ابهام‌زدایی واژگانی صفات چندمعنایی در ترجمهٔ ماشینی: بررسی پیکره- بنیاد

طیبه موسوی میانگاه[1] | **زهره ذوالفقار کندری**[2]

۱. دکتری زبان‌شناسی کاربردی و محاسباتی؛ دانشیار؛ دانشگاه پیام نور mosavit@pnu.ac.ir

۲. [پدیدآور رابط] دانشجوی دکتری زبان‌شناسی همگانی؛ دانشگاه پیام نور
zohreh.zolfaghar@istta.ir

مقالهٔ پژوهشی

دریافت: ۱۳۹۲/۱۱/۲۳
پذیرش: ۱۳۹۳/۰۹/۲۳

دورهٔ ۳۰ شماره ۳
ص‌ص. ۷۱۹-۷۳۵

چکیده: موضوع ابهام در معانی واژه‌ها و ساختارها و چگونگی برطرف‌کردن آن به‌ویژه هنگام ترجمهٔ ماشینی، ذهن بسیاری از محققان را در این حوزه به خود مشغول داشته و برای آن راهکارهای گوناگونی ارائه شده است. صفت‌ها و نام‌های مبهم با تعدد معانی خود دشواری‌هایی را در ترجمهٔ ماشینی به‌وجود می‌آورند. در برنامه‌های خودکار که ارادهٔ انسانی در انتخاب معادل دخالتی ندارد، این موضوع عیان‌تر است. ماشین نمی‌تواند بر اساس بافت به‌صورت خودکار بهترین معادل را انتخاب کند، حال آنکه، به کمک زبان‌شناسی رایانه‌ای و به‌کار گیری پیکره‌ها این امر ممکن است. در این مقاله بر آنیم تا به ابهام موجود در واژه‌ها بپردازیم. در پژوهش حاضر، به‌منظور اثبات این توانایی از میان دو مقولهٔ اسم و صفت، صفت‌ها را برای بررسی انتخاب کردیم. برای این کار کلیهٔ صفت‌های انگلیسی موجود در یک فرهنگ متوسط (فرهنگ هزاره) را به‌همراه معانی متعدد آنها استخراج کرده و ضبط کردیم، سپس آنها را در یک کشف‌اللغات موازی انگلیسی به فارسی قرار دادیم و جملاتی را که این صفت‌ها در آنها به کار رفته بودند، ضبط کرده و آنها را همراه با بافت و معنا استخراج نمودیم و فهرستی به‌صورت کشف‌اللغات تهیه کردیم. برنامه‌ای برای این کشف‌اللغات نوشته شد به‌گونه‌ای که به‌هنگام ترجمه، از میان معانی موجود بالاترین بسامد معنایی به‌همراه باهم‌آیی و بدون آن به‌عنوان معادل انتخاب شود. معادل‌های انتخاب‌شده را مترجمان انسانی نیز آزمودند و نتایج نشان داد که در بیش از ۵۰ درصد از موارد، معادل‌های انتخاب‌شده از سوی مترجمان با آنچه که برنامهٔ ابهام‌زدایی انتخاب کرده بود، یکسان و یا بسیار نزدیک بودند. نتایج حاصل از این پژوهش در امر ترجمهٔ ماشینی، بازیابی اطلاعات دوزبانه، ایجاد شبکه‌های واژگانی و آموزش زبان فارسی سودمند خواهد بود.

کلیدواژه‌ها: صفت‌های چندمعنایی؛ ابهام‌زدایی واژگانی؛ پیکرهٔ موازی؛ ترجمهٔ ماشینی؛ کشف‌اللغات

پژوهشنامهٔ پردازش و مدیریت اطلاعات
فصلنامه | علمی پژوهشی
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، ISC و Scopus
http://jipm.irandoc.ac.ir
پژوهشگاه علوم و فناوری اطلاعات ایران

۱. مقدمه

موضوع ترجمهٔ ماشینی و دشواری‌های آن در سال‌های اخیر بسیار مورد توجه قرار گرفته و از جنبه‌های گوناگون به آن پرداخته شده است. فلاحتی در بررسی یکی از ماشین‌های ترجمهٔ فارسی به نام «پدیده» به تفصیل به این موضوع پرداخته است. او یکی از دلایل عمدهٔ ناکارآمدبودن این ماشین را وجود چندمعنایی‌ها و حساس‌نبودن ماشین به بافت و مخاطب می‌داند (۱۳۸۳). به اعتقاد فخر احمد و همکارانش ابهام معنایی جدی‌ترین دشواری ماشین ترجمه است. ذهن انسان می‌تواند معادل مناسب را با درک بافت در زبان مقصد انتخاب کند. همچنین، ذهن انسان به‌صورت خودکار گروهی از کلمات، و نه یک واژه، را در نظر می‌گیرد تا به معنا برسد. برای برطرف‌کردن این موضوع در ماشین نیاز به حجم بالایی از داده‌ها به‌عنوان درون‌داد و برون‌داد داریم (Fakhr Ahmad et al. 2011, 456). به لحاظ شناختی، واژه‌های چندمعنایی مقولاتی هستند که مانند شبکه با مفاهیم درونی بسیار مرتبط می‌باشند، حال آنکه، در زبان‌شناسی پیکره‌ای آنچه که اهمیت دارد، آن است که مفاهیم گوناگون واژه‌ها چگونه به‌هم مربوط‌اند و بیشتر روی ویژگی توزیعی مفاهیم واژه تأکید شده است. یکی از حوزه‌های مرکزی در زبان‌شناسی شناختی بررسی واژه‌های چندمعنایی است. نیومن در مباحث معناشناسی پیکره‌ای، پیکره‌ها را منابع طبیعی داده‌های زبان‌شناسی شناختی می‌داند، زیرا پیکره‌ها به تمام منابع اصلی زبان‌شناسی شناختی چون استعاره، چندمعنایی، هم‌معنایی، پیش‌نمونه‌ها و تحلیل ساختاری مرتبط هستند و از این‌رو، بیش از هر منبع اطلاعاتی دیگری کاربرد را منعکس می‌کنند. زبان‌شناسی شناختی به‌ویژه به کشف روابط معنایی بر اساس روابط واژه‌ها بسیار می‌پردازد. او به نقل از راوین و لیکاک[1] (2000) بیان می‌کند که در حوزهٔ چندمعنایی بافت نقش بسیار مهمی را ایفا می‌کند (Newman 2011, 521-30). پستفسکی و بوگرائف یکی از دشوارترین مسائل پردازش زبان را ابهام‌های واژگانی می‌دانند. آنها به نقل از واینرایش[2] (۱۹۶۴) به ابهام تقابلی و ابهام تکمیلی اشاره می‌کنند که ابهام تقابلی همان هم‌نامی و ابهام تکمیلی همان چندمعنایی است که هر دو منجر به ابهام واژگانی می‌شوند. از نظر ایشان چندمعنایی به معنای آن است که یک عنصر واژگانی دارای دو یا چند معنای مرتبط با هم است و

