دانشگاه صنعتی امیرکبیر

دانشکده‌ی ریاضی و علوم کامپیوتر

پایان‌نامه ی کارشناسی ارشد

مدلی برای تحلیل و تولید متن وبلاگ فارسی

فرزانه سرافراز

استاد راهنما: دکتر محمد ابراهیم شیری

استاد مشاور: دکتر رضا عزمی

۱۳۸۵

فهرست مطالب

چکیده

هدف از این پروژه استفاده از پیکره‌ی متنی وبلاگ‌های فارسی برای استخراج اطلاعات زبان‌شناختی و یافتن مدلی برای تولید خودکار متن فارسی است.

مهم ترین ابزار مورد نیاز برای حل مسائل مختلف به روش تجربی در حیطه‌ی پردازش زبان طبیعی وجود پیکره‌ی زبانی بزرگ و متعادل است. در زبان فارسی با رشد کمّیِ وبلاگ‌های فارسی در سال‌های اخیر چنین پیکره‌ی متنی بزرگی در اینترنت وجود دارد. این پیکره با اینکه خصوصیات یک پیکره‌ی متعادل را ندارد، اما ویژگی‌هایی دارد که آن را برای تحلیل زبان‌شناختی مناسب می‌کند.

برای اینکه پیکره‌ی بسیار ناهمگن وبلاگ‌های فارسی را که در زمان‌ها و موقعیت‌های مختلف نوشته شده و نویسنده‌های متعدد دارد تبدیل به پیکره‌ی مناسبی برای پردازش ماشینی کنیم روی آن انواعی از پیش پردازش‌ها انجام می‌دهیم و سپس با روش‌های مختلف آماری ویژگی‌های زبان‌شناختی آن را بررسی می‌کنیم.

در پایان مدلی برای تولید خودکار متن مشابه پیشنهاد می‌کنیم و نتایج حاصل از تولید خودکار متن را ارزیابی می‌کنیم.

کلمات کلیدی:

پردازش زبان طبیعی، زبان‌شناسی محاسباتی، تولید زبان طبیعی، پیکره‌ی متنی، مدل زبانی آماری، n- نگاشت، وبلاگ.

فصل اول

مقدمه

فصل ۱: مقدمه

یکی از ویژگی‌های برجسته‌ی قرن بیست‌ویکم، به وجود آمدن چرخه‌ی اطلاعات زبانی از طریق شبکه‌های جهانی و اینترنت است. بر اثر عمومی شدن خدمات کامپیوتری و آشنایی بیشتر مردم با دانش کامپیوتر به سرعت بر حجم داده‌های نوشتاری و گفتاری زبان‌ها افزوده می‌شود. به عنوان مثال، مشترکان تلفن همراه در ایران هر روز بیش از ده‌میلیون پیام کوتاه به یکدیگر ارسال می‌کنند که با افزایش حجم پیام های کوتاه از ۱۶۰ نویسه به ۸۰۰ نویسه، حجم داده‌های زبانی فقط بر روی شبکه‌ی مخابراتی می‌تواند پنج برابر شود. این موقعیت سه نتیجه‌ی اساسی در بر داشته است. نتیجه‌ی اول، تأثیری است که بر واژگان و ویژگی‌های ساختی همه‌ی زبان‌ها و از جمله زبان فارسی وارد می‌شود. نتیجه‌ی دوم، فراهم شدن حجم عظیمی از داده‌های زبانی به صورت الکترونیکی است که می‌تواند منبع مهمی برای تجزیه و تحلیل زبانی باشد. نتیجه‌ی سوم، اهمیت روزافزون پردازش داده‌های زبانی به منظور دست‌یابی به سیستم های فن‌آوری زبان از قبیل نویسه‌خوان نوری[1] یا OCR، ترجمه‌ی ماشینی[2]، تولید زبان‌های گفتاری و نوشتاری[3]، و درک زبان‌های گفتاری و نوشتاری است. بعضی از صاحب‌نظران معتقدند [1] که وظیفه‌ی صیانت از زبان فارسی در محیط کامپیوتری به عهده‌ی دولت و به طور مشخص فرهنگستان زبان و ادب فارسی است، اما مدیریت و شناخت علمی اطلاعات زبانی در گرو تعامل زبان‌شناسان، ادبا، و متخصصان علوم کامپیوتر در رشته‌های مختلف دانشگاهی است. بنابراین غنای دانش ما از زبان فارسی در محیط کامپیوتری تابع توسعه‌ی زبان‌شناسی نظری و محاسباتی در کنار یکدیگر است.

