عنوان زیرپروژه:

بررسی ابعاد و لایه‌های ابهام در واژگان مشابه زبان فارسی

فهرست مطالب

عنوان **شماره صفحه**

مقدمه 1

فصل اول: **ترجمه ماشینی** 3

1-1- مقدمه ای بر ترجمه ماشینی 5

1-2-تاریخچه مختصر ترجمه ماشینی 5

1-3- انواع سیستم های ترجمه ماشینی 10

1-3-1- سیستم های مستقیم 10

1-3-2- سیستم های غیر مستقیم 12

1-3-2-1- روش انتقالی 12

1-3-2-1-1- انتقال واژگانی 14

1-3-2-1-1- انتقال ساختاری 15

1-3-2-2- روش زبان میانجی 15

1-3-3- روش های مبتنی بر قانون 16

1-3-4- روشهای مبتنی بر دانش 17

1-3-5- ماشین های ترجمه مبتنی بر پیکره(زبانی)..... 20

1-3-5-1- روشهای مبتنی بر نمونه..... 21

1-3-5-2- ترجمه ماشینی آماری..... 23

1-3-5-2-1- پیکره متناظر واژه ای..... 25

1-3-5-2-2- پیکره متناظر جمله ای..... 26

1-3-5-2-2- مراحل ساخت یک پیکره در ماشین ترجمه آماری..... 26

الف- پیش پردازش و علامت گذاری متون..... 27

ب- هم ترازی..... 27

ب-1- اندازه گیری رخداد کلمات و عبارات..... 29

ب-2- یک استراتژی در هم ترازی عبارات..... 30

ب-3-طبقه بندی و انتخاب عبارات..... 31

ب-4- هم ترازی در سطح عبارات..... 32

ب-5- هم ترازی در سطح کلمات..... 32

ب-6- پس پردازش..... 33

1-3-6-ترجمه ماشینی مبتنی بر اصل..... 34

1-3-7- روشهای ترکیبی..... 34

1-4- مکانیزم ،مؤلفه ها و اجزای سازنده سیستم های ترجمه ماشینی..... 35

1-4-1- واژگان 36

1-4-2- نحو و ساختار 37

1-4-3- معنا شناسی 38

1-5- ابهام زدایی معنی واژه 39

1-5-1- روش های مبتنی بر دانش یا مبتنی بر قاعده 40

1-5-2- استفاده از روشهای آماری در ابهام زدایی معنایی 40

1-5-2-1- استفاده از گنجواژه 41

1-5-2-1-1- محدود سازی طبقه 42

1-5-2-1-2- جمع بارها 42

1-5-2-1-3- انتخاب محتمل ترین طبقه 43

1-5-3- استفاده از موصوف برای دستیابی به معنای صفت 43

1-5-4-ترکیب آماری شاخص های چند گانه 45

1-5-4-1- مدل های تجزیه شدنی 45

1-5-4-2- ابهام زدایی مبتنی بر نمونه 46

1-5-5- استفاده از پیکره یک زبانه زبان مقصد 47

1-5-5-1- روش داگان و ایتای 47

1-5-5-2- روش هیونگ لی وجنگ سی پارک و چنگ کیم 48

الف-آشکار سازی معنی کلمه مبدأ و نگاشت کلمه..... 48

ب-نگاشت کلمه زبان مبدأ به مجموعه ای از کلمات زبان مقصد..... 49

ج- انتخاب واژه با استفاده از اطلاعات آماری زبان مقصد..... 50

3-5-5-1-روش موسوی و دلاور خلفی..... 52

فصل دوم: مقدمه ای بر کشف دانش..... 54

2-1- مقدمه ای بر کشف دانش 54

2-2- بررسی موضوع و مسائل پیرامون آن..... 54

2-3- تعاریف اصلی..... 54

2-3-1-طبقه بندی چند هدفه و چند طبقه بودن..... 54

2-3-2-کشف دانش در پایگاه داده ها..... 54

2-3-3- داده کاوی..... 54

2-3-4-عامل..... 55

2-4- هدف پروژه..... 55

2-5- مروری بر کشف دانش در پایگاه داده ها..... 55

2-5-1- دلایل توجه..... 55

2-5-2- فناوری های کلیدی مورد استفاده..... 57

فصل سوم: کشف دانش بوسیله الگوریتم بهینه سازی 58

3-1-کشف دانش بوسیله الگوریتم بهینه سازی 58

3-2- فعالیت طبقه بندی 58

3-3- مزیتهای الگوریتم بهینه سازی 59

3-3-1- الگوریتم بهینه سازی 59

3-3-2-حل مسئله 59

3-4- الگوریتم بهینه سازی 59

3-5- استفاده از الگوریتم در حل مسئله 63

3-6- الگوریتم و جزئیات یادگیری مسئله 64

3-7- سیستم یادگیری 65

فصل چهارم: کشف دانش چند طبقه مبتنی بر بهینه سازی 67

4-1-کشف دانش چند طبقه مبتنی بر بهینه سازی 67

4-2-دانش چند طبقه و نحوه نمایش آن 67

4-3- تحلیل هدف از نقطه نظر الگوریتم بهینه سازی 68

4-3-1- گام های دیگر طراحی 70

4-4- سلسله مراتب درالگوریتم 71

4-4-1- پی آمد ناشی از تخصصی شدن کار الگوریتم ها..... 71

4-4-2- تغییر فضای مسئله..... 72

4-5- بررسی چگونگی انتخاب عبارات..... 73

4-6- نحوه کار الگوریتم چرخ رولت در انتخاب عبارت ها..... 74

4-7- الگوریتم کشف دانش چند طبقه مبتنی بر ACO..... 74

4-7-1- ساخت قانون..... 80

4-7-2-هرس قانون..... 81

4-7-3-بروز رسانی فرومون..... 81

4-7-4-ساخت قوانین طبقه بندی..... 82

4-7-5-تابع هیوریستیک..... 83

4-7-6-هرس قوانین طبقه بندی..... 85

4-7-7-قوانین انتقال و بروز رسانی فرومون..... 86

فصل پنجم: ابهام زدایی واژه 89

5-1- ابهام زدایی واژه با استفاده 89

5-2-نتایج و بررسی کارایی الگوریتم ابهام زدایی از معنی واژه..... 94

5-3-نحوه طبقه بندی بوسیله قوانین کشف شده 94

5-4-مجموعه آموزشی 94

5-5-تنظیم پارامتر های الگوریتم 95

5-6-نتایج کسب شده 96

5-6-1-1-بررسی نتایج کسب شده 97

5-6-1-2- محاسن بکارگیری این روش در مقایسه با روشهای موجود 99

فصل ششم- نتیجه گیری 100

منابع و مراجع 103

پیوست 107

مقدمه

پیشرفت علوم و فن آوری و پیرو آن ، راحتتر شدن زندگی امروزی ، پیوندی ذاتی و ناگسستنی با دانش بشری دارد و در این میان بشر از همان ابتدای خلقت سعی بر این داشته تا با الهام از طبیعت پیرامون خود برای مشکلات و موانعی که برسر راه داشته راه چاره ای بیاندیشد.از طرفی حجم زیاد داده هایی که در زمینه های مختلف علوم بدست آمده اند که با رشدی تصاعدی و روز افزون ، در جریانی یک طرفه به سمت ما می آیند ، که این خود باعث بحرانی شدن مسئله دسترسی به دانش مورد نیاز گردیده است.این پدیده نیز به تبعه خود باعث بحرانی شدن فعالیتهایی مانند تصمیم گیری[1] و یا طرح ریزی[2] در حوزه های مرتبط با دانش می شود.(۱)

با این وجود، فناوری ابزارهای لازم برای مقابله با این مشکل رادر اختیار ما نهاده است. تکنیکهای سخت افزاری ، نرم افزاری و هوش مصنوعی به عنوان راههای مقابله با آن می توانند در نظر گرفته شوند . که تلفیق این تکنیکها با هم نیز خود باعث شکل گیری راههای جدید شده است.از این میان جستجو و کشف دانش در پایگاه داده ها[3] با مجموعه ای از روشهایی که امکان غربال کردن داده ها و بدست آوردن دانش را در اختیار ما می گذارد ، قدرت و انعطاف مناسبی را از خود نشان داده است.

از میان روشهای متعددی که برای کشف دانش در پایگاه داده ها وجود دارند ، روشی خاص مبتنی بر الگوریتم بهینه سازی ها[4] (2)(3)(4) مطرح گردیده است .

هدف از پروژه ای که این پروژه گزارش تحلیلی آن می باشد ، استفاده از کشف دانش چند طبقه با بکارگیری الگوریتم بهینه سازی در ابهام زدایی از معنی واژه است. این الگوریتم توانایی زیادی در حل مسائل ترکیبی[5]نشان داده است(3) لزوم انجام این پروژه مشخص می شود.

فصل اول بطور خلاصه به ترجمه ماشینی می پردازد و موضوع اصلی مورد تحقیق را مورد بررسی قرار میدهد.مقدمه ای برالگریتم بهینه سازی در فصل دوم مرور می گردد . کشف دانش بوسیله الگوریتم بهینه سازی و الگوریتم های بهینه سازی مبتنی بر آن و و مزایای این الگوریتم در فصل سوم مورد مطالعه قرار می گیرد. در فصل

[1] Decision making
[2] Plannig
[3] Knowledge Discovery in Data Base(KDD)
[4] Ant Colony Optimization(ACO)
[5] Combinatorial

چهارم کشف دانش چند طبقه مبتنی بر بهینه سازی به عنوان یک جزء از روشهای جستجوی دانش بررسی می گردد. و ابهام زدایی معنی واژه ونتایج پیاده سازی این الگوریتم بر روی یک پیکره فارسی در فصل پنجم مطالعه می شود.

فصل اول: ترجمه ماشینی

1-1- مقدمه ای بر ترجمه ماشینی

مکانیزه کردن ترجمه یکی از نخستین رؤیاهای بشریت بوده است . در قرن بیستم این رؤیا به شکل برنامه های کامپیوتری که قادر به ترجمه حجم زیادی از متون از یک زبان به زبان دیگر بودند به حقیقت پیوست. اما چون همیشه ،واقعیت کامل نیست. ماشینهای ترجمه ای وجود ندارند که با فشار چند تکمه هر متن به هر زبانی را دریافت و ترجمه کاملی به هر زبان دیگر بدون دخالت یا کمک انسان تولید کنند. این کمال مطلوبی است برای آینده ای نزدیک که اگر چه اساساً قابل دسترسی است، رسیدن به آن چندان قطعی نیست. آنچه که تاکنون به آن رسیده ایم ساخت برنامه هایی است که می توانند ترجمه های خامی از متون در موضوعات نسبتاً تعریف شده ای تولید کنند که بعداً مورد تجدید نظر قرار می گیرند تا متون ترجمه شده ای با کیفیت خوب و مناسب از نظر اقتصادی ارائه دهند یا اینکه به همان حالت اصلاح نشده به وسیله متخصصان آن موضوع صرفاً برای گرفتن اطلاعات خوانده و فهمیده شوند. در بعضی موارد با کنترلهای مناسب بر روی زبان متون ورودی، ترجمه های خودکاری با کیفیت بهتر که نیاز کمی به اصلاح دارند یا اصلاً چنین نیازی ندارند می توان تولید کرد.

اینها پیشرفتهایی هستند در آنچه که معمولاً ترجمه ماشینی نامیده می شوند، اما این مفهوم اغلب مبهم بوده و اشتباه تعبیر می شود. دریافت عمومی از ترجمه ماشینی به وسیله دو نقطه نظر افراطی تحریف شده است. از یک طرف کسانی هستند که متقاعد نشده اند که تجزیه زبان با مشکل همراه است چرا که حتی کودکان قادرند که براحتی زبان را فراگیرند. اینها کسانی هستند که معتقدند هر کس که یک زبان خارجی را بداند باید بتواند براحتی آن را ترجمه کند. بنابراین آنها قادر به فهم مشکلات کار و یا اینکه چه مقدار کار با موفقیت انجام شده است نیستند. از طرف دیگر افرادی معتقدند که چون ترجمه خودکار شکسپیر، گوته، تولستوی و دیگر ادیبان میسر نیست پس هیچگونه جایی برای هیچ نوع ترجمه مبتنی بر کامپیوتر وجود ندارد. آنها قادر نیستند سهمی راکه ترجمه هایی نه کامل می توانند در کارشان یا در پیشرفت کلی ارتباطات بین المللی داشته باشند ارزیابی نمایند (5).

در سالهای 1950 و 1960 رشته ترجمه ماشینی حوضه تحقیقی مهمی در زبانشناسی رایانه ای به حساب می آمد.

در شروع، هدف مورد انتظار ترجمه خودکار انواع متون و اسناد در کیفیتی معادل یا حتی بهتر از مترجمان انسانی بود. اما خیلی زود معلوم شد که این هدف در آینده ای قابل پیش بینی غیر ممکن است. اگر نتایج ترجمه بخواهد به شکلی قابل چاپ در آید، اصلاح و تجدید نظر برونداد ترجمه ماشینی بوسیله انسان ضروری است. در عین حال این نکته نیز قابل درک بود که برونداد ترجمه ماشینی به شکل اصلاح نشده و خام نیز می تواند برای بسیاری از مقاصد از جمله برای خواندن متن به منظور پی بردن به ایده و محتوای کلی آن به یک زبان ناآشنا مفید واقع گردد. اما این کاربرد دوم ترجمه ماشینی برای سالها مورد غفلت واقع گشت (6).

بیشترین حجم ترجمه در دنیا از متونی نیست که از موقعیت ادبی و فرهنگی بالایی برخوردار باشند. اکثریت قریب به اتفاق مترجمان حرفه ای دست اندرکار ترجمه اسناد و متون علمی و فنی، معاملات تجاری و بازرگانی، اساسنامه های مدیریتی، اسناد حقوقی، دستور العملها، کتابهای کشاورزی و پزشکی، پروانه های صنعتی، نشریه های تبلیغاتی، گزارشات روزنامه ای و غیره می باشند. مقداری از این کار پردرد سر و مشکل است . اما بیشتر آن کسل کننده و تکراری است ولی در عین حال نیاز به صحت و ثبات دارد. تقاضا برای چنین ترجمه هایی با سرعتی بیش از قابلیت حرفه مترجمی رو به افزایش است. مساعدت یک کامپیوتر جذابیتهای واضح و ضروری دارد. سودمندی عملی یک سیستم ترجمه ماشینی نهایتاً بوسیله کیفیت برونداد آن تعیین می شود. اما آنچه که به عنوان یک ترجمه خوب صرفنظر از اینکه بوسیله کامپیوتر باشد یا انسان تلقی شود مفهومی است که تعریف دقیق آن مشکل به نظر می رسد. در عین حال بیشتر آن وابسته به موقعیتهای ویژه ای که در آنها ترجمه صورت می گیرد و نیز دریافت کننده های ویژه ای که ترجمه برای آنها انجام می شود است . صداقت، صحت، وضوح ، سبک مناسب و سیاق همه از جمله ملاکهایی هستند که می توان از آنها استفاده کرد، اما باز هم اینها قضاوتهای ذهنی هستند. تا آنجایی که به ترجمه ماشینی مربوط می شود آنچه که در عمل مهم است این است که چه مقدار تغییرات بایستی صورت گیرد تا برونداد ترجمه به حد استاندارد قابل قبولی ازنظر مترجم یا خواننده انسانی رسانده شود. با چنین مفهوم دشوار و بی ثباتی چون ترجمه، محققان و سازندگان ترجمه ماشینی نهایتاً آرزو دارند که حداقل بتوانند ترجمه هایی تولید کنند که در موقعیتهای ویژه ای که مجبور به توضیح دادن اهداف تحقیق هستند مفید واقع شوند و یا کاربردهای مناسب دیگری برای ترجمه هایشان بیابند.

ترجمه ماشینی بخشی از حوزه وسیعتر" تحقیق محض" در پردازش زبان طبیعی مبتنی بر کامپیوتر در زبانشناسی

رایانه ای و هوش مصنوعی می باشد که مکانیزم های اساسی زبان و ذهن را بوسیله مدلسازی و شبیه سازی در برنامه های کامپیوتری مورد بررسی قرار می دهد. تحقیق در زمینه ترجمه ماشینی با پذیرش و بکارگیری جنبه های نظری و روشهای عملی در فرآیندهای ترجمه شدیداً با این امور سروکار دارد و به نوبه خود می تواند بینش ها و راه حلهایی را برای مشکلات و مسائل خاص خود بیابد. علاوه بر این ترجمه ماشینی می تواند بستر آزمایشی مناسبی را در مقیاس بزرگتر برای نظریه ها و روشهایی که بوسیله آزمایشات و تحقیقات در مقیاس کوچک در زبانشناسی رایانه ای و هوش مصنوعی انجام می شود فراهم نماید.

مانع اصلی بر سر راه ترجمه بوسیله کامپیوتر مانع کامپیوتری یا محاسبات نیست بلکه مانع زبانی یا زبانشناختی می باشد. مشکلات زبانی عبارتند از: ابهام واژگانی، پیچیدگی نحوی، تفاوتهای واژگانی بین زبانها، ساختارهای غیر دستوری و مستتر و به طور خلاصه استخراج معنای جملات و متون از تجزیه نشانه های نوشتاری و تولید جملات و متون به مجموعه دیگری از علامات زبانی با معنای معادل. در نتیجه ترجمه ماشینی به شدت متکی به پیشرفتهای تحقیقات زبانی به ویژه شاخه هایی از آن با درجات بالایی از فرمالیته بودن است و در حقیقت همیشه همینطور بوده و خواهد بود. اما ترجمه ماشینی نمی تواند نظریه های زبانی را مستقیماً به کار گیرد. زبانشناسان با شرح مکانیزم های زیر ساختی تولید و درک زبان سروکار دارند. آنها بر روی مشخصه های قطعی متمرکزند و نه کوشش برای توصیف یا توضیح در مورد هر چیزی. در مقابل، ترجمه ماشینی با متون واقعی سروکار دارد. سیستم های ترجمه ماشینی با دامنه وسیعی از پدیده های زبانی، پیچیدگی های اصطلاحات، غلط های املایی، واژه های جدید و جنبه های اجرایی که همیشه هم در رابطه با زبانشناسی نظری مجرد نیستند سروکار دارد.

بطور خلاصه، ترجمه ماشینی به خودی خود یک شاخه مستقلی از تحقیق محض نیست بلکه هرگونه ایده، روش و فنی را که ممکن است در توسعه سیستم های پیشرفته سهیم باشد خواه از زبانشناسی یا علم کامپیوتر، هوش مصنوعی و یا نظریه ترجمه باشد، اقتباس میکند. در واقع ترجمه ماشینی تحقیق کاربردی است، اما شاخه ای که از حجم مهمی از فنون و مفاهیم که به نوبه خود می توانند در حوضه های دیگر پردازش زبان مبتنی بر کامپیوتر به کار روند تشکیل شده است (5).

2-1- تاریخچه مختصر ترجمه ماشینی

استفاده از واژه نامه های خودکار برای غلبه بر موانع زبانی اولین بار در قرن هفدهم پیشنهاد شد. در قرنهای بعد پیشنهادهای (طرحهای) زیادی برای ایجاد زبانهای بین المللی (مانند اسپرانتو1 به عنوان مشهورترین آنها) ارائه شد لیکن طرحهای اندکی برای مکانیزه کردن ترجمه تا اواسط قرن 20 ارائه گردید. در سال 1933 دو ابداع به طور جداگانه در فرانسه و روسیه به ثبت رسید. یک فرانسوی - ارمنی بنام جورج آرتسرونی[2] یک وسیله ذخیره ای را بر روی نوار کاغذی طراحی کرده بود که می توانست برای پیدا کردن معادل هر واژه به زبان دیگر مورد استفاده قرار گیرد. یک نسخه اصلی از این کار ظاهراً در سال 1937 عرضه شد. طرحی که به وسیله پتر اسمیرنف ترویانسکی[۳] روسی ارائه شد در آن زمان مهم تر به نظر می رسید و سه مرحله برای ترجمه مکانیکی در نظر می گرفت. در مرحله اول ویرایشگری که فقط زبان مبدأ را می دانست تجزیه منطقی واژه ها به صورت اشکال اصلی آنها و نقشهای نحوی را انجام میداد. در مرحله دوم ماشین می بایستی زنجیره هایی از اشکال اصلی و نقشها را به زنجیره های معادل آنها در زبان مقصد تبدیل کند. و در نهایت ویرایشگر دیگری که فقط زبان مقصد را می دانست این برونداد را به اشکال عادی آن زبان تبدیل می نمود. ترویانسکی از زمان خودش جلو بوده و خارج از روسیه ناشناخته باقی مانده بود، حتی زمانی که امکان استفاده از کامپیوترها برای ترجمه اولین بار به وسیله وارن ویور[٤] و آندره دی بوت[٥] مورد بحث قرار گرفت. اما این ویور بود که در جولای 1949 برای اولین بار توجه عموم را به ایده ایده ترجمه ماشینی جلب کرده و روشهایی از قبیل: استفاده از فنون رمز نویسی[٦] دوران جنگ، تجزیه آماری، نظریه اطلاعاتی شنون[۷]، و جستجوی (کشف) مشخصه های زیربنایی منطقی و جهانی زبان را پیشنهاد نمود. ولی در مجموع لئون دوسترت[۸] در دانشگاه جورج تاون با همکاری IBM اولین پروژه عملی را که نتیجه آن اولین نمایش عمومی سیستم ترجمه ماشینی در ژانویه 1954 بود به ثمر رسانید. در این پروژه نمونه های به دقت انتخاب شده از جمله های روسی با استفاده از واژگان بسیار محدود در حدود 250 واژه و فقط شش قاعده دستوری به انگلیسی ترجمه شد. اگر چه این پروژه ارزش علمی کمی داشت و چندان موفق از آب در نیامد اما برای سرمایه

[1] - Esperanto
[2] -Gorge Artsrouni
[3] - Petr Smirnov-Troyanskii
[4] - Warren Weaver
[5] - Andrew D.Booth
[6] - cryptography
[7] - shannon's information theory
[8] -Leon Dostert

گذاری وسیع در زمینه تحقیقات ترجمه ماشینی در ایالات متحده محرک بسیار مؤثری بوده و مشوقی برای شروع پروژه های ترجمه ماشینی در هر کجای دیگر دنیا به ویژه در شوروی سابق بود (5).

بطور خلاصه می توان گفت که در این دوره ماشین های ترجمه نسل اول تولید شدند .این نسل از ماشین های ترجمه اغلب وابسته به فرهنگ لغت [1] بوده و تحلیل های گرامری سطح پایین داشتندو استفاده از خصیصه های معنایی از جمله ویژگیهای آنها بوده است.

در دهه بعد ترجمه ماشینی رشد بیشتری نمود. دراین دوره ماشین های ترجمه نسل دوم توسعه یافتند که این نسل از ماشین ها مبتنی بر قواعد نوشتاری [2] بوده و نمایش خلاصه ای از متن تبدیل یافته ارائه می نمودند و اغلب از رویکرد تبدیل استفاده می کردند . در این دوره که به دوره کمال گرایی [3] مشهور است تأکید تحقیقات بر روی جستجوی نظریه ها و روشهایی برای رسیدن به ترجمه های کامل بود و نه چیزی کمتر از آن، که این خود تا حد زیادی مخرب بود چون هیچ سیستم کاملاً خودکار که قادر به ترجمه هایی با کیفیت بالا باشد وجود نداشته است و شاید هنوز هم وجود نداشته باشد (7).

در یک ارزیابی از ترجمه ماشینی که در سال 1960 انجام شد بار- هیلر [4] بر این فرضیه متداول که هدف تحقیقات ترجمه ماشینی باید تولید ترجمه کاملاً خودکار با کیفیت بالابوده و نتایج ترجمه های آنها بایستی غیر قابل تشخیص از نتایج ترجمه های انسانی باشد خرده گرفت. او چنین بحث کرد که موانع معنایی بر سر راه ترجمه ماشینی اساساً می توانند تنها به وسیله وارد کردن مقدار زیادی دانش دایره المعارفی در مورد جهان واقعی برداشته شوند. توصیه او این بود که ترجمه ماشینی نباید زیاد بلند پروازی کند، و بایستی سیستم هایی را بسازد که استفاده مقرون به صرفه از تعامل بین انسان و ماشین نماید .(نظریه ای که او برای اولین بار تقریباً ده سال قبل از آن یعنی زمانی که ترجمه ماشینی هنوز در طفولیت بود ارائه کرد). در سال 1966 گزارشی توسط شرکتهای دولتی ترجمه ماشینی در ایالات متحده بنام ALPAC منتشر شد مبنی بر اینکه ترجمه ماشینی کندتر، ناصحیح تر و دو برابر گرانتر از ترجمه انسانی است و اینکه هیچ امیدی به ساخت ترجمه ماشینی مفید نمی رود. آنها دیگر نیازی نمی

[1] -Dictionary Driven
[2] -Rule Based
[3] - perfectionism
[4] -Bar-Hiller

دیدند تا بر روی تحقیقات ترجمه ماشینی سرمایه گذاری کنند و به جای آن توصیه کردند که کمک های ماشینی ای مانند فرهنگهای ماشین خوان برای مترجمان و حمایت مداوم از تحقیقات اساسی در زمینه زبانشناسی رایانه ای انجام شود. تأثیر گزارشALPAC بسیار زیاد و وسیع بود هر چند که این انتقاد درست نبود. انتشار و نفوذ این گزارش آنقدر عمیق بود که خط پایانی برای تحقیقات ترجمه ماشینی برای بیش از یک دهه در ایالات متحده بود و برای سالها بعد از آن احساس عموم را نسبت به ترجمه ماشینی خدشه دار نمود (5).

در دهه بعد تحقیقات زیادی در زمینه ترجمه ماشینی خارج از ایالات متحده انجام گرفت و منجر به ساخت سیستم هایی مانند سیستم میتیو[1] در مونترال برای ترجمه گزارشات آب و هوایی در رسانه های عمومی در سال 1976 و ساخت سیستم یوروترا[2]، یک سیستم چند زبانه براساس آخرین پیشرفتهای ترجمه ماشینی و زبانشناسی رایانه ای در سال های پایانی 1970 گشت (6).

سیستم ترجمه ماشینی در اواخر سالهای 1970 بالاخره به بازار تجارت راه یافت. یعنی زمانی که شرکت زیراکس کورپ[3] از این سیستم برای ترجمه هزاران صفحه از اسناد رایج در طی یک سال استفاده کرد در واقع به محصولات چند زبانی اجازه ورود به بازارها و نمایشگاههای جهانی را داد. در ابتدا کاربران در اندیشه اینکه این نرم افزار می تواند ترجمه های کامل را تولید نماید آن را خریداری می کردند لیکن امروزه همانطور که می دانیم ترجمه ماشینی بیشتر به عنوان یک ابزار برای ترجمه کامل بکار گرفته می شود تا یک مترجم کامل (8).

تا قبل از سالهای 1980 سیستم های ترجمه ماشینی عمدتاً بر اساس ترجمه مستقیم بوسیله فرهنگهای دو زبانه بودند که از تجزیه نحوی نسبتاً پایینی نیز برخوردار بودند. اما با پیشرفت در زمینه زبانشناسی رایانه ای روشهای مناسب تری در این زمینه اتخاذ شد و برخی سیستم ها از روشهای غیر مستقیم نیز در ترجمه بهره جستند. بدین ترتیب که متنهای زبان مبدأ به باز نمود مجردی از معنا تبدیل می شد و در پی آن برنامه هایی برای تشخیص ساختار واژه (صرف) وساخت جمله (نحو) و نیز برای حل مسئله ابهام ساخته می شد (5).

در خلال سالهای 1980طرح مبتنی بر انتقال با روشهای جدید نظریه چند زبانگی در هم آمیخت. مهمترین آنها تحقیق بر روی سیستم های مبتنی بر دانش به خصوص در دانشگاه ملون و پیتزبورگ بود که براساس ساخت سیستم

[1] - Meteo system
[2]- Eurotra system
3- Xerox Corp

های درک زبان طبیعی در محدوده جامعه هوش مصنوعی بوده است. بحث بر سر این است که ترجمه ماشینی بایستی فراتر از اطلاعات زبانی محض (نحو ومعنا شناسی) رود؛ ترجمه با درک مفهوم متن سروکار دارد و بایستی به دانش جهان واقعی رجوع نماید.

مهمترین توسعه قرن بیستم ظهور سیستم های ترجمه ماشینی تجاری بود. سیستم هایی از قبیل لوگوس[1] آمریکایی و متال[2] ژاپنی از جمله مهمترین آنها بودند. تقریباً تمام اینگونه سیستم های اجرایی به شدت وابسته به عمل پس ویرایش برای تولید ترجمه قابل قبول بودند. اما پیش ویرایش نیز متداول بود. بدین ترتیب که در بعضی از سیستم ها حضور یک اپراتور هنگام ورود متن برای مشخص کردن مرزهای واژه ها یا حتی نشان دادن محدوده عبارات و بندها لازم بود.

تا اواخر سالهای 1980 سیستم های با قابلیت ترجمه حجم زیادی از متون و کیفیت پائین عرضه شدند که اغلب آنها نیازمند پس ویرایشهای اساسی بودند. به منظور بهبود کیفیت ترجمه ماشینی تمهیدات زیادی در زمینه واژگان و نحو و نیز سبک متن قبل از ورود صورت گرفت و بدین ترتیب سیستم هایی چون سیسترن[3]، لوگوس ، متال و سیستم های مشابه بوسیله سازمانها و شرکتهای مهم بین المللی بکار گرفته شدند (10). احیای تحقیقات ترجمه ماشینی در سالهای 1980 و ظهور سیستم های ترجمه ماشینی در بازار تجارت سبب آگاهی عمومی روزافزونی در جهت اهمیت ابزار ترجمه شد. پیشرفتهای بیشتر در فناوری کامپیوتر، هوش مصنوعی و زبانشناسی نظری خطوط تحقیقات آینده را تعیین می نماید. اما اساسی ترین مشکل ترجمه مبتنی بر کامپیوتر مشکل آن با فناوری نیست بلکه با زبان، معنا، درک و تفاوتهای فرهنگی و اجتماعی جوامع بشری می باشد (5).

در واقع آغاز سالهای 1990 پیشرفتهای چشمگیری را در ترجمه ماشینی همراه با تغییرات اساسی در استراتژی از ترجمه مبتنی بر قوانین دستوری به ترجمه مبتنی بر پیکره های مبتنی بر نمونه ها (مثلا برنامه ریورسو[4]) به خود دید و زبان ، دیگر به عنوان یک وجود ساکنی که با قوانین ثابت و تغییر ناپذیر اداره می شد تلقی نمی گشت بلکه به عنوان یک پیکره فعال و متحرک که بسته به کاربرد و کاربر در طول زمان و بر طبق واقعیتهای فرهنگی و اجتماعی

1 - LOGOS
2 - METAL
3 - SYSTRAN
4- Reverso Program

تغییر پیدا می نمود در نظر گرفته می شد.

در این دوره نسل سوم ماشین های ترجمه بوجود آمدند. در این نسل ماشین های ترجمه از ترکیب رویکرد مبتنی بر قوانین نوشتاری نسل دوم با متد های مبتنی بر پیکره[1] بوجود آمده بودند. ویژگی های این نسل شامل تحلیلی دستوری در سطح وابستگی ارتباطی تحلیل معنایی در سطح تشخیص جملات و اطلاعات لغوی از فرهنگ لغات بدست آمده و همراه با مشخصه های دستوری در زبان مبداء نگهداری می شوند ، می باشد. در این نسل متدهای مبتنی بر نمونه وروشهای آماری بکار گرفته شدند ، همچنین دانش جمع آوری شده محدود به حوزه مشخصی گردید و از بازخورد[2] کاربران ماشین هم در جمع آوری آماری اطلاعات استفاده میشد . بطور خلاصه می توان گفت که اولین پیاده سازی ماشین ترجمه های مبتنی بر دانش[3] در این دوره بوده است.

در دوره حاضر همچنان بهره برداری تجاری ماشین های ترجمه نسل دوم مشهود بوده و آگاهی نسبت به این تکنولوژی افزایش یافته است بطوریکه به متدهای ارزیابی ماشین های ترجمه نیاز مبرم می باشد که اگر این نیاز مرتفع نگردد به زودی کلیه تلاشها در جهت پیشرفت این تکنولوژی با خطر بی اعتمادی روبرو می گردند و طراحی ها در این دوره ،جهت تحقق اهدافی مانند ترکیب ماشین های ترجمه با سیستم های حافظه ترجمه ،ماشین های ترجمه سخنگو و سیستم های همه منظوره و بالا بردن کیفیت ترجمه ها می باشند

امروزه بسیاری از سیستم های تجاری ترجمه ماشینی از حافظه ترجمه بهره می جویند و بسیاری از سیستم های حافظه ترجمه نیز به نوبه خود با روشهای جدیدتر ترجمه ماشینی پیشرفته تر می شوند. در واقع مشخصه عمده سالهای 1990 افزایش سریع استفاده از ترجمه ماشینی و ابزار ترجمه است.