1. Ravin & Leacock
2. Weinreich

هم‌نامی به معنای دو واژهٔ مجزاست که یـا هـم‌نویسـه هسـتند و یـا هـم‌آوا (Pestejovsky & Buguraev 1997, 2). از نظر زبان‌شناسان صورتگرا چندمعنایی در بافت یا در بـه‌کـارگیـری رفع ابهام می‌شود. ایوانز و گرین با اشاره به باور لیکاف بیان می‌کننـد کـه واژهٔ چنـدمعنایی به‌صورت مقوله‌ای بـا مفـاهیم مجـزا ذخیـره مـی‌شـود و یـک پدیـدهٔ ذهنـی اسـت. مفـاهیم به‌صورت رادیال ذخیره می‌شوند، یعنی به‌ترتیب و با داشتن یک پیش‌نمونه. این پیش‌نمونـه همان مفهوم پربسامد است که در بسیاری از مدل‌های ابهام‌زدایی از آن استفاده شـده اسـت و به هنگام صحبت دربارهٔ چندمعنایی نباید نقش بسیار مهم بافت را فراموش کنـیم (Evans & Green 2006, 340-342).

چن و همکارانش نیز به موضوع چندمعنایی‌ها در ترجمه اشاره کرده و بیان کـرده‌انـد که چندمعنایی‌ها چه در زبان مبدأ و چه در زبان مقصد، ترجمهٔ ماشینی و بازیابی اطلاعـات را با دشـواری مواجـه مـی‌کننـد. بـرای مثـال، در ترجمـهٔ پرسـش‌هـا بـه زبـان دیگـر، ابهـام اصلی‌ترین موضوع است. واژه‌ای که در زبان مبدأ وجود دارد، ممکن اسـت بـیش از یـک مفهوم در زبان مقصد داشته باشد. آنها از پُربسامدترین باهم‌آیی‌های واژه‌ها در دو پیکـره و روش آماری اطلاعات متقابل[1] برای حل مسئلهٔ ابهام در ترجمه استفاده کردند (Chen et al. 1999, 215-216).

گلین هنگام اشـاره بـه نقـایص شـبکه‌هـای واژگـانی، بـه در نظـرگـرفتن تمـایز میـان چندمعنایی دستوری یا نقشی از چندمعنایی مفهومی تأکید می‌ورزد. بـدان‌معنـا کـه ممکـن است گسترش معنایی یک واژه، حاصل تعامل آن با معنای مفهومی همراه با آن و یا معنای همراه با یک مقولهٔ دستوری باشد و یا به همین صورت یک نقـش اجتمـاعی معـین، ماننـد تلویح می‌تواند عامل گسترش معنایی باشد (Glynn 2009, 78). فیشـر و همکـارانش چنـد معنایی را به دشوار و آسان تقسیم می‌کنند. دشوار، آنهایی هستند که تشـخیص معـانی‌شـان تنها با نگاه‌کردن به واژه‌های اطراف امکـان‌پـذیر نیسـت و چنـدمعنایی‌هـای آسـان آنهـایی هستند که تشخیص مفهوم آنها به‌آسانی از روی بافت امکان‌پذیر است. از آنجا که ماشـین، چنین بافتی را در دسترس نـدارد، بنـابراین، امکـان ابهـام‌زدایـی وجـود نـدارد (Fišer et al.

---

۱. در نظریهٔ احتمالات و نظریهٔ اطلاعات، اطلاعات متقابـل (mutual information) بـرای دو متغیـر تصـادفی برابر است با مقدار وابستگی متقابل متغیرها. رایج‌ترین واحد اندازه‌گیری اطلاعات متقابل، بیت است.

2012, 30-32). از این‌رو، یکی از موضوعات اصلی در پردازش زبان طبیعی[1] (NLP) حل مسئلهٔ ابهام و یا ابهام‌زدایی از مفهوم واژه[2] (WSD) است. نجومیان دربارهٔ این فرایند بر این باور است که پردازشگر واژگانی انسان، هم از اطلاعات نحوی و هم از اطلاعات واژگانی و نیز نشانه‌های بافتی برای حل مسئلهٔ انواع ابهام‌ها استفاده می‌کند. لازم است که اطلاعات واژگانی را تا حد امکان برای فرایند ابهام‌زدایی فراهم کنیم. به‌دلیل ماهیت زبان‌های طبیعی هر مدل ابهام‌زدایی باید نشانه‌های بافتی را در نظر بگیرد. ابهام ممکن است ناشی از عناصر نحوی، معنایی یا واژگانی باشد. بنابراین، دشواری مدلی که می‌خواهیم ارائه دهیم، طراحی مکانیزمی است که اطلاعات را مانند انسان‌ها فیلتر کند و از بافت استفاده نماید. الگوریتم‌های متعددی بر مبنای روان‌شناسی زبان ارائه شده‌اند تا بتوانند چگونگی به‌کارگیری بافت توسط انسان را نمایش دهند (Nojoumiyan 2011, 90). در پژوهش حاضر که از یک پیکرهٔ دو زبانه (فرهنگ انگلیسی به فارسی متوسط هزاره) استفاده شده، پس از استخراج صفت‌های چندمعنایی، آنها را در کشف‌اللغات دو زبانهٔ انگلیسی به فارسی قرار داده و با توجه به بافت آنها (اسم همراه) ترجمهٔ ارائه‌شده توسط پیکره برای هر یک یادداشت شد. آن‌گاه برنامه‌ای با استفاده از این پایگاه داده‌ای نوشته شد تا بتوان جملاتی را که در آنها صفت قبل از اسم قرار نگرفته و یا با فاصلهٔ مشخصی از آن قرار دارد، ابهام‌زدایی نمود. سپس، برای ارزیابی دقت این برنامه، نتایج توسط مترجمان انسانی نیز ارزیابی گردید. نتایج نشان می‌دهد که مترجمان در بیش از ۵۰ درصد موارد، ترجمهٔ برنامهٔ ابهام‌زدایی را پذیرفتند و خود نیز آن معادل را به‌عنوان گزینهٔ مناسب انتخاب کردند. این برنامه با محدودیت‌ها و دشواری‌هایی روبه‌رو بود که به آنها در بخش ۵ خواهیم پرداخت.

