

Lexical Disambiguation of Polysemeous Adjectives in Machine Translation: A Corpus Based Study

Tayebeh Mosavi Miyangah¹ | Zohreh Zolfaghar
Kondori²

1. Associate Prof. of Applied and Mathematical Linguistics;
Payame Noor University; I. R. of Iran mosavit@pnu.ac.ir
2. [Corresponding Author] Ph.D Candidate of General
Linguistics Payame Noor University; I. R. Iran
zohreh.zolfaghar@istta.ir

Iranian Journal of
**Information
Processing &
Management**

Abstract: Ambiguous nouns and adjectives are words with more than one meaning. Having several meanings creates problems in MT particularly in automatic programs where human plays no role in choosing the equivalent target word and there is no context available to the machine. One of the ways to solve the problem is to provide the machine with proper and enough contexts; here by context we mean the collocations of a word. In order to show this we chose English adjectives for our study. We extracted polysemeous adjectives from a bilingual English-Persian Medium Dictionary (Hezareh) and checked them all in a bilingual parallel English-Persian Corpus to get the context and collocations for each adjective. Then we recorded all the adjectives with their associate words, context and the meaning provided by the corpus. Then a concordance was created out of these adjectives and their equivalents along with their collocations and a disambiguating program was written for it, so that it could choose the best equivalent for the target adjective according to the highest meaning frequency of the given adjective or its collocation. This program's reliability was tested by using five translators. Their choices were compared through a statistical method (Cronbach alpha) and their covariance was calculated by SPSS software. The results show that in more than %50 of the cases translators' choices were the same as the program's choices. The results of this study could be helpful in MT, bilingual information retrieval, word nets and even in teaching Persian to foreigners.

Keywords: Polysemeous Adjectives; Lexical Disambiguating; Parallel Corpus; Machine Translation; Concordance

Iranian Research Institute
for Information Science and Technology
(IranDoc)

ISSN 2251-8223

eISSN 2251-8231

Indexed by SCOPUS, ISC, & LISTA

Vol. 30 | No. 3 | pp. 719-735

Spring 2015

<https://doi.org/10.35050/JIPM010.2015.030>



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

طیبه موسوی میانگاه^۱ | زهره ذوالفقار کندری^۲

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

مقاله پژوهشی

دریافت: ۱۳۹۲/۱۱/۲۳
پذیرش: ۱۳۹۳/۰۹/۲۳

دانشگاه
مدرسه مطالعات

فصلنامه علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
(پرانداکت)

شاپا (چاپی) ۸۲۳-۲۴۰۱

شاپا (الکترونیکی) ۸۳۱-۲۴۰۱

نمایه در SCOPUS، JISC، LISTA

jipm.irandoc.ac.ir

دوره ۳۰ | شماره ۳ | صص ۷۱۹-۳۵

بهار ۱۳۹۴

<https://doi.org/10.35050/JIPM010.2015.030>

چکیده: موضوع ابهام در معانی واژه‌ها و ساختارها و چگونگی برطرف کردن آن به‌ویژه هنگام ترجمه ماشینی، ذهن بسیاری از محققان را در این حوزه به خود مشغول داشته و برای آن راهکارهای گوناگونی ارائه شده است. صفات و نام‌های مبهم با تعدد معانی خود دشواری‌هایی را در ترجمه ماشینی به وجود می‌آورند. در برنامه‌های خودکار که اراده 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 به‌کارگیری 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 مقوله اسم و صفت، صفات را برای بررسی انتخاب کردیم. برای این کار کلیه صفات 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 صفات 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 برنامه ابهام‌زدایی انتخاب کرده بود، یکسان یا بسیار نزدیک بودند. نتایج حاصل از این پژوهش در امر ترجمه ماشینی، بازیابی اطلاعات دوزبانه، ایجاد شبکه‌های واژگانی و آموزش زبان فارسی سودمند خواهد بود.