زبان فارسی مانند هر زبان طبیعی دیگری پیچیدگی‌هایی دارد که انجام عملیات محاسباتی و پردازش خودکار را با مسائل متعددی مواجه می‌کند. در این پایان‌نامه با پردازش متونِ زبان طبیعی و در واقع تنها با زبان نوشتاری سروکار

1 Optical Character Recognition
2 Machine Translation
3 Automatic Language Generation

داریم. این مسئله باعث می‌شود گرچه به جهت از دست دادن اطلاعات گویشی مانند لحن گوینده، آهنگ صدا، تأکید، و مکث با مشکلات و ابهاماتی مواجه شویم ولی در مقابل با شکل محدودتری از زبان کار کنیم.

در تلاش برای ساخت یک سیستم پردازش و درک متون فارسی با مسائل و مشکلاتی مواجه می‌شویم که بعضی از آن‌ها خاص زبان فارسی‌اند، بعضی بین فارسی و زبان‌های نزدیک آن مشترک‌اند، و بعضی از این مسائل در بیشتر زبان‌های طبیعی دیده می‌شوند. در فصل ۲ به برخی از این مسائل اشاره می‌کنیم. خواهیم دید که بعضی از این مسائل کار *تولید* زبان طبیعی را آسان‌تر می‌کنند.

۱-۱ پیکره

اصولاً دو روش اصلی متفاوت برای پردازش زبان طبیعی و به طور خاص تولید زبان طبیعی به صورت محض وجود دارد: روش تجربی[1] و روش تکوینی[2]. در سال‌های اخیر روش تجربی اهمیت زیادی پیدا کرده و توجه دانشمندان این رشته را به خود جلب نموده است[2] و [3].

مهم ترین ابزار مورد نیاز برای حل مسائل مختلف به روش تجربی در حیطه‌ی پردازش زبان طبیعی وجود پیکره‌ی زبانی بزرگ است. در زبان فارسی با رشد کمیِ وبلاگ‌های فارسی در سال‌های اخیر چنین پیکره‌ی متنی‌ای در اینترنت وجود دارد. این پیکره با این که خصوصیات یک پیکره‌ی متعادل را ندارد، اما ویژگی‌هایی دارد که آن را برای تحلیل زبان‌شناختی مناسب می‌کند.

۱-۲ وبلاگ

وبلاگ یک قالب روزنامه‌نگاری برخط است که به ترتیب زمانی عکس منتشر می‌شود، مرتباً به‌روز می‌شود، و نوعاً شامل تفکرات و تأملات شخصی، زندگی روزمره، مقالات، نظرات، و پیوند به صفحات وب است. [ویکی‌پدیا].

1 empirical approach
2 formalist approach

در سال‌های اخیر تعداد وبلاگ‌های فارسی افزایش چشمگیری پیدا کرده است. طبق آمار امروز زبان فارسی چهارمین زبان متداول وبلاگ در دنیا است و در حال حاضر بیش از ۷۵۰۰۰ وبلاگ به زبان فارسی در ایران و خارج از ایران نوشته می‌شود [4].

متن وبلاگ ویژگی‌هایی دارد که آن را از سایر پیکره‌های زبانی متمایز می‌کند؛ از جمله زبان غیر رسمی و محاوره‌ای، مفاهیم و موضوعات تکراری، و وجود شکلک و پیوند به صفحات وب در بین متن. این ویژگی‌ها باعث می‌شوند بررسی وبلاگ‌های فارسی چه از نظر زبان‌شناختی و چه از دیدگاه محاسباتی جالب توجه باشد.

۱-۳ تولید زبان طبیعی

سیستم های تولید زبان طبیعی[1] از داده‌هایی که کامپیوتر به آن دسترسی دارد متن زبان طبیعی (مثل فارسی یا انگلیسی) تولید می‌کنند. تولید زبان یک زیرشاخه‌ی مهم کاربردهایی مانند ترجمه‌ی ماشینی، مکالمه‌ی انسان و کامپیوتر، و توضیح و خلاصه کردن متن است[5].

