



ترجمه ماشینی زبان فارسی: راه کارها و موانع

امیر شهاب شهابی

دانشگاه آزاد اسلامی

دکتر عبدالحسین صرافزاده

استادیار دانشگاه تربیت معلم

در این مقاله، خلاصه پژوهش‌های انجام شده در زمینه پردازش زبان فارسی اعم از واکافت ساختوازی، نحوی و معنایی، نحوه بازنمایی معنای جملات و به کارگیری دانش بازنمایی شده، درک متن فارسی و ساختار دستور زبان فارسی و نیز نظریه گشتاری چامسکی در ساختار زبان فارسی و پژوهش‌هایی از این قبیل گردآوری گردیده‌اند. این مجموعه می‌تواند در ارائه طریق به پژوهشگرانی که هدفشان تحقیق در حوزه پردازش زبان فارسی است، کمک مؤثری باشد. در ضمن، به موانع موجود به خصوص در زمینه ابهام معنایی اشاره گردیده و در خاتمه راه حل مناسبی برای ترجمه ماشینی فارسی مورد بررسی قرار گرفته است.

مقدمه

ترجمه خودکار متون، جزو اولین کاربردهای غیر محاسبه‌ای کامپیوتر است. ایده ترجمه ماشینی اولین بار در بحثی بین وارن ویور و آندره بوث شکل گرفت (بوث، ۱۹۵۳). مدل آنان بر اساس جداولی طراحی شده بود که در آنها برای رمزگشایی، از تواتر کلمات و حروف استفاده می‌شد.

با پیدایش تکنیک‌هایی چون جست‌جوی دودوئی و روش تقلیل مسأله در الگوریتم‌های برنامه‌سازی و ایده ویرایش ثانویه در ترجمه، اعتقاد به عملی بودن ترجمه ماشینی قوت گرفت. در پایان دهه ۵۰ و اوایل ۶۰، COMIT به عنوان اولین زبان همه منظوره پردازش نماد مطرح شد و برنامه‌های کمکی واژه پرداز به مرحله پیاده‌سازی رسید. از سوی دیگر در زمینه زبان‌شناسی، نسبت به ساخت جملات زبان، دید آلفگوریتمیکی تحت عنوان دستور زبان گشتاری مطرح گردید (چامسکی، ۱۹۷۵).

نتیجه پروژه‌های دهه ۶۰ که بیشتر از آنها برای مقاصد سیاسی استفاده می‌شد، پیدایش دو شاخه اصلی در روش‌های ترجمه ماشینی بود. یکی از این روش‌ها استفاده از فرم‌های اصلاح شده

دستور زبان گشتاری و دیگری استفاده از اطلاعات آماری در مورد نقش کلمات در جمله و ترجمه آنها بود. در پایان دهه ۷۰، اعتقاد عمومی بر آن بود که ترجمه ماشینی باید پیچیده‌تر از جابه‌جایی و جایگذاری کلمات زبان مبدأ برای رسیدن به زبان مقصد باشد و ماشین باید به نوعی یک فهم از معنای هر دو نوشتار پیدا کند. این اعتقاد راه را برای استفاده از تکنیک‌های جدید هوش مصنوعی و به خصوص پردازش زبان طبیعی در ترجمه ماشینی هموار کرد.

در این مقاله با مرور بر ادبیات تحقیق، ابتدا به پژوهش‌هایی که در زمینه درک متن فارسی و چگونگی واکافت واژه‌ای، ساختوازی، نحوی و معنایی صورت گرفته، اشاره شده است و سپس اجزا و ارکان یک سامانه مترجم فارسی بیان و مهمترین رکن آن یعنی رفع ابهام معنایی از کلمات جمله‌های فارسی، مطرح گردیده است. در ضمن، روش‌های مختلف حل این مشکل به اختصار بررسی و در پایان به موانع پرداخته شده است. از آنجا که در زمینه ترجمه ماشینی زبان فارسی کمتر مطالعه شده است، این تحقیق سعی در تعمیم مطالعات به زبان فارسی دارد.