۲. ادبیات تحقیق

از میان الگوریتم‌ها و برنامه‌هایی که در سال‌های اخیر برای ابهام‌زدایی به کمک بافت طراحی شده‌اند، می‌توان به الگوریتم پیش‌بینی کینچ[3] اشاره کرد. در این الگوریتم از

1. natural language processing
2. word sense disambiguation
3. Kintsch

بردار نمایشی واژه‌های تولیدشده توسط پردازش معنایی پسین[1] استفاده شده است تا درک محمول‌ها و حتی استعاره‌ها امکان‌پذیر گردد. پردازش معنایی پسین نظریه و روشی است برای استخراج و نشان‌دادن معنای کاربردی- بافتی واژه‌ها به‌کمک محاسبات آماری به‌کار گرفته‌شده روی پیکره‌ای از متون؛ بدان معنا که مجموعهٔ همه بافت‌هایی که یک واژهٔ فرضی در آن به‌کار می‌رود یا نمی‌رود، مجموعه‌ای از محدودیت‌های متقابل را ایجاد می‌کند که تشابه معانی واژه‌ها با هم و مجموعهٔ آنها را با یکدیگر مشخص می‌کند (Landauer et al. 1998, 259). طبق نظر« گیلرمو و همکارانش» این روش دشواری‌هایی از جمله ارائهٔ معنای غالب، عدم داشتن دقت کافی، و تعریف سطح پایین برای واژه را دارد. اینها از دشواری‌های مدل‌های فضای برداری[2] هستند.آنها از الگوریتم پیش‌بینی برای گرفتن معنای یک واژهٔ چندمعنایی استفاده کردند. نتایج نشان داد که الگوریتم رایانه‌ای انسان- بنیاد می‌تواند مشخصه‌هایی را که تضمین‌کنندهٔ نمایش‌های دقیق‌تری از ساختارها هستند، بنمایاند (Guillermo et al. 2010). اخیراً کاربردهای برخی تکنیک‌های آماری و روش‌های برنامه‌ریزی قدرتمند (مانند برنامه‌ریزی شئ- بنیاد، روش‌های جدید نمونه‌سازی از اشیاء ریاضی مانند ماتریس‌ها) توانسته‌اند توانایی‌ها را در تسخیر پیکره‌های وسیع اطلاعات به‌صورت تقلیدی از نمودهای ذهنی نشان دهند. در مورد تحلیل معنایی پسین هم این موضوع صادق است.یعنی یک جبر خطی و روش پیکره- بنیاد که از سوی برخی محققان به‌عنوان یکی از کارآمدترین ابزار شبیه‌سازی اکتساب زبان در انسان است. «گیلرمو و همکاران» به روش‌های زیر نیز اشاره کرده‌اند:

الف. (Landauer & Dumais 1997): در این روش معنای پسین یک پیکره تحلیل شده و یک ماتریس ابعادی ساخته می‌شود که در آن در هر ردیف یک واژهٔ مجزا وجود دارد و هر ستون یک پاراگراف، جمله یا متن است. پس از برخی محاسبات زبانی روی این ماتریس، ماتریس اصلی کوچک می‌شود. در ماتریس‌های حاصل که فاصلهٔ معنایی نام

1. latent semantic analysis

۲. مدل فضای برداری (Space Vector Model) یک مدل جبری است برای نشان‌دادن اسناد متنی به‌صورت بردارهای تعریف‌کننده، مانند اصطلاحات نمایه‌ای. این مدل در محدودکردن اطلاعات، بازیابی اطلاعات، نمایه‌گذاری و رتبه‌بندی به‌لحاظ مرتبط‌بودن استفاده می‌شود و برای نخستین بار در نظام بازیابی اطلاعات هوشمند استفاده شد.

دارند، یک واژه یا ترکیبی از واژه‌ها با یک بردار نشان داده می‌شوند. برای برقرارکردن ارتباط معنایی میان دو واژه یا متن از کسینوس زاویهٔ میان آنها استفاده می‌شود. کسینوس نزدیک به یک نشان‌دهندهٔ ارتباط معنایی قوی است، در حالی که کسینوس نزدیک به صفر یا منفی، هیچ ارتباط معنایی را نشان نمی‌دهد. از این گذشته، این مدل از طول بردار استفاده می‌کند. اما این مدل و سایر مدل‌ها معانی را بدون بافت نشان می‌دهند. نتایج نشان دادند که این مدل برای اِعمال قاعده‌های هدفمند از برخی فرایندهای مدل شناختی و گرفتن نتیجه معتبر است.

ب. (FiŠer et al. 2012): آنها برای یافتن معادل ترجمه از پیکرهٔ قیاسی[1] استفاده کردند و بر خلاف مدل بالا که یک بردار بافتی[2] مجزا برای تمام صورت‌های ممکن یک سرواژه می‌سازد، ابتدا خود سرواژه را با برچسب‌های مفهومی شخص ثالث[3] ابهام‌زدایی نموده و آن‌گاه برای هر مفهوم متعلق به آن سرواژه یک بردار بافتی ساختند. آنها برای بهبود نتایج، مفاهیمی را که در دو برچسب مفهومی مختلف بودند، با هم ترکیب نمودند. طبق بررسی آنها در اغلب موارد کاندیداهای ترجمهٔ یک واژهٔ چندمعنایی همگی با پُربسامدترین مفهوم واژهٔ اصلی چندمعنایی مرتبط هستند. آنها می‌خواستند که برای سایر مفاهیم واژهٔ چندمعنایی نیز معادلی بیابند. برای این کار شیوه‌های توزیعی برای اکتساب مفهوم واژه که پیش‌تر معرفی شده بودند، به‌کار برده شدند.

---

۱. منظور از پیکرهٔ قیاسی مجموعه‌ای از متون است که به‌صورت مستقل به زبان‌های مورد نظر تنظیم و بر اساس مشابهت محتوایی، حوزه و نقش ارتباطی مرتب شده باشد. در چنین پیکره‌هایی متون از یک نوع هستند و محتوای یکسانی را پوشش می‌دهند. برای مثال، پیکره‌ای از حروف تعریف در فرانسه و انگلیسی.

