

به هر حال در امکان سنجی طرح با توجه به ملاحظات زبانشناختی و مهندسی تصمیم‌های زیر گرفته شد و در تقطیع و برچسب دهی داده‌ها اعمال شدند.

۱- تمامی اسم مصدرها به کلمات سازنده‌شان تجزیه شدند.

۲- تمامی فعل های مرکب به کلمات سازنده‌شان تجزیه شدند. البته به جز در مواردی که جزء غیر فعلی از نوع حرف اضافه باشد.

۳- بسیاری از حروف اضافه مرکب، حروف ربط (مرکب) و قیدها به یک کلمهٔ مستقل تبدیل شدند.

به این ترتیب حجم داده‌ها در مرحلهٔ اوّل طرح از ۵۰۷۵۳۰ کلمه به ۵۰۱۱۱۱ کلمه در پایان فعالیت تقطیع و برچسب‌دهی در مرحلهٔ دوم طرح کاهش یافت.

برنامه‌های مدل زبانی

مجمّوعهٔ برنامه‌های مدل زبانی تحت عنوان یک برنامهٔ رایانه‌ای "تولید کننده و مرورگر پرونده‌های آماری مدل زبانی" نوشته شد. این برنامه یک بستهٔ نرم‌افزاری برای استخراج مدل زبانی در دو حالت است:

حالت اول - مدل زبانی بدون در نظر گرفتن مقولات نحوی - معنایی کلمات.

حالت دوّم - مدل زبانی با در نظرگرفتن مقولات نحوی - معنایی کلمات .

منظور از مدل زبانی محاسبهٔ توزیع احتمال وقوع یک کلمه به شرط صفر، یک، دو و سه کلمه قبل از آن است. ساخت زبان فارسی با یک فرآیند تصادفی از نوع مارکف مدلسازی شده‌است.

ساختمان کلی بستهٔ نرم‌افزاری به صورت زیر است.

الف - درون داد بسته شامل پرونده‌های تقطیع و برچسب دهی شده‌است. کاربر می‌تواند پرونده‌های مورد نظر خود را وابسته به موضوعات یا مستقل از موضوعات انتخاب کند.

ب - پردازش‌های بسته شامل محاسبهٔ توزیع احتمال مشروط کلمات در چهار وضعیت یک کلمه‌ای (unigram) دو کلمه‌ای (bigram)، سه کلمه‌ای (trigram) و چهار کلمه‌ای fourgram) است. پردازش‌ها در حالت‌های اوّل و دوّم که قبلاً" گفته شد، انجام می‌شوند. در این مرحله کاربر می‌تواند برای استخراج مدل زبانی در حالت دوّم دو متغیر را تعریف کند:

۱- برچسب‌های خاص: این که کلمات با یا بدون کسرهٔ اضافه در نظر گرفته شوند. علاوه بر آن میزان پیشروی در زیرگروه‌های برچسب نحوی - معنایی چقدر باشد.

۲- برچسب‌های منتخب: برچسب‌هایی هستند که کاربر تعریف می‌کند.

ج - برون داد بسته شامل جدول‌های متعدد مدل زبانی در حالت اوّل و دوّم است، که در حالت دوّم می‌توان مشخصات جدول‌ها را بر حسب مقادیری که کاربر برای برچسب‌های خاص و منتخب تعریف می‌کند، تعیین کرد و از این رهگذر مدل‌های زبانی متعددی به دست آورد. در جدول‌ها می‌توان دو فعالیت انجام داد

۱- مرتب کردن جدول‌ها بر حسب ترتیب قاموسی کلمات و مقدار احتمال‌های محاسبه شده،

۲- جستجو در جدول‌ها بر حسب کلمات.

برنامه‌ها به زبان ویژوال بیسیک و در محیط ویندوز ۹۸ نوشته شدند.