پردازش زبان فارسی

در سال ۱۳۷۲، در آزمایشگاه پردازش زبان طبیعی دانشگاه صنعتی شریف، مهنوش شمس فرد (۱۳۷۳) سامانه‌ای به نام دنا طراحی و پیاده نمود. در این سامانه با به کارگیری نظریه وابستگی مفهومی شنک (۱۹۷۵)، جمله‌های فارسی به شبکه‌ای از مفهوم‌ها و روابط میان آنها تبدیل می‌شوند. گام‌های گوناگون پردازش جملات عبارت‌اند از واکافت واژه‌ای، واکافت ساختوازی، واکافت نحوی، واکافت معنایی و استنتاج.

پس از آن، سه پروژه دیگر با تمرکز بر بخش‌های گوناگون سامانه دنا ۱ تعریف و اجرا شدند. هر یک از این پروژه‌ها سعی در گسترش یکی از بخش‌های این سامانه داشتند. یکی از آنها پروژه‌ای است که روی پردازش نحوی جمله‌های نسبتاً پیچیده زبان فارسی از جمله گروه‌های اسمی و جمله‌های مرکب متمرکز بوده (مهدیه، ۱۳۷۶). پروژه دیگر به طراحی و پیاده‌سازی بخش واژگان و نیز ارکان ساختوازی پرداخته (رضانیا، ۱۳۷۶) و سرانجام سومین پروژه، بخش استنتاج گر را با عمق بیشتری مورد توجه قرار داده است (امامی، ۱۳۷۶).

برای درک جمله‌های زبان فارسی، پس از برداشتن گام‌های گوناگون در زمینه پردازش ساختوازی و نحوی و پیش از به کارگیری شیوه‌های استنتاج باید چارچوب مناسبی برای بازنمایی معنای جمله‌ها در نظر گرفته شود تا ضمن پردازش معنایی، جمله‌ها به این زبان بازنمایی تبدیل شوند. با توجه به اهمیت پردازش معنایی در یک سامانه درک متن و ضرورت کار بیشتر در این زمینه، پروژه‌ای برای بررسی شیوه‌های بازنمایی معنا و انتخاب یک شیوه مناسب برای سامانه دنا ۲ تعریف شد که در طی اجرای پروژه با بررسی برخی شیوه‌های بازنمایی معنا، روش گراف‌های مفهومی به دلیل تمرکز تعداد زیادی از سامانه‌های زبان طبیعی بر آن، به عنوان شیوه مناسب‌تر انتخاب شد (فضلی، ۱۳۷۷).

در تحقیق دیگری (شهابی، ۱۳۷۶)، ابتدا، جمله‌های فارسی در محدوده معنایی خاص (درخواست اشتغال به کار افراد برای یک سازمان) توسط سیستم طراحی شده اخذ می‌شود. در گام اول، پردازش روی شناخت واژه‌های آن و یا به عبارت دیگر واکافت