تا امروز ترجمه ماشینی همچنان به پیشرفت خود ادامه داده است. شرکتهای بزرگ بیشتر از آن استفاده کرده و فروش نرم افزارهای آن در بازار رو به افزایش است. این موقعیت سبب ایجاد سرویسهای ترجمه ماشینی شبکه ای (متصل به اینترنت) مانند آلتاویستا[4] که سرویسهای پست الکترونیکی سریع، صفحات وب و غیره را به زبان دلخواه ارائه میدهند شده است.(49)

1- Corpus_Based

2 - Feed back

3 -Knowledge-Base

4 - Altavista

1-3- انواع سیستم های ترجمه ماشینی

فرض عموم براین است که سیستم های ترجمه ماشینی به صورت یک جعبه سیاه جداگانه و مجزا وجود دارند. اما اکثر سیستم های ترجمه ماشینی خود از دو بخش عمده موتور ترجمه و یک یا چند واژه نامه تشکیل شده اند که طرق عملکرد این بخشها از سیستمی به سیستم دیگر متفاوت است. همچنین روشهای مختلفی که در ترجمه ماشینی اتخاذ می شود عمدتاً بر اساس منابع دانش مختلفی است که در کار ترجمه مورد استفاده قرار می گیرد.

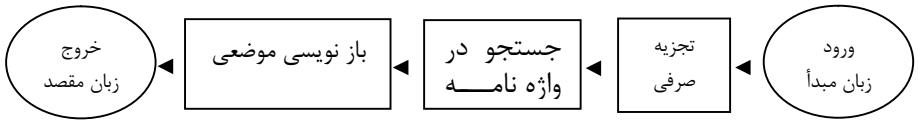
1-3-1- سیستم های مستقیم

سیستم های قدیمی اغلب از روش سیستم مستقیم که به سیستم های ترجمه ماشینی نسل اول معروف بودند استفاده می کردند. این روش از ابتدائی ترین روشهاست که از یک فرآیند یک مرحله ای استفاده نموده و جملات زبان مبدأ با توجه کمی به تفاوتهای ترتیب واژه ها به طور مستقیم و بدون مراحل واسطه ای به جملات زبان مقصد تبدیل می شوند. این روش بر اساس فرهنگ لغات غنی و تجزیه صرفی تنظیم شده است (9).

در رابطه با عملکرد سیستم های ترجمه ماشینی نسل اول باید گفت که کامپیوترهای موجود در اوخر سالهای 1950 و اوایل 1960 حتی در مقایسه با محقر ترین محاسبه گرهای الکترونیکی امروزی بسیار ابتدایی بودند. آنها فاقد زبانهای برنامه نویسی سطح بالا بوده و اکثرا برنامه نویسی ها بوسیله کد اسمبلی انجام می شد. به طور کلی سیستم های ترجمه ماشینی مستقیم نسل اول با مرحله ای شبیه آنچه که ما امروزه تجزیه صرفی می شناسیم آغاز می شد که در این مرحله تشخیص پایانه ها و تبدیل و تقلیل اشکال صرفی [1] مختلف به اشکال ستاکی غیر صرفی [2] انجام شده و نتیجه به عنوان ورودی برنامه جستجوگر در واژه نامه دو زبانه عمل می کرد. هیچگونه تجزیه ای از ساخت نحوی و یا روابط معنایی وجود نداشت. به عبارت دیگر تشخیص واژگانی تنها بر اساس تجزیه صرفی بوده و مستقیماً به واژه نامه دو زبانه منتهی می شد که در آنجا معادل های واژگانی زبان مقصد مشخص می گشت. البته برخی قواعد بازنویسی برای دستیابی به برونداد قابل قبول تر وجود داشت (مثلاً قواعدی که صفتها یا اجزای فعل را

[1] - inflected
[2] - uninflected

جابجا می کردند.) و پس از آن بلافاصله متن زبان مقصد تولید می شد. روش مستقیم را به طور خلاصه می توان به صورت شکل 1 نمایش داد.



شکل 1. سیستم ترجمه ماشینی مستقیم

محدویتهای بسیار زیاد این روش واضح است. این روش به عنوان ترجمه واژه به واژه با برخی انطباقات ترتیب واژه ای موضعی[1] شناخته می شود. کیفیت ترجمه آن همانند ترجمه شخصی است که دارای واژه نامه دو زبانه بسیار ابتدایی وساده ترین نوع دانش دستور زبان مقصد بوده و اشتباهات ترجمه ای مکرری در سطح واژگانی داشته و ساختهای نحوی بسیار نا مناسب در زبان مقصد که منعکس کننده ساختهای نحوی زبان مبدأ است را تولید می کند (10).

روش مستقیم روشی یکطرفه است یعنی اگر زبان مقصد بخواهد دوباره به زبان مبدأ ترجمه شود یک مبدل متفاوت دیگری باید مورد استفاده قرار گیرد. طبیعی است که برای این روش معایبی هم می توان بر شمرد. اولاً بدون یک تجزیه دستوری پیچیده تر ترتیب واژه ها در جمله زبان مقصد اغلب اشتباه خواهد بود. علاوه بر این در صورت وجود ابهام واژگانی در متن مبدأ، ترجمه نا درست حاصل خواهد شد. تجزیه رابطه بین قسمتهای مختلف جمله اغلب وجود ندارد و بنابراین منجر به ترجمه ضعیف یا نا مفهوم خواهد شد. روشهای جستجوی دو زبانه خام اغلب منجر به اشتباهاتی مانند ترجمه اشتباه معروف روسی " می تر بوام میرا" (ما جهان را تقاضا می کنیم) به جای "ما صلح می خواهیم" و بسیاری دیگر از اشتباهات می شود. روشهای مشابه که از قوانین ترتیب واژه ها نیز پیروی می کنند اگر چه ممکن است ترتیب اسم و صفت را رعایت کنند تا عبارت فرانسوی un chat noir (یک گربه

[1]-Local

سیاه) را درست ترجمه نماید اما زمانی که با عبارت " یک گربه خیلی سیاه " روبرو می شود ترجمه نادرست un tres chat noir را به جای un chat tres noir ارائه می نماید. بنابراین روشهای مستقیم در مواجه با ابهام واژگانی و حتی تفاوتهای نسبتاً ساده در نحو با شکست مواجه می شوند.

سادگی زبانشناختی و کامپیوتری این روش به سرعت شناخته شد. از نقطه نظر زبانشناختی آنچه که این روش فاقد آن بود تجزیه ساخت داخلی متن مبدأ به ویژه روابط دستوری بین قسمتهای اصلی جمله بود. فقدان پیچیدگی های کامپیوتری هم عمدتاً انعکاسی از ابتدایی بودن علم کامپیوتر در آن زمان بود (10).

2-3-1- سیستم های غیر مستقیم

شکست سیستم های ترجمه ماشینی نسل اول منجر به ظهور روشهای زبانشناختی پیشرفته تری برای ترجمه گشت. در واقع این روشها که به روشهای غیر مستقیم معروفند متن زبان مبدأ را با واسطه یک بازنمود میانجی به زبان مقصد تبدیل می کنند. این بازنمود میانجی نمایشی از معناست که می تواند اساس تولید متن مقصد را تشکیل دهد (5). به گفته هارولد سومرز[1] این ممکن است به عنوان بازنمودی از معنای متن باشد یا حتی با ابهام کمتر، نمایشی از ساخت نحوی متن باشد (11).

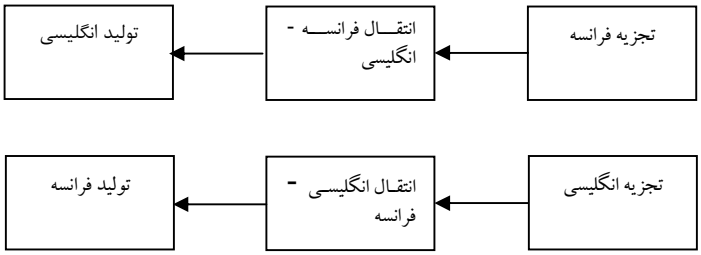
دو گونه اصلی روش غیرمستقیم عبارتند از : روش انتقالی و روش میان زبان،که در ذیل به توضیح آنها می پردازیم.

۱-۲-۳-۱- روش انتقالی

اولین گونه از روش غیرمستقیم روش انتقالی نامیده می شود. در واقع تمام سیستم های ترجمه درگیر نوعی انتقال هستند یعنی تبدیل متن یا نمایش مبدأ به متن یا نمایش مقصد. اصطلاح "روش انتقالی" به سیستم هایی اطلاق می شود که بخش های دو زبانی را بین بازنمودهای میانجی قرار می دهند اما این بازنمود ها وابسته به زبان هستند. نتیجه تجزیه، بازنمود مجردی از متن مبدأ است و درونداد، تولید بازنمود مجردی از متن مقصد می باشد. عملکرد بخش های انتقالی دو زبانه تبدیل بازنمودهای (میانجی) زبان مبدأ به بازنمود های (میانجی) زبان مقصد می باشد.

[1] - Harold Somers

شکل 2 این عملکرد را نشان میدهد. از آنجایی که این بازنمود ها بخش های مجزا (تجزیه، انتقال، تولید) را به هم ربط می دهند به آنها بازنمود های اتصالی نیز گفته می شود.



شکل 2. مدل انتقالی با دو جفت زبان

در روش انتقالی نحو زبان مبدأ - مشابه آنچه که قبلاً (به طور سنتی) بوسیله زبانشناسان انجام میگرفت - تجزیه می شود. سپس درخت نحوی به وجود آمده تبدیل به ساختی می‌شود که مناسب با زبان مقصد باشد. در نتیجه ترجمه ای در قالب ساخت زبان مقصد تولید می شود. یکی از معایب این روش نسبت به روش میان زبان این است که اضافه نمودن یک زبان جدید به سیستم نه تنها مستلزم دو بخش تجزیه و تولید است بلکه اضافه کردن بخش های انتقالی جدیدی را نیز می طلبد که تعداد آنها بسته به تعداد زبانهای موجود در سیستم تفاوت می کند. مثلاً در یک سیستم دو زبانه، زبان سوم نیاز به چهار بخش انتقالی جدید دارد. اما از طرف دیگر پیچیدگی نسبی بخش های تجزیه و تولید در یک سیستم انتقالی کمتر از روش میان زبان است چرا که بازنمود های میانجی درگیر در سیستم در روش انتقالی وابسته به زبان بوده و مشخصه هایی از متون مبدأ و مقصد می‌باشند.

اکثر سیستم های مبتنی بر انتقال، ترجمه را به کمک بازنمود میانجی انجام می دهند. اما نکته مهم در اینجا این است که این بازنمود ها وابسته به زبان هستند یعنی در آنها واژه های نمونه زبان مبدأ و مقصد مشاهده می شود. همچنین آنها ساخت زبانهای مبدأ و مقصد درگیر در ترجمه را، خواه به طور سطحی (صوری) خواه با عمق بیشتر، منعکس می‌نمایند، و این مشخصه یکی از تفاوتهای عمده روش انتقالی با روش میان زبان (که در بخش بعدی به معرفی آن می پردازیم) است. بدین ترتیب که سیستم های میان زبان سعی در ارائه بازنمود های میانجی دارند که مستقل از زبان باشند چه از نظر واژگان و چه از نظر ساخت (5).

در سیستم های مبتنی بر انتقال اساساً دو نوع مشکل وجود دارد، مشکلات واژگانی و مشکلات ساختی که در سطور زیر به بررسی آنها می پردازیم.

۱-۱-۲-۳-۱- انتقال واژگانی

انتقال واژگانی به معنای جایگزینی واحد واژگانی زبان مبدأ به وسیله واژه زبان مقصد است. واضح است که اگر تنها یک معادل برای واژه زبان مبدأ در زبان مقصد وجود داشته باشد مشکل خاصی به جز در مورد تفاوتهایی در ساخت نحوی وجود نخواهد داشت. لیکن این حالت تنها در حوضه ترجمه فنی که چنین تناظر واژگانی یک به یک وجود دارد مصداق پیدا می نماید. البته انتقال واژگانی در مورد ترجمه های تناظر چند به یک نیز مشکل زا نیست (مثلاً واژه های انگلیسیmilk و lion که هر دو در فارسی به معنای شیر هستند)، بلکه مشکل از انتقال واژگانی تناظرهای یک به چند ناشی می شود. (مثلاً واژه انگلیسی term که در فارسی به معناهای دوره، اصطلاح و واژه می باشد). در برخی موارد این امکان برای سیستم ترجمه ماشینی وجود دارد تا ترجمه های خاصی را اغماض کند و این به خاطر مهجور شدگی آن واژه ها یا به ندرت مورد استفاده قرار گرفتن آنهاست و یا می تواند به این دلیل باشد که واژه خاصی تنها در برخی از گونه های زبانی خاص دیده می شود. استفاده از واژه آکبند به جای واژه فارسی ناسوده در ترجمه ها مثالی از این مورد است. حتی واژه کامپیوتر به جای رایانه در بسیاری از موارد می تواند مورد استفاده قرار گیرد چرا که استفاده از واژه کامپیوتر برای فارسی زبانان راحت تر و جا افتاده تر از واژه رایانه است.

جایی که چند ترجمه مختلف برای یک واژه وجود داشته باشد طبیعی است که انتقال واژگانی نیاز به بررسی محیط یا بافت اطراف آن واژه پیدا می‌کند. مثلاً واژه انگلیسی know بسته به مفعول خود می تواند به شناختن یا دانستن یا معادل های دیگر در فارسی ترجمه شود. بدین ترتیب که ابتدا بایستی مفعول این فعل شناسایی شده و مشخصه های معنایی آن از قبیل جاندار یا بی جان بودن آن نیز تعریف گردد، سپس با استفاده از این بافت ترجمه

مناسب این فعل انجام شود. مثال دیگر می تواند واژه انگلیسی morphology باشد که بسته به موضوع متنی که واژه در آن یافت می شود ترجمه می شود. در متون پزشکی این واژه به معنای ریخت شناسی است در حالیکه در متون مربوط به زبانشناسی و دستور زبان از آن با نام صرف یاد می کنیم. یافتن ترجمه مناسب برای چنین واژه هایی (که با توجه به موضوع متن تعیین می شوند) برای کامپیوتر بسیار مشکل تر از مورد نخست به نظر می رسد هر چند برای ترجمه انسانی تشخیص موضوع متن کار ساده ایست (12).

2-1-2-3-1- انتقال ساختاری

انتقال ساختاری زمانی ضروری به نظر می رسد که ساختاری که زبان مبدأ دارد برای زبان مقصد نامناسب باشد. برخی از مشکلات انتقال ساختاری نسبت به برخی دیگر راحت تر برطرف می شوند به طوریکه می توان برای انتقال آنها از زبان مبدأ به زبان مقصد قواعدی تعریف کرد و به صورت فرمول در ترجمه از آنها استفاده نمود. برای مثال به جمله انگلیسی زیر و ترجمه فارسی آن توجه کنید.

- Ali sent his friend a present.

- علی برای دوستش یک هدیه فرستاد.

با توجه به ساختارSOV در زبان فارسی وSVO در زبان انگلیسی تشخیص فاعل، فعل و مفعول در جمله زبان مبدأ و تبدیل آن به ساختار مناسب در زبان مقصد چندان مشکل به نظر نمی رسد. لیکن مشکل اصلی از آنجایی ناشی می شود که تفاوتهای ساختاری زبانهای مبدأ و مقصد عمیق تر باشد. تبدیل جمله انگلیسی مجهول به جمله فارسی معلوم با همان مفهوم به دلیل قابل قبول نبودن بعضی ساختارهای مجهول در زبان فارسی می تواند مثال خوبی در این زمینه باشد (برای اطلاعات بیشتر رجوع شود به13).

هم روش مستقیم و هم روش انتقالی در ترجمه ماشینی تکیه بر برنامه ریزی جداگانه ای برای هر جفت زبان درگیر که در جهت مشخصی ترجمه می شوند دارند. این مشکل با اضافه شدن جفت زبانهای بیشتری به سرعت

بیشتر می شود. روش انتقالی در n زبان درگیر در سیستم ترجمه از n2 واحد انتقالی[1] ، n واحد تجزیه، و n واحد ترکیب استفاده می نماید. بنابراین یکی از معایب آن تعداد بسیار بالای قوانینی است که برای انجام آن استفاده می‌شود. مسئله دیگر حالت تحت اللفظی بودن ترجمه حاصل است. اگر چه روش انتقالی سبک واژه به واژه روش مستقیم را دنبال نمی کند اما پردازش معنایی کمی از متن مورد ترجمه را انجام می دهد. بنابراین با احتمال زیاد نمی تواند از عهده زبان عامیانه و مصطلح بر آید (14).

۲-۲-۳-۱- روش زبان میانجی[2]

اخیراً محققان کوشش کرده اند تا معماری ای ارائه نمایند که در آن متون زبان مبدأ بوسیله نوعی بازنمود معنایی مستقل از زبان، زیربنایی و خنثی (بی طرف) به زبان مقصد تبدیل شود (برای اطلاعات بیشتر نگاه کنید به آرنولد و دیگران 1994). این کار از نظر تئوری تأثیر زبان مبدأ بر روی متن برونداد (در زبان مقصد) را از بین می برد. در یک سیستم میان زبان صرف هیچ نوع قانون انتقالی وجود ندارد چرا که بازنمود های میانجی بایستی در همه زبانهای درگیردر سیستم مشترک باشند.

جدا بودن فرآیندها ی تجزیه و تولید در این روش یکی از معایب آن محسوب می شود. بدین مفهوم که این روش برای هر زبان مبدأ یک تجزیه گر و برای هر زبان مقصد یک تولید کننده نیاز دارد. تجزیه متن مبدأ نیاز به یک تجزیه معنایی عمیقی که خود نیز مستلزم داشتن دانش گسترده جهان بیرون است دارد. متأسفانه معنای واقعی یک جمله همیشه به دست نمی آید و علاوه بر این اگر هم یک متن به همان عمقی که تصور می شود تجزیه شود باز هم قسمت بیشتر سبک نویسنده زبان مبدأ القا نمی شود (15).

روش میان زبان مخصوصاً برای سیستم های چند زبانه کاربردی تر است و زبانهای مقصد هیچگونه تأثیری بر روی هیچ یک از فرآیندهای تجزیه ندارند. مزیت این روش در این است که اضافه کردن یک زبان جدید به سیستم مستلزم ایجاد تنها دو بخش جدید است: دستور زبان تجزیه و دستور زبان تولید، که در این حالت ترجمه از زبانهای دیگر به زبان جدید و از زبان جدید به زبانهای دیگر مقدور می باشد (5). لازم به ذکر است که روش میان

[1]- Transfer modules
[2] - interlingua

زبان توجه کمی به فقدان دانش جهان بیرون[1] که قبلاً به عنوان مانعی بر سر راه ترجمه ماشینی با کیفیت بالا تشخیص داده شده بود دارد.

اگر چه تنها سه روش اصلی در معماری ترجمه ماشینی وجود دارد که عبارتند از: روش مستقیم، روش انتقالی و روش میان زبان، شخص می تواند چندین بخش اطلاعاتی را اعمال نماید تا کارآیی سیستم ترجمه را افزایش دهد. چنین بخشهای اطلاعاتی شامل بخشهای مبتنی بر دانش، مبتنی بر آمار، مبتنی بر نمونه و غیره می باشند که در بخشهای بعدی به توصیف آنها خواهیم پرداخت.

3-3-1- روشهای مبتنی بر قانون[2]

در ماشین های ترجمه مبتنی بر قوانین ، پردازشی که صورت می گیرد ، تحلیل و نمایش معنی متن به زبان مبدا و تولید ترجمه معادل متن به زبان مقصد است . مسئله مورد بحث در اینجا روش بازنمایی است که باید از لحاظ ساختاری و لغوی کاملا واضح و صریح باشد. این باز نمایی در روشهای مبتنی بر قانون ، یا روشهای انتقالی و میان زبان هستند و یا روشهای مبتنی بر دانش. روشهای انتقالی و میان زبان همچنانکه که گفته شد بر این اساس هستند که موفقیت در ترجمه ماشینی عملی مستلزم تعریفی از یک سطح بازنمودی برای متون است که در عین حال که مجرد است به حد کافی صوری هم هست تا بتواند در مورد هر دو زبان درگیر در ترجمه قابل انطباق باشد.

رویکرد تبدیل شامل سه مرحله می باشد ، ابتدا پردازش روی نمایش بدست آمده در زبان مبدا و سپس تبدیل آن به زبان مقصد و در آخر ترکیب و بازنمایی به زبان مقصد .

رویکرد میان زبان شامل دو مرحله میباشد ،مرحله اول تحلیل روی نمایش یک زبان واسط و سپس بازنمایی و تبدیل آن به زبان مقصد.

3-3-1-4- روشهای مبتنی بر دانش[3]

تاکنون بحث عمدتاً برروی روشهای زبانشناختی مسائل انتقال در ترجمه بوده است، در حالیکه تفاوتهای

[1] - world knowledge
[2] - rule-based approaches
[3] - knowledge-based approaches

واژگانی ترجمه نیز به طور اجتناب ناپذیری انعکاسی از تفاوتهای زمینه های فرهنگی و تفاوتهایی در اختلافات تصوری از واقعیت یکسان می باشند. بنابراین برخی معتقدند که ترجمه بایستی بر اساس بازنمود های تصوری غیر زبانی یا به بیان دیگر بر اساس معنایی که از فرآیندهای درک متن مشتق می شود باشد. درک یک متن یعنی برقرار کردن ارتباط بین آنچه که در متن گفته می شود، (محتوای زبانی آن) و پدیده های (فعالیتها، وقایع، ماهیت) خارج از متن (واقعیت غیر زبانی). خوانندگان متون را با توجه به آنچه که از جهان واقعی یا جهانهایی که خود می توانند تصور کنند می دانند تفسیر می کنند. اگر فرض شود که مفاهیم و درک برای تمام سخنگویان زبانهای بشری مشترک و جهانی هستند، نتیجه این می شود که این بازنمود های تصوری بین زبانی هستند و اینکه می توانند به صورت بازنمود های میانجی در سیستم های ترجمه ماشینی عمل می نمایند و نیز اینکه با کمک پایه های دانش مناسب، متون می توانند با این بازنمود های بین زبانی تجزیه شده و نیز از چنین بازنمود هایی تولید شوند. پس می توان گفت که استفاده از دانش جهان واقعی در سیستم های ترجمه ماشینی میان زبان کاربرد بیشتری دارد. یعنی به کمک دانش جهان واقعی نه تنها می توان ابهامات متن مبدأ را رفع نمود بلکه می توان بازنمود های مستقل از زبانی را به وجود آورد. مثلاً واژه انگلیسی wear دارای هشت واژه معادل در زبان ژاپنی با معانی ذیل می باشد:

Kitu (generic) , haoru (coat or jacket) , haku (shoes or trousers) , Kaburu (hat) , hameru (ring or gloves) , shimeru (belt or tie or scarf) , tsukeru (brooch or clip) , kakeru (glasses or necklace)

با توجه به این مثال پایه دانشی یک سیستم میان زبان باید قادر باشد تا شانزده (یا بیشتر) جنبه مختلف پوشیدن انواع لباس و اشیاء را از یکدیگر تشخیص دهد و این کار را مسلماً بر اساس دانشی که از انواع مختلف پوشیدنیها (مانند کلاه، کت، جوراب و غیره)، تفاوتهای نژادی، فرهنگی، اجتماعی (لباس مراسم، لباس کار، لباس بچه، لباس گرمسیری، لباس سردسیری، لباس بارانی و غیره)، و فعالیتهای مربوط به آنها (دوختن، خریدن، پوشیدن، درآوردن، تعمیر کردن و غیره) دارد باید انجام دهد.

اصطلاح ترجمه ماشینی مبتنی بر دانش به سیستم مبتنی بر قاعده ای(قانونی) اطلاق می شود که قادر به نشان

دادن دانش کاربردی و معنایی وسیعی در یک حوزه باشد که شامل توانایی تعقل در مورد مفاهیم آن حوزه است. مشخصه عمده روش مبتنی بر دانش تأکید زیاد آن بر روی درک کامل (از نظر کارکردی) معنای متن مبدأ قبل از ترجمه آن به متن مقصد است. اساساً قضیه به این شکل است که ترجمه با کیفیت بالا نیاز به درک عمیق متن دارد و ارائه این مدل حوزه ای[1] برای این درک عمیق ضروری به نظر می رسد. یکی از نکات مهمی که در این رابطه بایستی به آن توجه نمود این است که پس ویرایش امری است وقت گیر و پر هزینه و بنابراین تلاشهایی برای تولید ترجمه با کیفیت بالا جریان طولانی ای را در بر خواهد گرفت. پس روش مبتنی بر دانش می تواند روشی باشد که درجاتی از درک متن را بر اساس حوزه تعریف شده ای از دانش در بر می گیرد. مدل حوزه ای که در این روش از آن استفاده می شود فقط تمام مفاهیم مربوط به یک حوزه خاص را نمایش می دهد و هیچگونه تعقل یا استنباط دیگری در مورد مفاهیم حوزه های دیگری که تعریف نشده اند بدست نمی دهد. نقش اساسی مدل حوزه ای رفع ابهام کامل از متن است و قسمت مهمی از آن مشخص نمودن محدودیتها یا قیدهایی که هر مفهوم در آن حوزه از نظر مفعولی که می گیرد و استدلالات آن دارد است (مثلاً اینکه تنها موجودات زنده می میرند، تنها انسانها فکر می کنند و ...).

به محض وارد کردن دانش تعریف شده ای در ترجمه ماشینی با کیفیت بالا به وسیله درک متن شخص ترغیب می شود تا منابع دانش را مرتب اضافه نماید. کاملاً روشن است که حل پیش مرجع[2] و حل دیگر ابهامات مرجعی نیاز به مراجعه به سطح ساختاری ای مافوق نحو و معنا شناسی جمله دارد. همینطور هم به دلایل سبکی برای افزایش پیوستگی متن، شخص نیاز به داشتن نمایش هایی از ساخت (ساختار) پاراگراف دارد. رسیدن به یک ترجمه با کیفیت واقعاً بالا به ویژه با انواع مختلف متون، ممکن است نیاز به شیوه مواجهه با استعاره[3]، کنایه[4]، کنش غیرمستقیم سخن[5]، نگرش سخنگو یا شنونده و غیره داشته باشد. در طول چند سال اخیر گروههای مختلفی در قسمتهای مختلف دنیا شروع به آزمایشاتی با الگو هایی که با دانش صریح یا اجزای قاعده ای که انواع بسیار متنوعی از اطلاعات را شامل می شدند نمودند. تمامی این روشها به هر شکلی که باشند ترجمه ماشینی مبتنی بر

1 - domain model
2 - anaphora resolution
3 - metaphor
4 - metonymy
5 - indirect speech act

دانش نامیده می شوند (17).

سیستم های مبتنی بر دانش ترجمه هایی با کیفیت بالا عرضه میدارند. با وجود این به خاطر نیاز به حجم زیاد دانشی که بتوان جملات زبانهای مختلف را به طور صحیح نمایش داد تولید این سیستم ها بسیار گران حاصل می شود. KANT و KANTOO سیستم های مبتنی بر دانشی در دانشگاه Carnegie meuon هستند که برای تولید اسناد چند زبانه (فرانسوی، اسپانیایی و انگلیسی) به کار می روند. این سیستم های میان زبانی شامل ابزار نگهداری واژگان و دانش هستند و ترجمه ها را تا حد امکان از نظر صحت بیمه نمایند.

پیچیدگی عملی این روش آن را به ناچار محدود به حوضه های بسیار محدودی با کاربردهای نسبتاً کم می نماید. همچنین لازم به ذکر است که سیستم های ترجمه ماشینی مبتنی بر دانش تاکنون در زمینه ابهام زدایی متون مبدأ متمرکز بوده اند و درگیر ایجاد بازنمود های مستقل از زبان جهانی برای فراهم کردن اطلاعات جهت تولید متن به زبان دیگری نیستند (5).

5-3-1- روشهای مبتنی بر پیکره (زبانی)[1]

همچنان که دیدیم استفاده از روش مبتنی بر قاعده و مبتنی بر دانش در ترجمه ماشینی مشکلات عدیده ای را می آفریند. هیچ زبان طبیعی ای تماماً با مجموعه ای از قواعد مطابقت نمی کند و ابهام های زیادی به وجود می آید. علاوه بر این ترجمه ضرورتاً با بیش از یک زبان درگیر است که در اینصورت پیچیدگی افزایش می یابد. به این دلایل زبان شناسان شروع به جستجوی روشهای دیگری برای سیستم های ترجمه ماشینی نموده اند. برخی از این روشها از پیکره یا بدنه متن استفاده می کند که به عنوان دریاچه ای از نمونه های آماده ای از کاربرد زبان عمل می کند. این روشها که به روشهای تجربی[2] معروفند اغلب از فنون آماری سطح پائینی استفاده می نمایند که یا به طور مستقیم به خود متون اعمال می شود و یا در تجزیه سطحی آنها به کار می رود. دلیل اینکه به این روشها

[1] - corpus-based approaches
[2] - empirical method

روشهای تجربی گفته می شود این است که در چنین روشهایی دانش زبانی ای که سیستم مورد استفاده قرار می دهد (از هر نوع آن که باشد) به طور تجربی از آزمایش متون واقعی استخراج می شود و نه از دانشی که به وسیله تعقل زبانشناسان به دست می آید.

این روشهای جدید هنوز مرحله کودکی شان را سپری می کنند. و این تا حدی به خاطر سخت افزار مورد نیاز برای ذخیره و دستیابی به حجم بسیار زیادی از داده های مورد نیاز در کار است که اخیراً شناسایی، فراهم و قابل دسترسی شده است. دلیل دیگر این است که ترجیح دادن دانش زبانی بر توانش زبانی که بوسیله چامسکی انجام شد و به عنوان اساسی برای تجزیه زبانی قرار گرفت مبین این نکته است که چنین کار تجربی ای به بهای توجه به روشهای نظری و صوری تر مورد غفلت واقع گشته است (14).

ماشین های ترجمه مبتنی بر پیکره به معنی هر نوع از ماشین های ترجمه مستقیم یا Direct می باشدکه بر مبنای پیکره های زبانی موازی دو زبانه به صورت اتوماتیک عمل می کنند. منظور از ماشین های ترجمه مستقیم به هر تکنیکی که ترجمه ، در آنها مستقیما از کلمات زبان مبدا به کلمات زبان مقصد، بدون استفاده از بازنمایی در یک ساختار زبانی واسط صورت می گیرد ، می باشد (16) .

در این نوع از ماشین های ترجمه از منبعی از اطلاعات شامل لغات ، ساختار گرامری و آمار و احتمالات که پیکره نام دارد جهت تحلیل، تبدیل و ترجمه استفاده می گردد .

در سطور زیر به معرفی برخی از این روشها می پردازیم:

1-3-5-1- روشهای مبتنی بر نمونه[1]

پیشرفتهایی در زمینه فناوری کامپیوتر که دسترسی سریعتر به حافظه ها و ذخایر اطلاعاتی بیشتر را فراهم کرده است جستجوی روشهایی مبتنی بر دسترسی به پیکره های عظیمی از متون به زبانهای مبدأ و مقصد را در رأس کار محققان قرار داده است. در این روش ترجمه اغلب با یافتن یا بازخوانی نمونه های مشابه و نیز با پیدا کردن و به خاطر سپاری اینکه چگونه عبارتی خاص از زبان مبدأ یا چیزی شبیه به آن قبلاً ترجمه شده است سروکار دارد. این

[1] - example-based approaches

روش که به " ترجمه به کمک قیاس[1]" نیز معروف است روشی است که در آن ترجمه به وسیله مقایسه برونداد با پیکره ای از نمونه های ترجمه شده و استخراج نزدیکترین همانندها و استفاده از آنها به عنوان مدلی برای متن هدف (متن مقصد) انجام می شود (11).

پیشنهاد روشهای مبتنی بر نمونه به طور کلی به عنوان شقی دیگر در مقابل روشهای مبتنی بر دانش مطرح گردید و نیز مکملی برای روشهای تجزیه ای، انتقالی و تولیدی مبتنی بر قاعده به حساب می آید. بانکهای داده ای زبانی نمونه ها می توانند از تجزیه ساختاری پیکره عظیمی از متون مبدأ و ترجمه آنها به زبان مقصد که بوسیله مترجمان انسانی انجام شده است استخراج شوند. نتیجه مجموعه های دو زبانه ای از عبارات به زبان مقصد و ترجمه آنها به زبان مبدأ در بانک داده ای است (5). بانک داده ای متن از پیش تجزیه شده در حافظه ترجمه[2] ذخیره می شود. روشهای زیادی برای ذخیره کردن جفتهای ترجمه ای در حافظه ترجمه وجود دارد. متن می تواند به صورت یک درخت تجزیه نحوی[3] کامل یا به صورت جملات کامل و یا حتی عبارات ذخیره شود.