امروز کاربرد اصلی سیستم های تولید زبان طبیعی کمک به نویسندگان نوشتارهای تکراری مانند نامه‌های اداری و گزارش‌های هواشناسی است. کاربرد دیگر آن به عنوان ابزار توضیح محاوره‌ای است که کار مبادله‌ی اطلاعات را با کاربرهای غیرمتخصص به ویژه در مهندسی نرم افزار و کاربردهای پزشکی آسان می‌سازد.

۱-۴ هدف پروژه، مراحل کاری، و ساختار این نوشتار

هدف از این پروژه استفاده از پیکره‌ی متنی وبلاگ‌های فارسی برای استخراج اطلاعات زبان‌شناختی و یافتن مدلی برای تولید خودکار متن فارسی است. این کار در پنج مرحله انجام می‌شود:

۱. جمع آوری پیکره و استخراج متن

1 Natural Language Generation

با استفاده از تنظیمات پیش رفته‌ی برنامه‌ی wget و یک مدول به زبان پیتون حجم زیادی وبلاگ فارسی را از اینترنت بارگیری و روی دیسک ذخیره می‌کنیم. سپس متن خام فارسی مربوط به مطالب وبلاگ را از اطلاعات اضافی بارگیری شده (مانند اطلاعات html) جدا می‌کنیم.

۲. پیش پردازش متن و تولید پیکره‌ی متنی

هر متنی پیش از آنکه برای بررسی و کار زبان‌شناختی مناسب شود احتیاج به پیش پردازش دارد. این پیش پردازش عمدتاً برای یک‌دست شدن متن از نظر املاء، رسم الخط، مجموعه‌ی نویسه‌های به کار رفته و مانند آن است.

۳. تحلیل پیکره

پیکره را به لحاظ اطلاعات آماری از جمله تعداد واژه‌های مختلف، بسامد رخداد هر واژه، و سایر اطلاعات زبان‌شناختی بررسی می‌کنیم.

۴. تولید خودکار متن مشابه

با استفاده از نتایج به‌دست آمده در مراحل بالا اثر اجرای روش‌های خودکار تولید متن را روی تولید متن فارسی مطالعه می‌کنیم.

۵. بررسی کیفیت متن تولیدشده

نهایتاً لازم است متن تولیدشده را از لحاظ کیفیت بررسی کنیم. یک‌نواختی متن، قابل فهم بودن، و اشتباهات دستوری آن را پیدا می‌کنیم. امکان استفاده از برنامه را روی پیکره‌های مختلف دیگر به جز وبلاگ می‌سنجیم.

ساختار این پایان‌نامه به شرح زیر است:

در فصل بعدی (فصل دوم) مختصری در مورد تاریخچه و ضرورت طرح مسأله بحث می‌کنیم. روش‌های فنی استفاده شده در این زمینه تا کنون را مطرح می‌کنیم و مفاهیم اولیه را معرفی می‌کنیم.

در فصل سوم به تفصیل در مورد مسائل عمده‌ی مطرح در پردازش زبان طبیعی بحث می‌کنیم. گزارشی از نتایج مطالعه در مورد مشکلات و شیوه‌های مختلف برخورد با آن‌ها را می‌آوریم و جدیدترین کارهای انجام شده در این زمینه را بررسی می‌کنیم.

در فصل چهارم به مسأله‌ی جمع آوری پیکره‌ی زبانی از روی شبکه‌ی جهانی اینترنت می‌پردازیم و جزئیات فنی واحد جمع آوری کننده‌ی پیکره را شرح می‌دهیم. در پایان این فصل نتایج حاصل از کار جمع آوری پیکره‌ی فارسی از روی وب را جمع بندی می‌کنیم.

در فصل پنجم به مسائل موجود در مرحله‌ی پیش پردازش پیکره‌های خام متن نوشتاری زبان طبیعی می‌پردازیم. روش‌های پیشنهاد شده‌ی قبلی در این زمینه را مطرح می‌کنیم، اشکالات و خوبی‌های هر یک را برمی‌شماریم، و راه حل به‌کاررفته در این پروژه را تشریح می‌کنیم.

در فصل ششم به مسائل نظری و مبانی علمی مسأله‌ی کلاسیک تولید زبان طبیعی می‌پردازیم. زیربناهای نظری و ریاضی آن را عنوان می‌کنیم. نتایج چاپ شده در تحقیقات کلاسیک این مبحث را بیان می‌کنیم، و چند نمونه از آن‌ها را با یافته‌های این تحقیق مقایسه می‌کنیم.