کلیدواژه‌ها: صفات چندمعنایی؛ ابهام‌زدایی واژگانی؛ پیکره موازی؛ ترجمه ماشینی؛ کشف‌اللغات



۱. مقدمه

موضوع ترجمه 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 ترجمه فارسی به نام «پدیده» به تفصیل به این موضوع پرداخته است. او یکی از دلایل عمده ناکارآمد بودن این ماشین را وجود چندمعنایی‌ها و حساس نبودن 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 به صورت خودکار گروهی از کلمات، و نه یک واژه، را در نظر می‌گیرد تا به معنا برسد. برای برطرف کردن این موضوع در ماشین نیاز به حجم بالایی از داده‌ها به عنوان 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 به هم 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 این رو، 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^۱ (2000) بیان می‌کند که در حوزه چندمعنایی بافت نقش بسیار مهمی را ایفا می‌کند (Newman 2011, 521-30). پستفسکی و بوگراف یکی از دشوارترین مسائل پردازش زبان را ابهام‌های واژگانی می‌دانند. آنها به نقل از واینرایش^۲ (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

2012, 30-32). از این رو، یکی از موضوعات اصلی در پردازش زبان طبیعی^۱ (NLP) حل مسئله ابهام و یا ابهام‌زدایی از مفهوم واژه (WSD)^۲ است. نجومیان درباره 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 مسئله انواع ابهام‌ها استفاده می‌کند. لازم است که اطلاعات واژگانی را تا حد امکان برای فرایند ابهام‌زدایی فراهم کنیم. به دلیل 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 (Nojourniyan 2011, 90). در پژوهش حاضر که از یک پیکره دو زبانه (فرهنگ انگلیسی به فارسی متوسط هزاره) استفاده شده، پس از استخراج صفات چندمعنایی، آنها را در کشف اللغات دو زبانه انگلیسی به فارسی قرار داده و با توجه به بافت آنها (اسم همراه) ترجمه 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 فاصله 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، ترجمه برنامه ابهام‌زدایی را پذیرفتند و خود نیز آن معادل را به‌عنوان گزینه 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^۳ اشاره کرد. در این الگوریتم از بردار نمایشی واژه‌های تولیدشده توسط پردازش معنایی پسین^۴ استفاده شده است تا درک

1. natural language processing

2. word sense disambiguation

3. Kintsch

4. latent semantic analysis

محمول‌ها و حتی استعاره‌ها امکان‌پذیر گردد. پردازش معنایی پسین نظریه و روشی است برای استخراج و نشان‌دادن معنای کاربردی-بافتی واژه‌ها به کمک محاسبات آماری به کار گرفته‌شده روی پیکره‌ای از متون؛ بدان معنا که مجموعه همه بافت‌هایی که یک واژه فرضی در آن به کار می‌رود یا نمی‌رود، مجموعه‌ای از محدودیت‌های متقابل را ایجاد می‌کند که تشابه معنایی واژه‌ها با هم و مجموعه آنها را با یکدیگر مشخص می‌کند (Landauer et al. 1998, 259). طبق نظر «گیلمو و همکاران» این روش دشواری‌هایی از جمله ارائه معنای غالب، عدم داشتن دقت کافی، و تعریف سطح پایین برای واژه را دارد. اینها از دشواری‌های مدل‌های فضای برداری^۱ هستند. آنها از الگوریتم پیش‌بینی برای گرفتن معنای یک واژه چندمعنایی استفاده کردند. نتایج نشان داد که الگوریتم رایانه‌ای انسان-بنیاد می‌تواند مشخصه‌هایی را که تضمین‌کننده نمایش‌های دقیق‌تری از ساختارها هستند، بنمایاند (Guillermo et al. 2010). اخیراً کاربردهای برخی تکنیک‌های آماری و روش‌های برنامه‌ریزی قدرتمند (مانند برنامه‌ریزی شی-بنیاد، روش‌های جدید نمونه‌سازی از اشیاء ریاضی مانند ماتریس‌ها) توانسته‌اند توانایی‌ها را در تسخیر پیکره‌های وسیع اطلاعات به صورت 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 به عنوان یکی از کارآمدترین ابزار شبیه‌سازی اکتساب زبان در انسان است. «گیلمو 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 واژه 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 فاصله معنایی نام دارند، یک واژه یا ترکیبی از واژه‌ها با یک بردار نشان داده می‌شوند. برای برقرار کردن ارتباط معنایی میان دو واژه یا متن از کسینوس زاویه میان آنها استفاده می‌شود. کسینوس