روش مبتنی بر نمونه به ویژه برای ترجمه عبارات اسمی پیچیده کاربرد بیشتری دارد. واژه های بسیاری در انگلیسی وجود دارد که دارای چندین معادل فارسی است. یافتن عباراتی که چنین واژهایی در آنها قرار دارد و معادل فارسی این عبارات در پیکره زبان فارسی می تواند کمک خوبی در جهت ترجمه با ثبات چنین عبارات پیچیده ای باشد. حتی وجود مثالهای مشابه به عبارات اسمی در پیکره و محاسبه بعد (اندازه) شباهت احتمالی آنها می تواند ترجمه چنین عبارات مشکل زایی را آسان تر نماید. ترجمه چنین عباراتی با استفاده از نمونه های موجود در پیکره زبانی به مراتب راحت تر از متوسل شدن به قواعد مبتنی بر مقوله ها و ساختارهای دستوری و غیره می باشد. روشهای مبتنی بر نمونه همچنین می تواند در ترجمه جمله هایی که از نظر ساختاری مشابه جملات از قبل ترجمه شده هستند بکار رود. در این حالت نیز بعد شباهت نیاز است اگر چه این بار محاسبه این بعد بر اساس توزیع عناصر کلیدی مانند واژه های دستوری یا بر اساس زنجیره های خاصی از مقوله های دستوری یا ترکیبی از اینهاست (5). در توصیف خصوصیات این نوع ترجمه سومرز سه ملاک خاص را برای آن در نظر می گیرد.

الف) ترجمه ماشینی مبتنی بر نمونه از پیکره دو زبانه استفاده می کند.

1 - translation by analogy
2 - translation memory
3 - Parse tree

ب) ترجمه ماشینی مبتنی بر نمونه از پیکره دو زبانه به عنوان پایه دانش اصلی خود استفاده می نماید.

ج) ترجمه ماشینی مبتنی بر نمونه از پیکره دو زبانه به عنوان پایه دانش اصلی خود در زمان اجرای برنامه استفاده می نماید (18).

سومرز اشاره می نماید که دو ملاک نخست بسیار کلی هستند ولی ملاک سوم به اندازه کافی تعریف را محدود به ترجمه ماشینی مبتنی بر نمونه می نماید به طوریکه مثلاً آن را از ترجمه ماشینی مبتنی بر آمار که دو ملاک نخست نیز در آن صدق می کند اما احتمالات مستخرج از پیکره از قبل محاسبه می شوند متمایز می سازد (19).

این نکته نیز باید خاطر نشان شود که درجه عملی بودن این روش قطعاً بستگی به مجموعه داده های خوب دارد. با این وجود یکی از مزیتهای این روش این است که همچنان که مجموعه نمونه ها کامل تر می شود کیفیت ترجمه نیز بدون نیاز به بروز کردن و اصلاح توصیفات واژگانی و دستوری به طور فزاینده ای بهبود می نماید. بعلاوه این روش می تواند بسیار مؤثر واقع شود چرا که در بهترین حالت هیچگونه نیازی به اجرای کاربرد قوانین پیچیده نیست و تنها کاری که بایستی انجام شود پیدا کردن نمونه مناسب و (گاهی اوقات) محاسبه فاصله است. اما گاهی اوقات پیچیدگی هایی هم وجود دارد. مثلاً یک نوع مشکل زمانی به وجود می آید که یک واحد تعدادی نمونه متفاوت دارد که هر کدام از اینها با یک قسمت از زنجیره مطابقت می نماید ولی قسمتهایی که با آن واحد مطابقت می کنند همپوشی داشته و کل زنجیره را نمی پوشانند. در چنین حالاتی محاسبه بهترین تطابق نیاز به بررسی تعداد زیادی احتمالات دارد (17). و نکته آخر اینکه روش مبتنی بر نمونه می تواند در هر یک از مدلهای اصلی: روش مستقیم، انتقالی یا میان زبان گنجانده شود. برخلاف روش مبتنی بر دانش که وابسته به تجزیه معنایی با درجه بالایی از تجرد است، این روش با تجزیه های روساختی نسبتاً سطحی و روشهای انتقال واژگانی ساده کار می کند، یعنی این روش می تواند به هر سطحی از انتقال، از تجزیه صرفی گرفته تا میان زبانی، اعمال شود (10).

2-5-3-1- ترجمه ماشینی آماری

طی چند سال اخیر توجه زیادی به استفاده از روشهای آماری در پردازش زبان طبیعی شده است. تجزیه داده

های آماری برای جمع آوری خودکار دانش لازم برای ترجمه ماشینی از متون دو زبانه متناظر[1] مورد استفاده قرار گرفته است. از نظر ترجمه ماشینی اصطلاح روشهای آماری به کاربرد فنون مبتنی بر آمار یا احتمالات در بخشهایی از کار ترجمه ماشینی (مانند ابهام زدایی معنایی واژه) اطلاق می شود. در سطور زیر سعی می شود تا روش مبتنی بر آمار محض در ترجمه ماشینی توصیف شود. این روش که برای اولین بار به وسیله براون معرفی شد (17) در بازشناسی گفتار[2] بسیار موفق بوده است که اگر چه نیاز به پیچیدگی های آماری زیادی دارد، اما تصور کلی آن کاملاً ساده به نظر می رسد. دو مفهوم کلیدی در این رابطه عبارتند از مدل زبانی[3] و مدل ترجمه[4]. مدل زبانی احتمالات زنجیره واژه ای (در حقیقت جملات) را به ما می دهد که این احتمالات به وسیله (Pr (S (برای یک جمله زبان مبدأ) و (Pr (T (برای هر جمله زبان مبدأ مفروض) نشان داده می شوند. درواقع (Pr (S احتمال وقوع زنجیره ای از واژه های زبان مبدأ و(Pr(T احتمال وقوع زنجیره ای از واژه های زبان مقصد می باشد. مدل ترجمه نیز احتمالاتی را به ما می دهد. (Pr (T| S که احتمال شرطی وقوع یک جمله زبان مقصد (T) در یک متن مقصد که ترجمه متنی است که شامل جمله زبان مبدأ (S) باشد است. حاصل این و احتمال وقوع S به تنهایی، یعنی (Pr (T|S) × Pr (S) احتمال وقوع جفت جملات مبدأ - مقصد که به صورت (Pr (T, S نوشته می شود را به دست میدهد. اما چگونه می شود احتمال وقوع یک زنجیره مبدأ (یا جمله مبدأ) یعنی (Pr (S را محاسبه نمود؟ این احتمال را می شود به صورت احتمال واژه نخست ضرب در احتمالات شرطی واژه های متعاقب آن تجزیه نمود، یعنی:

$$Pr(S3 \mid S1, S2), etc \ldots \times Pr(S2 \mid S1) \times Pr(S1)$$

در واقع احتمال شرطی (Pr (S2 | S1) احتمال وقوع S2 با دانستن اینکه S1 رخ داده است می باشد. مثلاً احتمال اینکه am و are و در یک متن رخ دهند تقریباً یکسان است. اما احتمال اینکه am بعد از I رخ دهد بسیار بالاست در حالیکه احتمال اینکه are بعد از I رخ دهد بسیار پائین است. برای اینکه بتوان همه چیز را به صورت قابل کنترل و محدود در اختیار داشت معمول چنین است که تنها یک یا دو واژه قبل را در محاسبه این احتمالات

[1] - parallel bilingual texts
[2] - speech recognition
[3] - language model
[4] - translation model

شرطی در نظر می گیرند (که به ترتیب به مدلهای گرام دوگانه و گرام سه گانه مشهورند). به منظور محاسبه این احتمالات زبان مبدأ (که مدل زبان مبدأ به وسیله تخمین پارامترها را تولید می کنند) به حجم بسیار زیادی داده یک زبانه نیاز می باشد چرا که اعتبار، کار آیی یا صحت مدل عمدتاً به حجم این پیکره داده ای بستگی خواهد داشت. جای دیگری که به حجم وسیعی از داده نیاز می باشد مشخص کردن پارامترهای مدل ترجمه می باشد که این نیاز به یک پیکره متناظر دو زبانه خواهد داشت. البته منابع نسبتاً کمی از چنین پیکره ای وجود دارد اگر چه گروه تحقیقی در IBM که عمدتاً مسئول توسعه چنین روشی بوده است موفق به تهیه سه میلیون جفت جملاتی از مذاکرات رسمی کانادایی[1] (فرانسوی - انگلیسی) که درآن هر جمله مبدأ متناظر با ترجمه اش در زبان مقصد می باشد به صورت یک پیکره متناظر از جملات شد (17).

پیکره های متناظر بر دو نوعند: پیکره های متناظر واژه ای و پیکره های متناظر جمله ای.

1-2-5-3-1- پیکره متناظر واژه ای

به مثال زیر توجه کنید :

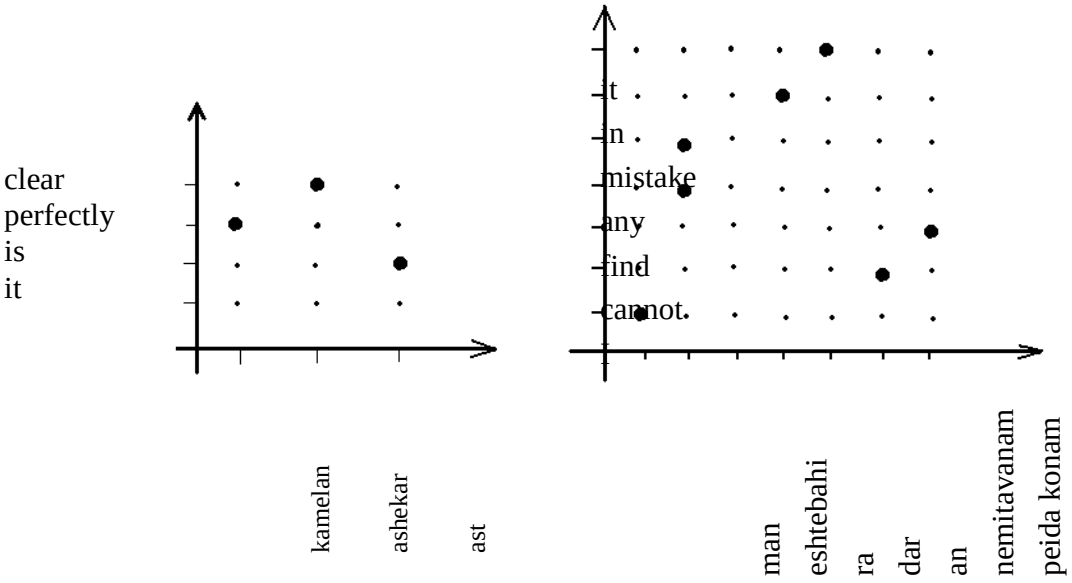
1 2 3 4
a) It is perfectly clear.

1 2 3 4 5 6 7
b) I cannot find any mistake in it.
a`) Ka:melan(3) a:sheka:r(4) ast(2).
b`) man(1) eshteba:hi(4,5) ra: dar(6) a:n(7) nemitava:nam(2) Peida: konam(3)

در این مثال زبان مقصد، انگلیسی درنظر گرفته شده است. شماره های جلوی واژه های زبان مبدأ متناظر با ترتیب قرار گرفتن واژه ها در جمله متناظر زبان مقصد می باشند. اگر متناظر یک واژه در زبان مقصد وجود نداشته باشد طبیعتاً هیچ شماره ای هم در جلوی آن واژه زبان مبدأ قرار نمی گیرد (مثل واژه :ra در جمله فارسی)، و چنانچه بیش از یک واژه برای واژه متناظر مقصد یا مبدأ وجود داشته باشد آن نیز مشخص می شود (مثل واژه

[1] - Canadian Hansard

eshteba:hi در جمله فارسی) سس پیکره های متناظر را به شکل دیگری هم (شکل 3) می توان نشان داد.



شکل 3. تناظر واژه به واژه

این نمودارها در بعضی از جفت زبانها تقریباً یکنواخت دیده می شود اما در زبانهایی مانند انگلیسی و فارسی که عدم تناظر در برخی واژه ها و هم در ترتیب واژه ها مشاهده می شود یکنواخت نیست.

این روشها بسیار امید بخش هستند از آن جهت که می توانند راه میان بری از جمع آوری دانشی باشد که همه کاربردهای پردازش زبان طبیعی را در بر می گیرد. اما متأسفانه روشهای ترجمه ماشینی آماری در جامعه ترجمه ماشینی به طور گسترده مورد استفاده قرار نگرفته است و این تا حدی به این دلیل است که ریاضیات درگیر در این روش برای محققان زبانشناسی کامپیوتری کاملاً و به طور دقیق جا نیفتاده است. دلیل دیگر هم این است که نرم افزارها و مجموعه داده های لازم در این روش به طور عموم در دسترس نیست. برای ایجاد زیر ساخت های نرم افزاری لازم جهت آزمایش و تحقیق در این زمینه کارهای زیادی بایستی انجام شود (12).

2-2-5-3-1- پیکره متناظر جمله ای

نوع دوم پیکره های متناظر بر اساس جمله هستند که در آنها جمله متناظر هر جمله زبان مقصد به زبان مبدأ داده شده است.

روش مبتنی بر آمار اگر چه مزیتهایی دارد از جمله اینکه در آن نیاز به دانش زبانی به حداقل می رسد اما از نظر کاربرد دارای محدودیتهایی است چرا که بکارگیری این روش مستلزم فراهم کردن پیکره های حجیم یک زبانه و دو زبانه است که همانطور که می دانیم اکثر زبانها فاقد چنین پیکره هایی می باشند (12).

3-2-5-3-1- مراحل ساخت یک پیکره در ماشین ترجمه آماری

- جمع آوری متن ها یا لغات از هر دو زبان که این متون در واقع ترجمه معادل یکدیگر باشند و همچنین این متون باید از ارزش ادبی و کیفیت خوبی در ترجمه برخوردار باشند .
- پیش پردازش[1] و علامت گذاری متون
- هم ترازی[2] و تعیین پارامترهای مدل و ساخت و آموزش مدل ها
- ارزیابی پیکره

الف- پیش پردازش و علامت گذاری متون

پیکره مجموعه ای متن به دو جفت زبان متفاوت می باشند این متن ها جمع آوری شده و معمولا در مورد حوزه خاصی می باشند ولی ممکن است فرمت یکسانی نداشته و قابل خواندن توسط ماشین نباشند و همچنین دسته بندی نشده باشند . پیش پردازش شامل دسته بندی و یکسان سازی فرمت می باشد و سپس همه متن ها در کنار هم قرار می گیرند به طور مثال متن هایی که از آرشیو روزنامه بدست می آیند به فرمت HTML می باشند و شامل طبقه های[3] مختلف هستند ابتدا این متون به فرمت متن[1] تبدیل می گردند و در دسته های مختلف تقسیم می شوند و

[1] Pre Processing
[2] Alignment
[3] Category

سپس همه در یک فایل بزرگ ذخیره می شوند . سپس روی متن های زبان مبدا و مقصد تفسیر[2]باید انجام گیرد و متن ها باید بخش بندی[3] شوند یعنی بخش های مختلف مانند پاراگراف ، جمله و کلمه باید مشخص گردند و پس از بخش بندی ، برچسب گذاری[4] باید روی جملات زبان مقصد و مبدا صورت گیرد و یا به بیان دیگر متن ها باید علامت گذاری[5] شوند .

معمول ترین زبان نشانه گذاری SGML [6] می باشد که امکان تعریف قاعده[7] را برای متون تامین می نماید (48).

چند مثال متن علامت گذاری شده به صورت زیر قابل مشاهده می باشند :

```
<p><s>He left.</s>
<s>.او آمد</s></p>
<utt speak=”Ali” Date=”1385-07-01”>.باید او باشد</utt>
```

ب- هم ترازی

هم ترازی به معنی مشخص نمودن ارتباط بین واحد های دو زبان (واحد می تواند کلمات ، عبارات یا زیر کلمه باشد) ، در یک پیکره زبانی موازی است و همچنین مشخص نمودن اینکه چه کلماتی از هر زبان در یک موقعیت از ورودی ، به هم مرتبط هستند ، می باشد .

در این زمینه رویکرد های مختلفی ارائه شده است ولی دو رویکرد که قابل توجه می باشند به شرح زیر می باشند :

---

[1] Text
[2] Annotation
[3] Segment
[4] Tag
[5] Mark
[6] Standard Generalized Markup Language
[7] Gramer

رویکرد اول بر پایه استفاده از نرم افزار GIZA++ و پیاده سازی مدل مخفی مارکو و مدل IBM می باشد. با اینکه این رویکرد تنها به بخش هم ترازی نمی پردازد و هم ترازی به نوعی متغیر مخفی در مدل ترجمه[1] محسوب می گردد ولیکن نتایج بسیار خوبی حتی در پیکره های کوچک داشته است اما از دو نقص ساختاری برخوردار است که تاثیر شایانی درکارائی[2] این نوع سیستم ها دارند. یک مسئله اینکه یک عبارت چندین متناظر در زبان مقصد می تواند داشته باشد یا بالعکس ، یعنی هم ترازی نامتقارن می باشد. این مشکل با روشهای متقارن سازی مختلفی که بر اساس ابتکارهای[3] مختلف انجام میشوند ،که اکثرا از نظر زبانی نا آگاهانه می باشند در تحقیق برای یافتن استراتژی که سیستمهای posterior phrase-based با دقت بالا را تولید کنند برطرف شده است . و مسئله دیگر اینکه به دلیل پیچیدگی مدلهای IBM و هزینه تخمین پارامترهای این مدل ، معرفی اطلاعات زبانی بسیار دشوار می باشد .

رویکرد دیگر بر اساس رخداد کلمات و احتمال اتصال کلمات به یکدیگر می باشد. سادگی و انعطاف پذیری در دریافت دانش و تقارن باعث شده که با وجود وابستگی به اطلاعات تجربی و استراتژی های tunning ، روشهای قابل توجهی محسوب گردند. بزرگترین عیب این روشها در اجبار آنها به هم ترازی یک به یک است که منجر به هم ترازی با دقت بالا و تعداد فراخوانی های کم می گردد و این عوامل باعث محدودیت در استفاده این روشها در ماشین های ترجمه عملی می شود.

تحقیقات سالهای اخیر بر روی ترجمه آماری بر بالا بردن کیفیت ترجمه با آموزش مدل ها که نه تنها بر مبنای لغات بلکه بر مبنای عبارات ، ترجمه صورت گیرد ،متمرکز بوده است .

در مدلهای GIZA ابتدا هم ترازی صورت می گیرد و سپس استراتژی های متقارن سازی اعمال می گردند تا به این ترتیب هسته اصلی هم ترازی که عبارات از آن ساخته شده اند بدست آید . در اینجا با کمبود دانش زبانی و آموزش مدل ترجمه در یادگیری احتمالات از یک منبع نویزی[4] روبرو هستیم .

در این رویکرد که معرفی میشود ، ازروی تعداد تکرر رخداد کلمات در پیکره ، مستقیما به هم ترازی عبارات[5]

[1] Translation model
[2] Performance
[3] Heuristic
[4] Noize

عبارات[1] می رسیم ،که بر این مشکل غلبه شده است . در اینجا قبل از اینکه تناظر بین کلمات را پیدا کنیم ،پیوند[2] های با دقت بالا بین عبارات را پیدا می کنیم و جستجو برای اتصال بین عبارت ها ، به مجموعه ای از عبارات ممکن محدود می گردد (48).

1-ب- اندازه گیری رخداد کلمات و عبارات

در این روش پارامتری به نام Dice-score برای هر جفت عبارت محاسبه میشود که میزان شایستگی هم ترازی بین دو عبارت یا کلمه را مشخص میکند که این مقدار از فرمول زیر بدست می آید :

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

رابطه 1-1- میزان شایستگی هم ترازی بین دو عبارت یا کلمه

a تعداد دفعاتی است که هر دو لغت یا عبارت با هم رخ داده اند ، b و c تعداد دفعاتی است که یکی رخ داده و دیگری اتفاق نیفتاده و d تعداد دفعاتی که نه اولی و نه دومی در مجموعه داده ها[3] رخ نداده اند ، می باشند.در پیاده سازی مقدار باهم آیی[4] یک کلمه یا عبارت که x بار در جمله به زبان مبدا و y بار در جمله به زبان ترجمه اتفاق افتاده است ، را برابر $\min(x,y)$ در نظر گرفته می شود .

با اینکه محاسبه این پارامتر ساده می باشد ولی در مواجهه با گرام دوگانه[5] یا گرام سه گانه[6] یا در واقع عبارات، بار محاسباتی ایجاد می نمایند ولی با این وجود در اکثر موارد اطلاعات قابل توجه و مفیدی را دارا می باشند.

مشکل اصلی عملی نبودن محاسبه تمام ترکیبات ، حتی اگر پیکره از تعداد لغات کمی تشکیل شده باشد ، است.

[1] phrase-alignment
[2] Link
[3] Data Set
[4] Co-occurrence
[5] Bi-gram
[6] Tri-gram

چاره ای که برای رفع این مشکل در نظر گرفته شده از الگوریتمی استفاده می نماییم که تا حد امکان از اندازه گیری این باهم آیی ها اطلاعات مفید جمع آوری نموده و بدین ترتیب از بین تمام عبارات ممکن مجموعه ای را انتخاب نماید و از آن استفاده کند . برای این انتخاب از اطلاعات زبانی و آمار الگوریتم ، می تواند استفاده نماید (48).

2-ب- یک استراتژی در هم ترازی عبارات

این استراتژی شامل چهار مرحله می باشد :

ابتدا از مجموعه تمام کلمات ممکن ، مجموعه کوچکی از عبارت های جالب[1] را در هر زبان انتخاب می کنیم این انتخاب با کمک اطلاعات زبانی انجام میشود و مجموعه ای از عبارات را تشکیل می دهد که از چند کلمه تشکیل می شوند ولی نقش معنایی واحد را ایفا می کنند.

سپس الگوریتم با دقت بالا ، هم ترازی این عبارات را انجام می دهد و این عبارات را به عبارت دیگر یا کلمه پیوند می کند و از مقدار با هم آیی استفاده می نماید.

بعد از انجام هم ترازی در عبارات، الگوریتمی جهت هم ترازی واژه ها اعمال می گردد که مزیت وجود این مرحله کاهش پیچیدگی منتج شده از مرحله قبل می باشد.

در مرحله آخر پس پردازش روی هم ترازی بدست آمده ، اعمال می گردد که در این مرحله رفع ابهام روی مواردی مثل پیوندهای آزاد و همچنین تصمیم گیری در سطح جمله صورت می گیرد (48).

3-ب- طبقه بندی وانتخاب عبارات

[1] Interesting

همان طور که اشاره شد تمام ترکیبات ، رخداد هم زمان عبارات مختلف برای هر زبان بسیار زیاد می باشد و تقریبا غیر ممکن می گردد پس اولین کار این است که این فضای بزرگ را به بخش کوچکی که همان ترکیبات جالب از دیدهم ترازی و ترجمه می باشند ، محدود می کنیم ."جالب" به عباراتی که به یک مفهوم اشاره دارند و یا از لحاظ معنایی به هم مرتبط می باشند گفته می شود و انتظار می رود که به یک کلمه در زبان مقصد هم تراز گردند یا اشاره نمایند و از طرفی طبقه بندی معنایی نمودن عبارات به اندازه گیری با هم آیی آنها کمک میکند(در یک مفهوم همان مقدار را به نمونه های مختلف نسبت می دهیم).

در انتخاب عبارات از اطلاعات زبانی کمک گرفته می شود و خصوصا دو معیار برای انتخاب داریم که از دانش تکمیلی استفاده می کنند. ابتدا افعال را مشخص می نماییم که برای این کار از ماشین های خودکار قطعی که قوانین ساده را پیاده سازی می کنند استفاده می کنیم.

ورودی این قوانین، کلمات[1] و یا ریشه یا اصل[2]، می باشند و عبارت حاصل را به فعل اصلی،نگاشت[3] می کنند به همین جهت طبقه بندی افعال در کاهش محاسبات در هم رخداد ها موثر است زیرا فرم عبارت هر چه باشد به فعل اصلی نگاشته می شود .

برای جلوگیری از اشتباه بعد از بدست آوردن ریشه یا فعل اصلی[4] این فعل با دو لیست از افعال ممکن در زبان مورد نظر مقایسه می گردند (48).

یک نمونه از این قوانین در زبان انگلیسی به صورت زیر هستند :

[1] POS Tag
[2] lemma
[3] Map
[4] Head

PrP + VB VB(L=do) + PrP + VB VB(L=be) + PrP	PrP + MD(L=will/would/...) + VB MD(L=will/would/...) + PrP + VB
PrP + VB(L=be) +VBG VB(L=be) + PrP +VBG	PrP + VB(L=have) {+ been} + VB{G} VB(L=have) + PrP {+ been} + VB{G}

PrP: Personal Pronoun
VB / MD / VBG: Verb / Modal / Gerund (PennTree Bank POS)
L: Lemma (or base form)
{ } / ( ): optionality / instantiation

شکل4. نمونه ای از قوانین طبقه بندی در زبان انگلیسی

معیار دیگر انتخاب اصطلاحات می باشد. مجموعه ای از عبارات مصطلح در دو زبان را داریم و قصد بر این است که یک کلمه معادل برای این عبارات پیدا نماییم و چون فرهنگ لغات مورد استفاده قرار نمی گیرد پس معیار انتخاب معنی نمی باشد .

عبارات دیگری که قابل طبقه بندی هستند اعداد و تاریخ یا زمان می باشد که مستقل از زبان بوده و در هر زبان قاعده خاص خود را دارند .

اگر اطلاعات زبانی در دسترس نباشند از اطلاعات آماری برای انتخاب مجموعه عبارات می توان استفاده کرد به طور مثال انتخاب بیشترین گرام دوگانه [1]یا گرام سه گانه یا به طور کل محتمل ترین ترکیب دو تایی یا سه تایی کلمات می تواند معیار انتخاب باشند .

باید توجه شود که در این مرحله تصمیم گرفته میشود که آیا اتصال روی یک عبارت صورت گیرد یا کلمات رها شوند تا در مرحله بعد به صورت کلمه به کلمه[2] متصل یا هم تراز گردند (48).

4-ب- هم ترازی در سطح عبارات

در این مرحله میزان هم رخدادی بین یک عبارت از یک زبان در مقابل همه عبارات یا کلمات انتخابی از زبان دیگر محاسبه می گردد و یک استراتژی رقابتی برای متصل نمودن ، استفاده می گردد که در این استراتژی یک

[1]Most frequent bigram
[2] word-to-word

حد آستانه در نظر گرفته میشود و اتصال بین دو عبارت یا بین عبارت و یک کلمه بر اساس بهترین درجه با هم آیی و بالاتر بودن آن از حد آستانه انتخاب می گردد.این استراتژی بر مبنای این اصل است که عبارات در ترجمه مناسب تر از کلمات می باشند و حد آستانه مقداری است که محتمل ترین پیوند را تضمین می نماید و بدین ترتیب تمام اتصالات ممکن صورت نمی گیرد بلکه آنهایی که بالاترین ملاک هم رخدادی را داشته باشند انتخاب می شوند .

هنگامی که پیوند بین دو عبارت انتخاب می شود ، می توان با روشهایی بین کلمات دو عبارت نیز اتصال پیدا نمود ، اگر لزومی داشته باشد . این کار بوسیله الگوریتم های هم ترازی کلمات می تواند انجام گردد در صورتی که محدود به جستجو در عبارات نیز باشد (48).

5-ب- هم ترازی در سطح کلمات

از الگوریتم تکراری استفاده می شود که در آن ابتدا هم ترازی اولیه با استفاده از با هم آیی کلمات انجام می گیرد و احتمال اتصالات محاسبه می شود سپس جستجو بر اساس اولین پیوند برتر[1] اعمال شده که ابتکار هایی که بکار می برد global aligned sentence link probability می باشد و جستجو در مراحل بعد با استفاده از اطلاعات گرامری نیز آگاهانه تر میشود و به دلیل بزرگی فضای جستجو استفاده از این ابتکار ها بسیار موثر است.

دو پارامتر که برای بهینه سازی کارایی الگوریتم استفاده می شوند

احتمال اینکه لغت به هیچ کلمه دیگری وصل نشود[2] که باعث قابل مقایسه بودن پاسخهای بطور کامل هم تراز شده[3] و یا آنهایی که اصلاً هم تراز نشده اند[4] و می گردد.

حداقل امتیازی[5] که یک پیوند قابل قبول واقع شود .

همچنین محدود کردن جستجو ، در پیوندهای ممکن در زبان مقصد ، به یک پنجره که به این ترتیب کارایی الگوریتم افزایش می یابد و همچنین ابهام کاهش می یابد زیرا از وجود کلمات تکراری جلوگیری می کند .

[1]Best first search
[2]Null probability
[3]Full aligned
[4] Partialy aligned
[5]Minimum score

پنجره در همسایگی قطری که از تقسیم طول هر دو جمله بدست می آید، است.این پنجره وابسته به هر دو زبان بکار رفته بوده و ثابت نیست . برای انگلیسی اسپانیایی به تجربه مشخص شده که 8 کلمه بهینه است (48).

6-ب- پس پردازش

در این مرحله تصمیم برای هم ترازی نهایی گرفته میشود و از اطلاعات در سطح جمله استفاده می شود.

کارهایی که انجام می گردد تصمیم راجع به کلمات پیوند نشده که آیا می تواند آنها را نیز پیوند کند یا خیر و جستجو برای پیوند های طولانی تر و بازبینی مجدد پیوند های کلمات و عبارات در تمام جفت جملات پیوند شده ، می باشند.

در حالت ایده آل این بخش به قسمتهای انتخاب عبارت و هم ترازی ، بازخورد[1] می دهد که در تصمیمات قبلی که برای link ها در نظر گرفتند با اطلاعات کلی تجدید نظر نمایند (48).

6-3-1- ترجمه ماشینی مبتنی بر اصل[2]

سیستم های ترجمه ماشینی مبتنی بر اصل روشهای تجزیه ای را که بر اساس نظریه اصول و پارامترهای دستور زبان زایای چامسکی[3] است به کار می گیرند. تجزیه گر نحوی یک ساخت نحوی کاملی که شامل اطلاعات واژگانی، عبارتی، دستوری و موضوعی باشد را تولید می‌نماید، و تأکید بیشتر بر روی توانمندی نمایش های مستقل از زبان و تجزیه های زبانی عمیق می باشد.

در ترجمه ماشینی مبتنی بر اصل دستور زبان به صورت یک مجموعه مستقل از زبان، اصول تعاملی خوش ساخت بودن و مجموعه ای از پارامترهای مستقل از زبان دیده می شود. بنابراین برای یک سیستمی که از n زبان استفاده می کند، n بخش پارامتری و یک بخش اصول لازم است. پس میتوان گفت این روش مناسب استفاده با معماری میان زبان می باشد. این روش برد وسیعی از پدیده های زبانی را در بر می گیرد، اما فاقد دانش عمیق

[1] Feed back
[2] - principle-based machine translation
[3] - principle & parameters Theay of Chamsky's Generative Grammer .

درباره حوضه ترجمه که سیستم های مبتنی بر دانش و مبتنی بر نمونه بکار می برند می باشد (15).

7-3-1- روشهای ترکیبی

دهه اخیر شاهد ظهور تعدادی از روشهای جدید و امیدوارکننده ای برای ترجمه ماشینی بوده است و به تازگی تحقیقات در این زمینه متوجه ایجاد و توسعه روشهای ترکیبی برای طراحی سیستم های ترجمه ماشینی که بیش از یک روش را اتخاذ می کنند شده است. این روشهای ترکیبی مزیت هر دو یا چند روش را با یکدیگر دارند. پژوهشگران ترجمه ماشینی امروزه آزمایشات و تحقیقات متعددی را با ترکیبهای مختلفی از روشهای مختلف انجام داده تا به مزیتهای بیشتری دست یابند. تاکنون ترکیب سازی روشهای ترجمه ماشینی اساساً برروی دو جنبه تأکید داشته است: جنبه فنی و کیفیت ترجمه پیشرفته. مثال قابل ملاحظه در این رابطه اجرای همزمان سه موتور ترجمه ماشینی مختلف و ترکیب سه برونداد آنهاست (7).

ایده کلی در روشهای ترکیبی استفاده از یک مدل زبانی برای تجزیه متن مبدأ و یک مدل غیر زبانی برای یافتن تفسیر مناسب است.