واژه‌ای صورت می‌گرفت که هر نشانه از جمله با حروف آن خوانده می‌شد و پس از اتمام کلمه، تشخیص داده می‌شد که چه کلمه‌ای است و آیا جزو کلمات بامعنی زبان است یا خیر؟ سپس، گام دوم یعنی واکافت ساختوازی شروع می‌شد. به این معنا که آیا کلمه‌های مرکب از ترکیب صحیحی تشکیل شده‌اند یا خیر؟ پس از آن، گام سوم یعنی واکافت نحوی صورت می‌گرفت. یعنی آیا کلمه‌های تشکیل دهنده جمله دستور زبان فارسی را رعایت کرده‌اند و آیا جمله مزبور متعلق به زبان می‌باشد یا خیر؟ این کار را سیستم به کمک دستور زبانی که برای جمله‌های فارسی در آن تعبیه شده است انجام می‌دهد. این دستور زبان کلیه جمله‌های فارسی را شامل می‌شود و در واقع پایانه‌های آن به کلمه‌های فارسی ختم می‌شود (باطنی، ۱۳۷۴). پس از اینکه جمله اصلی تشخیص داده شد، شاید ژرف‌ساخت نیاز به تبدیل داشته باشد؛ مثلاً برای به دست آوردن جمله‌های سؤالی یا مجهول و غیره احتیاج به نوعی گشتار است که ژرف ساخت را به روساخت تبدیل نماید (مشکوة الدینی، ۱۳۷۴) که سیستم با وجود گشتارهای موجود روی زبان با تبدیل ژرف ساخت به روساخت به جمله مزبور می‌رسد و واکافت نحوی پایان می‌یابد. گام بعدی، واکافت معنایی یا استنباط معنا از جمله مذکور است. در این مرحله از یک شبکه معنایی به عنوان بازنمایی دانش موجود در معنای جمله استفاده شده است.

در یک تحقیق دیگر (شریفی، ۱۳۷۶)، درک معنای جمله‌ها در ترجمه ماشینی به شکل دیگری صورت گرفته است و از هیچ یک از روش‌های فوق برای رسیدن به معنای جمله و نگهداری آن در یک پایگاه دانش استفاده نشده است. در این روش سعی شده است روشی جهت ماشینی کردن عمل ترجمه از یک زبان محاوره‌ای به زبان محاوره‌ای دیگر ارائه شود. برای رسیدن به این مقصود با در توجه به پیدایش شاخه‌های جدید هوش مصنوعی در زمینه شبیه‌سازی رفتار انسان و یادگیری ماشین، چگونگی پیاده‌سازی مترجم در قالب یک سیستم یادگیرنده ارائه شده است. این مترجم با استفاده از تجربیات قبلی خود در زمینه ترجمه تک تک کلمات و ترتیب قرار گرفتن آنها در جمله، ترجمه جملات با ساختارها و کلمات جدید را حدس می‌زند و به صورت پویا به



می‌رود، مرتفع می‌گردد (دلماش و زاورل، ۱۹۹۶؛ دمتریو و اتول، ۲۰۰۰؛ ویلکس، ۱۹۹۷). روال مذکور با تعیین نقش کلمه، ابهام ساختاری آن را مرتفع و در پیدا کردن معنای صحیح کلمات کمک شایانی می‌نماید. البته ابهام معنایی کلمات خیلی پیچیده‌تر از ابهام ساختاری آن است و برای رفع آن به روش‌های پیچیده‌تر نیاز است. ولی روش نشانه‌گذار بخش‌کننده جملات در پیدا کردن معنای دقیق کمک می‌کند و مراحل اولیه آن را انجام می‌دهد. در واقع، رفع ابهام از معنای یک کلمه بر می‌گردد به اینکه آن کلمه در چه جمله‌ای به کار رفته و با چه کلماتی هم‌نشین گردیده است؛ چرا که یک کلمه در یک جمله یک معنی می‌دهد، در حالی که همان کلمه در جمله دیگر معنای دیگری به خود می‌گیرد.

رفع ابهام از معنای کلمات جملات یک زبان، از سه راه کلی می‌تواند انجام پذیرد که هر راه خود روش‌های متعددی دارد: اول روش رفع ابهام با مربی یا سرپرست است که مبنای آن مجموعه آموزشی برجسب دار است. روش دوم رفع ابهام بر مبنای منابع لغوی مانند فرهنگ لغت و یا فرهنگ تساروس است (ویلکس و استیونس، ۱۹۹۷ و ۱۹۹۸). و بالاخره روش سوم رفع ابهام بدون مربی یا سرپرست است که در این حالت تنها مجموعه لغات و متون بدون برجسب در دسترس است (کیت و ویلکس، ۲۰۰۰).