۲. در زبان‌شناسی رایانه‌ای درک مفهوم واژه (word sense induction) دشواری حل‌نشدهٔ پردازش زبان طبیعی است که به تشخیص خودکار مفاهیم یک واژه مربوط می‌شود. به فرض اینکه برون‌داد یک درک مفهوم واژه مجموعه‌ای از معانی واژهٔ هدف است (فهرست مفاهیم)، این عمل به ابهام‌زدایی از مفهوم واژه مربوط می‌شود که در یک فهرست مفهومی از پیش تعریف‌شده قرار دارد و هدف آن حل ابهام واژه‌ها در بافت است. در این فرضیه واژه‌ها اگر در متون مشابه، بافت مشابه، و یا بافت نحوی مشابه ظاهر شوند، به هم شباهت دارند. وقوع هر یک از واژه‌های هدف در پیکره به‌صورت یک بردار بافتی (context vector) نشان داده می‌شود.

3. third party sense taggers

ج. (Pantel & Lin 2002): آنها در این روش خوشه‌های روی هم‌افتاده یا دارای اشتراک تولید می‌کنند، به‌گونه‌ای که یک واژهٔ چندمعنایی دارای چندین خوشه می‌شود که هر یک از آنها مفاهیم خود را می‌نمایاند. اما در این شیوه پایان‌دادن به ادغام مفاهیم به‌صورت خودکار دشوار است. به همین دلیل، آنها از رویکردی که بر مبنای شبکهٔ واژگانی «پرینستون»[1] بود، به‌عنوان فهرست مفاهیم استفاده کرده و در آن از الگوریتم‌های ابهام‌زدایی برچسب مفهومی عضو سوم استفاده نمودند؛ به‌گونه‌ای که بشود دفعات وقوع یک واژهٔ چندمعنایی را به چند گروه تقسیم کرد و برای هر یک بردار بافتی مستقل ساخت. سپس، مشخصه‌های بردار به زبان مقصد ترجمه می‌شوند و با تمام بردارها در زبان مقصد مقایسه می‌گردند، به‌گونه‌ای که مشابه‌ترین آنها پیدا شود، یعنی نزدیک‌ترین مفهوم را نشان دهد. روش آنها موفق به ابهام‌زدایی شد. هرچند یکی از یافته‌های مهم آزمایشات آنها این بود، که گرچه هنوز ابزار ابهام‌زدایی مفهوم واژه چندان دقیق نیست، اما اگر داده‌های کافی وجود داشته باشد، می‌توان از رویکردهای تصادفی استفاده کرد. علاوه بر این، می‌توان آنها را با هم ترکیب کرد و معانی عجیب را کنار گذاشت و دقت را بالا برد (Guillermo et al. 2010, 1707-1709).

«لاپاتا» صفت‌های چندمعنایی را که معنای‌شان بر اساس اسم‌هایی که تعریف‌کنندهٔ آنها هستند، تفاوت می‌کند (مانند fast)، بررسی می‌کند. او در بررسی خود صفت‌ها را از یک پیکرهٔ بزرگ می‌گیرد و یک مدل احتمال ارائه می‌دهد که می‌تواند یک نوع درجه‌بندی برای تفسیرهای مختلف ارائه کند. او اطلاعات واژگانی را به‌صورت خودکار و با استفاده از مطابقت میان نشانه‌های نحوی، سطحی و معنای واژگانی انجام داد. آن‌گاه نتایج را در برابر قضاوت‌های تجربی انسان مورد ارزشیابی قرار داده و نشان داد که درجه‌بندی معنایی این مدل با شم زبانی انسانی هماهنگ است. معانی که بسیار محتمل هستند، در این مدل از سوی آزمودنی‌ها هم محتمل درجه‌بندی می‌شوند. در نهایت، مقایسه میان مدل او و قضاوت‌های انسانی حاکی از آن است که این مدل در گرفتن تفسیرهای مرتبط، خوب عمل می‌کند. او از مدل نایو[2] هم استفاده کرد و هر دو را با هم

1. Princeton Word Net (PWN)

۲. نایو (naive) روشی رایج برای طبقه‌بندی متون است و دربارهٔ اینکه متون به کدام مقوله (ورزش یا سیاست و جز آن) تعلق دارند، تصمیم‌گیری می‌کند و این کار را بر اساس بسامد واژگان به‌عنوان مشخصه‌هایش انجام می‌دهد.

مقایسه نمود و اختلاف ضریب آنها نشان داد که مدل« لاپاتا »از مدل نایو بهتر عمل می‌کند (Lapata 2003, 1-8).

موسوی میانگاه برای حل مسئلهٔ چندمعنایی در واژه‌های فارسی به کمک آمارهٔ اطلاعات متقابل[1] از اطلاعات باهم‌آیی گرفته‌شده از مجموعه‌ای از واژه‌های زبان مبدأ استفاده می‌کند. برای این کار او از آمارهٔ اطلاعات متقابل میان جفت واژه‌ها استفاده می‌کند تا مناسب‌ترین معادل انگلیسی را برای واژهٔ مورد نظر بیابد. اطلاعات متقابل بر اساس باهم‌آیی واژه تعیین می‌شود و به‌عنوان ابزاری برای اندازه‌گیری ارتباط میان واژه‌ها استفاده می‌گردد. نتیجهٔ این محاسبات نیز در رفع ابهام واژه‌های چندمعنایی فارسی هنگام ترجمه به انگلیسی بسیار کارآمد بود (Mosavi Miangah 2007). نجومیان به مسئلهٔ ابهام و ابهام‌زدایی اشاره می‌کند. به نظر او گذاردن ابهام در سطوح پایین تحلیل، برای مثال در تحلیل‌کننده/ تولیدکننده صرفی[2] در یک نظام پردازش زبان طبیعی مشکل‌ساز است، زیرا این ابهام بعدها به سطوح بالاتر نحوی در فرایند می‌رسد. بنابراین، برای یک سیستم بسیار کارآمدتر خواهدبود اگر ابهام را در سطوح پایین برطرف کند. او مدل تحلیل‌گر-تولیدکننده‌ای را ارائه می‌دهد که می‌تواند برخی از ابهام‌ها، مانند مدخل‌های چندتایی را حل کند. برخی از هم‌نویسه‌ها به سطح پساپردازش[3] خواهند رسید. کشف زودهنگام باهم‌آیی‌ها و یا ترکیب در نمونه‌ساز[4] و یا تحلیل دیرتر توسط تحلیل‌گر به‌صورت خودکار از نمونه رفع ابهام خواهد کرد و باعث خواهد شد تا بار کمتری بر دوش پردازشگر پسین باشد. او با نگاه‌کردن به شواهد روان‌شناختی زبان در ابهام‌زدایی هم‌نویسه‌های هم‌آوا که در واژگان ذهنی صورت می‌گیرند و نقش بسامد معنای غالب و مغلوب واژهٔ مبهم، مدلی را برای ابهام‌زدایی مفهوم هر واژه بر اساس «هر مفهوم برای یک باهم‌آیی» پیشنهاد می‌دهد که از سوی یاروووسکی[5] (۱۹۹۳) ارائه شد. او از مدل برجسته‌سازی، شبیه به چیزی که فراست و بنتین[6] (۱۹۹۲) استفاده کرده بودند، بهره می‌برد. این روش به سایر مواردی که واژهٔ مبهم با بسامدهای مختلف اتفاق می‌افتد، ارجحیت دارد، زیرا می‌تواند سطوح