روش مبتنی بر نمونه در ترجمه جملاتی که قبلاً در حافظه ترجمه نمایش داده شده اند بسیار خوب عمل می نماید. روش مبتنی بر آمار قادر به تولید جملات با صحت بالا می باشد، اماعموماً در برخورد با اصطلاحات، با هم آیی ها و توابع منقطع (با فاصله طولانی دور از هم) خیلی موفق نیست. با ترکیب این دو روش یک سیستم ترکیبی مبتنی بر نمونه و مبتنی بر آمار بدست می آید که ابتدا در حافظه ترجمه جملات منطبق را جستجو می نماید، اگر هیچ انطباق نزدیکی پیدا نشد آنگاه تجزیه آماری و تفسیر جملات مورد استفاده قرار می گیرد.

به هرحال آزمایشات و نیز تجربیات در زمینه ترکیب سازی در روشهای ترجمه ماشینی به شدت رو به افزایش است. انواع ترکیب سازیهای انجام شده در این زمینه عبارتند از:

الف ـ ترکیب اطلاعات آماری در سیستمهای مبتنی بر قاعده با استفاده از تکنیکهای مختلف (21 و 20)

ب ـ استخراج واحدهای ترجمانی جدید از متون دوزبانه و تبدیل آنها به سیستم های ترجمه ماشینی مبتنی بر قاعده (22)،

ج ـ ترکیب حافظه های ترجمه با ترجمه ماشینی مبتنی بر قاعده (9)،

د ـ ترکیب ترجمه ماشینی مبتنی بر نمونه با ترجمه ماشینی مبتنی بر قاعده (23)،

ه ـ ترکیب حافظه ترجمه با ترجمه ماشینی مبتنی بر نمونه (24).

علیرغم وجود نتایج عملی و تنوع تکنیکهای استفاده شده ، مایه بسی تعجب است که هنوز هیچ نظریه ترجمه ماشینی ترکیبی ارائه نشده است.

4-1- مکانیسم، مؤلفه‌ها و اجزای سازنده سیستم‌های ترجمه ماشینی

در ساختار سیستم‌های ترجمه ماشینی سه مؤلفه اصلی وجود دارد و هر مؤلفه نیز به‌نوبه خود از مجموعه اجزایی تشکیل می‌شود که در راستای تحقق نقش یا نقشهای آن سه مؤلفه‌اصلی عمل می‌کنند. مؤلفه‌های فوق‌الذکر در تناظر با سه سطح که علم زبان‌شناسی نوین برای‌هر زبان طبیعی و زنده قائل است قرار دارند. این سه سطح عبارتند از:

الف - واژگان

ب- نحو و ساختار

ج- معناشناسی

باید همواره این نکته را درنظر داشت که تفکیک این سطوح مبین عدم ارتباط بین مختصات‌واژگان، نحو و ساختار و معناشناسی متن نیست، بلکه هر سه سطح در سلسله مراتب چندبعدی‌خود در حکم تار و پود یکدیگرند.

به عبارت ساده‌تر این تفکیک صوری و اعتباری است، و در راستای نوعی شبیه‌سازی‌فرایند ترجمه، در ذهن مترجم است، در واقع ترجمه ماشینی از طریق این شبیه‌سازی متحقق‌می‌شود. از نظر اهمیت، اولویت، سهولت پردازش و یا مسأله‌زایی آنها نیز نمی‌توان بین سه مؤلفه‌یاد شده تفاوتی قائل شد، بلکه هر سه آنها، از دیدگاههای یاد شده هم‌سطح‌اند. شاید در وهله اول‌تصور شود که در پردازش‌ها، به علت انتزاعی‌تر بودن مؤلفه معناشناسی، این سطح پیچیده‌تر ودر نتیجه مسأله‌زاتر باشد، ولی باید توجه داشت که در عمل چنین نیست، هر سطح حتی سطح‌واژگان نیز مسائل و مشکلات خاص خود دارد چه ابهام‌ها در هر سطح می‌تواند ظاهر شود.اگر در هر سطح به گونه‌ای منطقی با مسائل و قضایا برخورد شود از میزان اغماض و پیچیدگی‌و ابهام‌زایی در سطوح بعدی کاسته می‌شود. دلیل امر وجود تعامل و ارتباط ظریف بین عناصراین سه سطح است. در سطور زیر سعی می‌شود تصویری کلی از مکانیسم و نقش هر یک ازمؤلفه‌های مورد بحث ارائه شود.

1-4-1- واژگان

نخستین مرحله پردازش در ترجمه ماشینی در سطح واژگان انجام می‌گیرد. پردازش‌مورد نظر عبارتست از تجزیه و تحلیل مرفولوژیکی واژه‌های واحد ترجمه. در این مرحله عناصرموجود در سطح واحد ترجمه یک به یک تفکیک، پردازش و مقوله‌گذاری می‌شوند. واژگان هرزبان مختصات و دشواریهای خاص خود را دارد و بالقوه می‌توانند مسأله‌زا باشند. برخی از این‌مشکلات به زبانی خاص اختصاص ندارند; ازجمله مصادیق این مضمون می توان به مثال های زیر توجه داشت: در زبان فارسی واژه‌های ساده و پرکاربرد "مرد"و" آرد" ، در زبان انگلیسی واژه‌های" run " و" correct" و در زبان فرانسه واژه‌های "livre" " و" cours" می‌توانند به مقوله‌های گرامری متعدد، تعلق داشته باشند. و این امر جدا و سوای ازپدیده چند معنایی است که در هر زبان امری عادی تلقی می‌شود. عدم پردازش صحیح این‌گونه‌موارد، می‌تواند موجد مسائل و مشکلات لاینحل در مراحل و سطوح بعد می‌گردد. شایان ذکر آن‌که، زبان انگلیسی از این منظر جزو خطازاترین زبانها محسوب می‌شود ولی باید همواره به خاطرداشت که هم‌نویسه‌ها تنها عامل خطازا در امر پردازش واژگان نیستند، در بعضی زبانها مشکلات‌خاص و مختص آن زبان وجود دارد که در عرصه پردازش زبان مجال ظهور و بروز یافته، منشأابهام و خطا می‌شوند. انفصال و ناپیوستگی عناصر واحد ترجمه در زبان انگلیسی، امکان ایجاد خطا و انتساب آنها به سازه‌های غیر، از جمله عوامل ابهام‌زا و خطازا در مرحله پردازش واژگان‌محسوب می‌شود بنابراین باید با لحاظ همه موارد مسأله‌آفرین، تدابیری اندیشید که در این‌سطح- به ظاهر ساده- میزان خطا به حداقل ممکن برسد و راه برای انجام پردازش و تجزیه وتحلیل‌های زبان‌شناختی بعدی، بهتر هموار گردد. در جهت نیل به این هدف در زبان انگلیسی اقدامات متعدد انجام گرفته است . از جمله در سال 1991 پروژه‌ای عظیم به مدیریت کنسرسیومی متشکل از چند دانشگاه و نهاد علمی و صنعتی، من‌جمله دانشگاه اکسفورد و دانشگاه لانکاستر، مرکز تحقیقات کتابخانه ملی بریتانیا، شورای علمی و مهندسی، آکادمی انگلیس، به منظور تدوین پیکره زبانی انگلیسی BNCو مقوله‌گذاری واژه‌های انگلیسی شروع و در سال 1996 خاتمه یافت. پیکره زبانی براساس صدمیلیون واژه تدوین شد و از طریق آن امکان کددهی و مقوله‌گذاری فراهم آمد و در نهایت ، حدود 98% مقوله گذاری صحیح میسر گشت .(25)

2-4-1- نحو و ساختار

تجزیه[1] در اصطلاح زبان شناسی رایانه ای[2] یعنی اجرای برنامه تجزیه و تحلیل‌نحوی متن و یا به عبارت کلی‌تر تجزیه و تحلیل گرامری جمله از طریق کامپیوتر. و تجزیه کننده[3] به‌سیستمی اطلاق می‌شود که با استعانت از آن واحد ترجمه از منظر ساختار و نحو تجزیه وتحلیل می‌شود 0به نحوی که کامپیوتر می‌تواند از طریق آن به سهولت عناصر تشکیل دهنده‌واحد ترجمه را بازشناسی کند، نقش آنها و روابط درونی و فیمابین آن عناصر و نوع قانون‌حاکم بر آنها را تشخیص دهد. در ترجمه ماشینی دستیابی به اطلاعات یاد شده ضروری واجتناب‌ناپذیر است، چه علاوه بر برگردان عناصر معنائی و واژگان، ترجمه ساختار نحوی واحدترجمه، از زبان مبدأ به مقصد نیز باید انجام شود. در واقع برگردان قالب و ساختار جمله از زبان‌مبدأ به زبان مقصد ، شرط اصلی و اساسی ترجمه است ، و این امر جز با تجزیه و تحلیل گرامری واحد ترجمه میسر نمی شود.انجام این مهم از طریق سیستم‌های تجزیه و تحلیلگر رایانه‌ای صورت می‌گیرد. گفته شده‌است که مهمترین رکن ماشین ترجمه تجزیه و تحلیلگر نحوی یعنی تجزیه کننده است. هر چند این‌طرز نگرش خالی از اغراق نیست ولی بیانگر اهمیت دومین مؤلفه اصلی در ساختار سیستم‌ترجمه ماشینی است. این اهمیت، انگیزه آن بوده که تاکنون صدها نوع تجزیه کننده در سطح جهان‌ابداع شود، تنوع مکانیسم‌های بکار رفته در ساختار تجزیه و تحلیلگرهای نحوی واقعا اعجاب‌انگیز است. تقریبا هیچ دانشگاهی و مؤسسه آموزش عالی در سطح جهان یافت نمی‌شودکه دارای مرکز تحقیقات زبان‌شناسی یا رایانه‌ای باشد و حداقل یک تجزیه و تحلیلگر نحو در آن‌ابداع نشده باشد. نقش اصلی تجزیه و تحلیلگر نحوی درساختار ترجمه ماشینی عبارت است از:

1- تشخیص دستوری بودن جمله

2- شناخت عناصر سازنده جمله و شیوه ترکیب آنها به منظور تشکیل واحدهای بزرگتر،عبارت، بند و...

3-تعیین نقش هر یک از واحدها در جمله

4- شناخت دقیق روابط معنایی بین عناصر سازنده جمله

[1]-Parsing
[2]- Computational linguistics
[3] -Parser

علاوه بر نقش اصلی، نقشهای فرعی- بنا به مقتضای کار و سلیقه ابداع کننده - نیز برای تجزیه وتحلیلگرنحوی‌درنظرگرفته‌می‌شود.

شیوه کار تجزیه کننده ها متفاوت است ولی معمولا براساس دو نوع شیوه کارشان، به دودسته تقسیم‌بندی می‌شوند که عبارتند از:تجزیه کننده های نزولی وتجزیه کننده های صعودی با آنکه هیچکدام از این دو نوع‌تجزیه کننده ها کامل و بی‌نقص نیستند ولی معمولا تجزیه کننده نزولی بیشتر مورد توجه بوده ، که خود نشانه موفقیت‌آمیز بودن، تجزیه و تحلیل نحوی‌است.(25)

3-4-1- معناشناسی

مشکل چندمعنایی بودن واژه‌ها، شناخته شده‌ترین ویژگی زبانهای طبیعی است و این امر زاییده قانونی است که در زبان‌شناسی تحت عنوان اقتصاد کلامی[1] اشتهار یافته است.

مترجم به هنگام برخورد به واژه چند معنا، با تکیه بر توانش زبانی[2] خود و بافت و موضوع متن می‌تواند معادل واقعی واژه را تشخیص دهد و برگزیند. رایانه فاقد چنین‌توانشی است. بنابراین طراح ماشین ترجمه ناگزیر است مقدماتی فراهم سازد و در الگوریتم ، به‌تمهیداتی دست یازد که از طریق آن بتواند خلأ موجود را ترمیم کند. این تمهیدات باید از نخستین‌مراحل پردازش یعنی از مرحله واژگان و نحو و ساختار آغاز گردد و تا آخرین مرحله یعنی‌مرحله تجزیه و تحلیل معناشناختی متن تداوم یابد. مثالهای زیر می‌تواندروشنگر مضمون باشد، کلمه "prime" در زبان انگلیسی یک مقوله دستوری، چهار مقوله معنایی‌و 23 معنا دارد. واژه "Suspect" نیز به سه مقوله دستوری (اسم و فعل و صفت) تعلق دارد ودارای شش مقوله معنایی و 21 معناست،" minister" نیز دو مقوله گرامری، شش مقوله معنایی و11معنادارد.

قاعدهŠ ماشین ترجمه می‌تواند 1280 معادل فارسی را برای تعبیر" suspect The prime " انتخاب کند. ولی اگر سیستم در پردازش واژگان و ساختارها دقیق عمل کرده باشد تعدادانتخابها به 32 و در نهایت اگر در مرحله پردازش معنا، منطقی با قضایا برخورد کند و فرهنگ‌معناشناختی تدوین شده براساس اصول زبان‌شناسی رایانه‌ای در اختیار داشته باشد، سیستم‌منحصراً تنها یک گزینه را انتخاب خواهد کرد. براساس همین محاسبه تعداد معادل‌های

[1]- Economie Langagiere
[2] - Competence

محتمل‌فارسی" prime minister " در سیستم‌های ماشین ترجمه عادی 176 خواهد بود. ولی اگر درمراحل قبلی، براساس معیارهای دقیق، پردازش‌ها صورت گرفته باشد، تعداد معادل‌ها به 112 ودر نهایت به تنها انتخاب صحیح دست خواهد یافت. واژه سه حرفی" run" را درنظر می‌گیریم، این‌واژه دو مقوله دستوری و 58 مقوله معنایی و بیش از 140 معنا دارد. تنها، سیستمی قادر خواهدبود که معنای مناسب این واژه را برگزیند که با استعانت از الگوریتمی قوی و فرهنگ‌معناشناختی جامع بتواند مؤلفه‌های معناشناختی را در طول پردازش and و or کند.(25)

5-1- ابهام زدایی معنی واژه

ابهام واژگانی اشاره به حالتی دارد که یا یک واحد واژگانی متعلق به طبقات دستوری یا اجزاء کلام مختلف با معانی مختلف باشد و یا یک واحد واژگانی بیش از یک معنی داشته و معانی مختلف آن متلعق به یک طبقه دستوری باشند (26). در اینجا هدف از ابهام زدایی معنی واژه رفع ابهام از حالت دوم است یعنی موردی که در آن معانی مختلف واژه متعلق به طبقه دستوری یکسانی باشند. ابهام زدایی معنی واژه اصطلاحی است که به استخراج معنی صحیح و مقتضی از واژه هایی که چند معنایی هستند اطلاق می شود و این کار غیر ممکنی نیست چرا که اگرچه واژه های ممکن است معانی زیادی داشته باشند، طبیعتاً تنها یکی از آن معانی مناسب بافتی که واژه مورد نظر در آن ظاهر می شود است. برای مثال واژه" bat "می تواند به معنای یک جانور شب زی، وسیله ای برای بازی بیسبال و یا چشمک زدن باشد. هدف ابهام زدایی معنی واژه تعیین یکی از این معناها بر اساس بافت زبانی که این واژه در آن ظاهر شده می باشد. در بسیاری از سیستم هایی که کار ابهام زدایی معنی واژه را به عهده دارند سیستم برچسب گذاری اجزاء کلام وجود دارد که مقوله دستوری واژه ها را مشخص می نماید. مثلاً در مورد واژه "bat " در صورتیکه سیستم برچسب گذاری اجزاء کلام تشخیص دهد که این واژه اسم است آنگاه سیستم ابهام زدایی معنی واژه کار خودرا شروع کرده و از میان دو معنای اسمی این واژه یعنی جانور شب زی و وسیله برای بازی بیسبال یکی را انتخاب می نماید.

ابهام زدایی معنی واژه یک امر ضروری در ترجمه ماشینی بوده و کاربرد زیادی درآن دارد چرا که سیستم ترجمه ماشینی بایستی قادر باشد تا مثلاً معنی واژه "bank "را در یک متن انگلیسی تشخیص داده و تعیین نماید

که آیا آن را به "بانک" یا به "ساحل رودخانه" در فارسی ترجمه نماید. از طرف دیگر سیستم های بازیافت اطلاعات نیز بایستی تنها متون و اسنادی را که معنی مناسب واژه داده شده برای جستجو را در برمی گیرند جستجو کرده و نمایش دهند. مثلاً اگر واژهٔ جستجو یا کلیدی "بانک" باشد پس سیستم می بایستی تمامی نمایش های مربوط به "ساحل رودخانه"را حذف نماید.

روشهای موجود و متداول برای ابهام زدایی معنی واژه به سه نوع تقسیم می شوند که عبارتند از: روشهای مبتنی بر قاعده، روشهای مبتنی بر پیکره یا مبتنی بر آمار و روشهای ترکیبی.

1-5-1- روشهای مبتنی بر دانش یا مبتنی بر قاعده

این روشها منابع دانش واژگانی را همراه با اطلاعاتی در مورد اولویتهای گزینشی، تداعی مفاهیم و طبقات دستوری بکار می گیرد و در قالب قواعدی در اختیار گرفته و بدین وسیله از بین معانی مختلف واژه یکی را برمی گزینند. چنین سیستم هایی نتایج خوبی را به دنبال دارند ولی ساختن قواعد گران تمام می شود و علاوه بر این قواعد موجود در سیستم ممکن نیست که بتوانند تمام موقعیتها و بافتهای ممکن موجود را پوشش دهند (12).

1-5-2- استفاده از روشهای آماری در ابهام زدایی معنایی

مانند دیگر کاربردهای پردازش زبان طبیعی، در سیستم های ابهام زدایی معنی واژه مبتنی بر آمار نیز استفاده از پیکره های زبانی که به طور خودکار یا دستی حاشیه نویسی و برچسب گذاری شده اند رو به افزایش است. مدلهای مبتنی بر آمار که از روشهای نظارتی برای نسبت دادن معانی به واژه های مبهم استفاده می نمایند نیاز به یک پیکره زبانی که از نظر معنای واژه ها برچسب گذاری شده است دارد. معنایی با بیشترین احتمال که براساس داده های پیکره ای محاسبه می شود به وسیله سیستم انتخاب می شود. برای کاربردها یی چون ابهام زدایی معنی واژه حجم بسیار زیادی از متون لازم است تا این اطمینان حاصل شود که تمامی معانی مختلف یک واژه چند معنایی در پیکره زبانی وجود دارد. برای تعیین معنای صحیح یک واژه مبهم با استفاده از پیکره های زبانی دو روش ممکن وجود دارد: استفاده از اطلاعات توزیعی[1] و استفاده از واژه های بافتی[1]. اطلاعات توزیعی توزیع فراوانی

[1] - Distributional information

معناهای یک واژه است در حالیکه واژه های بافتی به محدوده واژه های طرف راست و طرف چپ یک واژه خاص اطلاق می شود.

خط مشی آماری ،ما را از نیاز به الگوریتم های ماشینی و دانش زبان شناسی گسترده برای حل مشکلات زبان شناسی بی نیاز می کند.

از آنجایی که استفاده از روشهای آماری در ابهام زدایی معنی واژه بسیار مؤثر بوده و به سرعت در حال پیشرفت و فراگیر شدن است در بخش بعدی به معرفی چند سیستم مبتنی بر آمار که هدف آنها ابهام زدایی معنی واژه است اشاره می شود (12).

1-2-5-1-استفاده از گنجواژه[2]

در این روش می توان 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 مستقیما از تعریف معنای کلمات در فرهنگ لغت استفاده می شود. نوع دوم که نشان می دهد چگونه از اطلاعات طبقه بندی شده موجود در فرهنگ لغت می توان طبقه بندی معنایی یک کلمه را با توجه به متنی که آن کلمه در آن وجود دارد، به دست آورد . (یارافسکی[3] 1992) و سوم اطلاعاتی است که از ترجمه یک کلمه به کمک یک فرهنگ لغت دو زبانه به دست می آید و استفاده از معنای کلمه در زبان مقصد برای رفع ابهام از معنای کلمه .(27)

یاروفسکی [4] سیستمی را پیشنهاد می کند که معنی واژه در هر نوع متن انگلیسی را با استفاده از مدلهای آماری طبقات اصلی دستوری از گنجواژه Roget ابهام زدایی می نماید. معانی مختلف یک واژه به طور طبیعی با طبقات مختلف گنجواژه مطابقت دارند از این رو انتخاب محتمل ترین طبقه راه سودمندی است برای مجزاکردن و تشخیص معناهای مختلف. محتمل ترین طبقه دستوری یک واژه بوسیله جستجوی صد واژه دوروبر (مجاور) نمایندگان هر طبقه تعیین می شود. این واژه های نماینده تعیین شده و در طی محاسبه بوسیله امتحان با صد واژه مجاور هر نمونه از یک طبقه بارگذاری می شوند. این سیستم نیاز به پیکره زبانی مربی[5] بزرگی (که نیازی به

---

[2] - Context words
[3] - Thesaurus
[3] - Yarawsky
[4]- Yarowsky
[5]- training corpus

برچسب گذاری آن نیست) و یک گنجواژه (یا هر نوع منبع داده ای دیگر که بتواند مجموعه ای از طبقات معنایی و نیز واژه ها در هر طبقه را تعریف نماید) دارد. هیچ نوع بخش سیستمی دیگری به جز تجزیه اجزاء کلام و استخراج شکل ریشه ای واژه ها مورد نیاز نیست. اطلاعاتی که از داده های مربی به دست می آید شامل فهرستی از واژه های نماینده برای هر طبقه دستوری و بار[1] آنهاست. تمام واژه های موجود به اشکال ریشه ای خود تقلیل می یابند (یعنی اطلاعات صرفی نادیده گرفته می شود) تا با افزایش تعداد رخدادها بتوان به آمار دقیق تری دست یافت.

روش محاسبه بار به صورت لگاریتم برجستگی واژه[2] برای آن مقوله تعریف می شود . برجستگی یعنی Pr (w | Cat) / Pr (w) ، به عبارت دیگر احتمال اینکه یک واژه در یک بافت از یک مقوله داده شده ظاهر شود تقسیم بر احتمال وقوع آن واژه در پیکره زبانی به طور کلی (برای واژه های مفید ، برجستگی احتمال دارد که بزرگتر از یک بوده و بنابراین بار آن یا لگاریتم برجستگی مثبت خواهد بود.). این سیستم از صد واژه مجاور واژه مورد نظر استفاده می کند. تمام واژه ها در بافت بدون توجه به مقوله دستوریشان مورد توجه قرار می گیرند. اطلاعات نحوی و صرفی دخالتی در کار نداشته و تمام واژه ها به شکل ریشه ای خود در می آیند. برای محاسبه محتمل ترین طبقه یک واژه مراحل زیر اجرا می شوند:

1-1-2-1-5- محدود سازی طبقه

تنها طبقاتی که شامل این واژه می شوند به عنوان احتمالات (طبقات احتمالی) در نظر گرفته می شوند. این کار به طور کلی مفید است اما در صورتیکه در گنجواژه از قلم افتادگی هایی وجود داشته باشد می تواند سبب بروز مشکلاتی نیز گردد. یاروفسکی مثالی از واژه "suit" می زندکه معنای آن در بازی ورق (که به معنای خال است) در گنجواژه آورده نشده بود و بنابراین اشتباه شماره گذاری شده بود. وقتی که واژه به صورت طبقه محدود نشود (به طوریکه تمام طبقات موجود باشند) همه نمونه ها به طور صحیح نشانه گذاری خواهند شد.

2-1-2-1-5-جمع بارها

[1]- weight
[2]- word's salience

با در نظر گرفتن هر طبقه ممکن به صورت انفرادی صد واژه بافتی مجاور برای هر واژه نماینده آن طبقه جستجو می شوند. بار هر نماینده‌ای که پیدا شد به جمع کل اضافه می شود. (اگر برای یک نماینده بیش از یک بار پیدا شود بار آن نیز بیش از یک بار اضافه می شود). این نوعی تکمیل سازی قانون بیز[1] برای $Pr(Cat \mid w)$ می باشد که برای هر w مربوطه در بافت:

$$\underset{\Sigma}{\text{Argmax}} \quad \text{Log} \quad \frac{Pr(w \mid Cat)\ Pr(Cat)}{Pr(w)}$$

W in context

رابطه 2-1- جمع بارها

احتمال واژه $Pr(w)$ می تواند حذف شود چون تغییری در نتیجه ماکزیمم سازی نخواهد داد. (البته اگر نتایج بخواهند به صورت بین واژه ای مقایسه شوند وجود آن ضروری است.). احتمال پیشین برای تمام طبقات یعنی $Pr(Cat)$ مساوی فرض می شود از اینرو این نیز در سیستم مورد استفاده قرار نخواهد گرفت.

3-1-2-5-1- انتخاب محتمل ترین طبقه

طبقه ای که بالاترین مجموع بار را داشته باشد به عنوان طبقه واژه مورد نظر انتخاب می شود.

این سیستم در مقایسه با تلاشهای قبلی خوب عمل می نماید مخصوصاً به این خاطر که این یک سیستم واقعی است، محدود به واژگان خاصی نمی شود و در قلمرو نسبتاً وسیعی کار می کند (یک دایره المعارف). برای ارزشیابی این سیستم از تعدادی واژه‌های مبهم که قبلاً برای سیستم های ابهام زدایی دیگر مورد استفاده قرار گرفته شده استفاده شد و صحتی بین 72% تا 99% بدست آمد (28).

3-1-5- استفاده از موصوف برای دستیابی به معنای صفت

[1]- Bayesian rule

جاستسن [1] و کتز [2] یک روش اصولاً زبانی را برای نوعی ابهام زدایی که در آن تنها از سرنخهای از نظر معنایی یا نحوی مرتبط استفاده می شود ارائه می دهند.در این روش از روشهای آماری برای سازماندهی و تجزیه داده ها در پیکره های مربی مورد استفاده قرار می گیرند اما مرحله ابهام زدایی از تکنیکهای آماری استفاده نمی نماید. سیستم واقعی ای که آنها توصیف می کنند صفتها را تنها با استفاده از اسم هایی که بوسیله آن صفتها تعریف می شوند ابهام زدایی می نماید. در ابتدا سیستم به طور کامل خودکار است یعنی با استفاده از تکنیکهای آماری آموزش داده می شوند تا موصوفهایی را که معنی صفت هایشان را به نوعی نشان می دهند جدا کنند. برای اینکه بتوان پوشش وسیعتری را در اختیار داشت، سیستم بعداً با استفاده از قواعدی که با دست کدگذاری شده اند و نه‌اطلاعات آماری تکمیل می شود. جاستسن و کتز تمام آزمایشات خود را بر روی پنج صفت انگلیسی مبهمی که بیشترین فراوانی را دارند یعنی صفتهای , hard ,light ,old, right و short انجام دادند.

برای انجام آزمایش یک پیکره مربی ابهام زدایی شده لازم است. برای پنج صفت فوق این کار می تواند به طور خودکار با استفاده از خاصیت متضادهای معنایی خاص انجام پذیرد یعنی معانی مختلف واژه old می تواند بوسیله متضادهایش یعنی young و new تعریف شود. هنگامیکه یک صفت و یکی از متضادهایش در یک جمله ظاهر شوند معمولاً یک اسم (اسم یکسانی) را توصیف می کنند و صفت مبهم هم تقریباً همیشه معنی عکس متضادش را دارد. بنابراین پیکره مربی می تواند بوسیله انتخاب جملاتی از پیکره بزرگی که جفت های متضاد در آن ظاهر شده اند و تجزیه آنها به منظور تضمین اینکه صفتها به اسم یکسانی برمی گردند تولید شود. پس یک تجزیه گر نحوی نیز ضروریست بدین منظور و به منظور تعیین اینکه کدام اسم بوسیله صفتی که در جملاتی که در حال ابهام زدایی هستند تعریف می شود . تجزیه گر صرفی نیز لازم است تا اطلاعات صرفی را از اسمها حذف نماید . این سه جزء اصلی یعنی پیکره مربی ، تجزیه گر نحوی و تجزیه گر صرفی بوسیله جاستسن و کتز به طور دستی شبیه سازی شدند .

نتیجه ای که از پیکره مربی بدست آمد مجموعه ای از اسمهای شاخص برای هر معنای جداگانه یک صفت بود بطوریکه وقتی این اسم ها بوسیله آن صفت تعریف می شوند آن صفت همان معنای خاص خود را نشان میداد .

[1]- Justeson
[2]- Katz

برای تعیین اینکه چه اسمهایی شاخص های یک معنی صفت باشند برای هر اسمی تعداد با هم آیی (با هم آیی هر معنای صفت با آن اسم) از زیر پیکره ابهام زدایی شده استخراج می شود . اطلاعات صرفی هم از اسم ها جدا می شود ، از اینرو تعداد اسم های مفید موجود جهت استنتاج آماری افزایش می یابد . به جهت اینکه زیر پیکره ابهام زدایی شده انتخاب تصادفی ای از یک پیکره واقعی نبوده است (چرا که این زیر پیکره تنها جملاتی با جفتهای متضاد را شامل می شود) سو گیری آماری[1] در آن مشاهده می شود . جاستسن و کتز فرمولی برای تعیین سطح اطمینان میدهند که این سوگیری را مدنظر قرار میدهد و در آخر مقاله شان آن را به عنوان پیوست ارائه میدهند . اسم هایی به عنوان شاخص مورد استفاده قرار می گیرند که معنای ویژه ای با سطح اطمینان 95% از خود نشان می دهند .

برای ابهام زدایی یک صفت ابتدا جمله بایستی تجزیه شود و اسمی که آن صفت آن را تعریف می کند مشخص گردد . اطلاعات صرفی باید از این اسم جدا شوند . اگر صورت پایه اسمی که تعریف می شود (یا نوع اسم طبقاتی مانند اسامی افراد و اسامی اشیاء نیز استفاده می شوند) شاخصی برای یکی از معناهای آن صفت باشد در آن صورت آن معنا انتخاب می شود و در صورتیکه شاخصی برای هیچ یک از معناهای آن صفت نباشد هیچ معنایی به آن تعلق نمی گیرد.

به علت پوشش محدودی که استفاده از این روش دارد (حتی اگر به جای آن اسمهایی که از نظر آماری شاخص های مهمی به شمار میروند تمام اسمها شاخص فرض شوند) جاستسن و کتز بعدها سیستم متفاوت دیگری را پیشنهاد دادند که از صفتها (نسبت ها) ی نحوی و معنایی به جای اسمهای واقعی برای اسمهای مورد نظر استفاده می نماید و نیز از قوانین دست نوشته شده ای برای تخصیص معنا براساس آن صفتها (نسبت ها) استفاده می نماید.

(مثلاً برای light نسبت not dark ⟵ colour+ استفاده می شود).

این روش از داده های آماری استفاده نمی کند بنابراین به جزئیات آن در اینجا نمی‌پردازیم. این سیستم در مورد هر صفت بر اساس مجموعه ای از صد جمله به طور تصادفی انتخاب شده از کل پیکره زبانی ایکه شامل آن صفت می شود آزمایش شد. جملاتی که دارای صفتی که معنای آن بوسیله یک متضاد قابل توضیح نبود بودند

[1]- statistical bias

مستثنی شده و خارج شدند و این سیستم صحتی معادل 100% برای اسمهای شاخص مهم با سطح پوشش 22/6% و صحتی معادل 97% برای تمامی اسمها با سطح پوشش 50/1% و صحتی معادل 99/1% برای صفتها (نسبت ها) ی نحوی و معنایی که از قوانین دست نوشته استفاده نموده بودند با سطح پوشش 73/6% بدست آورد (29).

4-5-1- ترکیب آماری شاخص های چندگانه

دو روش زیر شاخص های چندگانه را از نظر آماری ترکیب می نمایند تا معنای واژه بدست آید یعنی به جای تکیه بر روی یک نوع شاخص از چند نوع شاخص استفاده می نمایند .

4-5-4-1- مدلهای تجزیه شدنی[1]

بروس[2] و ویب[3] مدلی را توصیف می کنند که از مشخصه های بافتی چندگانه استفاده نموده و نیازی به فرضیات آزمایش نشده ای در مورد شکل مدل ندارد، یعنی مشخصه هایی که ترکیب می شوند مستقل فرض نمیشوند (یا از نظر بافتی بسته به معنای واژه مستقل فرض نمی شوند) و برخی مدلهای قبلی این فرضیه را ساخته اند . فرض اینکه تمام مشخصه ها در حقیقت به طریقی وابسته هستند بدان مفهوم است که داده های مربی بسیار بیشتری برای ساختن یک مدل خوب از یک واژه نیاز است. به جای آن، روش بروس و ویب مدلها را با تمام وابستگی های ممکن با استفاده از داده های مربی امتحان می نماید تا بفهمند کدام مدل انطباق بهتری دارد و کدام پارامترهارا می توان متقابلاً مستقل فرض نمود. این مدل از چندین مشخصه که به دلیل ماهیت اطلاعی4 بودنشان انتخاب می شوند استفاده می کند. اگر مدلی که فرض می نماید ارزش مشخصه ای مستقل از معنای واژه است انطباق ضعیفی داشته باشد آن مشخصه ماهیت اطلاعی دارد. بهترین مدل آنها از مشخصه های زیر استفاده می نماید: مفرد / جمع، برچسب اجزاء کلام برای واژه مورد نظر و نیز دو واژه از هر دو طرف، و مشخصه های ویژه با هم آیی که برای هر واژه ای W که بخواهد ابهام زدایی شود متفاوت است.