الف) روش رفع ابهام با مربی

در این روش یک مجموعه لغات ابهام زدایی شده برای آموزش در دسترس است. این مجموعه، نمونه کلماتی است که هر کدام یک کلمه مبهم W می‌باشند و هر معنی آنها با یک برجسب معنایی متناسب با متنی که در آن قرار می‌گیرند تعریف شده‌اند (SK).

این کار باعث می‌شود که بتوان یک طبقه‌بندی آماری با سرپرست روی کلمات انجام داد. وظیفه اصلی این روش، ایجاد یک روتین طبقه‌بندی کننده است که کلمات جدید را بر اساس متن‌هایی که قبلاً تعریف شده‌اند (CI) در طبقه مناسب خود قرار می‌دهد. در رفع ابهام با سرپرست روش‌های مختلفی وجود دارد که از جمله آنها می‌توان استفاده از طبقه‌بندی بیس را نام برد (دپو

تکمیل دانش خود در این زمینه می‌پردازد. از آنجا که در این روش ترجمه، برای طراحی الگوریتم‌های ارائه شده جهت ایجاد پایگاه دانش و استنتاج دانش از آن از هیچ گونه اطلاعات پیش فرضی در زمینه چگونگی ساختار جملات زبان‌های مبدأ و مقصد و نقش کلمات در جمله استفاده نشده است، مترجم حاضر مستقل از زبان‌های مبدأ و مقصد، توانایی انجام عمل ترجمه از هر زبان به زبانی دیگر را دارد.

ترجمه ماشینی

برای اینکه بتوان جمله‌ای را از یک زبان ترجمه و به زبان دیگری تبدیل نمود، ابتدا به یک روال تشخیص واژه‌ها یا واکافت واژه‌ای و واکافت ساختاروی نیاز می‌باشد، به طوری که کلمات ساده و مرکب آن زبان از ورودی تشخیص داده شود. سپس باید ترکیب کلمات از نظر نحوی صحیح باشد و جمله متعلق به آن زبان را ایجاد نماید. حال برای اینکه این جمله به زبان دیگری برگردان شود، باید نقش و معنای هر کلمه مشخص و با توجه به دستور زبان مقصد و ترجمه کلمه با توجه به نقش آن، جمله زبان مقصد تولید شود. بنابراین دیده می‌شود که مهمترین بخش، مشخص کردن نقش و معنای کلمات می‌باشد. نقش کلمات با توجه به محل قرار گرفتن آن در جمله مشخص می‌گردد، اما مسئله مهم معنای کلمات می‌باشد؛ چرا که تعدادی از کلمات دارای چند معنی می‌باشند که این ابهام باید با روش مناسبی رفع گردد (مانینگ و شوتس، ۲۰۰۰). در ترجمه ماشینی ابهام هم در معنای کلمات زبان مبدأ وجود دارد (از نظر درک معنی آن) و هم در زبان مقصد و همین باعث می‌گردد روند ترجمه دچار مشکل شود. در ادامه، بیشتر به روش‌های حل این مشکل و موانعی که در زبان فارسی وجود دارد، پرداخته می‌شود.

رفع ابهام معنایی کلمات

یکی از اولین مشکلاتی که هر سیستم پردازش زبان طبیعی آن درگیر است، مسئله ابهام معنایی و ساختاری کلمات است. بخش عمده‌ای از این ابهام به کمک روالی به نام نشانه‌گذار بخش‌کننده جملات که برای تعیین نقش کلمات در جمله به کار