1 Mutual Information Statistics
2. morphological analyzer/ generator
3. post- processing
4. tokenizer
5. Yarovsky, D, 1993
6. Frost, R. & Sh. Bentin

فعال‌سازی معنای پربسامدتر را از کم‌بسامدترین آنها دریافت کند. فهرست بسامدها از یک پیکرهٔ بزرگ گرفته می‌شود و به‌سادگی به واژگان افزوده می‌گردد. او آزمایش‌های متعددی را برای مشخص‌کردن چگونگی تأثیر برجسته‌سازی و رابطهٔ آن با بسامد انجام داد. نتایج حاکی از تأثیر بسامد و میانجی‌گری بخش آوایی قبل از رسیدن به معنا بود. طبق نظر «سیمپسوم و برگیس»[1] هنگامی که یک واژه اتفاق می‌افتد، ابتدا همهٔ معانی‌اش بازیابی می‌شوند. سپس، از بافت برای انتخاب معنای مناسب استفاده می‌گردد. در حقیقت، بافت، تنها ابزار تشخیص معنای واژهٔ چندمعنایی است. زبان‌شناسی پیکره‌ای محققان را قادر ساخته است تا واژه‌ها و باهم‌آیی‌ها را در مطالعات کمی و کیفی ابهام‌زدایی مفهوم واژه بررسی کنند (Nojoumiyan 2011, 107-121). فخر احمد و همکارانش با اشاره به دو روش نمونه- بنیاد و آماری برای رفع ابهام بیان می‌کنند که در روش نمونه- بنیاد مجموعهٔ بزرگی از نمونه‌ها ذخیره می‌شوند. اما در روش آماری رفع ابهام به کمک احتمالات آماری که پارامترهای آن از تحلیل یک پیکرهٔ دو زبانه حاصل می‌شوند، صورت می‌گیرد. این کار بر اساس بسامد واحدهای زبانی گوناگون مانند واژه‌ها، تکواژها، حروف و جز آن صورت می‌گیرد. در این روش نیازی به پیکرهٔ بزرگ نیست و به‌جای واژهٔ مبهم سایر واژه‌های مبهم در بافت ابهام‌زدایی می‌شوند. کاربردی‌بودن و دقت برنامهٔ آنها در مقایسه با مدل‌های دیگر خوب بود و مزیت برتر آن عدم نیاز به پیکرهٔ بزرگ است (Fakhrahmad et al. 2011, 457-8).

از این‌رو، می‌توان گفت که پیکره‌های زبانی قابلیت ایجاد پایگاهی از داده‌ها را برای مدل‌های گوناگون ابهام‌زدایی می‌توانند فراهم آورند. بنابراین، امکان تهیهٔ یک برنامهٔ ابهام‌زدایی به کمک پیکره‌ای بزرگ می‌تواند امکان‌پذیر باشد. پیکره می‌تواند آن چیزی را که ماشینِ ترجمه فاقد آن است، یعنی بافت را، در اختیار ماشین قرار دهد و به کمک برنامه‌ای که به‌کمک احتمال وقوع طراحی شده باشد، به ترجمه‌ای طبیعی‌تر و نزدیک‌تر به ترجمهٔ انسانی دست یابد. ماشین‌ها در پردازش‌های خود با مشکل نبودِ بافت مواجه هستند که این دشواری جهت ابهام‌زدایی باید برطرف شود. بنابراین، در صورتی که بتوان پیکرهٔ

1. Simpson and Burgess

کاملی برای کشف‌اللغات (کانکوردنس)[1] تهیه کرد، می‌توان به معادل‌های دقیق‌تری دست یافت.

۳. روش تحقیق

برای انجام پژوهش حاضر که در نوع خود پژوهشی تجربی به‌شمار می‌رود و به‌منظور نشان‌دادن تأثیر بافت، لازم بود ابتدا پیکره‌ای از صفت‌های چندمعنایی در زبان انگلیسی و نیز ترجمهٔ آنها به فارسی ارائه شود. برای این منظور، فرهنگ انگلیسی به فارسی هزاره انتخاب شد که فرهنگی متوسط به شمار می‌رود و تعداد مدخل‌های اصلی آن بیش از ۵۰۰۰۰۰ مدخل است که بالغ بر ۳۰۰۰۰۰۰ معادل فارسی برای آنها ارائه شده است. از میان این مدخل‌ها صفت‌های چندمعنایی که تعداد آنها بیش از ۵۰۰ مورد بود، استخراج شدند. در مرحلهٔ بعد، برای ارائهٔ بافت مناسب برای این مدخل‌ها لازم بود آنها را در پایگاه داده‌ها در یک کشف‌اللغات موازی قرار دهیم و باهم‌آیی‌ها و معانی آنها را برای واردکردن در پایگاه ضبط کنیم. از میان ۵۰۰ صفت مبهمی که از فرهنگ هزاره استخراج شد، تنها برای ۲۹۸ صفت در کشف‌اللغات مورد نظر (موسوی میانگاه ۲۰۰۹) که دارای ۸۰۰۰۰ مدخل بود، جمله یافت شد. صفت‌های موجود به‌همراه باهم‌آیی و معنای ارائه‌شده در پیکره ضبط شدند و به‌صورت یک کشف‌اللغات موازی در اختیار برنامه‌نویس قرار گرفتند. جدول شماره ۱ بخشی از کشف‌اللغات تهیه شدهٔ مورد نظر را نشان می‌دهد.