[1]- Decomposable
[2]- Bruce
[3]- Wiebe
[4]-Informative

بروس و ویب مدل خود را بر روی واژه interest امتحان کردند زیرا در مدلهای قبلی این واژه یکی از سخت ترین واژه ها برای ابهام زدایی بود. مجموعه ای از داده ها شامل 2369 کاربرد واژه interest به عنوان اسم به طور دستی برچسب گذاری شد. 600 تا از اینها برای امتحان کنار گذاشته شد و بقیه برای آزمایش مورد استفاده قرار گرفتند و در کل، سیستم صحتی معادل 79% بدست آورد. بروس و ویب مجموعه دادهای خود را در اختیار عموم قرار دادند به طوریکه کارهای آینده بتواند مستقیماً با کار آنها مقایسه شود (30).

2-4-5-1- ابهام زدایی مبتنی بر نمونه

نگ[1] و لی[2] سیستمی را توصیف می نمایند که اطلاعات بدست آمده از مشخصه های چندگانه را نیز ترکیب می نماید. اما سیستم آنها از مدلهایی استفاده می نماید. این مدلها مجموعه هایی از ارزشهای مشخصه ای هستند که نمونه هایی از واژه مورد نظر در داده های مربی را نمایش داده و سپس با نمونه واژه خواسته شده در متن با استفاده از معیار فاصله مقایسه می شود.

در نسخه اولیه این سیستم، واژه، معنایی که نزدیک ترین مدل دارد را به خود می گیرد و در نسخه اصلاح شده بعدی ، K تا ازنزدیک ترین نمونه ها پیدا شده و سپس متداول ترین معنی از این نمونه ها استفاده می شوند. مشخصه های مورد استفاده در این سیستم شبیه مشخصه های سیستم بروس و ویب هستند، یعنی شامل اطلاعات مربوط به اجزاء کلام واژه مورد نظر و سه واژه از هر طرف، مشخصه صرفی (مفرد / جمع برای اسمها)، و کلید واژه هایی معادل با هم آیی های بروس و یب. نتایج بر روی داده interest که سیستم بروس و یب آن را امتحان کردند و نیز بر روی داده های پیچیده تری که شامل 121 اسم و 70 فعل بود عبارت بودند از صحتی معادل 89% در مورد داده اولی و 75/2% و 58/7% برای داده های دومی که به ترتیب از پیکره های زبانی WSJ و Brown استخراج شده بودند. این نتایج بدست آمده در مقایسه با الگوریتم دیگری به نام " نایوبیز3 " که از همان مجموعه مشخصه ها استفاده می نمود بهتر به نظر می رسد (31).

[1]- Ng
[2]- Lee
[3]- Naïve-Bayes

5-1-5- استفاده از پیکره یک زبانه زبان مقصد

سیستم های مختلفی بر این اساس پیشنهاد شده است که به چند نمونه از آنها اشاره کرده و آنها را مورد بررسی قرار می دهیم:

1-5-5-1- روش داگان[1] و ایتای[2]

داگان و ایتای سیستمی را پیشنهاد می کنند که واژه های مبهم زبان مبدأ را با استفاده از داده های آماری مستخرج از پیکره زبانی زبان مقصد و یک واژه نامه دو زبانه (لیستی از واژه های معادل به زبان مبدأ و مقصد) ابهام زدایی می نماید. با استفاده از تجزیه گر زبان مبدأالگوریتم آنها قادر به تشخیص ارتباطات نحوی بین واژه ها بوده و از فرهنگ دو زبانه برای منطبق کردن آن ارتباطات با امکانات مختلف (با این فرض که بیش از یک ترجمه ممکن برای برخی از واژه ها وجود دارد) در زبان مقصد استفاده می کند. سپس این امکانات مختلف در زبان مقصد بر طبق آمار بدست آمده از تعداد دفعات وقوع آنها در پیکره مربی انتخاب می شوند (32).

داگان و ایتای الگوریتم پیشنهادی خود را با استفاده از یک سیستم کامل مورد ارزیابی قرار ندادند زیرا برخی از عناصر مورد نیاز (مانند تجزیه گر و واژه نامه هایی برای زبان مقصد) در دسترس نبودند. تنها دو آزمایش انجام شد: یکی با استفاده از جملات عبری و دیگری با استفاده از جملات آلمانی. صحت بدست آمده برای سیستم، معادل 91% و 78% به ترتیب برای زبان عبری و زبان آلمانی بود.

سیستم داگان و ایتای تنها از واژه هایی که در قالب ارتباطات نحوی خاصی (مانند ارتباط بین فعل و مفعول و یا ارتباط بین فعل و فاعل) هستند استفاده می نماید. بنابراین اطلاعاتی را که می توان از دیگر واژه ها در بافت منطقه ای بدست آورد نادیده می گیرد. در برخی موراد که چندین ارتباط در رابطه با یک واژه وجود دارند ممکن است تصمیم گیری نهایی به دلیل انتخاب ارتباط نادرست یعنی ارتباطی که فراوانی های آن باعث انتخاب گزینه نادرست به جای گزینه درست می شود به سبب اینکه نمونه های بیشتری از آن در پیکره دیده شده است اشتباه باشد. می

[1]- Dagan
[2]- Itai

توان برای حل این مشکل ارتباطات نحوی را بار گذاری نمود بطوریکه ارتباطات نزدیک مانند ارتباط بین یک فعل و مفعول را جزء ارتباطات مهم تر قرار داد.

2-5-5-1- روش هیونگ لی[1] وجنگ – سی پارک[2] وگیل چنگ کیم[3]

در این روش آنها نیز از یک مدل آماری مبتنی برپیکره استفاده کردند و یک روش انتخاب واژه ای سه مرحله ای را پیشنهاد دادند. این مراحل عبارتند از :آشکار سازی معنی از کلمات مبداء ، نگاشت معنی به کلمه، و انتخاب مناسب ترین جزء واژه ای زبان مقصد.

یک کلمه می تواند معانی بسیاری داشته باشد و هر کدام از معانی می تواند در کلمات زبان مقصدبسیاری نگاشت شود. دانشی و آگاهی لازم ، برای هر مرحله از یک دیکشنری ماشین خوان و پیکره یک زبانه زبان مقصداقتباس شده است. بوسیلهٔ عملکرد انتخاب واژه ای در سه مرحله و اقتباس آگاهی ضروری برای هر مرحله از منابع موجود ،این سیستم کلمهٔ مناسب برای ترجمه را با قابلیت تعمیم و نیرومندی بالا انتخاب می کند. در این روش اولین مرحلهٔ آشکارسازی کلمه از انتخاب واژه در جملات زبان مبداء با استفادهٔ قوانین برای آشکار سازی معنی کلمه می باشد. قوانین زبان مبداء برای ابهام زدایی از معنی واژه[4] از یک دیکشنری قابل خواندن ماشینی اقتباس می شوند و در دومین مرحله مفهوم انتخاب شده از کلمهٔ زبان مبداء در مجموعه ای از کلمات زبان مقصدنگاشت می شود. هم چنین اطلاعات نگاشت شده از معنی کلمهٔ زبان مقصد می تواند از یک دیکشنری قابل خواندن ماشینی اقتباس شود. یک کلمهٔ زبان مقصدکه بصورت مناسب در جملهٔ زبان مقصدواقع شده است در

[1] - Hyun Ah Lee
[2] - Jong C. Park
[3] - Gil Chang Kim
[4] - Word Sense Disambiguation(WSD)

سومین مرحله انتخاب می گردد. برای انتخاب یک جزء واژه ای زبان مقصد ، روابط اجزای واژه ای با یکدیگر در جملات زبان مقصد، با روابط صرفی و وزن های رابطهٔ صرفی مورد استفاده قرار می گیرد.

ترتیب کلمات در زبان مقصد و روابط نحوی شان از یک مجموعهٔ پیکره ها یک زبانهٔ زبان مقصدبصورت تجزیه و تحلیلی درختی فراهم شده اقتباس می شود. (33)

الف -آشکار سازی معنی کلمه مبداء و نگاشت کلمه

در اولین مرحله، معانی مختلف کلمهٔ زبان مبداء با استفاده از دانش موجود در فرهنگ لغت مشخص می شود. در اینجا بر اساس معنی واژه در هر طبقه(یاروفسکی 1993)، از طبقه های کلمه مبداء ، برای آشکار سازی معنی آن استفاده می شود و برای آشکار سازی دقیق تر ، از اطلاعات رابطه ای صرفی بین کلمه مبداء و طبقهٔ آن نیز استفاده می گردد. طبقات و رابطهٔ صرفی شان در یک فرهنگ MT شرح داده شده است. به جدول زیر توجه نمایید.

جدول 1. قسمتی از دانش WSDبرای کلمات Wear,Old

Table 1: Part of WSD knowledge for 'wear' and 'old'

	relation	collocation	word sense	mapped target words
wear	object	coat, glasses	wear[1]-1 (put on)	*ip-ta, ssu-ta, sin-ta, cha-ta*
		expression	wear[1]-3 (express)	*tti-ta, cis-ta*
old	modifiee	man, book	old[1]-2 (not young)	*nai tun, nonyen-uy, nulk-un*
		shoe, car	old[1]-3 (not new)	*nalk-un, hen*

فرض کنید دوجمله مانند جملات زیر را به کامپیوتر داده تا آن را به زبان کره ای ترجمه نماید:

(1) 'He wears an angry expression.'
(2) 'She wears old shoes.'

در جمله اول معنی" Wear " پوشیدنی می باشد(یک عبارت خاصی درباره پوشاندن سطح) اما در جمله دوم معنی" Wear " پوشیدنی بر بدن می باشد. همچنین معنی "Old " در جمله دوم "غیر جدید " می باشد ، اما در عبارت "Oldest student " کلمه "Old " به معنی "پیر" می باشد.

جدول فوق یک قطعه از دانش رفع ابهام از معنی کلمه موجود در فرهنگ لغت برای افعال Wear وOld را نشان میدهد.

منظور از Wear در جمله دوم "کفش ها " می باشد. بوسیله محاسبه کردن تشابه کفش و طبقه اش در جدول بالا می توان تعیین کرد که Wear در جمله دوم معنی $Wear^1 - 1$ را دارد.در اینجا از پیکره Wordnet برای محاسبه شباهت استفاده شده است.و دانش ابهام زدایی از معنای واژه به کمک دیکشنری ، بطور اتوماتیک و و از یک دیکشنری ماشین خوان اقتباس شده است.

ب- نگاشت کلمه زبان مبدأ به مجموعه ای از کلمات زبان مقصد

در مرحله دوم مفهوم کلمه زبان مبداء انتخاب شده و به مجموعه ای از کلمات زبان مقصدکه با آن معادل بوده ولی استفاده های متفاوت و تفاوتهای جزیی دارند نگاشت می شود.ستون آخر جدول بالا اطلاعات نگاشتی از مفاهیم واژه مبداء در واژه های زبان مقصد را نشان می دهد.همچنین این اطلاعات نگاشتی از یک دیکشنری ماشین خوان اقتباس شده است.(33)

ج - انتخاب واژه با استفاده از اطلاعات آماری زبان مقصد

سومین مرحله ، مرحله ای است که در بین واژه های زبان مقصد موردی را انتخاب می کند. یک کلمه مقصدکه بصورت رایج در متن زبان مقصدمشابه رخ می دهد باید یک کلمه مقصدمناسب از مفهوم کلمه مبداء باشد. بافت زبان مقصدمی تواند بوسیلهٔ جمع آوری کلمات ساخته شود. ازاطلاعات آماری جمع آوری شده زبان مقصد برای انتخاب کلمه زبان مقصد مناسب استفاده می شود.

به جدول های 2و3 توجه نمایید:

جدول 2 برخی از کلمات را نشان می دهد که بصورت رایج همزمان با واژه مبهم در پیکره زبان مقصد رخ می دهند.

جدول 2. کلماتی که بصورت رایج همزمان با واژه مبهم در پیکره مقصد رخ می دهند

Table 2: Sense of the object of wear[1]-1 and its appropriate target word

	body-wear	head-wear	face-wear	hand-wear	foot-wear	arm-wear
wear[1]-1	*sin-ta*	*ssu-ta*	*kki-ta*	*kki-ta*	*sin-ta*	*cha-ta*

جدول 3. بخشی از جمع تکرار کلمات مقصد $Wear^1-1$

Table 3: Part of cooccurrence statistic information

	"clothes"		"hat"	"glasses"	"glove"	"shoes"			"watch"
	os	*uypok*	*moca*	*ankyeng*	*cangkap*	*kwutwu*	*sinpal*	*sin*	*sikyey*
sin-ta	36	1	-	-	-	-	-	-	-
ssu-ta	1	-	8	11	-	-	-	-	-
kki-ta	-	-	-	2	2	-	-	-	-
sin-ta	-	-	-	-	-	2	5	1	-
cha-ta	-	-	-	-	-	-	-	-	1

و جدول 3 بخشی از جمع تکرار کلمات مقصد $Wear^1-1$ و دیگر کلمات موضوع را نشان میدهد. با کمک این جدول می توان فهمید که یک طبقه در برخی از رابطه های صرفی یک شاخص خوب برای انتخاب واژه ای خاص می باشد.

با استفاده از اطلاعات آماری زبان مقصد،تعداد هریک از ترکیب ها ی نحوی در پیکره زبان مقصد مشخص می شود.(دایگان و ایتای 1994) . دایگان و ایتای از ارتباطات صرفی و نحوی زبان مبداء برای ایجاد مجموعه چند تایی صرفی از رابطهٔ صرفی زبان مقصدبکار میبرند. اما در اینجا یک چندتایی صرفی با روابط صرفی زبان مقصد ساخته

می شود زیرا سومین مرحلهٔ انتخاب واژه بعد از انتقال ساختاری رخ می دهد. دو کلمه ای های زبان مقصددر مجموعهٔ چند تایی صرفی عوامل مجموعهٔ کلمات زبان مقصدهستند که در مرحلهٔ دوم نگاشت شده است.

احتمال (i و a) p را از انتخاب یک کلمه مقصدai به عنوان کلمه ای برای ترجمه ارزیابی می کنیم. بیایید فرض کنیم که مفهوم کلمهٔ انتخابی a می باشد. ما می خواهیم یک کلمه را از a1 و a2 و ... ak انتخاب کنیم که مجموعه ای از کلمات هدف مفهوم a می باشد. بیاییئد مجموعه از گره هایی[1] که به ساختار جمله زبان مقصد به گره معنی a متصل هستند با(a) Θ از یکدیگر تفکیک کنیم .اگر یک عنصر از مجموعه (a) Θ را بصورت گره (g ، m، a) مشخص کنیم به این معنی که گره a می تواند به mکلمه زبان مقصد ترجمه شود، a1 و a2 و ... ak ، و گره با رابطهٔ صرفی g به گره با معنی a مرتبط شده است بنابراین این گره مجموعه ای ازمعانی a را نمایش میدهد و (g, a_i, a_1) یک مجموعهٔ چند تایی صرفی روی معانی a_i می باشد. فرض کنید که $f_{(a_i, a_1, g)}$ تعداد مجموعهٔ چند تایی صرفی (g, a_i, a_1) در یک پیکره ها زبان مقصدمی باشد. و (g و a) w را بصورت وزن اهمیت رابطه نحوی gدر انتخاب معنی کلمه a از زبان مقصد در نظر می گیریم. سپس احتمال (a_i) p^ را که به مفهوم احتمال انتخاب a_i به عنوان معنی از کلمه مقصدa از فرمول زیر محاسبه شود:

$$n(a_i) = \sum_{(a,m,\gamma)\in\Theta(a)} \frac{\sum_{l=1}^{m} f(a_i, a_l, \gamma)}{m} \cdot w(a,\gamma) \qquad (1)$$

$$\hat{p}(a_i) = \frac{n(a_i)}{\sum_{j=1}^{k} n(a_j)} \qquad (2)$$

رابطه 3-1- احتمال انتخاب a_i بعنوان معنی از کلمه مقصد a

$n(a_i)$ تعداد ظهور کلمه مقصد ai در مجموعه های پیکره می باشد. با استفادهٔ مجموعهٔ $(\alpha, m, \gamma) \in \Theta(a)$ ، معادلهٔ (1) برای جمع کلی تعداد همهٔ گره هایی است که به گره با معنیa مرتبط شده

[1] - Nodes

است و بوسیلهٔ محاسبهٔ تعداد کلی فرکانس ها با استفاده از رابطهٔ نحوی (a_g , w) در (1)، ما می توانیم درجهٔ اهمیت رابطهٔ صرفی g رادر انتخاب کلمه مقصدمفهوم a تخمین بزنیم.

جدول 4 احتمالهای بدست آمده را نشان می دهد وبا استفادهٔ فرکانس ها در جدول 4 و معادلات میتوان فرکانس $n(\sin - ta) = \frac{1+5+2}{3} \times 1.0 + \frac{0}{1} \times 0.3 = 2.667$ را بدست می آوریم و احتمال $p\hat{}(\sin - ta) = 0.748$ را بدست می آوریم سپس ما آستانهٔ دینامیکی (پویایی) مشابه را برای P1/p2 از داگان و اتیای (1994) برای انتخاب بهترین کلمه برای ترجمه استفاده می کنیم. چون $p\hat{}(\sin - ta) / p\hat{}(ssu - ta)$ بزرگتر از مقدار آستانه ای در جمله 2 می باشد ما می توانیم کلمه مقصدمناسب را با استفاده ازاین روش انتخاب کنیم. (33)

3-5-5-1- روش موسوی و دلاورخلفی

در مدل آماری روش موسوی و دلاورخلفی به جای استفاده از مدل بیشترین احتمال[1] از اعداد تصادفی[2] و آزمون نکویی برازش[3] استفاده نمودند. آنها معتقدند که استفاده از اعداد تصادفی برای انتخاب بهترین معادل زبان مقصد برای یک واژه مبهم از زبان مبدأ در سیستم ترجمه ماشینی نتایجی نزدیکتر به واقعیت در بر داشته و نسبت به هنگامی که انتخاب بر اساس محتمل ترین حالت صورت می گیرد دقیق تر و علمی تر است.

فرض کنید جمله ای مانند جمله زیر را به کامپیوتر داده تا آن را ترجمه نماید:

She has a recently published article on biology.

از آنجا که واژه article واژه ای چند معنایی است در دیکشنری انگلیسی به فارسی چند معادل مختلف برای آن پیدا می کنیم از جمله آنها "مقاله ، ماده و حرف تعریف". حتی اگر کامپیوتر بتواند با استفاده از روشهای مختلف برچسب گذاری اجزاء کلام طبقه ی دستوری این واژه را نیز مشخص کند باز هم ابهام به قوت خود

[1]- Most likelihood Model
[2]- random numbers
[3]- goodness of fit test

باقیست چراکه هر سه این معادل ها متعلق به طبقه ی دستوری یکسانی هستند، یعنی طبقه اسم. در اینجاست که ابهام زدایی معنی واژه (word sense disambiguation) معنا پیدا می کند.

اساس این روش بر این است که با توجه به جدول شماره 1 در مقاله نزدیک ترین اسم، صفت ، یا فعل به اسم مورد نظربوسیله کامپیوتر تشخیص داده شده و معنی یا معناهای مختلف آن مشخص می شود. سپس فراوانیهای هم رخدادهای مختلف این دو واژه در پیکره یک زبانه زبان فارسی محاسبه می گردد. آز آنجا که واژه published چندمعنایی نیست و واژه article دارای سه معنای مختلف می باشد تعداد هم رخدادهایی که باید محاسبه شوند $3\times1=3$ خواهد بود به شکل جدول 4 می باشد.

جدول4. فراوانی هم رخداد ها در پیکره

فراوانی در پیکره	هم رخداد ها
X	**۱- مقاله چاپ شده**
Y	**۲- ماده چاپ شده**
Z	**۳- حرف تعریف چاپ شده**

با استفاده از فرمول:

$$P(tw_i) = \frac{f_i}{\sum_{j=1}^{n} f_i}$$

رابطه 4-1- احتمال وقوع هم رخدادها

و با استفاده از جدول توزیع تجربی و آزمونهای نکویی برازش توزیع احتمال هندسی X،Y و Z را بدست آورده و در نهایت برای هر گزینه یک احتمال وقوع مشخص می شود. در مدل بیشترین احتمال همیشه گزینه ای با بیشترین احتمال به عنوان گزینه درست انتخاب می شود. اما در مدل ما (اعداد تصادفی) این شانس به گزینه هایی

با احتمالات کمتر هم داده می شود تا با احتمال کمتری انتخاب شوند. در مثال بالا اگر احتمال وقوع 1 (مقاله چاپ شده) بیشتر از همه مثلاً 87 درصد باشد و احتمال وقوع 2 (ماده چاپ شده) مثلاً 7 درصد باشد ، نباید در همه موارد 1 بعنوان گزینه درست انتخاب شود بلکه باید این امکان را به 2 نیز داد که در 9 درصد موارد انتخاب شود.

این روش در مورد پیکره زبانی نسبتاً بزرگی در حوضه روانشناسی در زبانهای انگلیسی و فارسی به ترتیب به عنوان زبانهای مبدأ و مقصد آزمایش شد. الگوریتم این روش توانست مشکل ابهام 604 واژه از 764 واژه مبهم یک پیکره زبانی انگلیسی را با استفاده از متن معادل آن در زبان فارسی حل نماید که بدین ترتیب صحت این مدل به 79% رسید (12).

فصل دوم: کشف دانش

2-1- مقدمه ای بر کشف دانش

انسان همیشه برای الهام گرفتن به جهان زنده پیرامون خود نگریسته است. مشاهده ساختارهای اجتماعی آنان نتیجه مبادلات چند جانبه ای است که به ظاهر ساده اند. این مبادلات به کمک ارتباطهای شیمیایی صورت می گیرد که مورچه ها با شاخکهای خود این علائم را دریافت می کنند. فرومون ماده ای شیمیایی است که مورچه پس از عبور از خود برجا می گذارد.آزمایشی ساده این پدیده را ثابت می کند. پژوهشگران با قراردادن دو مسیر کوتاه و بلند در برابر مورچه ها برای رسیدن به غذا، پی بردند پس از مدتی همه مورچه ها از مسیر کوتاه تر عبور می کنند که علت آن تمرکزبیشترفرومون درآن مسیراست.هر چه میزان تمرکز فرومون در مسیر کوتاه تر بیشتر باشد میزان تردد در آن بیشتر می شود.مورچه ها در طبیعت با این ماده می توانند کوتاه ترین مسیر را بین لانه خود و چندین منبع غذایی پیدا کنند.الگوریتم کلونی مورچه برای اولین بار توسط دوریگو [1] و همکارانش به عنوان یک راه حل چند عامله[2] برای مسائل مشکل بهینه سازی مثل فروشنده دوره گرد[3] ارائه شد. عامل هوشند[4] موجودی

[1] Dorigo
[2] Multi Agent
[3] TSP:Traveling Sales Person
[4] Intelligent Agent

است که از طریق حسگرها قادر به درک پیرامون خود بوده و از طریق تاثیر گذارنده ها می تواند روی محیط تاثیر بگذارد. الگوریتم کلونی مورچه الهام گرفته شده از مطالعات و مشاهدات روی کلونی مورچه هاست. این مطالعات نشان داده که مورچه ها حشراتی اجتماعی هستند که در کلونی ها زندگی می کنند و رفتار آنها بیشتر در جهت بقاء کلونی است تا درجهت بقاء یک جزء از آن. یکی از مهمترین و جالبترین رفتار مورچه ها، رفتار آنها برای یافتن غذا است و بویژه چگونگی پیدا کردن کوتاهترین مسیر میان منابع غذایی و آشیانه. این نوع رفتار مورچه ها دارای نوعی هوشمندی توده ای است که اخیراً مورد توجه دانشمندان قرار گرفته است.

همانطور که می دانیم مسئله یافتن کوتاهترین مسیر، یک مسئله بهینه سازیست که گاه حل آن بسیار دشوار است و گاه نیز بسیار زمانبر. بعنوان مثال مسئله فروشنده دوره گرد(TSP). در این مسئله فروشنده دوره گرد باید از یک شهر شروع کرده، به شهرهای دیگر برود و سپس به شهر مبدا بازگردد بطوریکه از هر شهر فقط یکبار عبور کند و کوتاهترین مسیر را نیز طی کرده باشد. اگر تعداد این شهرها n باشد در حالت کلی این مسئله از مرتبه ! (n-1) است که برای فقط 21 شهر زمان واقعاً زیادی می برد:

۲۰! = ۲/۴۳۳*۱۰^۱۸*۱۰ms = ۲/۴۳۳*۱۰^۱۶ S = ۱/۷*۱۰^۱۳ روز

با انجام یک الگوریتم برنامه سازی پویا برای این مسئله ، زمان از مرتبه نمایی بدست می آید که آن هم مناسب نیست. البته الگوریتم های دیگری نیز ارائه شده ولی هیچ کدام کارایی مناسبی ندارند. ACO الگوریتم کامل و مناسبی برای حل مسئله TSP است.

2-2- بررسی موضوع و مسائل پیرامون آن

در این فصل ابتدا به منظور پیدا کردن درک یکسانی از مطالبی که در ادامه می آید به توضیح یک تعریف اصلی پرداخته می شود. سپس مسئله اصلی و هدف مورد نظر در این پروژه بیان می گردد.

2-3- تعاریف اصلی

2-3-1-طبقه بندی چند هدفه و چند طبقه بودن[1]

در صورتیکه نمونه هایی که برای آموزش بکار می روند ، دارای n صفت ومقدار باشند که در واقع از این مقادیر برای تقسیم طبقه ا استفاده می شود ($m>1$ و $n>0$) ، در این صورت فعالیت طبقه بندی روی این مجموعه ی آموزشی را طبقه بندی چند هدفه می گوییم.(34)

2-3-2-کشف دانش در پایگاه داده ها[2]

بعنوان استخراج اطلاعاتی مهم از میان داده ها، بطوریکه این اطلاعات قبلاً ناشناخته بوده و بطور بالقوه مفید باشند (17)تعریف گردیده است. تعریف متداول دیگری که برای کشف دانش در پایگاه داده آمده ،عبارتست از :پردازش نه چندان ساده تشخیص الگوهایی نهایتاً قابل فهم در داده ها ، بطوریکه معتبر، تازه و بالقوه مفید باشند.(35)

2-3-3- داده کاوی[3]

کاربرد الگوهایی خاص برای استخراج الگو از میان داده ها تعریف گردیده است. بعبارت دیگر داده کاوی گامی از گامهای موجود در پردازش عمومی به منظور کشف دانش در پایگاه داده محسوب می گردد. (35)

معادل های دیگری هم برای داده کاوی وجود دارد ، منجمله: برداشت دانش[4] ، کاوش دانش[5] ، استخراج دانش[6] ، و باستانشناسی داده[7]. (36)(35)

2-3-4- عامل[8]

[1] Multi Objective Classification
[2] Knowledge Discovery in Data Base(KDD)
[3] Data-Mining
[4] Information Harvesting
[5] Knowledge mining
[6] Knowledge extraction
[7] Data Archaeology
[8] Agent

موجودیتی خودگردان و محاسباتی که بتوان آنرا بعنوان ادراک کننده محیطش از طریق احساسگرها[1] و تأثیر گذار بر روی محیطش از طریق وسایل تأثیر دانست (37) . یا سیستم کامپیوتری که در محیطی قرار گرفته و قادر به رفتار و فعالیتی خودگردان در آن محیط به منظور نیل به اهداف مورد طراحی اش باشد(38).

2-4- هدف پروژه

هدف از این پروژه بکارگیری الگوریتم یادگیری ماشین بر اساس الگوریتم بهینه سازی اجتماع مورچه ها[2] به منظور ابهام زدایی از واژه های چند معنایی است.

2-5- مروری بر کشف دانش در پایگاه داده ها

از آنجا که در این پایان نامه می خواهیم با استفاده از یکی از نمونه های موجود از الگوریتم های کشف دانش به ابهام زدایی از معنی واژه بپردازیم ، لازم است ابتدا مروری بر کشف دانش در پایگاه داده ها داشته باشیم و بطور مختصر به بررسی دلائل و فناوری های موجود در این زمینه بپردازیم.

2-5-1- دلائل توجه

دلائل توجه به کشف دانش در پایگاه داده ها را می توان در عوامل زیر خلاصه نمود: (36)(35)

- تعداد و اندازه پایگاه های داده در سازمانها با سرعت بسیار بالایی در حال رشد است. زمانی پتابایت و ترابایت ، حجمهای غیر قابل تصور محسوب می شدند در صورتیکه اکنون در حوزه های متفاوتی مانند فروش و حراج ، علوم زمین شناسی، مراقبت های بهداشتی وبیولوژی مولکولی (مانند پروژه ژنوم شناسی)و... به واقعیت تبدیل گردیده اند.
- سازمانها متوجه وجود دانش هایی با ارزش بالا در میان داده هایشان گردیده اند که در صورت کشف،مزیت رقابتی برایشان دربر خواهد داشت.
- بعضی از فناوری های مورد نیاز که توانایی عمل بر روی حجم های بالای داده ها را داشته باشند اخیراً به اندازه کافی رشد کرده اند.

[1] Sensors
[2] Ant Colony Optimization(ACO)

1-5-2- فناوری های کلیدی مورد استفاده

جستجوی دانش در پایگاه داده ، زمینه ای مربوط به رشته ها و فناوری های مختلف علمی است که هسته آن محل تلاقی یادگیری ماشین ، آمار و پایگاه داده هاست.(39) فناوری هایی که پیشرفت آنها کاربرد جستجوی دانش در پایگاه داده ها را امکان پذیر ساخته عبارتند از: (36)

- مدیریت پایگاه های داده
- هوش مصنوعی
- آمار
- یادگیری ماشین
- شناسایی الگو
- عاملهای هوشمند

فصل سوم: الگوریتم بهینه سازی

3-1-کشف دانش بوسیله الگوریتم بهینه سازی

برای کشف دانش گامهای متعددی بایستی صورت گیرد و در هر کدام چندین تکنیک برای اجرا موجود است. بر همین اساس، در گام مربوط به داده کاوی و شناسایی الگو از تکنیکهای متفاوتی می توان استفاده نمود که بعضی از آنها شامل طبقه بندی ، خوشه بندی و مدل سازی مرتبط[1]، تقلیل[2] و ...مورد اشاره قرار گرفته است.

هر کدام از این تکنیکها خود می توانند بعنوان وظیفه[3] و مسئله ای جداگانه برای حل توسط الگوریتم کشف دانش دیده می شود (35). بنابراین اولین گام برای طراحی الگوریتم کشف دانش ، مشخص ساختن وظیفه ایست که الگوریتم می خواهد آنرا اجرا نماید.

3-2- فعالیت کلاس بندی[4]

هدف از این فعالیت تخصیص هر مورد[5] از اشیاء، رکوردها و یا نمونه ها[6] به یک کلاس از میان مجموعه ای از طبقه های از پیش معین شده است . این فعالیت بر اساس مقادیری از بعضی صفات مشخصه (بنام صفات مشخصه پیشگو) برای آن مورد انجام می گیرد. دانشی که نهایتاً توسط کلاس بندی کشف می گردد و در ادامه این بحث مورد اشاره می باشند به شکل قوانین تصمیم بیان می گردد. شکل کلی این بیان دانش به صورت زیر می باشد:

[1] Dependence modeling
[2] Regression
[3] task
[4] Classification
[5] Case
[6] Instance

IF <conditions> THEN <class>

مقدم قانون[1] ، شامل مجموعه ای از شرط هاست که معمولاً بوسیله ی عملگر منطقی عطف[2] بهم متصل گردیده اند . در ادامه این متن ما هر شرط را بعنوان یک عبارت[3] مورد اشاره قرار می دهیم. بنابراین مقدم قانون، حاصل عطف منطقی عبارت هایی به شکل زیر است:

If term1 AND term2 AND…

هر عبارت متشکل از یک سه تایی <صفت مشخصه، عملگر، مقدار> است.مانند <جنسیت = مونث >. بخش پی آمد[4] ، مشخص کننده کلاس پیش بینی شده برای مواردی است که صفات پیشگو ، همه عبارت ها ی مشخص شده در مقدم قانون را برآورده نمایند. از دید کشف دانش ، این نحوه بازنمایی دانش این مزیت را دارد که توسط کاربر قابل درک است . البته این قابلیت درک تا زمانی است که قوانین کشف شده و تعداد عبارتهای موجود در مقدم قوانین طولانی نباشند.