۱. مدل فضای برداری (Space Vector Model) یک مدل جبری است برای نشان‌دادن اسناد متنی به صورت بردارهای تعریف‌کننده، مانند اصطلاحات نمایه‌ای. این مدل در محدود کردن اطلاعات، بازیابی اطلاعات، نمایه‌گذاری و رتبه‌بندی به‌لحاظ مرتبط بودن استفاده می‌شود و برای نخستین بار در نظام بازیابی اطلاعات هوشمند استفاده شد.

نزدیک به یک نشان‌دهنده 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 اعمال قاعده‌های هدفمند از برخی فرایندهای مدل شناختی و گرفتن نتیجه معتبر است.

ب. (Fišer et al. 2012): آنها برای یافتن معادل ترجمه از پیکره قیاسی^۱ استفاده کردند و بر خلاف مدل بالا که یک بردار بافتی^۲ مجزا برای تمام صورت‌های ممکن یک سرواژه می‌سازد، ابتدا خود سرواژه را با برجسب‌های مفهومی شخص ثالث^۳ ابهام‌زدایی نموده و آن‌گاه برای هر مفهوم متعلق به آن سرواژه یک بردار بافتی ساختند. آنها برای بهبود نتایج، مفاهیمی را که در دو برجسب مفهومی مختلف بودند، با هم ترکیب نمودند. طبق بررسی آنها در اغلب موارد کاندیداهای ترجمه یک واژه چندمعنایی همگی با پُربسامدترین مفهوم واژه اصلی چندمعنایی مرتبط هستند. آنها می‌خواستند که برای سایر مفاهیم واژه چندمعنایی نیز معادلی بیابند. برای این کار شیوه‌های توزیعی برای اکتساب مفهوم واژه که پیش‌تر معرفی شده بودند، به کار برده شدند.

ج. (Pantel & Lin 2002): آنها در این روش خوشه‌های روی هم‌افتاده یا دارای اشتراک تولید می‌کنند، به گونه‌ای که یک واژه چندمعنایی دارای چندین خوشه می‌شود که هر یک از آنها مفاهیم خود را می‌نمایاند. اما در این شیوه پایان‌دادن به ادغام مفاهیم به صورت

۱. منظور از پیکره قیاسی مجموعه‌ای از متون است که به صورت 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 حل‌نشده پردازش زبان طبیعی است که به تشخیص خودکار مفاهیم یک واژه مربوط می‌شود. به فرض اینکه برون‌داد یک درک مفهوم واژه مجموعه‌ای از معانی واژه 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 به صورت یک بردار بافتی (context vector) نشان داده می‌شود.

3. third party sense taggers

خودکار دشوار است. به‌همین دلیل، آنها از رویکردی که بر مبنای شبکه‌واژگانی «پرینستون»^۱ بود، به‌عنوان فهرست مفاهیم استفاده کرده و در آن از الگوریتم‌های ابهام‌زدایی برچسب مفهومی عضو سوم استفاده نمودند؛ به‌گونه‌ای که بشود دفعات وقوع یک واژه 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. هر چند 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» صفات چندمعنایی را که معنای‌شان بر اساس اسم‌هایی که تعریف‌کننده آنها هستند، تفاوت می‌کند (مانند fast)، بررسی می‌کند. او در بررسی خود صفات را از یک پیکره 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^۲ هم استفاده کرد و هر دو را با هم مقایسه نمود و اختلاف ضریب آنها نشان داد که مدل «لاپاتا» از مدل نایو بهتر عمل می‌کند (Lapata 2003, 1-8).