۱. کشف‌اللغات یا کانکوردنس فهرستی از واژه‌هاست (کلمات کلیدی) که از بخشی از یک زبان معتبر (یک پیکره) گرفته شده است، در مرکز صفحه نمایش داده می‌شود و واژه‌ها را با بخش‌هایی از بافتی که در آن به‌کار رفته‌اند، نشان می‌دهد. کانکوردنس‌ها امروزه برای تهیهٔ فهرست واژگان و فرهنگ‌ها بسیار استفاده می‌شوند.

جدول ۱. بخشی از کشف‌اللغات مورد نظر

Collocation					
id	**Adjective Collocation**		**Meaning**	**Translation**	**Frequency**
5	acute	acute urgency	فوری	اهمیت فوری	1
6	acute	acute eye	کم‌سو	چشم کم‌سو	1
7	acute	acute suffering	بسیار	رنج بسیار	1
8	affected	affected by something	تحت تأثیر چیزی قرارگرفتن	تحت تأثیر	4
9	affected	affected workers	مبتلا	کارگران مبتلا	2
10	affected	affected areas	صدمه‌خورده	مناطق زلزله‌زده	5

برنامهٔ مورد نظر باید می‌توانست این صفت‌ها را در صورت وقوع با باهم‌آیی موجود در پایگاه ترجمه کند و در صورتی که واژه را در بافت جدیدی ببیند، آن را بر اساس بسامد معناهای ارائه‌شده، ترجمه نماید. یعنی باید قادر به ارائهٔ بالاترین بسامد معنایی برای صفت مورد نظر باشد. پس از تکمیل برنامه، لازم بود میزان دقت نیز مورد بررسی قرار گیرد.

۳-۱. برنامهٔ طراحی‌شده

این برنامه با استفاده از زبان برنامه‌نویسی C#.net و بانک اطلاعاتی اکسس طراحی و برنامه‌نویسی شده است. روش کار الگوریتم بدین‌صورت است که ابتدا با مبنا قراردادن کلمهٔ اول، سعی در یافتن رکورد متناسب در پایگاه داده بر اساس کلمهٔ دوم می‌کند و در صورتی که جستجوی الگوریتم نتیجه‌ای در بر نداشته باشد، از بین زیرمجموعهٔ استخراج‌شده به سراغ سایر معادل‌ها می‌رود. در صورتی که تعداد نتایج یافت‌شده بیش از یک معادل باشد، بر اساس بیشترین تعداد تکرار فیلتر می‌شود. نکتهٔ مهم در این الگوریتم رعایت تعداد فاصله بین کلمات می‌باشد که بر اساس نظر محققان تا ۴ فاصله در نظر گرفته شده است؛ بدین مفهوم که در صورت وجود فاصله تا ۴ کلمه، بین کلمهٔ مبنا و کلمهٔ هم‌آیند خودش، الگوریتم کماکان به جستجو ادامه می‌دهد و تمام حالات ممکن را استخراج کرده و در پیکرهٔ موجود جستجو می‌نماید. در ادامه، بخشی از الگوریتم نشان داده شده است.

```
////////////////////////////////
/*public override list<result> retrunsearch(string search)
{
   string[] search = pQuery.Split(',');
   List<result> myresult = new List<result>();

   // Build WHERE
   for (int i = 1; i < search.Length; i++)
      where += " And '%" + search[i] + "%'";

   // Now search
   OleDbCommand sqlcmdCommand0 = new OleDbCommand("select
Distinct name from table1 where     search like '%" + search[0] + "%' " + where + "
order by name", sqlcon);
   sqlcmdCommand0.CommandType = CommandType.Text;
   OleDbDataReader sdaResult0 = sqlcmdCommand0.ExecuteReader();
   while (sdaResult0.Read())
   {
      result restult1 = new result();
      result1.name = sdaResult0.String(0);
      result.add(result1);
   }
   sdaResult0.Close();

   return result;
}*/
////////////////////////////////
```

با اجرایی‌شدن این برنامه تعداد ۸۶ صفت از همین کشف‌اللغات انتخاب شدند. برای هر یک از این صفت‌ها جمله‌ای از پیکرهٔ ملی بریتانیا (BNC[1]) به‌صورت تصادفی انتخاب شد. این انتخاب بر اساس داده‌های موجود در پایگاه صورت گرفت. آن‌گاه، این جمله‌ها (تنها ترکیب صفت و اسم همراه آن) در برنامهٔ ابهام‌زدایی قرار گرفتند و معادل‌های ارائه‌شده برای آنها ضبط شد. سپس، همان جملات به‌همراه دو گزینه (یکی گزینهٔ انتخاب‌شده از سوی برنامه و دیگری گزینه‌ای که معنای دوم یا سوم واژه به‌شمار می‌رفته است) به ۵ مترجم داده شدند. مترجمان، گزینه‌ای را انتخاب کردند و انتخاب‌های مترجمان با انتخاب‌های برنامه مقایسه شد. برای مثال، واژهٔ affected در پیکرهٔ ملی بریتانیا جستجو شد و جملاتی برای آن مشاهده گردید و از میان آنها دو جمله به‌صورت تصادفی انتخاب شدند و قسمتی از جمله که صفت مورد نظر به‌همراه اسم بود، به برنامه داده شد و

1. http://corpus.byu.edu/bnc/

معادل آن ضبط گردید. آن‌گاه، همان جمله همراه با دو پاسخ در اختیار مترجمان انسانی قرار گرفتند. این دو پاسخ، یکی پاسخ ماشین و و مورد دیگر یکی از معادل‌های موجود در پیکره بودند.

1. Any in case of bleeding tries to raise the *affected* part of the body.

الف: صدمه‌خورده ب: تحت تأثیر

2. He was sentimentally *affected* by the ideals of revolution.

الف: صدمه‌خورده ب: تحت تأثیر

این کار برای ۸۶ جمله در پیکرهٔ ملی بریتانیا انجام و جملات گردآوری شد و در نهایت، پاسخ مترجمان با پاسخ‌های ماشین مقایسه و ضریب همبستگی آنها تعیین گردید.