در فصل هفتم ساختار سیستم پیشنهادی را به‌طور کامل توضیح می‌دهیم، و نتایج اجرای سیستم را در همه‌ی مراحل اجرای بخش های مختلف آن از ابتدا تا انتها بیان می‌کنیم.

در فصل هشتم سیستم را بر اساس هوش انسانی ارزیابی می‌کنیم و از مجموعه‌ی کارهای انجام شده نتیجه‌گیری می‌کنیم. در پایان برای توسعه‌ی سیستم و کارهای آینده پیشنهاداتی ارائه می‌کنیم.

فصل دوم

تاریخچه ، ضرورت ، و روش انجام تحقیق

فصل ۲: تاریخچه، ضرورت، و روش انجام تحقیق

۲-۱ پردازش زبان طبیعی چیست

زبان اصلی‌ترین شاهد در اثبات هوشمندی انسان است. انسان از زبان برای بیان نیازهای اولیه و سوداهای متعالی خود، و برای انتقال دانش فنی و گزارش خیال‌پردازی‌هایش استفاده می‌کند. با زبان ایده‌ها و افکار ماورای زمان و مکان پخش می‌شوند. در زیر مثال‌هایی از غنای زبان فارسی و انگلیسی می‌بینید:

1. Overhead the day drives level and grey, hiding the sun by a flight of grey spears. (William Faulkner, *As I Lay Dying*, 1935)

2. یار آن نیست که مویی و میانی دارد بنده‌ی طلعت آن باش که آنی دارد (حافظ، قرن هشتم هجری)

3. سعی کنید از دستگاه در محیط‌های خشک استفاده کنید و از استفاده در محیط‌های خیلی گرم، خیلی سرد، دودآلود و مرطوب خودداری کنید. (راهنمای پخش کننده‌ی موسیقی)

4. When using the toaster please ensure that the exhaust fan is turned on. (sign in dormitory kitchen)

5. گرامر پیوندی به عنوان فرمالیسمی برای بازنمایی زبان‌های مستقل از متن به علت لغوی بودن و در نتیجه کاهش پیچیدگی دستور زبان و [...] می‌تواند به عنوان بدیلی برای گرامرهای با ساختار گروهی بکار رود. ([بارفروش])

6. Amiodarone weakly inhibited CYP2C9, CYP2D6, and CYP3A4-mediated activities with Ki values of 45.1-271.6 μM (Medline)

7. Iraqi Head Seeks Arms (spoof headline,

http://www.snopes.com/humor/nonsense/head97.htm)

8. The earnest prayer of a righteous man has great power and wonderful results. (James 5:16b)
9. Twas brillig, and the slithy toves did gyre and gimble in the wabe (Lewis Carroll, *Jabberwocky*, 1872)
10. There are two ways to do this, AFAIK :smile: (internet discussion archive)

به علت پیچیدگی و غنای زبان، تحقیق در مورد آن، بخشی از بسیاری از رشته‌های علمی به جز زبان‌شناسی است که از معروف‌ترین آنها می‌توان از ترجمه، نقد ادبی، فلسفه، مرهم شناسی، و روان‌شناسی نام برد. در میان رشته‌های نابدیهی‌تری که با زبان سروکار دارند حقوق، علم هرمنوتیک، علوم دیوانی، علوم تربیتی، باستان‌شناسی، رمزنگاری تحلیلی، و آسیب‌شناسی گفتاری به چشم می‌خورند. هر یک از این رشته‌ها از متدلوژی منحصر به خود برای جمع آوری مشاهدات، تولید نظریه‌ها و آزمون آنها بهره می‌برند. با این همه تمام اینها در جهت تعمیق درک ما از این بیانیه‌ی هوش و خرد انسانی گام برمی‌دارند.

میزان اهمیت زبان در علوم و هنر به ذخایر فرهنگی آن زبان ربط دارد. هر کدام از حدود ۷۰۰۰ زبان طبیعی در نوع خود غنی است، در تاریخ شفاهی‌ای که به دنبال دارد، در ساختارهای دستوری خود، و در تفاوت‌های ظریف معنایی کلمات. زبان‌ها در ارتباط با یکدیگر تکامل پیدا می‌کنند. پیشرفت‌های فن‌آوری کلمات جدیدی به زبان اضافه می‌کند مانند *وبلاگ* و حتی ساخت‌های زبانی را تغییر می‌دهد مانند پیشوندهای - *e* و - *cyber*.