3-3- مزیتهای الگوریتم بهینه سازی

همانطور که گفته شد «تبخیر شدن فرومون» و «احتمال-تصادف» به مورچه ها امکان پیدا کردن کوتاهترین مسیر را می دهند. این دو ویژگی باعث ایجاد انعطاف در حل هرگونه مسئله بهینه سازی می شوند. مثلا در گراف شهرهای مسئله فروشنده دوره گرد، اگر یکی از یالها (یا گره ها) حذف شود الگوریتم این توانایی را دارد تا به سرعت مسیر بهینه را با توجه به شرایط جدید پیدا کند. به این ترتیب که اگر یال (یا گره ای) حذف شود دیگر لازم نیست که الگوریتم از ابتدا مسئله را حل کند بلکه از جایی که مسئله حل شده تا محل حذف یال (یا گره) هنوز بهترین مسیر را داریم، از این به بعد مورچه ها می توانند پس از مدت کوتاهی مسیر بهینه(کوتاهترین) را بیابند.

[1] If Part
[2] And
[3] Term
[4] Then Part

3-3-1- الگوریتم بهینه سازی

از آنجا که این الگوریتم الهام گرفته از فعالیت مورچه های واقعی در محیط طبیعی است ابتدا به بررسی رفتار مورچه های واقعی می پردازیم و بعد از آن به بیان الگوریتم می پردازیم.

3-3-2- حل مسئله

در یک کولونی از حشرات اجتماعی مانند مورچه ها ، زنبورها و موریانه ها هر حشره بطور معمول وظائف خاص خود را بصورت مجزا از سایر اعضاء کولونی انجام میدهد ، اما فعالیتهای انجام شده بوسیله ی هر حشره به شکلی با سایراعضاءآن کولونی مرتبط می باشد بطوریکه در مجموع کولونی ایجاد شده توانایی و ظرفیت حل مسائل پیچیده را از طریق همکاری با هم دارا هستند (40).

فعالیتهای مهم و حیاتی مانند انتخاب و جمع آوری مصالح مورد نیاز ، پیدا کردن و ذخیره سازی غذا که نیاز به طرح ریزی[1] پیچیده ای دارند ،همگی بوسیله ی کولونی های حشرات و بدون هیچ گونه نظارت خارجی و یا کنترل کننده مرکزی حل می گردند.این رفتار انتخابی که در حشرات و پرندگان اجتماعی دیده می شود تحت عنوان "هوش جمعی " شناخته می شود.

آنها هنگام انتخاب بین دو مسیر بصورت احتمالاتی مسیری را انتخاب می کنند که فرومون بیشتری داشته باشد یا بعبارت دیگر مورچه های بیشتری قبلاً از آن عبور کرده باشند. حال دقت کنید که همین یک تمهید ساده چگونه منجر به پیدا کردن کوتاهترین مسیر خواهد شد .

فرض کنید مورچه ها روی مسیر AB در حرکت اند (در دو جهت مخالف) اگر در مسیر مورچه ها مانعی قرار دهیم مورچه ها دو راه برای انتخاب کردن دارند. اولین مورچه از A می آید و به C می رسد، در مسیر هیچ فرومونی نمی بیند بنابر این برای مسیر چپ و راست احتمال یکسان می دهد و بطور تصادفی و احتمالاتی مسیر CED را انتخاب می کند. اولین مورچه ای که مورچه اول را دنبال می کند زودتر از مورچه اولی که از مسیر CFD رفته به مقصد می رسد. مورچه ها در حال برگشت و به مرور زمان یک اثر بیشتر فرومون را روی CED حس می کنند و

[1] Planning

آنرا بطور احتمالی و تصادفی (نه حتماً و قطعاً) انتخاب می کنند. در نهایت مسیر CED بعنوان مسیر کوتاهتر برگزیده می شود. در حقیقت چون طول مسیر CED کوتاهتر است زمان رفت و برگشت از آن هم کمتر می شود و در نتیجه مورچه های بیشتری نسبت به مسیر دیگر آنرا طی خواهند کرد چون فرومون بیشتری در آن وجود دارد. در تصویر سمت چپ مورچه ها را در حال حرکت در مسیری مستقیم از لانه به سمت منبع غذایی مشاهده می شوند . تصویر سمت راست بیانگر چیزی است که در طبیعت بطور معمول اتفاق می افتد ، وآن قرارگیری یک شئ (مانع) در سر راه مورچه هاست . بمنظور دور زدن این مانع ، مورچه ها در ابتدا بصورت اتفاقی به چپ یا راست می پیچند (50%-50%).

نکته بسیار با اهمیت این است که هر چند احتمال انتخاب مسیر پر فرومون تر توسط مورچه ها بیشتر است ولی این کماکان،یک احتمال است و قطعیت ندارد. یعنی اگر مسیر CED پرفرومون تر از CFD باشد به هیچ عنوان نمی شود نتیجه گرفت که همه مورچه ها از مسیرCED عبور خواهند کرد بلکه تنها می توان گفت که مثلا 90% مورچه ها از مسیر کوتاهتر عبور خواهند کرد. اگر فرض کنیم که بجای این احتمال قطعیت وجود می داشت، یعنی هر مورچه فقط و فقط مسیر پرفرومون تر را انتخاب میکرد آنگاه اساساً این روش ممکن نبود به جواب برسد. اگر تصادفاً اولین مورچه مسیر CFD (مسیر دورتر) را انتخاب می کرد و ردی از فرومون بر جای می گذاشت آنگاه همه مورچه ها بدنبال او حرکت می کردند و هیچ وقت کوتاهترین مسیر یافته نمی شد. بنابراین تصادف و احتمال نقش عمده ای در ACO بر عهده دارند.

نکته دیگر مسئله تبخیر شدن فرومون بر جای گذاشته شده است. برفرض اگر مانع در مسیر AB برداشته شود و فرومون تبخیر نشود مورچه ها همان مسیر قبلی را طی خواهند کرد. ولی در حقیقت این طور نیست. تبخیر شدن فرومون و احتمال، به مورچه ها امکان پیدا کردن مسیر کوتاهتر جدید را می دهند. رفتار مورچه های واقعی را از میان حشرات دیگر و بخصوص توانایی آنها را در پیدا کردن کوتاهترین مسیر بین منبع غذایی و لانه شان[1] در محیط هایی که دائم در حال تغییرندو بدون داشتن دانش بصری ، مورد بررسی قرار می دهیم تا با الهام از آنها در محیط های کامپیوتری مسائل مورد نظر در زمینه ی کشف دانش حل گردد.(29)

[1] Nest

در هنگامی که مسیری بین منبع غذایی و لانه مورچه موجود باشد و مورچه ها هم در اطراف یک شئ موجود حاضر باشند ، آنها خیلی زود سعی در دور زدن آن مانع می کنند .در ابتدا هر مورچه راهی را به سمت چپ یا راست مانع با توزیع احتمال مساوی 50% -50% انتخاب می کند . از آنجا که همه مورچه ها با سرعتی نسبتاً یکسان حرکت می کنند و از خود مسیرِفرومونی با نرخی ثابت از نظر میزان تولید فرومون باقی می گذارند ، در نتیجه آن مورچه هایی که بصورت اتفاقی مسیر کوتاهتر منتهی به غذا را طی می کنند و در نتیجه سریعتر مسیرشان را نسبت به سایر مورچه ها می پیمایند ودر نتیجه میزان فرومون هم در مسیر کوتاهتر سریعنر انباشته می گردد.چون مورچه ها تعقیب مسیرهایی با میزان فرومون بیشتر را ترجیح می دهند بنابراین بیشتر مورچه ها به مسیر کوتاهتر همگرا می شوند .

4-3-الگوریتم های بهینه سازی

یک الگوریتم بهینه سازی عموماً سیستمی بشکل عامل پایه[1] است که رفتار مورچه های طبیعی را شبیه سازی می کند و شامل مکانیزم هایی برای همکاری [2] و تطبیق [3] می باشد .در مراجع (3) و(2) از این نوع سیستم بعنوان روشی فوق ابتکاری [4] بمنظور حل مسائل ترکیبی [5] بهینه سازی ارائه گردیده است.

دو نوع کلی از الگوریتم بهینه سازی اجتماع مورچه ای وجود داردکه عبارتند از : سیستم کولونی مورچه ای [6] وسیستم مورچه ای[7] ، و در ادامه هرجا که سخن از ACO پیش می آید ، منظور سیستم کولونی مورچه ای است .

اثبات شده است که این روش فوق ابتکاری ، هم مقاوم در برابر خطا[8] و هم دارای تنوع[9] است و به همین دلیل نیز بر روی محدوده ی وسیعی از مسائل بهینه سازی ترکیبی متفاوت ، بطور موفق بکار رفته است(4).الگوریتم های بهینه سازی مبتنی بر کولونی مورچه ، مبتنی بر ایده های اصلی زیر هستند:

[1] Agent base
[2] Cooperation
[3] Adaptation
[4] Meta-heuristic
[5] Combinatorial
[6] Ant Colony System (ACS)
[7] Ant System
[8] Robust
[9] Versatile

- هر مسیر طی شده بوسیله ی یک مورچه برای یک مسئله ی فرضی با راه حلّ کاندیدی برای آن مسئله در ارتباط است .
- وقتیکه یک مورچه راهی را طی می کند ، میزان فرومون باقی گذاشته شده در آن مسیر با کیفیت راه حلّ کاندید متناظربرای آن مسئله (هدف) خاص متناسب خواهد بود.
- هنگامیکه مورچه ای بین دو مسیر یا بیشتر ، قرار است مسیری را انتخاب می کند ، مسیر با مقدار فرومون بیشتر با احتمال بیشتری بوسیله ی آن مورچه مورد انتخاب می گیرد.

در نتیجه ، نهایتاً مورچه ها به مسیر کوتاهتر همگرا خواهند شد ، که در این حالت مانند مورچه های واقعی برای بهینه بودن و یا نزدیک به بهینه[1] بودن راه حل یافته شده برای مسئله ی هدف ، امید زیادی وجود دارد .

مراحل طراحی و تطبیق الگوریتم ACO برای هر مسئله به شرح زیر است:(20)

- نمایش درست و مناسبی از مسئله ی مورد نظر، ودرصورت امکان نمایش به شکل گراف و یا ساختاری مشابه آن که بتواند بسادگی توسط مورچه ها پوشش داده شود ، مناسب خواهد بود.
- انتساب تابع هیوریستیک به هر کدام از انتخابهای مورچه در گام تولید یک راه حل
- روشی برای اطمینان پیدا کردن از ساخته شدنِ راه حل های معتبر
- در نظر گرفتن قانونی برای بروز رسانی مقدار فرومون (t)
- پایه گذاری راهی برای مقدار دهی اولیه ی فرومون
- قانون انتقال [2] مبتنی بر احتمالات بر اساس تابع هیوریستیک (h) ومقدار فرومون (t)، که بایستی مکرراً در ساخت قوانین بکار رود.

3-5- استفاده از الگوریتم در حل مسئله

برای پیاده سازی کامپیوتری رفتار مورچه های واقعی و تطبیق آن با مسئله ی کشف و استخراج دانش ، لازم است که هر مورچه را بعنوان یک عامل در نظر بگیریم . بنابراین هر مورچه ی مصنوعی بعنوان عاملی که رفتار مورچه ی واقعی را تقلید می کند ، تعریف می گردد(41)

[1] Near-Optimum
[2] Transition Rule

البته مورچه های مصنوعی و طبیعی دارای وجوه تمایز زیر هستند(42) :

- مورچه های مصنوعی دارای حافظه هستند.
- مورچه های مصنوعی کاملاً کور نیستند.
- مورچه های مصنوعی در محیطی که زمان گسسته دارد ،زندگی می کنند.

با توجه به مطالب فوق مورچه های مصنوعی با مورچه های طبیعی دارای اشتراک هایی نیز میباشند که عبارتند از:

- مورچه های مصنوعی با بکارگیری یک روش مبتنی بر احتمال ، مسیر های با مقدار فرومون بیشتر را، به مسیر های دیگر ترجیح می دهند .
- مسیر های کوتاهتر ، دارای نرخ بالاتری از نظر میزان رشد فرومون می باشند.
- مورچه ها از سیستمی غیر مستقیم برای برقراری ارتباط پایه ی میزان فرومون باقی گذاشته در هر کدام از مسیر ها را استفاده می کنند.

بنابراین یک سیستم کولونی مورچه ای (ACS) بصورت تکراری یک حلقه ی اصلی را انجام می دهد که در آن دو روال را فراخوانی می کند:

1. روال ساخت : که چگونگی ساخت و تغییر راه حل های مسئله ی در حال حل را مشخص می سازد.
2. روال بروز رسانی فرومونها در مسیر های مختلف : که ارتباط غیر مستقیم بین مورچه هارا فراهم می نماید .

6-3- الگوریتم و جزئیات یادگیری

در الگوریتم ACO هر مورچه بصورت افزایشی [1] راه حلی را برای مسئله ی هدف می سازدو تغییر می دهد. در مورد کشف دانش ، مسئله ی هدف ، کشف قوانین کلاس بندی است. هر قانون کلاس بندی شکل زیر را دارد:

If <term1 AND term2 AND …> THEN <Class>

هر عبارت ، سه تایی < صفت مشخصه ، عملگر، مقدار> است.

[1] Incremental

(<value, operator, attribute>)که value ی آن مقداری است که متعلق به دامنه ی صفت مشخصه . المان عملگر در هر سه تایی یکی از عملگر های نسبی[1] می تواند باشد.

نسخه ای از الگوریتم کشف دانش بوسیله ی کولونی مورچه ها که در ادامه مورد بررسی قرار می گیرد برای کار با صفات تطبیق داده شده است و بنابراین عضو عملگر در سه تایی فوق همواره "=" خواهد بود.

3-7- سیستم یادگیری

در این قسمت سیستم مورچه معرفی شده و از مسئله ی فروشنده دوره گرد بعنوان معیار استفاده می گردد. در مسئله فروشنده دوره گرد ، یک فروشنده سفر خود را از یک شهر آغاز کرده و پس از یک سفر کامل دوباره به شهر خود باز می گردد و از هر شهر فقط یکبار عبور می کند و در ضمن باید از همه ی شهرها هم عبور کند . هدف یافتن کوتاهترین مسیر برای این سفر می باشد .

d تعداد شهرها می باشد و dij فاصله اقلیدسی بین دو شهر i و j برابر است با:

$$dij = x_i - x_j)^2 + (y_i - y_j)^2]^{0.5} \quad [($$

رابطه 1-3- dij فاصله اقلیدسی بین دو شهر i و j

[1] Relational operator

یک نمونه از TSP بوسیله ی یک گراف (N,E) که N مجموعه ی شهرها و E مجموعه ی یالها (فاصله های) میان شهر ها می باشد .الگوریتم مورچه برای حل مسئله ی فروشنده ی دوره گرد بکار رفته است .

b_i (t) , i= 1,2,3,…,d را تعداد جمعیت (مورچه ها)در شهر i – اُم در زمان t و $m\sum_{i=1}^{d} b_i(t)$ کل جمعیت مورچه ها می باشد . هر مورچه یک نماینده ی ساده با ویژگیهای زیر است:

1. آن یک شهررا برای رفتن انتخاب می کند که تابعی از فاصله شهر و مقدار اثر فرومون موجود در آن مسیر می باشد.
2. برای وادار کردن مورچه ها جهت انجام سفرهای منطقی ، سفر به شهر هایی که یکبار از آن عبور کرده اند ممنوع می باشد .
3. هنگامی که یک مورچه یک سفر کامل را انجام می دهد مقداری فرومون بر روی هر مسیر i و j باقی می گذارد.

tij ،شدت اثر فرومون در مسیرهای (i,j) در زمان t می باشد . هر مورچه در زمان t شهر بعدی را انتخاب می کند که در زمان t+1 در آن خواهد بود . بنابراین اگر یک تکرار از الگوریتم سیستم مورچه m حرکت انجام شده بوسیله m مورچه در فاصله زمانی (t,t+1) در نظر گرفته شود ، بعد از d تکرار الگوریتم (یک سیکل) هر مورچه یک سفر کامل انجام داده است . در این نقطه شدت اثر فرومون بر اساس رابطه ی $tij(t+1) = r.tij(t) + \Delta til$ محاسبه می شود و r یک ضریب است بطوریکه ($r-1$) میزان تبخیر فرومون را در فاصله زمانی (t,t+1) را نشان می دهد و $\Delta tij = \sum_{k=1}^{m} \Delta tij^k$ که Δtij^k مقدار اثر فرومون در هر واحد طول می باشد که بر روی مسیر (i,j) بوسیله ی k- اُمین مورچه در فاصله ی (t, t+1) بر جای گذاشته می شود . قابل ذکر است که سه نوع الگوریتم مورچه وجود دارد :

1. Ant densit
2. Ant quantityy
3. Ant cycle

در دو الگوریتم اول مقدار فرومون در هر تکرار تعدیل و بروز رسانی می شود اما در الگوریتم سوم بعد از یک سیکل عمل تعدیل صورت می گیرد .

در الگوریتم اول داریم :

$$\Delta t_{ij}^{\,k}(t,t+1)=\begin{cases}Q_1 & \text{اگر k اُمین مورچه در فاصله زمانی (t,t+1) از i به j می رود}\\ 0 & \text{در غیر این صورت}\end{cases}$$

که Q_1 یک مقدار ثابت است که در واحد طول در مسیر (i,j) بر جای گذاشته می شود.

در الگوریتم دوم :

$$\Delta t_{ij}^{\,k}(t,t+1)=\begin{cases}Q_2/d_{ij} & \text{اگر k اُمین مورچه در فاصله زمانی (t,t+1) از i به j می رود}\\ 0 & \text{در غیر این صورت}\end{cases}$$

که Q_2 یک مقدار ثابت و dij فاصله ی بین شهر i و j میباشد.

در الگوریتم سوم :

$$\Delta t_{ij}^{\,k}(t,t+1)=\begin{cases}Q_3/L^k & \text{اگر k اُمین مورچه از مسیر های (i,j) در سفرش استفاده کند}\\ 0 & \text{در غیر این صورت}\end{cases}$$

Q_3 مقدار ثابت و dij طول سفر k -اُمین مورچه می باشد.

ضریب باید دارای ارزش کوچکتر از یک باشد تا از انباشتگی نامحدود فرومون جلوگیری کند .(46)

فصل چهارم :کشف دانش چند طبقه

4-1-کشف دانش چند طبقه مبتنی بر بهینه سازی

قبل از پرداختن به کارهای انجام شده توسط این بخش از پروژه ، لازم است که درباره ی دانش چند طبقه و نحوه ی نمایش آن وسایر مطالب مربوط به مقدمات کار به زبانی مشترک دست یابیم. از اینرو ابتدا بخشی جداگانه برای معرفی نحوه ی نمایش دانش چند طبقه آورده می شود .

4-2- دانش چند طبقه و نحوه نمایش آن

همانطور که می دانیم هرف از کشف دانش چند طبقه ،استخراج دانش با طبقه ای چند بعدی است .شکل5 نمایش دهنده ی چندطبقه بودن یک قانون است.

IF $term_{11}$ AND $term_{12}$ …..AND $term_{1M}$ THEN Class 1
IF $term_{21}$ AND $term_{22}$ …..AND $term_{2M}$ THEN Class 2
……..
IF $term_{i1}$ AND $term_{i2}$ …..AND $term_{iM}$ THEN Class i
……..
IF $term_{k1}$ AND $term_{k2}$ …..AND $term_{kM}$ THEN Class k

شکل 5. تعبیر یک قانون چند طبقه

در این شکل هر قانون چند طبقه به معنی آرایه ای از قوانین تک طبقه فرض شده است. مقدار k معادل تعداد طبقه های موجود در مجموعه ی آموزشی است . مقدار M برابر تعداد صفات مشخصه برای هر نمونه آموزشی است.به عبارت ساده یک قانون چند طبقه معادل آرایه ای از قوانین یک طبقه ای فرض شده است.

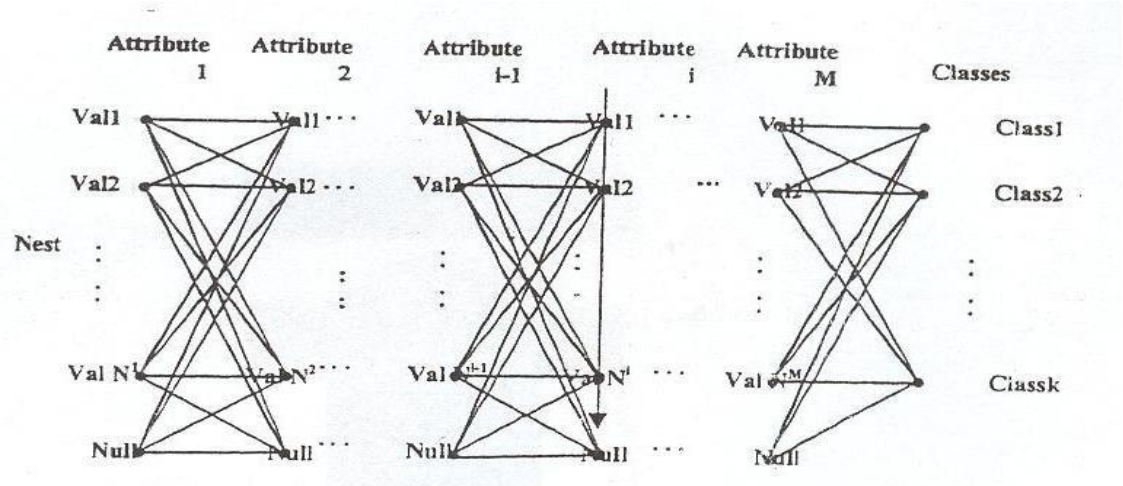
تعدادقوانین یک طبقه ای در هر قانون چند طبقه ، معادل تعداد طبقه های موجود در مجموعه ی آموزشی است.

این ظاهر شدن قوانین از همه ی طبقه ها باعث می شود از داشتن دانشِ مربوط به نمونه هایی که تعدادشان در مجموعه ی آموزشی کم است ؛ اطمینان یابیم. توجه به این نکته لازم است که قوانین یک طبقه موجود در قانون چندطبقه ای نمی توانند بخشهای مقدم معادل هم داشته باشند.

3-4- تحلیل هدف از نقطه نظر الگوریتم بهینه سازی

هدف کشف دانش چند طبقه (پیدا کردن معنی مناسب واژه از بین معانی مختلف واژه)با استفاده از این الگوریتم است. توجه به گامهای استاندارد در ساخت هر الگوریتم جدید مبتنی بر ACO اهمیت زیادی داشته و بر سرعت طراحی می افزاید.

ساده ترین ساختار برای پیمایش توسط مورچه ها ، ساختار گراف است . از اینرو ابتدا مسئله را با این ساختار نمایش می دهیم . شکل 3 در زیر نمایش دهنده فضای مسئله به شکل گراف است .



شکل6. نمایش فضای مسئله با استفاده از گراف

در شکل6 هر گره گراف معادل یک عبارت[1] در نظر گرفته شده و هر بخش عمودی از گره ها (ستون) معادل مجموعه مقادیر متعلق به دامنه ی یک صفت مشخصه است. مقدار Null در دامنه ی هر صفت مشخصه نشان دهنده ی انتخاب نشدن آن صفت مشخصه است. بنابراین هر مسیر از لانه به یک کلاس ، معادل یک قانون یک طبقه ای برای انتخاب آن طبقه می توتند تلقی شود.

توجه به این نکته لازم است که برای جواب قابل قبول قانون چند طبقه ،وجود فقط یک مسیر از لانه به کلاس مورد نظر کافی نیست و بایستی مسیر های به تمام طبقه ها را یافت ؛ یعنی K مسیر در هر بار.

با توضیحات فوق می توان مسئله ی کشف دانش چند طبقه را معادل کشف K قانون یک طبقه ای با فرض وجود محدودیتهای زیر دانست :

(k) قانون یک طبقه ای کشف شده بایستی تمام (k) کلاس مجموعه آموزشی را داشته باشند ، یعنی (k) قانون یک طبقه ای کشف شده بایستی کلاس تکراری در بخش نتیجه قوانین نداشته باشند.

بخش مقدم (k) قانون یک طبقه ای کشف شده نبایستی معادل هم باشند .

از نظر گراف ، راه حل مسئله کشف دانش چند طبقه ، بصورت داشتن k مسیر از لانه به k طبقه (هدف) به صورتی که این مسیر ها تکراری نباشند و اهداف تکراری هم نداشته باشند، تعریف می گردد.حال که توانستیم مسئله را بصورت یک طبقه ای در نظر بگیریم بنابر این می توان از روش مربوط به کشف دانش یک طبقه ای کمک گرفت . البته اینکار بایستی با توجه به محدودیتهای اعمالی فوق صورت گیرد. اما بایستی به این نکته توجه داشت که برای هر بار بدست آوردن دانش یک طبقه ای در فضای ترسیمی شکل 2 ، گراف بایستی پیمایش گردد. پس کشف دانش چند طبقه معادل پیمایش های متوالی و یا موازی به تعداد k بار خواهد بود.

اگر فضای را بصورت k لایه بر روی هم در نظر بگیریم یعنی k-تا از شکل 5 بر روی هم ، می توانیم با تغییراتی اندک از روش گفته شده قبلی برای کشف دانش یک طبقه ای با استفاده از این الگوریتم در کشف دانش چند طبقه ای یاری بگیریم.

[1] Term

البته همانطوری که گفته شد می توان صرفنظر از چند لایه ای شدن ، بر اساس شکل 5 و بصورت متوالی این مسئله را حل کرد.در این حالت مسئله را بصورت راه حل های جداگانه در نظر می گیریم که در هر دور از حل ، یک قانون یک طبقه ای کشف می گردد.ولی این نوع تحلیل با کشف کشف قانون چند طبقه بر اساس توضیح گفته شده قبلی برای چند طبقه ای بودن ، چندان مناسبتی ندارد زیرا برای چند طبقه ای بودن لازمه ی عمل کشف قانون چند طبقه ، ظاهر شدن قانونهای یک طبقه ای از همه یطبقات بصورتی یکجا است . لذا در این پروژه فضای مسئله به شکل k لایه ای در نظر گرفته می شود. k لایه ای بودن از نظر تداخل نداشتن فرومونها برای هر طبقه با طبقه ی دیگر لازم بنظر می رسد در غیر این صورت مسئله را باید در k بار و بصورت جداگانه حل نمود و برای هر بار حل ، مسیر ها را فرومون دهی اولیه نمود.

سایر گامهای طراحی عبارتند از: انتساب تابع هیورستیک ، پیدا کردن قانون مناسبی برای بروز رسانی فرومون (t)و...

بدلیل آنکه می خواهیم از روش ACO در کشف دانش یک طبقه ای استفاده نماییم لذا بایستی نمایش فضای مسئله را بصورت یک طبقه ای تبدیل نماییم . این تبدیل از دو جنبه حائز اهمیت است :

1. از حیث انطباق بیشتر مسئله
2. درک بهتر مسئله کشف دانش چند طبقه ، زیرا ترسیم و تصور گراف شکل 5 بصورت k لایه بر روی هم چندان هم ساده نیست.

برای انجام دادن تبدیل ،بایستی هدف از چند لایه ای شدن را تحلیل نمود . تبدیل شکل بالا به حالتی k- لایه ای باعث می شود که بتوان مورچه هایی را تصور نمود که در لایه های جداگانه ، بطور موازی برای رسیدن به اهداف مختلف (k- هدف) فعالیت داشته باشند . این فعالیت بر روی لایه های جداگانه باعث می شود که هر گروه از مورچه در هر لایه ، را بصورت عواملی متخصص در رسیدن به هدف خاص آن لایه در نظر گرفت.

پس برای تبدیل فضای مسئله یعنی تبدیل گراف k- لایه ای شکل بالا لازم است که تعریف جدیدی از مورچه داشته باشیم که با دید قبلی ما از مورچه کمی متفاوت است . بعبارت دیگر لازم است مورچه ها را متخصص در نظر بگیریم تا تبدیل فضای مسئله ، بطور کامل درک شود.

4-3-1- گامهای دیگر طراحی

بدلیل آنکه مسئله ی جند طبقه ای به حل مسئله های یک طبقه ای همزمان تبدیل شد، بنابراین بقیه ی گامهای طراحی مانند تابع هیوریستیک ، قانونی برای بروز رسانی فرومون (t) و ...همان هایی خواهند بود که در مسائل یک طبقه ای مطرح میشود که در ادامه و در هنگام توضیح الگوریتم به آن می پردازیم. این تبدیل از دو لحاظ دارای اهمیت است:

1. از حیث انطباق بیشتر مسئله جاری با مسئله حل شده قبلی
2. درک بهتر مسئله کشف دانش چند طبقه ، زیرا ترسیم و تصور گراف شکل 5 بصورت k- لایه بر روی هم چندان ساده نیست.

برای انجام دادن تبدیل، بایستی هدف از چند لایه ای شدن شکل 5 را تحلیل نمود. تبدیل شکل 2 به حالتی k-لایه ای باعث می شود که بتوان مورچه هایی را تصور نمود که در لایه های جداگانه، بطور موازی برای رسیدن به اهداف مختلف (k-هدف) فعالیت داشته باشند.این فعالیت بر روی لایه های جداگانه باعث می شود که هر گروه از مورچه در هر لایه را بصورت عواملی متخصص در رسیدن به هدف خاص آن لایه در نظر گرفت.

پس برای تبدیل فضای مسئله یعنی تبدیل گراف K – لایه ای شکل 5 به شکلی معادل شکل 7 که در الگوریتمهای دیگر استفاده شد، لازم است که تعریف جدیدی از مورچه را داشته باشیم که با دید قبلی که از مورچه داشته ایم ،کمی متفاوت است. بعبارت دیگر لازم است ،مورچه ها را متخصص در نظر بگیریم تا تبدیل فضای مسئله ، کاملاً درک گردد.

4-4- سلسله مراتب در الگوریتم

فرض تخصصی شدن کار مورچه های مصنوعی با فعالیت مورچه ها در طبیعت سازگار است .زیرا همانطور که می دانیم مورچه ها در طبیعت دارای سلسله مراتب و تخصصهای خاص خود هستند. بعنوان مثال مورچه ها را می توان به مورچه های کارگر، مورچه های سرباز و ملکه تقسیم بندی نمود. که البته هر کدام از این دسته ها دارای زمینه ی تخصصی خاص خود هستند.(43)(44)(45). هرچه میزان پیچیدگی و تعداد مورچه های کولونی بیشتر

شود ، تخصصهای مورچه ها هم زیادتر می گردد. تعداد مورچه ها می تواند بین حداقل 30 تا میلیون مورچه را در برگیرد.(44) اینکه آیا نحوه ی ارتباط برای مورچه ها از یک تخصص با مورچه های دیگری از تخصص های دیگر بصورتی متفاوت یا یکسان است ،مطلبی است که به جانور شناسی مربوط است و ما فعلاً با این مسئله کاری نداریم .

مشخصات مورچه های متخصص مصنوعی بصورت زیر است :

1. مورچه های مصنوعی از یک تخصص با مورچه های دیگر از تخصص های دیگر هیچگونه ارتباطی ندارند.
2. روش ارتباطی مورچه ها در یک تخصص بر اساس فرومون و بر اساس قواعد گفته شده می باشد.
3. تخصص مورچه های مصنوعی امری ذاتی است و از ابتدای ایجاد آنها با آنها بوجود می آید . بعبارت دیگر نمی توان تخصص یک مورچه را در خلال زندگی آن تغییر داد.

4-4-1- پی آمد ناشی از تخصصی شدن کار مورچه ها

با فرض تخصصی شدن کار مورچه ها ، هر گروه تخصصی ، بر روی لایه ای جداگانه و مخصوص به تخصصشان کار می کنند. امکان تبادل اطلاعات بین مورچه های متعلق به تخصصهای متفاوت متصور نیست ، زیرا اطلاعات راهنما در ساخت قانون یک کلاس برای ساخت قانون کلاس دیگر مفید نیست . بنابراین جدا نمودن لایه ها در اصل سعی بر جداسازی اطلاعات مبادله شده بین مورچه های هر تخصص را دارد بدون آنکه بر روی هم تأثیر منفی داشته باشند.

پی آمد دیگری که بایستی به آن توجه نمود ثابت ماندن تخصص هر مورچه ی متخصص در دوره زندگی آنست .