٤. تحلیل داده‌ها و بیان یافته‌ها

طبق آنچه که در روش پژوهش توضیح دادیم، با مقایسهٔ معادل‌های انتخاب‌شده از سوی مترجمان و معادل‌های داده‌شده از سوی برنامهٔ ابهام‌زدایی معلوم شد که این برنامه در صورت رفع دشواری‌ها می‌تواند برنامه‌ای کارآمد برای ابهام‌زدایی در ترجمه‌های ماشینی باشد. برای تعیین دقت برنامه پایایی آزمون انجام‌شده برای مترجمان و ماشین را به کمک روش آلفای کرونباخ و تعیین ضریب توافق میان آنها انجام دادیم. پایائی دلالت بر آن دارد که ابزار اندازه گیری در شرایط یکسان تا چه اندازه نتایج یکسانی به‌دست می دهد. دامنهٔ ضریب اعتبار از صفر تا یک است. برای تعیین پایائی ابزار اندازه گیری شیوه‌های مختلفی وجود دارد؛ از جمله روش اجرای دوباره (بازآزمائی)، روش موازی (همتا)، روش تنصیف (دونیمه‌کردن)، روش کودر، ریچاردسون و روش آلفای کرونباخ. مشهورترین ضریب اعتبار از طریق یک بار اجرای آزمون توسط کرونباخ ارائه شده است که به ضریب آلفای کرونباخ معروف است. این روش برای محاسبهٔ هماهنگی درونی ابزار اندازه گیری از جمله پرسشنامه به‌کار می‌رود. در این ابزار، پاسخ هر سئوال می تواند مقادیر عددی مختلف را اختیار کند (بازرگان ۱۳۸۳؛ ۲۰۸). نرم‌افزار spss یکی از نرم‌افزارهای متداول برای تعیین پایایی با یکی از روش‌های فوق (و معمولاً روش آلفای کرونباخ) می‌باشد.

در این مطالعه به‌منظور آزمون درجهٔ صحت نرم‌افزار ساخته‌شده، به طراحی پرسشنامه‌ای پرداخته‌ایم تا از طریق جواب‌دهی انسانی به بررسی قابلیت اعتماد پرسشنامه و

نهایتاً مناسب‌بودن نرم‌افزار دست یابیم. در این راستا، پرسشنامهٔ طراحی‌شده با ۸۶ سؤال در اختیار ۵ مترجم انسانی قرار داده شد و داده‌های حاصل از پرسشنامه، توسط نرم‌افزار آماری spss 15 و از طریق تهیهٔ جداول توافقی آماره‌های توصیفی، بررسی وابستگی متغیرها و آزمون روایی آلفای کرونباخ برای ۶ گروه مورد نظر (نرم‌افزار و ۵ مترجم)، مورد تحلیل و بررسی قرار گرفته‌اند.

نتایج حاکی از آن است که در ۵۴ درصد از سؤالات، واریانس بین گروه‌ها صفر می‌باشد (۲۷ مورد مطابقت کامل، ۷ مورد عدم مطابقت کامل، ۱۲ مورد تساوی تعداد پاسخ صحیح و غلط موجود بین پاسخ نرم‌افزار و پاسخ‌های ۵ مترجم). بنابراین، به‌صورت معمول از بررسی انجام‌شده در آزمون کرونباخ حذف شده‌اند (شمارهٔ سؤالات مندرج در جدول ۲).

جدول ۲. سؤالات حذف‌شده

q3, q5, q7, q11, q12, q19, q20, q21, q22, q23, q24, q27, q28, q29, q30, q31, q32, q33, q38, q39, q42, q46, q48, q49, q50, q53, q54, q56, q57, q58, q59, q61, q63, q66, q68, q69, q72, q73, q75, q77, q78, q79, q80, q82, q83, q85

برای سایر سؤالات پرسشنامه، آزمون آلفای کرونباخ انجام شده و نتایج به شرح جدول ۳ می‌باشد:

جدول ۳. آمارهای روایی

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
0.953	0.969	40

در جدول آزمون آماره‌های روایی[1] برای ۴۰ سؤال باقی‌مانده، مقدار آلفای کرونباخ و آلفای کرونباخ بر پایهٔ مقادیر استانداردشده، به‌ترتیب برابر ۰/۹۵۳ و ۰/۹۶۹ می‌باشد (بالاتر از ۰/۷) که بیانگر همبستگی بالای متغیرها در شش گروه آزمون‌شده می‌باشد.

1. reliability statistics

۵. محدودیت‌های پژوهش

طبق تحلیل فوق، این برنامه در ۵۴ درصد موارد، موفق به ارائهٔ ترجمهٔ قابل قبول شد. عدم دقت بالاتر این برنامه ناشی از چند موضوع است. نخست آنکه کشف‌اللغاتی که برای یافتن بافت صفت‌ها استفاده شد، یک پیکرهٔ ۸۰۰۰۰ جمله‌ای بود (تنها کشف‌اللغات موازی در دسترس نگارنده) که تنها برای ۲۹۸ مورد از صفت‌های استخراج‌شده، جمله ارائه می‌کرد. این امر خود باعث پایین‌بودن دامنهٔ صفت‌ها در کشف‌اللغات تهیه‌شده می‌باشد. در صورتی که نگارنده می‌توانست پیکرهٔ کامل‌تری در اختیار داشته و برای همهٔ معانی موجود برای هر یک از صفت‌ها باهم‌آیی‌های کافی داشته باشد، به یقین به معادل‌های بسیار دقیق‌تری دست می‌یافت. نکته دیگر آنکه، اگر بخواهیم این برنامه با دقت بالاتری ترجمه کند، باید تا آنجا که امکان دارد باهم‌آیی‌های صفت‌های مبهم را از پیکره‌های گوناگون استخراج کنیم؛ به‌گونه‌ای که حتی کم‌بسامدترین‌ها نیز در کشف‌اللغات وجود داشته باشند. در این صورت، برنامه می‌تواند معنای دقیق‌تری ارائه دهد و در صورتی که باهم‌آیی مورد نظر را نداشته باشد، آن‌گاه بر اساس بسامد وقوع معنا، بالاترین بسامد معنایی برای صفت مبهم انتخاب می‌شود. بنابراین، داشتن پیکره‌ای بزرگ برای تهیهٔ کشف‌اللغات و نیز باهم‌آیی‌های هرچه بیشتر در آن می‌تواند برنامهٔ موفق‌تری ارائه دهد.