با هر موج پیشرفت در فن‌آوری محاسبات با چالش جدیدی در تحلیل زبان روبه‌رو می‌شویم. زبان‌های ابتدایی ماشین به زبان‌های برنامه‌نویسی پیچیده و سطح بالا تبدیل شدند که مترجم ها و کامپایلرها به صورت خودکار آنها را تجزیه می‌کنند. از پایگاه‌های داده با عبارت‌های زبانی مانند SELECT age FROM employee سؤال می‌کنیم. ابزارهای محاسباتی مدرن مجهز به رابط‌های کاربر چند وضعیتی شده‌اند که متن، گفتار، و اشارات را درک و تحلیل می‌کنند. با این مقدمات ساختن سیستم های جدید پردازش زبان طبیعی نیاز به تحلیل گسترده و سطح بالای زبان‌شناختی دارد.

امروزه بزرگ‌ترین چالش تحلیل زبان طبیعی در اثر گسترش انفجاری متن و مولتی‌مدیا در شبکه‌ی جهانی وب است. بسیاری از مردم بخش بزرگ و درحال بزرگ‌ترشدنی از وقت کار و تفریح خود را به استفاده از این دنیای اطلاعاتی می‌پردازند. *از چه مکان‌های توریستی‌ای در خلیج فارس با بودجه‌ی محدود می‌توانم بازدید کنم؟ صاحب‌نظران متخصص در مورد دوربین دیجیتال Canon چه می‌گویند؟ در هفته‌ی گذشته چه پیش بینی‌هایی در مورد بازار آهن شده است*؟ جواب دادن به این پرسش ها و بسیاری پرسش های مشابه دیگر نیاز به مجموعه‌ای از کارهای پردازش زبان از جمله بازیابی اطلاعات، استنتاج، و خلاصه‌سازی دارد. ابعاد این کارها معمولاً محاسبات سطح بالایی می‌طلبد.

همانطور که دیدیم *پردازش زبان طبیعی*[1] (NLP) اهمیت ویژه‌ای در زمینه‌های علمی، اقتصادی، اجتماعی، و فرهنگی دارد. نظریه و روش‌های پردازش زبان طبیعی به سرعت در حال پیشرفت و تحول هستند. در مراکز علمی رشته‌هایی از علوم انسانی، علوم کامپیوتر، و هوش مصنوعی تا زبان‌شناسی مشغول مطالعه و تحقیق بر روی جنبه‌های مختلف پردازش زبان طبیعی هستند. در صنعت گروه‌هایی که روی محاوره‌ی انسان و کامپیوتر، تحلیل اطلاعات تجاری، و توسعه‌ی نرم افزار کار می‌کنند از جمله‌ی کسانی‌اند که روی پردازش زبان طبیعی کار می‌کنند[6].

۲-۲ تاریخچه‌ی مختصر پردازش زبان طبیعی

چالش بزرگ علم کامپیوتر از آغاز تا کنون ساختن ماشین هوشمند بوده است. معیار اساسی هوشمندی ماشین نیز همیشه توانایی زبانی آن، یا *آزمون تورینگ*[2] بوده است. آیا ممکن است که یک *سیستم محاوره*[3] که متناسب با ورودی کاربر پاسخ می‌دهد آنقدر طبیعی رفتار کند که کاربر تفاوت آن رفتار را با رفتار مخاطب انسانی که از رابط یکسان استفاده می‌کند را متوجه نشود؟ امروزه تحقیقات گسترده‌ای در زمینه‌های مختلف ترجمه‌ی ماشینی در حال انجام است و سیستم های تجاری زیادی بر پایه‌ی این پیشرفت ها توسعه یافته‌اند. محاوره‌ی زیر یک نمونه از این کاربرد تجاری

1 Natural Language Processing
2 Turing Test
3 dialog system

است:

بفرمایید.

آیا می‌دانید فیلم خانه‌ی دوست کجاست در چه ساعتی پخش می‌شود؟

کدام سالن سینما؟

سینما فلسطین.

خانه‌ی دوست کجاست در سینما فلسطین نمایش داده نمی‌شود، اما در سئانس های ۳:۰۰، ۵:۳۰، ۸:۰۰، و ۱۰:۳۰ در سینما سپیده پخش می‌شود.

و یا مثال زیر که مدتی است که در ایران کاربرد دارد و هنوز کاملاً خودکار نشده است.