موانع ترجمه ماشینی زبان فارسی

همان طور که در بالا توضیح داده شد، مهمترین بخش یک سامانه مترجم ماشینی، تشخیص نقش کلمات در جمله و تشخیص معنای صحیح آنها با توجه به متنی است که آن کلمه در آن قرار گرفته است. برای تشخیص معنای صحیح کلمات از سه روش می‌توان استفاده نمود: در روش اول برای آموزش سامانه به یک مجموعه کلمات فارسی ابهام زدایی شده نیاز می‌باشد تا از روی آن بتوان عمل رفع ابهام برای کلمات جدید را در متن‌های مختلف انجام داد که متأسفانه هنوز این مجموعه به صورت یک فرهنگ فارسی قابل خواندن برای ماشین تهیه و تدوین نگردیده است. یعنی این اطلاعات به صورت دستی وجود دارند، اما نسخه رایانه‌ای از آنها با یک استاندارد معین برای تعریف هر کلمه و معنی و نقش‌های مختلف آن وجود ندارد. بدیهی است برای این کار ابتدا باید یک مرکز زبان‌شناسی معتبر، استاندارد ذخیره‌سازی اطلاعات و انواع آن را تدوین نماید و سپس نسخه‌ای رایانه‌ای از این اطلاعات جهت استفاده‌های بعدی تهیه شود.

در روش دوم نیز سامانه مستقیماً از یک فرهنگ لغت استفاده می‌کند، باز هم در مورد زبان فارسی این مشکل وجود دارد که هنوز فرهنگ‌های لغت فارسی رایانه‌ای با یک استاندارد مشخص تهیه نشده‌اند. این در حالی است که در زبان انگلیسی فرهنگ‌های متعددی مانند LDOCE, Wordnet یا Roget's International Thesaurus به صورت MRD یا نسخ قابل خواندن توسط رایانه وجود دارند که سامانه‌های مختلف مترجم ماشینی از آنها با یکی از روش‌های رفع ابهام مذکور بهره می‌جویند.

از روش سوم رفع ابهام نیز همان طور که دیده شد صرفاً برای لغات خاص و محدود فنی یا پزشکی استفاده می‌شود، نه برای کلیه لغات فارسی و چون متأسفانه هنوز استفاده از اصطلاحات فنی فارسی متداول و رایج نشده است، این روش رفع ابهام در زبان فارسی زیاد موفق نخواهد بود. از این رو برای اینکه بتوان از یک مترجم قوی و بدون مشکل فارسی بهره‌مند شد، ابتدا باید مقدمات برشمرده را فراهم نمود و این خود تلاش گروهی متشکل از زبان‌شناسان، مترجمان، ادیبان و متخصصان رایانه را می‌طلبد که سرآغازی خواهد شد برای رسیدن به این هدف ملی.

و دلمانس، ۲۰۰۰) و نیز نظریه اطلاعات که رفع ابهام بر مبنای آن نظریه صورت می‌گیرد (مانینگ و شوتس، ۲۰۰۰).

ب) روش رفع ابهام بر مبنای فرهنگ لغت

اگر اطلاعاتی در مورد طبقه‌بندی معنایی یک کلمه وجود نداشته باشد، در این صورت می‌توان از مشخصات عمومی معنای یک کلمه در فرهنگ لغت استفاده نمود. در این روش، تا کنون از سه نوع اطلاعات استفاده شده است. نوع اول روش لسک است که مستقیماً از تعریف معنای کلمات در فرهنگ لغت استفاده می‌شود. نوع دوم که نشان می‌دهد چگونه از اطلاعات طبقه‌بندی شده موجود در فرهنگ لغت می‌توان طبقه‌بندی معنایی یک کلمه را با توجه به متنی که آن کلمه در آن وجود دارد، به دست آورد. (یارافسکی، ۱۹۹۲) و سوم اطلاعاتی است که از ترجمه یک کلمه به کمک یک فرهنگ لغت دو زبانه به دست می‌آید و استفاده از معنای کلمه در زبان مقصد برای رفع ابهام از معنای کلمه.