موسوی میانگه برای حل مسئله چندمعنایی در واژه‌های فارسی به کمک آماره اطلاعات متقابل^۳ از اطلاعات باهم‌آیی گرفته‌شده از مجموعه‌ای از واژه‌های زبان مبدأ

1. Princeton Word Net (PWN)

۲. نایو (naive) روشی رایج برای طبقه‌بندی متون است و درباره 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.

3. Mutual Information Statistics

استفاده می‌کند. برای این کار او از آماره اطلاعات متقابل میان جفت واژه‌ها استفاده می‌کند تا مناسب‌ترین معادل انگلیسی را برای واژه مورد نظر بیابد. اطلاعات متقابل بر اساس باهم‌آیی واژه تعیین می‌شود و به‌عنوان ابزاری برای اندازه‌گیری ارتباط میان واژه‌ها استفاده می‌گردد. نتیجه این محاسبات نیز در رفع ابهام واژه‌های چندمعنایی فارسی هنگام ترجمه به انگلیسی بسیار کارآمد بود (Mosavi Miangah 2007). نجومیان به مسئله ابهام و ابهام‌زدایی اشاره می‌کند. به نظر او گذاردن ابهام در سطوح پایین تحلیل، برای مثال در تحلیل‌کننده/ تولیدکننده صرفی^۱ 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 خواهد بود 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 پسا-پردازش^۲ خواهند رسید. کشف زودهنگام باهم‌آیی‌ها و یا ترکیب در نمونه‌ساز^۳ 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 نگاه کردن به شواهد روان‌شناختی زبان در ابهام‌زدایی هم‌نویسه‌های هم‌آوا که در واژگان ذهنی صورت می‌گیرند و نقش بسامد معنای غالب و مغلوب واژه مبهم، مدلی را برای ابهام‌زدایی مفهوم هر واژه بر اساس «هر مفهوم برای یک باهم‌آیی» پیشنهاد می‌دهد که از سوی یارووسکی^۴ (۱۹۹۳) ارائه شد. او از مدل برجسته‌سازی، شبیه به چیزی که فراست و بنتین^۵ (۱۹۹۲) استفاده کرده بودند، بهره می‌برد. این روش به سایر مواردی که واژه مبهم با بسامدهای مختلف اتفاق می‌افتد، ارجحیت دارد، زیرا می‌تواند سطوح فعال‌سازی معنای پربسامدتر را از کم‌بسامدترین آنها دریافت کند. فهرست بسامدها از یک پیکره بزرگ گرفته می‌شود و به‌سادگی به واژگان افزوده می‌گردد. او آزمایش‌های متعددی را برای مشخص کردن چگونگی تأثیر برجسته‌سازی و رابطه آن با بسامد انجام داد. نتایج حاکی از تأثیر بسامد و میانجی‌گری بخش آوایی قبل از رسیدن به معنا بود. طبق

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

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 صفات چندمعنایی در زبان انگلیسی و نیز ترجمه آنها به فارسی ارائه شود. برای این منظور، فرهنگ انگلیسی به فارسی هزاره انتخاب شد که فرهنگی متوسط به‌شمار 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 صفات چندمعنایی که تعداد آنها بیش از ۵۰۰ مورد بود، استخراج شدند. در مرحله بعد، برای ارائه بافت مناسب برای این مدخل‌ها لازم بود آنها را در پایگاه داده‌ها در یک کشف‌اللغات موازی قرار دهیم و باهم آیی‌ها و معانی آنها را برای وارد کردن در پایگاه ضبط کنیم. از میان ۵۰۰ صفت مبهمی که از فرهنگ هزاره استخراج شد، تنها برای ۲۹۸ صفت در کشف‌اللغات مورد نظر (موسوی میانگانه ۲۰۰۹) که دارای ۸۰۰۰۰ جفت جمله انگلیسی فارسی بود، جمله یافت شد. صفات موجود به‌همراه باهم آیی و معنای ارائه‌شده در پیکره ضبط شدند و به‌صورت یک کشف‌اللغات موازی در اختیار برنامه‌نویس قرار گرفتند. جدول شماره ۱ بخشی از کشف‌اللغات تهیه شده مورد نظر را نشان می‌دهد.