راهنمای یکصدوبیست‌وهفت، بفرمایید.

ببخشید، شماره‌ی بیمارستان میلاد را می‌خواهم.

یادداشت بفرمایید... هشتادوهشت صفرشش بیست‌ودو پنجاه الی دو.

سیستم های تجاری امروز کاملاً محدود به دامنه‌های خیلی محدود مانند نمونه‌های بالا هستند. امکان ندارد بتوانیم از سیستم های بالا انتظار داشته باشیم بتوانند راننده‌ای را به سمت نشانی نزدیک‌ترین رستوران راهنمایی کنند مگر اینکه از قبل اطلاعات کافی با قالب‌بندی مناسب در اختیار سیستم قرار گرفته باشد و پرسش و پاسخ‌های مرتبط با موضوع به سیستم معرفی شده باشد. توجه کنید که به نظر می‌رسد سیستم های بالا هدف کاربر را می‌فهمند: کاربر زمان نمایش یک فیلم را می‌پرسد و سیستم به درستی تشخیص می‌دهد که کاربر می‌خواهد فیلم را ببیند. این استنتاج آنقدر برای انسان عادی است که معمولاً حتی متوجه آن نمی‌شویم؛ اما برای اینکه سیستم های پردازش زبان طبیعی خوب رفتار کنند لازم است که این قابلیت در آنها پیاده‌سازی شده باشد. بدون این قابلیت، سیستم در جواب «آیا می‌دانید فیلم

خانه‌ی دوست کجاست در چه ساعتی پخش می‌شود؟» خیلی ساده جواب می‌دهد «بله.» که چندان جواب به درد خوری نیست.

با وجود پیشرفت‌های اخیر، سیستم های پردازش زبان طبیعی که تا کنون ساخته شده‌اند عمدتاً قادر به انجام برهان عقلانی[1] یا ساختن پایگاه دانش بر اساس دریافت‌های جهانی نیستند. می‌توان نشست و منتظر ماند تا این مسائل پیچیده‌ی هوش مصنوعی حل شوند، اما تا آن وقت باید با محدودیت‌های جدی توانایی‌های برهان و دانش سیستم های پردازش زبان طبیعی کنار آمد. از این رو، از همان آغاز معرفی این شاخه از دانش، یک هدف عمده‌ی تحقیقات پردازش زبان طبیعی پیشبرد توانایی‌های محاوره‌ی زبان طبیعی بدون نزد یک شدن به حیطه‌ی توانایی‌های برهان و دانش آن بوده است[7].

ایده‌ی اولیه‌ی اینکه می‌توان به زبان طبیعی از دید محاسباتی نگریست از یک برنامه‌ی تحقیقاتی در سال ۱۹۰۰ میلادی پدید آمد[8]. هدف این تحقیقات که دانشمندانی چون فرگه[2]، راسل[3]، ویتگنشتاین[4]، تاسکی[5]، لامبک[6]، و کارنپ[7] روی آن کار کرده‌اند در اصل ساختن برهان ریاضی با استفاده از دستگاه‌های منطقی بود. برای این منظور، توجه دانشمندان به زبان به عنوان یک سیستم [8] مناسب برای پردازش خودکار جلب شد. بعدها پیشرفت در سه شاخه‌ی مهم پایه‌های علم پردازش زبان طبیعی را بنا گذاشت. اولین این شاخه ها *نظریه‌ی زبان‌ها*[9] بود. در نظریه‌ی زبان ها، زبان به صورت مجموعه‌ای از رشته‌ها که توسط یک کلاس از اتومات‌ها پذیرفته می‌شود تعریف می‌شود؛ مانند زبان‌های مستقل از متن و اتومات pushdown. این تعریف رسمی قواعد نحوی لازم برای پردازش محاسباتی زبان را در اختیار ما قرار می‌دهد.