ج) رفع ابهام بدون مربی

در این روش بدون اینکه اطلاعاتی در زمینه معنای لغات و کلمات به کار رفته در متن موجود باشد، ابتدا کلمات از نظر معنایی خوشه‌بندی و سپس بررسی می‌شود که هر کلمه جدید به کدام خوشه نزدیک‌تر است و به آن تعلق دارد و با توجه به اینکه به هر خوشه معنای خاصی اختصاص داده شده است، آن معنی برای آن لغت در نظر گرفته می‌شود و در ترجمه مورد استفاده قرار می‌گیرد. در دو روش قبلی، برای رفع ابهام به یک سری اطلاعات اولیه از معنای لغات نیاز بود، اما شرایطی، اگر چه نادر، پیش می‌آید که هیچ‌گونه اطلاعاتی از معنای کلمات در دسترس نیست؛ مثلاً در مورد اصطلاحات فنی و یا پزشکی که در فرهنگ‌های لغت عمومی پیدا نمی‌شود. در این حالات خاص است که روش رفع ابهام بدون سرپرست یا مربی مناسب است و در به دست آوردن معنای صحیح کلمه با توجه به متن کمک زیادی می‌نماید (کیت و ویلکس، ۲۰۰۰؛ مانینگ و شوتس، ۲۰۰۰).



نتیجه گیری

فرایندهای ترجمه ماشینی عبارت‌اند از: واکافت واژه‌ای، واکافت ساختوازی، واکافت نحوی، واکافت معنایی، تشخیص نقش کلمات در جمله، تشخیص معنای صحیح کلمات (با توجه به جمله حاوی و رفع ابهام از کلمات دارای معنی مشابه و یافتن معنی معادل کلمه در زبان مقصد و تبدیل جمله زبان مبدأ به جمله زبان مقصد با توجه به دستور زبان مقصد و در صورت نیاز اعمال گشتارهای مناسب برای تبدیل جملات مقصد به شکل روساختی آن). در این راستا مهمترین بخش یک مترجم که دقت در آن باعث بالا رفتن دقت کل عمل ترجمه می‌گردد، رفع ابهام معنایی از کلمات جملات زبان مبدأ و نیز رفع ابهام از ترجمه آنها به زبان مقصد می‌باشد. برای رفع ابهام معنایی کلمات سه روش کلی وجود دارد که عبارت‌اند از: (الف) رفع ابهام با مربی، (ب) رفع ابهام بر مبنای فرهنگ لغت، و (ج) رفع ابهام بدون مربی.

در روش اول مجموعه‌ای از لغات ابهام‌زدایی شده وجود دارد که سامانه با آن لغات آموزش داده می‌شود و اگر در حین عمل ترجمه با لغت جدیدی برخورد کرد، با توجه به اطلاعات آموزش دیده تشخیص می‌دهد که کلمه مزبور با توجه به متنی که در آن قرار گرفته در چه طبقه‌ای قرار می‌گیرد و معنای درست آن کدام است و آن را در ترجمه به کار می‌برد. در روش دوم اساساً برای هر کلمه از کل فرهنگ لغت استفاده می‌شود و با توجه به نقش

کلمه و دیگر مشخصات آن با کمک فرهنگ لغت معنی صحیح آن متناسب با جمله‌ای که در آن قرار گرفته است، به دست می‌آید. بالاخره روش سوم که روش رفع ابهام بدون مربی یا سرپرست است، مخصوص کلماتی است که در فرهنگ‌های لغت معمولی وجود ندارند (اصطلاحات فنی یا پزشکی و ...) و باید از فرهنگ‌های لغت تخصصی و یا دایرة المعارف‌ها بهره جست.