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 قرار گرفتن	تحت تأثیر	4
9	affected	affected workers	متلا	کارگران متلا	2
10	affected	affected areas	صدمه‌خورده	مناطق زلزله‌زده	5

برنامه مورد نظر باید می‌توانست این صفات را در صورت وقوع با باهم آیی موجود

در پایگاه ترجمه کند و در صورتی که واژه را در بافت جدیدی ببیند، آن را بر اساس بسامد معنای ارائه‌شده، ترجمه نماید. یعنی باید قادر به ارائه بالاترین بسامد معنایی برای صفت مورد نظر باشد. پس از تکمیل برنامه، لازم بود میزان دقت نیز مورد بررسی قرار گیرد.

۳-۱. برنامه طراحی شده

این برنامه با استفاده از زبان برنامه‌نویسی C#.net و بانک اطلاعاتی اکسس طراحی و برنامه‌نویسی شده است. روش کار الگوریتم بدین صورت است که ابتدا با مبنا قراردادن کلمه اول، سعی در یافتن رکورد متناسب در پایگاه داده بر اساس کلمه دوم می‌کند و در صورتی که جستجوی الگوریتم نتیجه‌ای در بر نداشته باشد، از بین زیرمجموعه استخراج‌شده به سراغ سایر معادل‌ها می‌رود. در صورتی که تعداد نتایج یافت‌شده بیش از یک معادل باشد، بر اساس بیشترین تعداد تکرار فیلتر می‌شود. نکته 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 کلمه مبنا و کلمه هم‌آیند خودش، الگوریتم کماکان به جستجو ادامه می‌دهد و تمام حالات ممکن را استخراج کرده و در پیکره 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 result1 = new result();
result1.name = sdaResult0.String(0);
result.add(result1);
}
sdaResult0.Close();