٦. نتیجه‌گیری

در این پژوهش بر آن بودیم که هنگام ترجمهٔ ماشینی برای ابهام‌زدایی صفت‌های مبهم، یعنی صفت‌های با بیش از یک معنا، نرم‌افزاری کارآمد ارائه دهیم. در انسان‌ها، معنای این صفت‌ها از طریق بافت زبانی در زمان کاربرد رفع ابهام می‌شود، اما این شرایط در ترجمهٔ ماشینی وجود ندارد و در این حالت ماشین از آن‌رو که قادر به تشخیص و درک بافت نیست، نمی‌تواند رفع ابهام کند. برای حل این موضوع، نرم‌افزار نیز باید دارای بافت باشد. ولی، این موضوع، چون زایایی و خلاقیت و عدم ثبات در تولید و به‌کارگیری در ماشین وجود ندارد، امکان‌پذیر نیست. بنابراین، سعی شد تا حد امکان بافت فراهم شود، اما برنامه نیز به‌گونه‌ای طراحی شد که بتواند بر اساس بسامد بالا به معنای مورد نظر و یا نزدیک‌ترین معنا دست یابد. برای این منظور، از یک پیکرهٔ انگلیسی به فارسی (فرهنگ هزاره) همهٔ صفت‌های چندمعنایی را استخراج کردیم. آن‌گاه، این صفت‌ها را در یک

کشف‌اللغات موازی قرار دادیم و هر یک را با بافت موجود ضبط کردیم. برنامه به‌گونه‌ای تهیه شد که این صفت‌ها را اگر با بافت همراهش می‌دید، همان معنی را ارائه می‌کرد و در غیر این صورت بالاترین بسامد معنایی را به‌عنوان معادل انتخاب می‌کرد. پس از طراحی برنامه، نتایج انتخاب‌های برنامه با نتایج انتخاب‌های مترجمان انسانی مقایسه شد و با استفاده از بررسی پایایی این مقایسه مشخص شد که برنامهٔ مورد نظر از دقت کافی برخوردار است و دارای روایی ۰/۷ می‌باشد؛ یعنی با خطای ناچیز می‌تواند معادل مناسب برای صفت‌های چندمعنایی را انتخاب کند.

این نرم‌افزار در صورتی که از پیکره‌ای بزرگتر بهره ببرد و کشف‌اللغات بزرگتری برای به‌دست آوردن بافت‌ها وجود داشته باشد، نرم‌افزاری مناسب جهت دستیابی به معادل مناسب در هنگام ترجمه خواهد بود. همچنین، این نرم‌افزار می‌تواند در امر آموزش برای زبان‌آموزان و مترجمان و حتی افرادی که به یادگیری زبان فارسی به‌عنوان زبان دوم مشغول هستند، بسیار کارآمد باشد.

۷. فهرست منابع

بازرگان، عباس. ۱۳۸۳. ارزشیابی آموزشی، مفاهیم، الگوها و فرایندهای عملیاتی، تهران: سمت

حق‌شناس، علی‌محمد، حسین سامعی، و نرگس انتخابی. ۱۳۸۱. *فرهنگ معاصر هزاره*. تهران: فرهنگ معاصر.

فلاحتی، محمدرضا. ۱۳۸۳. انسان مترجم و ترجمهٔ ماشینی: بررسی موردی مشکلات ماشین ترجمهٔ انگلیسی به فارسی «پدیده». *فصلنامهٔ کتابداری و اطلاع‌رسانی* (۷) ۲.

Fakhrahmad, S. M., A. R. Rezapour, M. Zolghadri Jahromi, and M. H. Sadreddini. 2011. *A New Word Sense Disambiguation System Based on Deduction*. In proceedings of the World Congress on Engineering, July 6 - 8, 2011, London, U.K.

Fišer, D., N. Ljubešić, and O. Kubelka. 2012. *Addressing polysemy in bilingual lexicon extraction from comparable corpora*. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, pp: 3031-3035.

Frost, R. and Sh. Bentin. 1992. Processing Phonological and Semantic Ambiguity: Evidence.From Semantic Priming at Different. *SOAS, Journal of Experimental Psychology: Learning,Memory, and Cognition* 18 (1): 58-68.

Glynn, Dylan. 2009. Polysemy, syntax and variation, a usage-based method for cognitive semantics. In *New Directions in Cognitive Linguistics*, edited by Vuvvyan Evans, Stephnie Pourcel. John Benjamins Publishing Company, University of Bangor, UK, 77-104.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of to run. In *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, edited

by Stefan Th. Gries & Anatol Stefanowitsch, 57-99. Berlin & New York: Mouton de Gruyter.

Guillermo, Jorge-Botana, José A. León, Ricardo Olmos, and Yusef Hassan-Montero. 2010. Visualizing polysemy using LSA and the predication algorithm. *Journal of American Society for Information Science and Technology* 61 (8): 1706–1724.

Hsin-Hsi Chen, Guo-Wei Bian, and Wen-Cheng Lin. 1999. Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. *International Journal of Computational Linguistics and Chinese Language Processing* 4 (2): 21-38.

Landauer, T. K., P. W. Foltz, and D. Laham.1998. Introduction to latent semantic analysis. *Discourse Processes* 25: 259-284.

Lapata, M. 2001. *A corpus-based account of regular polysemy: The case of context-sensitive adjectives*. In proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Carnegie Mellon University Pittsburgh, PA, USA, 2-7 June 2001, pp: 1-8.

Mosavi Miangah, T. 2007. *Solving polysemy problem of Persian words using mutual information statistics*. In: Corpus Linguistics Conference 2007, July 27-30 2007, University of Birmingham, UK.

Mosavi Miyangah, T. 2009. Constructing a Large Scale English-Persian Parallel Corpus. *META* 54 (1): 181-188.

Newman, J. 2011. Corpora and cognitive linguistics. *Revista Brasileira de Linguística Aplicada* 11 (2): 521-55.

Nojoumian, Payman. 2011. Towards the Development of an Automatic Diacritizer for the Persian Orthography based on the Xerox Finite State Transducer, PhD Thesis, University of Ottawa.

Pantel, Patrick and Dekang Lin. 2002. *Discovering word senses from Text.* In proceedings of ACM S16KDD conference on knowledge, discourse and data mining. Edmonton, AB, Canada — July 23 - 25, 200.

Pestejovsky, James and Branimir Boquraey. 1997. *Lexical Semantics: the problem of polysemy.* New York: Oxford University Press. pp: 1-15.

Simpson, G. B., and C. Burgess.1985. Activation and Selection Process in the Recognition of Ambiguous Words. *Journal of Experimental Psychology: Human Perception and Performance* 11 (1): 28-39.

Yarowsky, D. 1993. *One Sense per Collocation.* In proceedings of ARPA Human LanguageTechnology Workshop, Princeton, NJ, pp.: 266-271.