1 common-sense reasoning
2 Frege
3 Russel
4 Wittgenstein
5 Tarsky
6 Lambek
7 Carnap
8 formal system
9 formal language theory

دومین شاخه، *منطق صوری*[1] بود. منطق صوری شیوه‌ی نمادی‌ای برای استفاده از زبان طبیعی در بیان اثبات قضایای منطقی به دست می‌داد. حساب نمادی در منطق صوری سینتکس زبان و قواعد استنتاج را تعریف می‌کند. اگر چنین دستگاه حسابی داشته باشیم، و با سینتکس و سمنتیکس مشخص و تعریف‌شده، آنگاه می‌توانیم مفاهیم را با عبارات زبان طبیعی متناظر کنیم. برای مثال، عبارت «مریم علی را دید.» به فرمول

```
saw(m,a)
```

ترجمه می‌شود. ما (صراحتاً و تلویحاً) برای فعل «دید» یک رابطه‌ی دوتایی فرض می‌کنیم و «مریم» و «علی» را آرگومان‌های آن قرار می‌دهیم. همین مطلب برای عبارت‌های کلی‌تر هم که از سورهای مختلف استفاده می‌کنند برقرار است. مثلاً عبارت «همه‌ی پرنده‌ها پرواز می‌کنند.» را در نظر بگیرید. این عبارت در منطق صوری به شکل زیر بیان می‌شود:

```
∀x: bird(x) → fly(x)
```

این روشِ استفاده از منطق، راه را برای ماشینی که از روش‌های مکانیکی استفاده می‌کند باز کرد تا بتواند استنتاج منطقی‌ای را که بخشی از فهم زبان طبیعی است تولید و درک کند.

سومین پیشرفت، *اصل ترکیب*[2] بود. این اصل می‌گوید که معنای یک عبارت مرکب اساساً از معنای اجزایش و چگونگی ترکیب این اجزاء به دست می‌آید. این اصل تناظر مفیدی میان سینتکس و سمنتیکس برقرار می‌کند، و بطور خاص این امکان را مطرح می‌کند که معنای عبارات مرکب پیچیده می‌توانند به صورت بازگشتی محاسبه شوند. اگر بدانیم که *گزاره‌ی p نادرست است* را به شکل

```
not(p)
```

و *مریم علی را دید* را به شکل

```
saw(m,a)
```

1 symbolic logic
2 principle of compositionality

نشان می‌دهند، آنگاه می‌توان عبارت «مریم علی را دید درست نیست.» را به شکل زیر نمایش داد:

```
not(saw(m,a))
```

این نوع گرامر امروزه کاربردهای زیادی دارد که از جمله‌ی این کاربردها می‌توان از گرامرهای استفاده‌شده در برنامه‌های پردازش زبان طبیعی که در زبان برنامه‌سازی منطق prolog پیاده‌سازی شده‌اند نام برد.

۳-۲ انتخاب زبان پیتون برای انجام محاسبات

زبان برنامه‌نویسی پیتون[1] یک زبان کدنوشته[2]ی ساده و در عین حال بسیار قدرتمند با کارایی فوق‌العاده برای پردازش زبان طبیعی است. پیتون را می‌توان به رایگان از وبگاه www.python.org بارگیری کرد. در اینجا یک برنامه‌ی پنج‌خطی به زبان پیتون را می‌بینید که متنی را از ورودی می‌گیرد و همه‌ی کلماتی را که به ing ختم می‌شوند نمایش می‌دهد:

```
import sys                              # load the system library
for line in sys.stdin.readlines():      # for each line of input
    for word in line.split():           # for each word in the line
        if word.endswith('ing'):        # does the word end in 'ing'?
            print word                  # if so, print the word
```

با نگاه سریع به این برنامه چندین ویژگی مهم زبان پیتون روشن می‌شود. اول اینکه در پیتون قطعه کدهای تو در تو با فاصله از هم جدا می‌شوند؛ یعنی مثلاً آزمون ing برای هر کلمه انجام می‌شود. دوم اینکه پیتون یک زبان شیءگرا[3] است و اشیاء متدهایی دارند که برای انجام عملیات مختلف صدا زده می‌شوند. سوم اینکه درست مانند زبان C++ و جاوا متدها با نقطه از اشیاء جدا می‌شوند و صفر، یک، یا چند آرگومان دارند. از همه مهم تر اینکه برنامه‌ی نوشته شده به زبان پیتون بسیار خوانا است و کدنوشته‌های پیتون معمولاً بی‌نیاز از مستندات هستند. حتی اگر با برنامه‌نویسی آشنایی چندانی نداشته باشید احتمالاً به راحتی می‌توانید حدس بزنید که برنامه‌ی بالا چه می‌کند. برتری دیگر زبان پیتون پشتیبانی خوب آن از نوع داده‌ی رشته است. با پیتون به راحتی می‌توانید کل یک کتاب چندصد صفحه‌ای را در یک

1 Python programming language
2 scripting language
3 object-oriented