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

```

با اجرایی شدن این برنامه تعداد ۸۶ صفت از همین کشف اللغات انتخاب شدند. برای هر یک از این صفات جمله‌ای از پیکره ملی بریتانیا (BNC^۱) به صورت تصادفی انتخاب شد. این انتخاب بر اساس داده‌های موجود در پایگاه صورت گرفت. آن گاه، این جمله‌ها (تنها ترکیب صفت و اسم همراه آن) در برنامه ابهام‌زدایی قرار گرفتند و معادل‌های ارائه شده برای آنها ضبط شد. سپس، همان جملات به همراه دو گزینه (یکی گزینه انتخاب شده از سوی برنامه و دیگری گزینه‌ای که معنای دوم یا سوم واژه به شمار می‌رفته است) به ۵ مترجم داده شدند. مترجمان، گزینه‌ای را انتخاب کردند و انتخاب‌های مترجمان با انتخاب‌های برنامه مقایسه شد. برای مثال، واژه affected در پیکره ملی بریتانیا جستجو شد و جملاتی برای آن مشاهده گردید و از میان آنها دو جمله به صورت تصادفی انتخاب شدند و قسمتی از جمله که صفت مورد نظر به همراه اسم بود، به برنامه داده شد و معادل آن ضبط گردید. آن گاه، همان جمله همراه با دو پاسخ در اختیار مترجمان انسانی قرار گرفتند. این دو پاسخ، یکی پاسخ ماشین 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 پیکره ملی بریتانیا انجام و جملات گردآوری شد و در نهایت، پاسخ مترجمان با پاسخ‌های ماشین مقایسه و ضریب همبستگی آنها تعیین گردید.

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

۴. تحلیل داده‌ها و بیان یافته‌ها

طبق آنچه که در روش پژوهش توضیح دادیم، با مقایسه معادل‌های انتخاب‌شده از سوی مترجمان و معادل‌های داده‌شده از سوی برنامه‌ی 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. پایایی دلالت بر آن دارد که ابزار اندازه‌گیری در شرایط یکسان تا چه اندازه نتایج یکسانی به دست می‌دهد. دامنه ضریب اعتبار از صفر تا یک است. برای تعیین پایایی ابزار اندازه‌گیری شیوه‌های مختلفی وجود دارد؛ از جمله روش اجرای دوباره (بازآزمایی)، روش موازی (همتا)، روش تنصیف (دو نیمه کردن)، 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 محاسبه همابستگی درونی ابزار اندازه‌گیری از جمله پرسشنامه به کار می‌رود. در این ابزار، پاسخ هر سؤال می‌تواند 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 به منظور آزمون درجه صحت نرم‌افزار ساخته‌شده، به طراحی پرسشنامه‌ای پرداخته‌ایم تا از طریق جواب‌دهی انسانی به بررسی قابلیت اعتماد پرسشنامه و نهایتاً مناسب بودن نرم‌افزار دست یابیم. در این راستا، پرسشنامه طراحی شده با ۸۶ سؤال در اختیار ۵ مترجم انسانی قرار داده شد و داده‌های حاصل از پرسشنامه، توسط نرم‌افزار آماری spss 15 و از طریق تهیه 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، به صورت معمول از بررسی انجام‌شده در آزمون کرونباخ حذف شده‌اند (شماره سؤالات مندرج در جدول ۲).

جدول ۲. سؤالات حذف شده

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 می باشد:

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^۱ برای ۴۰ سؤال باقی مانده، مقدار آلفای کرونباخ و آلفای کرونباخ بر پایه مقادیر استاندارد شده، به ترتیب برابر ۰/۹۵۳ و ۰/۹۶۹ می باشد (بالاتر از ۰/۷) که بیانگر همبستگی بالای متغیرها در شش گروه آزمون شده می باشد.

۵. محدودیت‌های پژوهش

طبق تحلیل فوق، این برنامه در ۵۴ درصد موارد، موفق به ارائه ترجمه قابل قبول شد. عدم دقت بالاتر این برنامه ناشی از چند موضوع است. نخست آنکه کشف اللغاتی که برای یافتن بافت صفات استفاده شد، یک پیکره ۸۰۰۰۰ جمله‌ای بود (تنها کشف اللغات موازی در دسترس نگارنده) که تنها برای ۲۹۸ مورد از صفات استخراج شده، جمله ارائه می کرد. این امر خود باعث پایین بودن دامنه صفات در کشف اللغات تهیه شده می باشد. در صورتی که نگارنده می توانست پیکره کامل تری در اختیار داشته و برای همه معانی موجود برای هر یک از صفات باهم آبی‌های کافی داشته باشد، به یقین به معادل‌های بسیار دقیق تری دست می یافت. نکته دیگر آنکه، اگر بخواهیم این برنامه با دقت بالاتری ترجمه کند، باید تا آنجا که امکان دارد باهم آبی‌های صفات مبهم را از پیکره‌های گوناگون استخراج کنیم؛ به گونه‌ای که حتی کم‌بسامدترین‌ها نیز در کشف اللغات وجود داشته باشند. در این

1. reliability statistics

مشغول هستند، بسیار کارآمد باشد.

فهرست منابع

- بازرگان، عباس. ۱۳۸۳. ارزشیابی آموزشی، مفاهیم، الگوها و فرایندهای عملیاتی، تهران: سمت
- حق شناس، علی محمد، حسین سامعی، و نرگس انتخابی. ۱۳۸۱. فرهنگ معاصر هزاره. تهران: فرهنگ معاصر.
- فلاحتی، محمدرضا. ۱۳۸۳. انسان مترجم و ترجمه ماشینی: بررسی موردی مشکلات ماشین ترجمه انگلیسی به فارسی «پدیده». فصلنامه کتابداری و اطلاع رسانی (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, Stephanie 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.
- Nojournian, 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 Language Technology Workshop, Princeton, NJ, pp.: 266-271.