4-4-2- تغییر فضای مسئله

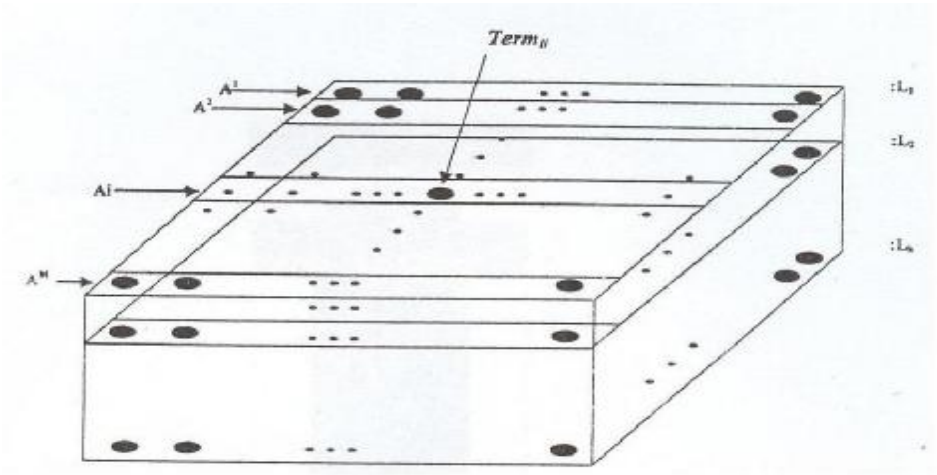
همانطور که گفته شد جدا نمودن لایه ها، در اصل سعی بر جداسازی اطلاعات مبادله شده بین مورچه های هر تخصص را دارد . اطلاعات مبادله شده میان مورچه ها در محاسبه ی احتمال انتسابی به هر عبارت متبلور می شود .

هر احتمال انتسابی حاوی اطلاعاتی از فرومون و تابع هیوریستیک بصورت توأم است . بنابراین بایستی جداسازی و لایه ای نمودن بر روی احتمالهای عبارات صورت گیرد.

اکنون برای چند لایه ای نمودن شکل 2 کافیست k تا جدول شکل 4 را تصور نمود .

در این حالت ، در هر حالت ، در هر لایه ، هر عبارت دارای احتمال انتسابی متفاوتی با لایه های دیگر است . این تفاوت ناشی از تفاوت اطلاعات مبادله شده در هر لایه است .

شکل زیرنمایش جدول خصیصه و مقدار برای گراف شکل بالا را در حالت k- لایه ای نمایش می دهد .بعبارت دیگر فضای مسئله کشف دانش k- طبقه بصورت k- جدول خصیصه و مقدار نمایش داده شده است، تا بتواند حالت حجمی بیابد.



شکل 7. جدول خصیصه و مقدار برای گراف k- لایه ای شکل3

توضیح آرگومانهای شکل 7 بصورت زیر است :

A^i نمایش دهنده ی صفت مشخصه ی I- اُم است .

($term_{ij}$) که معنی آن عبارتی با شرط $A_i = V_{ij}$ است ، که A_i صفت مشخصه i- اُم است و V_{ij} مقدار j- اُم از دامنه A_i .

مقدار L_k نشان دهنده لایه ی k- اُم از جدول است .

مقدار k برابر تعداد طبقه های موجود در مجموعه ی آموزشی است.

بایستی به این نکته توجه داشت که هر لایه از شکل 7 فضای کاری یک دسته از مورچه های متخصص را تشکیل می دهد .یعنی مورچه های متخصص در بدست آوردن کلاس i-اُم در لایه i- اُم مشغول می شوند. همزمان با آنها سایر مورچه ها از سایر تخصصها در حال کار بر روی لایه های متناظر با تخصصشان می باشند .این همزمانی فعالیت مورچه ها از مهارتهای متفاوت برای کسب اهدافی متفاوت در لایه های مختلف به معنی توان پردازشی موازی بالاتر برای الگوریتم کشف دانش چند طبقه مبتنی بر کولونی مورچه هاست.

5-4- بررسی چگونگی انتخاب عبارات و نمایش مسیر های مورچه ها

همانطور یکه در الگوریتم زیر خواهیم دید ، مورچه ها در حلقه ی Repeat-Until فعالیت می کنند و اصلی ترین فعالیت آنها ، ساخت قانون است. این فعالیت به صورت انتخابهای مکرر عبارت ها از میان عبارتهای موجود ی که قبلاً انتخاب نشده اند و اضافه کردن عبارت انتخابی به قانون در حال شکل گرفتن بیان می شود.

شکل زیر نمایش دهنده ی عبارت های محتمل برای انتخاب توسط مورچه هاست . بعبارت دیگر فضای کلیه ی حالات ممکن برای انتخاب توسط مورچه ها را نمایش می دهد:

Attribute1: value1(P_{11}) , value2(P_{12}), …,value $N^1(P_{1N}{}^1)$

Attribute2: value1(P_{21}) , value2(P_{22}), …,value $N^2(P_{2N}{}^2)$

……

……

AttributeM: value1(P_{M1}) , value2(P_{M2}), …,value $N^M(P_{MN}{}^M)$

شکل5. نمایش دهنده عبارت های محتمل

پارامترهای شکل 5 در زیر تعریف گردیده است :

Attribute j : صفت مشخصه j-اُم در قانون.

N^i : تعداد مقادیر دامنه برای صفت مشخصه i-اُم(Attribute j).

Value k : مقدار k-اُم از دامنه ی صفت مورد نظر.

P_{ab} : احتمال انتخاب شدن مقدار b-اُم از صفت مشخصه a-اُم .بعبارت دیگر احتمال انتخاب عبارت $(term_{ab})$ که به شکل $A_a = V_{ab}$ نمایش داده می شود.

توجه به این نکته لازم است که تعداد حالات محتمل برای انتخاب شدن بوسیله ی یک مورچه در ابتدای کار ساخت قوانین ، عددی معادل $(N^1+1)*(N^2+1)*...*(N^M+1)$ (حاصل ضرب تعداد مقادیر دامنه بعلاوه یک برای هر صفت مشخصه) است.

هر مورچه می تواند در هر بار یک مقدار از مقادیر مربوط به یک صفت مشخصه را بردارد، بعد از آن دیگر حق برداشتن هیچ مقداری از آن صفت را ندارد. انتخاب شدن عبارت ها و اضافه شدن آنها به قانون در حال ساخت در مراجع (41)و(42) بعنوان ایجاد مسیر جزئی[1] توسط مورچه ها در نظر گرفته شده است.

6-4- نحوه کار الگوریتم چرخ رولت در انتخاب عبارت ها

حال به بررسی مکانیزم انتخاب عبارت ها در الگوریتم می پردازیم . در روش انتخابی عبارات بر اساس چرخ رولت ، هر عبارت بر اساس احتمال انتسابی به آن ، فضایی از محیط چرخ را به خود اختصاص می دهد . بعد از تقسیم بندی محیط چرخ برای آن عددی تصادفی تولید می کنیم (که معادل چرخاندن چرخ است). اگر عددایجاد شده (شاخص) با هر یک از بخشهای مختلف بوجود آمده بر روی محیط تلاقی کرد ،آنرا بعنوان عبارت مورد نظر انتخاب می کنیم .

[1] Partial path

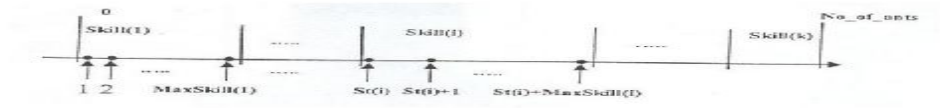
مزیت این روش در سادگی پیاده سازی و رعایت عدالت است .یعنی این روش به عبارتهای مختلف بر اساس میزان احتمال انتسابی به آنها ، شانس انتخاب شدن می دهد .

7-4- الگوریتم کشف دانش چند طبقه مبتنی بر ACO

با توجه به مطالبی که در رابطه با دانش چند طبقه و معادل بودن ان با کشف چندین قانون یک طبقه ای گفته شد الگوریتم زیر کشف دانش چند طبقه مبتنی بر الگوریتم بهینه سازی کولونی مورچه ها را نمایش می دهد.

قبل از پرداختن به الگوریتم لازم است نحوه ی دسترسی به هر مورچه متخصص را مشخص نماییم .

هر مورچه در کولونی با اندیس منحصر بفرد و مربوط به خود شناسایی می گردد . هر مهارت از مورچه ها از شماره اندیس خاصی در کولونی شروع می شود و به شماره اندیسی در کولونی ختم می گردد.شماره مورچه های متخصص از یک مهارت بصورت متوالی است . به شکل8 توجه نمایید:



شکل 8. محیط چرخ رولت برابر عدد یک است. احتمالات مربوط به عبارت ها ، محیط چرخ رولت را به بخشهایی تقسیم نموده است. عددی اتفاقی یکی از بخشها (segment) روی این محیط را انتخاب کرده است. این عدد اتفاقی مانند یک شاخص بر روی محیط چرخ رولت عمل می کند.

در این شکل k تخصص و مهارت جداگانه وجود دارد . مقدار k مساوی است با تعداد طبقه های موجود در مجموعه ی آموزشی . تخصص i-اُم از این کولونی در میانه ی جمعیت در شکل 4 مشخص شده است . St(i) اندیس اولین مورچه در مهارت i- اُم است.MaxSkill(i) حداکثر مورچه های ممکن متصور برای مهارت i-اُم را نمایش می دهد . اندیس آخرین مورچه از تخصص i-اُم برابر با(St(i) + MaxSkill(i خواهد بود.

درادامه الگوریتم کشف دانش چندطبقه ی مبتنی بربهینه سازی کولونی مورچه ای را در زیر می آوریم (47)

1. TrainingSet = {all training cases};
2. Initialize population of each skill in colony;
3. DiscoveredRuleList = []; /*rule list is initialized with an empty list*/
4. While (TrainingSet > Max_uncovered_cases)
5. j = 1; /* convergence test index */
6. s = 1; /* Skill index */
7. Initialize all trails with the same amount of pheromone;
8. MultiClassRule = []; /*a multi class rule buffer is initialized with an empty list */
9. Evaluate maximum skills(i.e. MaxSkill) of TrainingSet;
10. While (s<MaxSkill)
11. Calculate starting ant index (St(s)) for s-th skill;
12. t = St(s);
13. Calculate ending ant index (i.e. EndIndex = MaxSkill(s) + St(s) for s-th skill;
14. Repeat
15. Antt starts with an empty rule and incrementally constructs a classification rule Rt by adding one term at a time to the current rule;
16. Prune rule Rt;
17. Update the pheromone of all trails by increasing pheromone in the trail followed by Antt (proportional to the quality of Rt) and decreasing pheromone in the other trails(simulating pheromone evaporation);
18. IF (Rt is equal to Rt-1) /*update convergence test*/
19. Then j = j+1;
20. Else j = 1;

21. End IF
22. t = t+1;
23. Until (t≥ No_of_ants) OR (j≥No-rules_converg) OR (t≥EndIndex)
24. Generalize and Identify intervals for continuous attributes;
25. Updare all partial paths by means of discovered intervals;
29". Choose the best Rbest among all rules Rt constructed by all ants of s;
27. add rule Rbest to MultiClassRule as a single objective rule;
28. s = s+1
29. END While
30. Add single objective rules from MultiClassRule to DiscoveredRuleList;
31. TrainingSet = TrainingSet – {set of cases correctly covered by MultiClassRule};
32. END While

توضیح این روش بصورت خط به خط در زیر آورده شده است:

خط 1 : تمامی نمونه های آموزشی بایستی به مجموعه ی آموزشی انتقال یابد (خواندن مصادیق آموزشی)

خط 2: بایستی جمعیت هر گروه از مهارتهای مورچه های متخصص مشخص گردد. بعبارت دیگر این گام معادل مشخص ساختن ساختار تخصصی یک کولونی است .

خط 3: خالی کردن و آماده سازی مجموعه ای که برای نگه داری دانش ذخیره شده استفاده می گردد تا آمادگی لازم را برای فرایند داده کاوی جدید حاصل شود . دانش استخراجی به شکل لیستی از قوانین نگهداری می گردد.

خط 4: شروع فرآیند کشف دانش بصورت تکراری . بدنه اصلی کشف دانش در بین خط های 4 تا 32 قرار دارد که بصورت یک حلقه while پیاده سازی شده است .که این حلقه قوانین کلاس بندی را کشف می کند . در این الگوریتم می توان دو سطح متمایز را تشخیص داد: سطح مورچه ها (عامل ها) و سطح کولونی (سیستم چند عامله) . قسمتهای داخل حلقه (Repeat –Until) با سطح مورچه ها و راه حل های جزئی آنها از مسئله در ارتباط است وبقیه ی قسمتهای داخلی حلقه ی while با سطح کولونی (سیستم چند عامله) در ارتباط است .

هر قانون کشف شده به لیستی از قوانین کشف شده اضافه می گردد و مصادیق آموزشی که بدرستی توسط این قانون پوشش داده شده اند (یعنی آن مواردی که مقدم قوانین را ارضاء می کند و کلاس پیش بینی شده در بخش پی آمد آن قانون را دارد) از مجموعه ی آموزشی حذف می گردند. این پردازش تا هنگامی که تعداد مصادیق آموزشی پوشش نیافته از حد آستانه ای که توسط کاربر مشخص می گردد و Max_uncovered_cases نامیده می شود ، بیشتر باشد ، بصورت تکراری ادامه می یابد.

هر تکرار از حلقه Repeat-Until (سطح عوامل) متشکل از سه گام است: ساخت قانون ، هرس قانون [1] و بروز رسانی فرومون ، که به جزئیات آن در ادامه خواهیم پرداخت.

خط 5: مقدار دهی اولیه به متغیری خاص جهت آزمون همگرایی مورچه ها.

خط 6: مقدار دهی اولیه به متغیری خاص بعنوان اندیس مهارت و تخصص مورچه ها. متغیرS ،شماره ی تخصص جاری را که در آن فعالیت ساخت قانون صورت می گیرد مشخص می سازد.

خط 7: مقدار دهی اولیه فرومون به تمام فضای حالت و مسیر های ممکن برای مورچه ها در همه ی تخصص ها. در این گام ابتدا جدول ساخته شده ،برای کار توسط مورچه ها مورد مقدار دهی اولیه ی فرومون قرار می گیرد. این مقدار دهی اولیه ی فرومون توسط رابطه ی4-1 صورت می گیرد.

$$t_{ij}(t=0) = \frac{1}{\sum_{i=1}^{a} b_i}$$

رابطه 1-4. تابع مقدار دهی اولیه برای هر $term_{ij}$ در ابتدای شروع بکار t=0

[1] Rule Pruning

در هر لایه بصورت جدا از سایر لایه ها صورت می گیرد (زیرا فرض بر مبادله نکردن اطلاعات بین مهارتهای متفاوت است). سپس بر اساس "میزان اطلاعات" موجود در هر عبارت که در گام پیش پردازش از رابطه ی 1برای هر کدام از مقادیر عبارت ها بدست آمده و فرومون احتمال نسبی به هر عبارت محاسبه می گردد.

خط8: یک قانون چند طبقه بعنوان بافر مقدار دهی اولیه می گردد که بتوان قوانین یک طبقه ای که مورچه ها از هر تخصص بدست می آورند را در خود نگهداری نماید.

خط 9: مقدار دهی اولیه به متغیر MaxSkill ، این متغیر حداکثر مهارتهای موجود را مشخص می سازد . مقداردهی اولیه به این متغیر بر اساس محاسبه تعداد طبقه های موجود در مجموعه ی آموزشی انجام گیرد.

خط10: شروع حلقه While داخلی که نمایش دهنده و کنترل کننده کار بر روی مهارتهای مختلف است . با عوض شدن مقدارs در این حلقه ، لایه کاری و تخصص مورد بررسی روی آن عوض می گردد.

خط11: محاسبه مقدار اندیس اولین مورچه در مهارت s-اُم . این اندیس اولیه St(s) است .

خط12: مقدار دهی اولیه به شمارنده t . این شمارنده هر مورچه در کولونی را صرفنظر از تخصصش مشخص می سازد. بنابراین برای دسترسی به مورچه هایی از تخصص s-اُم بایستی شماره مورچه آغازین از آن تخصص را به t انتساب داد.

خط13: اندیس نهایی مورچه های متخصص از مهارت s-اُم را مشخص می سازد. این مقدار کرانی راEndIndex می نامیم که از جمع کردن تعداد کلی جمعیت مورچه ها در تخصصs -اُم با اندیس آغازین مورچه ها بدست می آید.

$$EndIndex= MaxSkill(s)+St(s)$$

خط14: ایجاد حلقه ای برای شبیه سازی کار مورچه های متخصص در یک لایه خاص (عوامل از یک مهارت) را بعهده دارد .ناحیه قرار گرفته در بین 14 تا 23 سطح عوامل متخصص در الگوریتم را نشان میدهد.

خط 15**و**16**و**17: ساخت قوانین (بوجود آوردن مسیر ها)، هرس قوانین و بروز رسانی فرومون در عبارت های بکار رفته بر روی تخصص s-اُم از مهارتهای موجود کولونی در حال انجام است. بعبارت دیگر تمام فعالیت های ذکر شده بر روی لایه s-اُم در سیستم در حال انجام است.

خط 18 تا 21: آزمون همگرائی برای مورچه های متخصص از مهارت s-اُم را بر عهده دارد.این خطوط از الگوریتم تعداد قوانین کشف شده یکسان را در داخل تخصص s-اُم (لایه جاری) مورد شناسایی قرار می دهند . حاصل این آزمون در آرگومان j از الگوریتم ذخیره می گردد تا در خطوط بعدی مورد استفاده قرار گیرد.

خط 22: اندیس مورچه ی متخصص بعدی از مهارت s-اُم شکل می گیرد.

خط 23: شرایط کرانی برای خاتمه ی کار مورچه های متخصص (عوامل متخصص)از مهارت s-اُم بررسی می گردد. در این شرط،.حداکثر مورچه های متخصص برای مهارت s-اُم در کولونی که آنرا بصورت MaxSkill(s) نمایش می دهیم، بایستی مورد بررسی قرار گیرد.بعبارت دیگر بایستی کنترل کرد که مورچه های متخصص از مهارت s-اُم دقیقاًبه تعداد مشخص شان در کولونی ایجاد و مورد استفاده قرار گیرند.

خط 24: این خط متعلق به سطح تخصصی مورچه هاست. بعبارت دیگر قوانین کشف شده توسط مورچه های متخصص توسط سطح تخصصشان (لایه کاری) مورد تعمیم قرار می گیرد.

خط 25: وظیفه ی این خط بروز رسانی قوانین جزئی (مسیر های جزئی) کشف شده توسط مورچه هاست. این خط از الگوریتم نیز مربوط به سطح تخصصی مورچه هاست . بعبارت دیگر قوانین کشف شده توسط مورچه های متخصص توسط سطح تخصصشان (لایه کاری) مورد تعمیم قرار می گیرد .

خط 27: این بخش الگوریتم به انتخاب بهترین قانون اکتشافی (بهترین قانون جزئی کشف شده) می پردازد.اضافه نمودن بهترین قانون کشف شده در مهارت s-اُم (لایه s-اُم) به قانون چند طبقه بافر[1] انجام می گیرد.

خط 28: اضافه نمودن به شمارنده مربوط به مهارتها، این اضافه نمودن لایه کاری و تخصص مورد کار را عوض می کند.

خط 29: پایان سطح تخصص .

خط 30: قانون چند طبقه اکتشافی که حاوی بهترین قانون های یک طبقه ای کشف شده در هر تخصص است ، بایستی به مجموعه ی قوانین اضافه گردد.

خط 31: مجموعه ی آموزشی بایستی بروز رسانی گردد، به نحوی که مجموعه ی نمونه هایی که توسط قانون چند طبقه ی کشف شده پوشش داده شوند، از مجموعه ی آموزشی حذف گردند . حذف نمونه های پوشش یافته

[1] Multi Class Rule

باعث می گردد که در هر چرخه از حلقه While ،کولونی با فضای مسئله ی جدیدی روبرو گردد. در نتیجه حتی امکان دارد که تعداد طبقه های موجود در مجموعه ی آموزشی در طی این بروز رسانی تغییر کند که این تغییرات چرخه های بعدی الگوریتم را متأثر می کند.

1-7-4- ساخت قانون:

در ابتدا مورچه ی t-اُم (Ant_t) با قانونی خالی شروع بکار می کند که یعنی قانونی با هیچ عبارتی در مقدمش و در هر لحظه یک عبارت به آن اضافه می کند ، بعبارت دیگر مسیری جزئی برای هر مورچه در حال شکل گرفتن است . قانون جزئی ساخته شده جاری برای هر مورچه با مسیر جزئی جاری طی شده بوسیله ی آن مورچه در ارتباط است . انتخاب هر عبارت برای اضافه شدن به قانون جزئی جاری با انتخاب مسیر حرکت بطوریکه مسیر جزئی ادامه یابد مشابه است . انتخاب عبارت برای اضافه شدن به قانون جزئی جاری بستگی به دو مطلب دارد : تابع هیورستیکی وابسته به مسئله ی (h) و مقدار فرومون (t) که منتسب به هر عبارت است . مورچه ی t –اُم به اضافه کردن عبارت ها بصورت یکی یکی به قانون جزئی جاری مربوط به خودش در هر دفعه اجرا حلقه ادامه می دهد تا اینکه یکی از شرایط زیر برقرار شود:

- هر عبارتی که به قانون اضافه می گردد باعث می شود آن قانون موارد کمتری را پوشش دهد ، هرگاه این تعداد کمتر از حد آستانه ای که توسط کاربر تعریف گردیده است که در الگوریتم آنرا تحت نام Min_cases_per_Rule می شناسیم (حداقل تعدادی از نمونه ها که بوسیله ی قانون پوشیده شود) کار خاتمه می یابد.

همه ی صفات مشخصه قبلاً بوسیله ی آن مورچه استفاده شده باشد بنابراین صفات مشخصه ی بیشتری برای اضافه کردن به مقدم قانون وجود نداشته باشد . قابل توجه است که هر صفت مشخصه بیشتر از یکبار نمی تواند در هر قانون ظاهر گردد.

4-7-2- هرس قانون

در گام دوم این حلقه ، قانون R_t که بوسیله ی مورچه ی Ant_t ساخته شده است به منظور حذف عبارت های بی ربط ،بریده[1] می گردد که بعداً مورد بحث دقیق تر قرار خواهد گرفت .

در اینجا فقط باید اشاره شود که ساخته شدن عبارت های غیر مرتبط در قانون جاری براساس متغیر های اتفاقی در روال انتخاب عبارت ها و یا بر اثر دید مقطعی تابع هیورستیک که فقط یک صفت مشخصه را در هر دفعه مورد توجه قرار می دهد و از بقیه ی صفات مشخصه صرفنظر می کند ، ممکن است حاصل شده باشد.

4-7-3- بروز رسانی فرومون

سومین گام در این حلقه است و میزان فرومون را در هر مسیر مورد بروز رسانی قرار می دهد که فرومون مسیری را که بوسیله ی مورچه ی Ant_t طی شده بر اساس کیفیت انتساب داده شده به قانون R_t بایستی افزایش یابد و میزان فرومون سایر مسیر ها بایستی کاهش یابد که این کاهش شبیه سازی حالت تصعید فرومون در محیط طبیعی است .

سپس مورچه بعدی شروع به ساخت قانون خودش می کند که از مقادیر جدید فرومون بعنوان راهنما ی جستجوی مربوط به خودش استفاده می کند ، این فرایند(حلقه ی Repeat Until) تا رسیدن به یکی از شروط زیر ادامه می یابد:

- تعداد قوانین ساخته شده مساوی یا بزرگتر از حد آستانه ای No_of_ants که توسط کاربر مشخص می گردد باشد.
- مورچه ی جاری Ant_t قانونی ساخته باشد که دقیقاً مشابه قانون ساخته شده توسط (No_rules_converg-1) مورچه ی قبلی باشد در حالیکه No_rules_converg نشان دهنده ی تعداد قوانینی است که برای آزمون همگرائی مورچه ها بکار رفته است .

وقتی که حلقه ی Repeat-Until کامل شد یعنی در سطح کولونی (سیستم چند عامله) بهترین قانون در میان قانون های ساخته شده بوسیله ی تمام مورچه ها انتخاب و به فهرست قوانین کشف شده اضافه می گردد و سیستم

[1] Prune

چرخه ی جدیدی از حلقه ی While را بوسیله ی مقدار دهی اولیه به تمام فرومونها در تمام مسیرها آغاز می کند .

لازم بذکر است که در تعریف استاندارد از ACOدر مرجع [21] جمعیتی از مورچه ها بعنوان مجموعه ای از مورچه ها تعریف گردیده که دائماً در حال ساخت راه حل هایی در بین دو بروز رسانی فرومون هستند . بر اساس این تعریف در هر جرخه از حلقه ی While الگوریتم با جمعیت یک مورچه کار می کند زیرا بروز رسانی فرومون بعد از ساخت قانون توسط مورچه است . بنابراین می توانیم بگوییم هر حلقه ی While از الگوریتم ، یک مورچه ی منفرد دارد که تکرار های زیادی را انجام می دهد . توجه به این مطلب ضروری است که تکرارهای متفاوت از حلقه While با جمعیت های متفاوتی در ارتباط است ، زیرا هر جمعیت از مورچه ها شروع به حل مسئله ی جداگانه ای می نمایند زیرا مجموعه ی آموزشی هر کدام تفاوت می کند (به دلیل حذف موارد پوشش یافته توسط قوانین کشف شده) البته در متن ،t-اُمین تکرار از آن مورچه ، مورچه ی t-اُم ، Ant_t نامیده شده است که به منظور توصیف ساده تر الگوریتم است .

4-7-4- ساخت قوانین طبقه بندی

ازدیدگاه کشف دانش ، هسته ی این الگوریتم در گام اول حلقه ی Repeat –Until است که در آن مورچه ی جاری بصورت تکراری در هر دفعه تکرار یک عبارت به قانون جزئی خودش اضافه می نماید . در صورتیکه هر عبارت را بصورت $term_{ij}$ فرض کنیم که معنی آن عبارتی با شرطی به فرم $A_i = V_{ij}$ است که A_i صفت مشخصه ی i-اُم از دامنه ی A_i . در این حال احتمال آنکه عبارت مذکور $term_{ij}$ برای اضافه شدن به مسیر جزئی جاری انتخاب گردد از رابطه ی زیر بدست می آید:

$$P_{ij} = \frac{h_{ij}.t_{ij}(t)}{\sum_{i=1}^{a} X_i . \sum_{j=1}^{b_i} (h_{ij}.t_{ij}(t))}$$

رابطه 2-4. احتمال انتخاب هر عبارت برای اضافه شدن به مسیر جزئی جاری

که در این رابطه شرح متغیر ها بصورت زیر است:

h_{ij} مقدار تابع هیوریستیک وابسته به مسئله برای عبارت($term_{ij}$) است . هرچه مقدار h_{ij} بیشتر باشد عبارت ($term_{ij}$) برای کلاس بندی کاربردی تر است و در نتیجه بایستی احتمالش برای انتخاب شدن بیشتر باشد . تابعی که مقدار هیوریستیک وابسته به مسئله را مشخص می سازد بر اساس تئوری اطلاعات پایه ریزی گردیده و در ادامه مورد بحث قرار می گیرد.

$t_{ij}(t)$میزان فرومون انتساب داده شده به عبارت ($term_{ij}$) در تکرار t-اُم است که با مقدار فرومون موجود جاری در مکان i و j از مسیری که مورچه ی جاری پیموده ،مرتبط است .هرچه کیفیت قانون ساخته شده بوسیله ی یک مورچه بهتر باشد ، به میزان فرومون اضافه شده به آن قسمت از مسیر که توسط مورچه ملاقات شده اضافه می گردد.بنابراین هر چه که زمان می گذرد بهترین قطعات از مسیر طی شده – یعنی بهترین عبارت هایی (زوجهای خصیصه – ارزش) که به قانون اضافه گردیده – مقدار فرومون بیشتر و بیشتری خواهند داشت که باعث افزایش احتمال انتخاب آنها خواهد شد.

a تعداد همه ی صفات مشخصه است.

x_iمجموعه ایست که مانند پرچم عمل می کند و به ازای هر صفت مشخصه ، عضو دارد. اگر صفت مشخصه ی A_i تا بحال بوسیله ی مورچه ی جاری استفاده نشده باشد عضو متناظرش در این مجموعه یک است ، و در غیر اینصورت صفر خواهد بود.

b_i تعداد مقادیر دامنه ی i- اُمین صفت مشخصه را نشان می دهد .

عبارت ($term_{ij}$) برای اضافه شدن به قانون جزئی جاری متناسب با احتمال بدست آمده از فرمول فوق انتخاب می گردد، البته با توجه به دو محدودیت زیر:

1. صفت مشخصه ی A_i نمی تواند قبلاً در قانون جزئی جاری انتخاب شده باشد که این مطلب برای ارضای این محدودیت است که مورچه ها بایستی "بخاطر داشته باشند" که کدامیک از عبارت ها (زوج های خصیصه-ارزش) در قانون جزئی جاری اتفاق افتاده اند.

2. یک عبارت ($term_{ij}$) نمی تواند به قانون جزئی جاری اضافه گردد اگر این اضافه شدن باعث شود که آن قانون از حدی که قبلاً مشخص شده تعدادی مصادیق کمتری را مورد پوشش قرار دهد که این حد آستانه راMin_cases_per_rule می نامیم.

وقتیکه مقدم قانون کامل گردید ، الگوریتم بخش پی آمد را انتخاب می کند (یعنی کلاس پیش بینی شده توسط قانون انتخابی است) بطوریکه کیفیت قانون ایجاد شده بیشینه گردد .این مطلب بوسیله ی نسبت دادن کلاس غالب [1] در میان موارد پوشش داده شده توسط قانون به بخش پی آمد قانون صورت می گیرد.

5-7-4-تابع هیوریستیک

برای هر عبارت ($term_{ij}$) که توانایی اضافه شدن به قانون جاری را داشته باشد ،الگوریتم ،مقدار تابع هیوریستیک h_{ij} را که تخمینی از کیفیت این عبارت است را با توجه به توانایی آن در بهبود دادن صحت پیش بینی [2] قانون محاسبه می کند . این تابع هیوریستیک بر پایه ی تئوری اطلاعات (40) محاسبه می گردد. بعبارت دقیق تر مقدار h_{ij} برای($term_{ij}$) دربرگیرنده ی میزان انتروپی (مقدار اطلاعات) انتساب داده شده به آن عبارت است . برای هر ($term_{ij}$) که به شکل $A_i = V_{ij}$ است (که در آن A_i صفت مشخصه ی i – اُم است و V_{ij} مقدارj – اُم ازدامنه ی A_i میزان انتروپی از رابطه ی 3-4 محاسبه می گردد: (30)

$$H(W/A_i = V_{ij}) = -\sum_{w=1}^{k} (P(w/A_i = V_{ij}).\log_2 P(A_i = V_{ij}))$$

رابطه ی 3-4، میزان انتروپی عبارت $term_{ij}$

که در این رابطه :

W صفت مشخصه ی کلاس (یعنی صفت مشخصه ای که دامنه ی آن شامل طبقه هایی برای پیش بینی است .

K تعداد طبقه هاست .

[1] Majority class
[2] Predictive accuracy

$P(w/A_i = V_{ij})$ احتمال تجربی مشاهده ی کلاس w است بشرط مشاهده ی عبارت $term_{ij}$ (یعنی همان $A_i = V_{ij}$).

نکته : برای محاسبه ی احتمال تجربی کافیست که رابطه ی $freq_{ij}^{w} / |T_{ij}|$ محاسبه گردد که در آن $|T_{ij}|$ تعداد همه ی مصادیق در پارتیشن T_{ij} است (پارتیشنی شامل آن نمونه ها و مواردی که صفت مشخصه A_i مقدار V_{ij}را دارد)و $freq_{ij}^{w}$ تعداد مصادیقی است که در پارتیشن T_{ij} کلاس w دارند.(38)

هرچه مقدار انتروپی $H(W/ A_i = V_{ij})$ بیشتر باشد نشان دهنده ی آن است که طبقه ها بطور یکنواخت تری[1] توزیع شده اند و بنابراین بایستی مورچه ی جاری با احتمال کمتری آن عبارت ($term_{ij}$) را برای قانون جزئی خودش مورد انتخاب قرار دهد .