اما مشکل اصلی بر سر راه رفع ابهام از کلمات فارسی این است که اگر از دو روش اول استفاده شود، به یک نسخه رایانه‌ای از مجموعه لغات همراه با نقش‌های مختلفی که می‌توانند به خود بگیرند و معانی مختلفی که با توجه به نقش‌های مختلف و یا هم‌نشینی با کلمات مختلف به خود می‌گیرند، نیاز می‌باشد که متأسفانه این نسخه رایانه‌ای و همچنین استاندارد ذخیره‌سازی این دسته اطلاعات وجود ندارد که باید با کار گروهی متشکل از زبان‌شناسان، ادیبان و متخصصان رایانه ایجاد و جهت سیستم‌های مترجم زبان فارسی صورت بگیرد. استفاده از روش سوم رفع ابهام نیز خاص لغات فنی است که در زبان فارسی هنوز کاملاً جا نیفتاده است. بنابراین با توجه به این که دقت در رفع ابهام از معنای کلمات دقت مترجم را بالا می‌برد، هنوز نمی‌توان سامانه‌های مترجم فارسی مناسب را تهیه و از آنها بهره‌برداری نمود.

منابع

- امامی و م. بررسی مسایل درک متن فارسی و پیاده سازی نمونه‌هایی از آن، پایان نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۶.
- باطنی، م. توصیف ساختمان دستوری زبان فارسی، انتشارات امیرکبیر، ۱۳۷۴.
- رضانیا، ک. پی‌ریزی طرح کلی واژگان پیاده‌سازی واکاوی ساختوازی برای زبان فارسی، پایان نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۶.
- شریفی، م.، نوایی، م. ص. و دهقان نیری، ه. مترجم یادگیرنده ماشینی، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، ۱۳۷۶.
- شمس فرد، م. درک متن فارسی، پایان نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۳.
- شهابی، ا.ش. درک متن فارسی و تبدیل آن به پایگاه داده رابطه‌ای، پایان‌نامه کارشناسی ارشد، دانشکده تحصیلات تکمیلی، دانشگاه آزاد اسلامی - واحد تهران جنوب، ۱۳۷۶.



- فضلی، ا. بازنمایی معنای جمله‌های فارسی به کمک گراف مفهومی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۷.
 مشکوة‌الدینی، م. دستور زبان فارسی بر پایه نظریه گشتاری، انتشارات فردوسی مشهد، ۱۳۷۴.
 مهدیه، ط. بازنمایی جملات فارسی، پایان نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۶

Booth, A.D. (1953), *Machine Translation, Computer and Automation*, Vol. 2, no. 4,

Chomsky N., (1975), *Reflection on Language*, New York: Pantheon Books.

Daelemans, W., Zavrel, J. Berck P., Gillis S. (1996), *MBT: A Memory – Based Part-of-speech Tagger-Generator, Proceeding at 4th Workshop on Very Large Corpora*, Copenhagen.

Demetriou, G., Atwell, E.s., (2000), A domain Independent Semantic Tagger for the study of meaning Associations in English text. *IWCS-4 Program*.

De Pauw, g., Daelemans, W. (2000), The Role of Algorithm Bias Vs information Source in Learning Algorithms for morphosyntactic Disambiguation, *Proceeding of conll-2000 and LLL-2000*, Pages 19-24, Lisbon, Portugal.

Kit, C., Wilks, Y. (1999), *Unsupervised Learning of Word Boundary With Description Length Gain*. Computational Natural Language Learning.

Manning, C.D. Schutze, H. (1999), *Foundations of Statically Natural Language Processing*, MIT Press, Cambridge, MA.

Schank, R.C. (1975), *Conceptual Information Processing*, North Holland, Publishing Company, Amsterdam.

Wilks, Y., Stevenson, M. (1998), Word Sense Disambiguation Using Optimized Combinations of Knowledge Sources, *Proceedings of COLING/ ACL*.

Wilks, Y., Stevenson, M. (1997), Combining Weak Knowledge Sources for Sense Disambiguation. *Proceeding of the third conference of Recent Advances in Natural language Processing Conference*, pp. 1-7.

Wilks, Y. in N. Ide (Ed.) (1997), *Senses and Texts, Computers & G. humanities*, vol. 31, pp. 77-90.

Yarowsky, D. (1992), Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, *COLING 14*, 454-460.