استخراج مدل زبانی

برنامه‌ها روی ۵۰۱۱۱۱ کلمه اجرا شدند. پس از ۲۴ ساعت کار کامپیوتر معلوم شد زمان اجرای برنامه حدّاقل سه هفته به طول می‌انجامد. با توجه به این که از همین برنامه، البته با بعضی تفاوت‌های ماهوی، می‌بایست برای استخراج مدل زبانی با حجم داده‌های بسیار بالا استفاده شود، تصمیم گرفته شد که الگوریتم بعضی از برنامه‌های بستهٔ نرم‌افزاری که اجرای آنها زمان‌بر است تغییر کند و به زبان ویژوال سی نوشته شوند. این فعالیت انجام شد و زمان اجرای برنامه برای استخراج جدول‌ها به حداکثر دو ساعت کاهش یافت، که بسیار رضایت‌بخش بود.

پس از اجرای برنامه روی ۵۰۱۱۱۱ کلمه و استخراج جدول‌ها مشخص شد ۱۲۵ کلمه برچسب دهی نشده‌است. به این ترتیب خطای مرحله تقطیع و برچسب‌دهی ۰/۰۰۲۴ کلمه بیشتر از حجم داده‌ها در قرارداد طرح است. علاوه بر آن باید توجه داشت که تأثیر ۱۲۵ کلمه بر روی مدل زبانی از ۰/۰۰۲۴ نیز کمتر خواهد بود.

حجم واژگان دادگان ۳۰۱۴۱ کلمه است که با ۳۹۰ برچسب نحوی - معنایی تقطیع شده‌اند. علاوه بر آن، فراوانی ۲۹۴۰۸ کلمه از ۳۰۱۴۱ کلمهٔ واژگان کمتر از ۱۰۰ است. بنابراین نمی‌توان از مدل زبانی در مرحلهٔ امکان سنجی بهره‌برداری مؤثر کرد، مگر آن که مدل زبانی بر حسب موضوع استخراج شود که آن هم به علت حجم کم داده‌های موضوعی کارآیی خوبی نخواهد داشت.

با اجرای برنامه، پروندهٔ جدول‌های مدل زبانی به عنوان استخراج مدل زبانی به دست آمدند. حجم فایل جدول‌ها با پیشروی ۴۵،۳ مگابایت است. پرونده‌ها به

مثابه دادگانی هستند که می‌توان از آنها برای مدل زبانی در بازشناسی گفتار پیوسته فارسی استفاده کرد.

تشکر و قدردانی:

از همکار عزیزم آقای دکتر علی درزی به خاطر رهنمودهای مؤثر در اجرای طرح کمال تشکر و قدردانی را دارم. از دیگر همکار عزیزم آقای مهندس حسین رازی زاده که زحمت نوشتن نرم‌افزار تقطیع و برچسب‌دهی را کشیدند، بی‌نهایت متشکرم. از ادارهٔ امور پژوهش‌های کاربردی معاونت پژوهشی دانشگاه تهران که اعتبار این طرح پژوهشی را تأمین کردند، تشکر و قدردانی می‌نمایم.

منابع:

۱- اسلامی، محرم: شناخت نوای گفتار زبان فارسی و کاربرد آن در بازسازی و بازشناسی رایانه‌ای گفتار، پایان نامهٔ دکتری، گروه زبان شناسی همگانی، دانشکدهٔ ادبیات و علوم انسانی، دانشگاه تهران، ۱۳۷۹.

۲- شفائی، احمد: مبانی علمی دستور زبان فارسی، تهران: چاپخانهٔ خوشه، چاپ اول، ۱۳۶۳.

۳- شفیعی کدکنی، محمدرضا: موسیقی شعر، تهران: انتشارات آگاه، چاپ ششم، ۱۳۷۹.

۴- صادقی، علی‌اشرف و ارژنگ، غلامرضا: دستور سال دوم، آموزش متوسطه عمومی، فرهنگ و ادب، چاپ کیهانک، ۱۳۶۳.

5- Joao P.N.: A large vocabulary continuous speech Recognition Hybrid system for the portugese language, ICSLP'98 proceedings, CD ROM. Australia, Sydney, 1998.

6- Ren-Yuan Lyu, et.al: Golden Mandarin (3), IEEE proceedings, pp.57-60, 1995.

7- Zue, Victor, et.al: the Summit speech Recognition System: Phonological modelling and Lexical Access, IEEE Proceedings, pp. 49-52, 1990.