برای فرمول 1 ، نرمال بودن و بهنجاری میزان تابع هیوریستیک ایده آل است و بمنظور پیاده سازی این نرمال سازی از این حقیقت استفاده می کنیم که مقدار تابع $H(W / A_i = V_{ij})$ در محدوده $0 \le H(W / A_i = V_i) \le (\log_2 k)$ تغییر می کند که در آن k ، تعداد طبقه ها را نشان می دهد . بنابراین تابع هیوریستیک نرمال شده بر اساس تئوری اطلاعات به شکل فرمول4-4 است :

$$h_{ij} = \frac{\log_2 k - H(W / A_i = V_{ij})}{\sum_{i=1}^{a} X_i . \sum_{j=1}^{b_i} (\log_2 k - H(W / A_i = V_{ij}))}$$

رابطه 4-4. تابع هیوریستیک نرمال شده برای عبارت $term_{ij}$

که a و x_i و b_i همان معانی گفته شده در رابطه ی قبل را دارند.

توجه به این نکته لازم است که مقدار انتروپی $H(W / A_i = V_{ij})$ یک عبارت ($term_{ij}$) ، صرفنظر از محتویات قانونی که عبارت درآن بکار رفته است ، همواره یکسان است. بنابراین به منظور صرفه جویی در زمان محاسبه ،

[1] Uniformly

$H(W/A_i=V_{ij})$ برای همه ی عبارات ($term_{ij}$) در گام پیش پردازش محاسبه می گردد. در رابطه ی 3 فقط دو نکته مهم بعنوان تذکر وجود دارد:

1. مقدار V_{ij} از صفت مشخصه ی A_i در مجموعه ی آموزشی حادث نشده باشد ، بنابراین $H(W/A_i=V_{ij})$ بیشینه ی خود را خواهد داشت که مقدار $\log_2 k$ خواهد بود که این مطلب معادل انتساب کمترین قدرت پیش بینی ممکن به عبارت ($term_{ij}$) است.

2. اگر همه ی موارد متعلق به یک کلاس باشند بنابراین $H(W/A_i=V_{ij})$ (مقدارش مساوی صفر خواهد شد که این مطلب مرتبط با انتساب دادن حداکثر قدرت پیش بینی به عبارت ($term_{ij}$)است.

در الگوریتم کشف دانش مبتنی بر کولونی مورچه ، اندازه انتروپی با بروز رسانی فرومون بصورت توأم با هم بکار می رود .کاربرد توأم اندازه یا انتروپی و بروزرسانی فرومون باعث مقاوم شدن این الگوریتم در برابر خطا می گرددو همچنین باعث می شود که الگوریتم ،کمتر در دامهای بهینه ی محلی در فضای جستجو گرفتارآید، زیرا بازخورد ارائه شده بوسیله ی بروز رسانی فرومون به تصحیح بعضی از خطاهای رخ داده شده توسط دید محدود اندازه ی انتروپی کمک می نماید.توجه به این نکته ضروری است که اندازه ی انتروپی ، اندازه ی هیوریستیک ومحلی است که تنها به یک صفت خاصه درهرلحظه توجه می نمایدو بنابراین به مسئله ی تعامل [1] بین صفت های مشخصه ، حساس است .

در مقابل ، بروز رسانی فرومون تمایل به تطبیق بهتری با تعامل بین صفات مشخصه دارد ،زیرا بروز رسانی فرومون مستقیماً بر پایه ی کارائی قانون بعنوان یک موجودیت استوار است.یعنی فرومون ، مستقیماً تعامل میان تمام صفات مشخصه ی موجود در قانون را در نظر می گیرد.

6-7-4- هرس قوانین طبقه بندی

هرس ی قطع قوانین [2] تکنیکی عمومی در زمینه ی کشف دانش است (29). هدف اصلی از قطع قوانین ، حذف عبارتهای بی ربطی است که بطور زائد به قانون اضافه گردیده است .قطع قوانین ، توانایی بالقوه ی افزایش قدرت

[1] Interaction
[2] Rule Pruning

پیش بینی قوانین را داراست که اینکار را از طریق ممانعت بیش از حد [1] قانون انجام می دهد . علت دیگر برای قطع قوانین آنست که اینکار سادگی قانون را افزایش می دهد، زیرا هرچه قانون کوتاهتر باشد دارای درک بالاتری توسط کاربر خواهد بود.

بعد از اتمام کار یک مورچه در ساخت قانون مربوط به خودش ،پردازش مربوط به قطع قانون فراخوانی می گردد. این راهکار برای پردازش قطع قانون مشابه راهکاری است که توسط مرجع (2) بکار گرفته شده ، اما ملاک کیفیت استفاده برای قانون خیلی باهم متفاوت هستند.

ایده ی اصلی در قطع قوانین ، حذف یک عبارت از قانون در هر دفعه است که بصورت مکرر مادامیکه این پردازش کیفیت قانون را افزایش دهد، اجرا می گردد.

7-7-4- قوانین انتقال و بروز رسانی فرومون

همانطوریکه گفته شد ، هر عبارت ($term_{ij}$)، با قطعه ای از مسیری که توسط مورچه طی گردیده ، در ارتباط است .

در هر تکرار حلقه While از الگوریتم (در سطح کولونی) همه ی عبارات ($term_{ij}$) به مقدار اولیه ای از فرومون مقدار دهی می شوند .مورچه ها ترجیح می دهند به سوی معنایی حرکت کنند که بوسیله ی مسیر هایی با مقدار بالای فرومون متصل شده باشند . بخشی از فرومون بر روی همه ی مسیرها بخار می شود و سپس هر مورچه مقداری فرومون بر روی مسیر ها باقی می گذارد. ومسیر هایی که به سفرهای خوب تعلق دارند مقدار بیشتری فرومون دریافت می کنند و این فرایند تکرار می شود.به عبارتی این شبیه یک برنامه ی تقویت یادگیری است که در آن راه حل های بهتر ، تقویت بالاتری دریافت می کنند

فرومون موجود بر روی این مسیر ها نقش یک حافظه ی بلند مدت توزیع شده را بازی می کنند .

پس بعد از اینکه همه عبارات به مقدار اولیه ای از فرومون مقدار دهی شدند ، اولین مورچه ، جستجوی خودش را در حالی شروع می کند که همه مسیر ها مقدار دهی اولیه شده اند . دراین الگوریتم همه ی مسیر ها یک مقدار

[1] Over fitting

فرومون را دارا هستند ،که این مقدار اولیه فرومون که در هر موقعیت مکانی از مسیر بر جای گذاشته شده با تعداد مقادیر همه ی صفات مشخصه نسبت عکس داردکه با رابطه 6 تعریف گردیده است.

هر جائیکه یک مورچه ساخت قانون خودش را تمام کرده باشد ، آنگاه مقدار فرومون در همه ی قسمتهای تمام مسیرها باید بروزرسانی گردد.این بروز رسانی فرومون بوسیله دو ایده اصلی پشتیبانی مگردد که عبارتند از:

1. مقدار فرومون انتساب داده شده به هر عبارت که در قانون بوسیله مورچه یافت گردیده به نسبت کیفیت قانون افزایش یابد.

2. مقدار فرومون انتسلبی به هر عبارت که در قانون وقوع نیافته باشد کاهش یابد. در طبیعت فرومون باقی مانده در مسیرها بر اساس عوامل محیطی تضعیف و تصعید می گردد که این نرخ خود مسئله مهمی در پویایی الگوریتم بهینه سازی مورچه هاست.

افزایش مقدار فرومون انتساب داده شده به هر عبارت که در قانون بدست آمده توسط یک مورچه استفاده شده باشد ، مرتبط با افزایش دادن احتمال انتخاب شدن آن عبارت توسط سایر مورچه ها در آینده است که بایستی مرتبط با کیفیت آن قانون باشد. کیفیت یک قانون که با Q نمایش داده می شود بوسیله رابطه کلی زیر تعریف شده است(45)

$$Q= Sensitivity* Specificity$$

رابطه 4-5. کیفیت قانونی که با Q نمایش داده می شود

که این رابطه را به صورت زیر باز می کنیم:

$$Q = \frac{TP}{TP + FN} . \frac{TN}{FP + TN}$$

رابطه 4-6. محاسبه درجه کیفیت یک قانون

که آرگومانهای آن بصورت زیر تعریف می گردند:

TP (True positive) : تعداد مصادیقی که بوسیله قانون پوشش یافته باشند و طبقه پیش بینی شده توسط قانون را داشته باشند.

FP (False positive) : تعداد مصادیقی که بوسیله قانون پوشش یافته باشند ولی طبقه پیش بینی شده توسط قانون را نداشته باشند.

FN(False negative) : تعداد مصادیقی که بوسیله قانون پوشش نیافته باشند اما طبقه پیش بینی شده توسط قانون را داشته باشند.

TN(True negative) : تعداد مصادیقی که بوسیله قانون پوشش نیافته باشند و طبقه پیش بینی شده توسط قانون را نداشته باشند.

مقدارQ در محدوده ی $0 \le Q \le 1$ تغییر می کند. هر چه مقدار Q بیشتر باشد کیفیت قانون بالاتر است. بروز رسانی فرومون برای هر عبارت بر اساس رابطه 7-4 برای همه ی عبارتهایی که در قانون واقع شده باشند انجام می گیرد:

$$t_{ij}(t+1) = t_{ij}(t) + t_{ij}(t).Q, \forall i, j \in R$$

رابطه 7-4 - مشخص کننده نحوه ی بروز رسانی فرومون برای هر عبارت

که در آن R مجموعه عبارت هایی است که در قانون ساخته شده بوسیله مورچه در تکرار t-ام (مورچه t- ام) از الگوریتم وجود دارند. بنابراین مقدار فرومون های همه ی عبارت های موجود در قانون کشف شده بوسیله مورچه جاری ، به نسبت مقدار فرومون جاری آن عبارت ها و کیفیت اخذ شده نهایی از قانون افزایش می یابند.

همانطوری که قبلاً اشاره شد میزان فرومون انتساب داده شده به هر عبارت که در قانون کشف شده ی مورچه ی جاری حضور ندارد بایستی به منظور شبیه سازی تصعید و تبخیر فرومون در کولونی ها ی مورچه های واقعی ، کاهش یابد.

تأثیر تصعید فرومون برای عبارتهای استفاده نشده بوسیله ی نرمال سازی هر مقدار فرومون t_{ij} بدست می آید. این نرمال سازی بوسیله تقسیم مقدار هریک از مقادیر t_{ij} بر جمع t_{ij} به ازای همه ی i,j ها انجام می گیرد. برای

دیدن چگونگی تصعید فرومون ، به این نکته توجه داشته باشید که فقط میزان فرومون آن عبارت هایی که بوسیله مورچه استفاده شده توسط رابطه ی 12 افزایش می یابد.بنابراین در زمان نرمال سازی میزان فرومون یک عبارت استفاده نشده بوسیله ی تقسیم مقدار جاری آن برکل حاصل جمع فرومون های همه ی عبارت ها ، پیاده سازی می شودکه باعث افت مقدار فرومون برای آن عبارت استفاده نشده خواهد شد.

فصل پنجم:ابهام زدایی معنی واژه

5-1- ابهام زدایی واژه با استفاده از کولونی مورچه

برای ابهام زدایی از معنی واژه ، یکی از اقدامات اصلی و اولیه در ابهام زدایی با استفاده از این روش ، تهیه و آماده سازی یک پیکره مناسب می باشد. پیکره ای که در اینجا از آن استفاده شده است بخش کوچکی از پیکره خانم دکتر طیبه موسوی میانگاه است. این پیکره حاوی 4 میلیون و هشتصدو شصت واژه بوده اما بدلیل اینکه برای اعمال الگوریتم مجبور هستیم که پیکره موجود را به شکل جفت های خصیصه – ارزش بکار ببریم ، تنها از قسمت کوچکی از این پیکره که حاوی 100 واژه مبهم است استفاده شده است .

برای مثال جمله زیر با واژه مبهم article را در نظر می گیریم:

Some of the articles of the constitution need to be amended.

قسمتی ازبروندادِ این پیکره برای این واژه با چهار معنی مختلف که به شکل جفت های خصیصه –ارزش تبدیل شده است را در جدول 5 می بینیم:

جدول 5. قسمتی از پیکره به شکل جفت های خصیصه –ارزش

ردیف	واژه مبهم	معنی واژه مبهم	واژه همبسته	معنی واژه همبسته	میزان همبستگی دوواژه درپیکره (AS)
1	Article	مقاله	constitution	قانون اساسی	0.0019
2	Article	مقاله	amend	اصلاح کردن	0.00049
3	Article	مقاله	constitution	ساختمان	0.00029
4	Article	مقاله	constitution	تشکیل	0.00012
5	Article	ماده	constitution	قانون اساسی	0.54
6	Article	ماده	constitution	ساختمان	0.00028
7	Article	ماده	constitution	تشکیل	0.0052
8	Article	ماده	amend	اصلاح کردن	0.006

ردیف	واژه مبهم	معنی واژه مبهم	واژه همبسته	معنی واژه همبسته	میزان همبستگی دوواژه درپیکره (AS)
9	Article	کالا	constitution	قانون اساسی	0.0005
10	Article	کالا	constitution	ساختمان	0.00014
11	Article	کالا	constitution	تشکیل	0.00033
12	Article	کالا	amend	اصلاح کردن	E-5.89 05

13	Article	حرف تعریف	constitution	قانون اساسی	NaN
14	Article	حرف تعریف	constitution	ساختمان	NaN
15	Article	حرف تعریف	constitution	تشکیل	NaN
16	Article	حرف تعریف	amend	اصلاح کردن	NaN
17	Article	قانون اساسی	amend	اصلاح کردن	0.029
18	constitution	ساختمان	amend	اصلاح کردن	0.0008
19	constitution	تشکیل	amend	اصلاح کردن	0.0001
20	constitution	قانون اساسی	amend	ترمیم کردن	0.00019
21	constitution	ساختمان	amend	ترمیم کردن	0.01
22	constitution	تشکیل	amend	ترمیم کردن	0.000002
23	constitution	قانون اساسی	amend	بهبود یافتن	0.0022
24	constitution	ساختمان	amend	بهبود یافتن	0.00029
25	constitution	تشکیل	amend	بهبود یافتن	0.00013

حال با توجه به مطالبی که در بخش 1-4 مطرح کردیم ، می توان قوانین را بصورت زیر داشت:

مقدم : اگر واژه مبهم Article باشد و در حالت دستوری اسم باشد و واژه همبسته آن با فاصله حداکثر 8 واژه از دوطرف در معنی "قانون اساسی" بکار رفته باشد

تالی : آنگاه معنی واژه مبهم "ماده" خواهد بود.

نکته دیگری که برا ی ساده کردن کار در نظر گرفته شده است ، فرض بر این است که حالت دستوری واژه از قبل تشخیص داده شده است ، یعنی میدانیم که واژه مبهم در حالت دستوری اسم قرار دارد و حال با اعمال کولونی مورچه می خواهیم تشخیص دهیم که کدامیک از معانی مختلف این واژه در حالت اسم ، به عنوان معنای درست برای آن واژه تشخیص داده می شود.

اکنون به سراغ روش کار می رویم.

در این روش ،ابتدا در پیکره زبان فارسی کلماتی که بصورت دوتایی هستند را جستجو می نماییم .وسپس با هم آیی آنها را در جدولی بنام جدول با هم آیی ذخیره می کنیم . بدین معنا که جدولی را بصورت زیر ایجاد می کنیم.

جدول 6. با هم آیی دو واژه

کلمه فارسی B \ کلمه فارسی A		
	a	b
	c	d

که در این جدول a،تعداد جملاتی در پیکره ی زبان فارسی است که در آن جملات، کلمه ی A وکلمه B با هم در آن جمله ظاهر شده اند ، b، تعداد جملاتی است که در آن جملات ،کلمه A به تنهایی و بدون کلمه B دیده شده است ؛ c ، تعداد جملاتی است که در آن کلمه A به تنهایی و بدون کلمه B دیده شده است و بالاخره d تعداد جملاتی است که در آنها نه کلمه ی A دیده شده و نه کلمه B .

حال برای هر جدول ، میزان همبستگی[1] آن دو کلمه را با استفاده از فرمول زیر محاسبه میکنیم :

$$AS = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

رابطه 1-5. فرمول محاسبه امتیاز با هم آیی بین دو کلمه

[1]- Association Score

در این صورت همواره AS عددی بین صفر و یک خواهد بود. بعداً از این عدد برای مقدار دهی اولیه فرومون روی یال های انتهایی استفاده می کنیم.(48)

اکنون در برخورد با یک جمله دارای چند واژه ی مبهم، ابتدا گرافی برای آن واژه های مبهم، بشکلی که کلیه معانی مختلف آن واژه ها در جمله را پوشش دهد ایجاد کرده (به مثال توجه نمایید) در مرحله ی بعد با توجه به تعداد معانی مختلف هر واژه برای آن جمله مورچه ی متخصص تعریف می کنیم و برای هر تخصص نیز یک ماکزیمم سطح فرومون تعریف می کنیم یعنی اگر میزان فرومون از a تا b بود پس در حیطه تخصص مورچه های متخصصِ x قرار دارد و....) . و با مراجعه به پیکره ی زبان مقصد احتمال وقوع هر معنی را بدست آورده و یالهای مربوطه را مقداردهی اولیه ی فرومون می نماییم . در مرحله ی بعد ،با استفاده از جدول وابستگی و میزان همبستگی محاسبه شده ،یالهای مربوطه را نیز مقدار دهی کرده و با مراجعه به پیکره ی زبان مقصد احتمال وقوع هر معنی را بدست آورده و یالهای مربوطه را مقداردهی اولیه ی فرومون می نماییم .

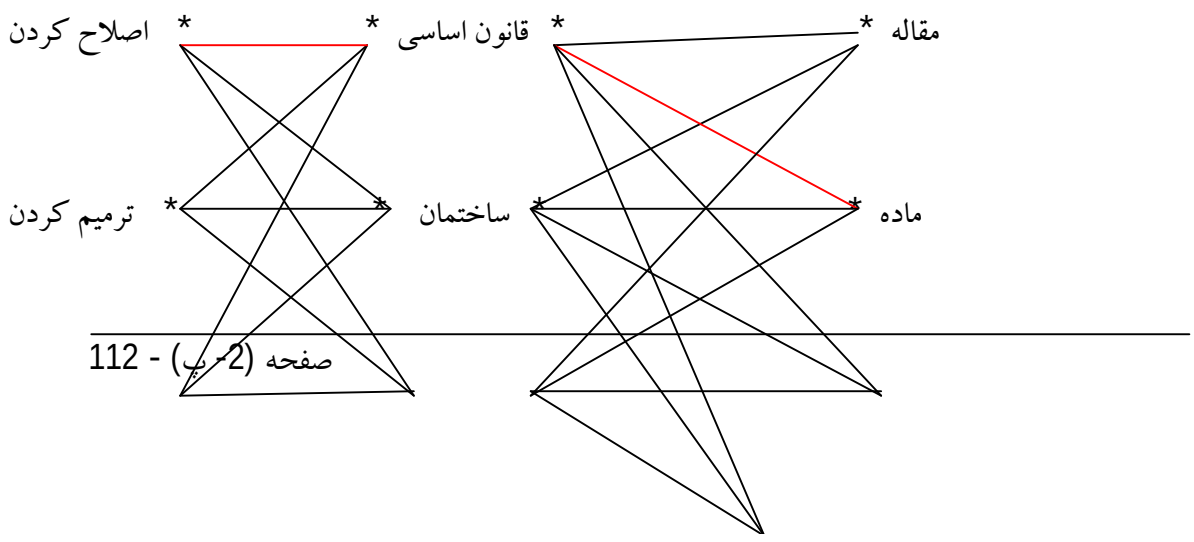
همان جمله فوق را در نظر می گیریم که در آن علاوه بر واژه article واژه های constitution و amend نیز مبهم می باشند:

Some of the articles of the constitution need to be amended.

article: مقاله / ماده / کالا / حرف تعریف

constitution: قانون اساسی / ساختمان / تشکیل

amend: اصلاح کردن / ترمیم کردن / بهبود یافتن



کالا * * تشکیل * بهبود یافتن

حرف تعریف *

شکل 9. مثالی از نحوه ایجاد گراف برای واژه مبهم

بدین ترتیب بعد از محاسبه ی مقدار AS برای هر یک از یالهای گراف فوق را از روی پیکره، و اعمال این الگوریتم می بینیم که مورچه در مسیر یالهایی که در شکل فوق با رنگ قرمز نشان داده شده اند، همگرا می شوند.

گام بعدی در این روش ،در نظر گرفتن رابطه ایست که به کمک آن بتوان میزان فرومون را در هر مسیر مورد بروز رسانی قرار داد.یعنی رابطه ای که قادر باشد فرومون مسیری را که بوسیله ی مورچه ی Ant_t طی شده بر اساس کیفیت انتساب داده شده به قانون R_t افزایش داده و میزان فرومون سایر مسیر ها کاهش یابد واین افزایش وکاهش شبیه سازی حالت تولید بیشتر ویاتصعید فرومون در محیط طبیعی است .

همانطور که قبلاً توضیح داده شد ، هسته ی این الگوریتم حلقه تکراری است که در آن مورچه ی جاری بصورت تکراری در هر دفعه تکرار یک عبارت به قانون جزئی خودش اضافه می نماید . و دیدیم که احتمال آنکه عبارت مذکور $term_{ij}$ برای اضافه شدن به مسیر جزئی جاری انتخاب گردد از رابطه ی2-4 بدست می آید.

. تابعی که مقدار هیوریستیک وابسته به مسئله را مشخص می سازد بر اساس تئوری اطلاعات پایه ریزی گردیده است. پس از اعمال الگوریتم ، به واژه ی مبهم، از بین معانی مختلف ، معنی مناسبی متناظر می شود.

2-5- نتایج و بررسی کارایی الگوریتم کولونی مورچه در ابهام زدایی معنی واژه

الگوریتم پیاده سازی شده توسط این پروژه ، که نتایج آن در این فصل ارائه می گردد همگی در محیط++VC ویرایش 6 برروی Windows XP Proftional تهیه و اجرا گردیده اند. این الگوریتم بر روی سخت افزار Pentium IV پیاده سازی و تست شده اند.

در ابتدا لازم است شرایط آزمون را مشخص سازیم.شرایطی مانند نحوه ی کلاس بندی بوسیله قوانین کشف شده ، تنظیم پارامترهاو... که در زیر ارائه می گردند.

5-3- نحوه کلاس بندی بوسیله قوانین کشف شده

در الگوریتم ذکر شده به منظور کلاس بندی کردن مجموعه مصادیقِ تست جدید[1] که در خلال آموزش دیده نشده بودند ، قوانین کشف شده به ترتیبی که کشف شده اند بکار می روند. یعنی اولین قانونی که مورد جدید را پوشش دهد برای کلاس بندی آن قانون انتخاب می گردد. امکان دارد که هیچ قانونی در لیست قوانین کشف شده برای پوشش مورد جدید وجود نداشته باشد ، در این حالت مورد جدید را تحت قانون همیشگی [2] کلاس بندی می کنیم .یعنی کلاس غالب را برای مجموعه موارد آموزشی کشف نشده ، مورد پیش بینی قرار می دهیم.

5-4- مجموعه آموزشی

کارایی الگوریتم کولونی مورچه بر روی مجموعه آموزشی به عنوان پیکره ای حاوی 100 واژه مبهم مورد آزمایش قرار گرفته است .مشخصه های این مجموعه آموزشی در جدول زیر خلاصه شده است.اولین ستون جدول نام مجموعه داده هاست. بقیه ستون ها به ترتیب عبارتند از: تعداد مصادیق (نمونه های موجود) و تعداد کلاس های مربوط به مجموعه داده ها.

[1] Test Set
[2] Defult

جدول 7. مشخصه های مجموعه آموزشی

مجموعه داده ها	تعداد مصادیق	تعداد کلاس ها
Sub_Coppus article	287	6
Sub_Corpus Constitution	263	3
Sub_Corpus Conflict	300	4
Sub_Corpus Figur	230	4

5-5- تنظیم پارامترهای الگوریتم

تعداد مورچه ها (No_of_ants) : در این الگوریتم تعداد مورچه ها بصورت غیر مستقیم تأثیر دارد . بعبارت دیگر وقتی تعداد مورچه ها را اضافه می نماییم در اصل به تعداد مورچه های زیر مجموعه یکی از (حداقل یکی و حداکثر همه) مهارتها افزوده می گردد.بعبارت دیگر MaxSkill(i) (به ازائ i نامشخص) رشد خواهد داشت . با رشد کردن تعداد جمعیت متخصص در یک مهارت از مهارتهای موجود در کولونی در اصل شانس یافتن بهترین قانون یک طبقه ای در آن مهارت افزوده می گردد. که باز هم با این افزایش سیستم به سمت دقت بالاتر و کند شدن می رود.

حداقل تعداد مصادیق در هر قانون (Min_cases_per_ rule): هر قانون یک طبقه ای بایستی حداقل ، این تعداد از مصادیق را پوشش دهدکه باعث تقویت نظر تعمیمی قانون کشف شده گردد.این مطلب باعث می گردد که از عدم تطابق قانون با مجموعه داده ها جلوگیری شود.

تعداد بیشینه موارد کشف نشده در مجموعه آموزشی (Max_uncovered_cases): پردازش کشف قوانین بطور تکراری اجراء می گردد و شرط کرانی آن همین مقدار است. بعبارت دیگر این پردازش ، تا هنگامی که تعداد مصادیق مجموعه آموزشی که بوسیله هیچکدام از قوانین کشف شده پوشش نیافته اند کوچکتر از این حد آستانه شود ادامه می یابد. یعنی تعداد نمونه های مجموعه آموزشی اگر از این مقدار کمتر گردد کل عمل ساخت قوانین متوقف می گردد.

تعداد قوانین استفاده شده برای آزمون همگرایی مورچه ها(No_rules_convrg): در این الگوریتم اگر مورچه ای از مهارت i-اُم قانونی یک طبقه ای ساخته باشد که دقیقاً مشابه با قانون یک طبقه ای ساخته شده توسط (No_rules_convrg-1) تعداد از مورچه های متخصص قبلی در مهارت s -اُم (لایهs -اُم) باشد ، سیستم نتیجه می گیرد که مورچه های مهارت s-اُم به یک قانون همگرا شده اند(به یک مسیر درلایه s- اُم). در این حالت تکرار فعلی از حلقه Repeat_Until (سطح مورچه ها ی متخصص از مهارت s-اُم)متوقف می گرددو چرخه دیگری از حلقه while داخلی را شروع می کند .که معادل شروع بکار بر روی لایه جدید یا مهارت جدیدی است. در نتایج بعدی مقادیر زیر بعنوان حدود آستانه برای الگوریتم انتخاب شده اند:

Min_casees_per_rule=10
Max_uncovered_case=10
No_rules_converg=10
No_of_ants=3000

در این الگوریتم ، در همه حالات تعداد کلی مورچه ها بصورت معادل بین تخصص های مختلف تقسیم شده است . مثلاً اگر تعداد کلاس ها 3 تا باشد و تعداد مورچه ها در کولونی برابر 3000 عدد باشد آنگاه تعداد مورچه ها برای هر یک از تخصص ها برابر 1000 فرض شده است.

$$MaxSkill(0)=\ldots=MaxSkill(i)=\ldots=MaxSkill(k)=(No\_of\_ants/k)$$

رابطه 2-5- تعداد مورچه ها در تخصص های مختلف

در این رابطه k تعداد کلاس های موجود در مجموعه آموزشی است.

6-5- نتایج کسب شده

نتیجه کسب شده از پیاده سازی الگوریتم کولونی مورچه روی پیکره زبان مقصد در جدول زیر آورده شده است . این جدول نشان دهنده صحت پیش بینی قوانین کشف شده حاصل از الگوریتم کولونی مورچه به کمک کشف دانش چند طبقه روی پیکره زبان مقصد است. اعداد بعد از علامت (±) انحراف از معیار [1] می باشد که بعد از هر کدام از صحت های پیش بینی دیده می شود.

[1] Standard Deviation

جدول 8. صحت پیش بینی قوانین کشف شده حاصل از الگوریتم کولونی مورچه
به کمک کشف دانش چند طبقه روی پیکره زبان مقصد

صحت پیش بینی به درصد	مجموعه داده ها
83%±3	Sub_Coppus article
86±2/9	Sub_Corpus Constitution
87±۲	Sub_Corpus Conflict
۸۶	Sub_Corpus Figur

4-1-5- بررسی نتایج کسب شده

این بخش به ارزیابی پیچیدگی محاسباتی[1] الگوریتم کشف دانش چند طبقه مبتنی بر کولونی مورچه ها می پردازد.این تحلیل به سه بخش تقسیم گردیده است:

1- پیش پردازش
2- یک چرخه از حلقه While بیرونی
3- کل چرخه مربوط به حلقه While

پیچیدگی محاسباتی پیش پردازش: پیچیدگی زمانی این گام O(n*a) است که در آن n تعداد مصادیق و a تعدادصفات مشخصه است.

پیچیدگی محاسباتی یک چرخه حلقه بیرونی While : هر چرخه بیرونیWhile از الگوریتم کولونی مورچه با مقدار دهی اولیه فرومون شروع می شود .این فعالیت بر روی هر لایه از جدول (شکل 3) انجام می دهد و بنابراین دارای پیچیدگی از مرتبهO(a*v*s) می باشد که در آن a تعداد صفات مشخصه ، v تعدادمقادیر به ازاء یک صفت مشخصه وs تعداد مهارتهای موجود(طبقه ها) در کولونی است.

[1] Computational Complexity

حال اگر به حلقه درونی While توجه کنیم ، هر بار اجراء حلقه درونی While با اجراء چرخه Repeat-Until و بخش تعمیم و بروز رسانی معادل است پس ابتدا به بررسی چرخه Repeat-Until می پردازیم.

یک تکرار از چرخه Repeat-until شامل فعالیت های یک مورچه متخصص خواهد بود. مهمترین گام انجام گرفته برای هر مورچه متخصص در این حلقه عبارتست از:

1. ساخت قانون
2. ارزیابی قانون
3. هرس قانون
4. بروز رسانی فرومون

بنابراین هر یک تکرار از چرخه Repeat-Until دارای پیچیدگی زمانی از مرتبه k^3O(k*a*v+n*) است که در آن k تعداد شرط های انتخابی توسط یک مورچه متخصص ،a تعداد صفات مشخصه ، وn تعداد نمونه های آموزشی (مصادیق) است.

به منظور بدست آوردن پیچیدگی محاسباتی مربوط به اجراء کامل حلقهRepeat-Until نتیجه قبلی بایستی در z_s ضرب گردد،که تعداد مورچه های تخصص s-اُم می باشد.بنابراین اجرا کامل حلقه Repeat-Until مرتبه زمانی آن O z_s *[k*a*v+n* k^3* می باشد.

بخش تعمیم و بروز رسانی فرومون دارای پیچیدگی زمانی از مرتبه(O(c* z_s است که z_s برابر با تعداد مورچه های تخصصs -اُم و c تعداد خوشه های ایجاد شده برای آن صفت خواهد بود.

پیچیدگی زمانی یکبار اجرای حلقه درونی While برابر خواهد بود با حاصل جمع مطالب فوق یعنی:

$$O(z_s*[k*a*v+n*k^3]+v^2+s*z_s)$$

که می توان آنرا بصورت زیر خلاصه نمود:

$$O(z_s*[k*a*v+n*k^3+s]+v^2)$$

پیچیدگی زمانی اجرای کلی حلقه درونی While برابر خواهد بود با حاصل ضرب مرتبه فوق در تعداد کلی مهارتها(طبقه ها) در کولونی یعنی:

$$O(s^* z_s *[k*a*v+n*k^3+s]+v^2)$$

چون در این الگوریتم تعداد مورچه از هر مهارت با تعداد مورچه های سایر مهارتها مساوی در نظر گرفته شده است بنابراین حاصل ضرب تعداد مهارتها در تعداد مورچه ها ی هر مهارت ، معادل تعداد کل مورچه های کولونی است.یعنی:

$$Z=s^* z_s$$

بنابراین می توان پیچیدگی زمان اجرای کلی حلقه بیرونی While را بصورت زیر ساده نمود:

$$O(z^*[k*a*v+n*k^3+s]+v^2)$$

پیچیدگی زمانی یکبار اجرای حلقه بیرونی While برابر خواهد بود با مجموع پیچیدگی های مقدار اولیه و حلقه درونیWhile :

$$O(z^*[k*a*v+n*k^3+s]+v^2+a+v*s)$$

<u>پیچیدگی محاسباتی کل چرخه While</u>: به منظور بدست آوردن پیچیدگی محاسباتی مربوط به کل چرخه بیرونی While بایستی نتیجه قبلی را در r ضرب کنیم که r تعداد قوانین چند طبقه کشف شده است .این مقدار بسیار متغیر است و بر اساس مجموعه داده ها بسیار متغیرمی باشد. بنابراین داریم :

$$O(r^*\{z^*[k*a*v+n*k^3+s]+v^2+a+v*s\}+a*n)$$

بایستی این مطلب را مورد توجه قرار داد که این پیچیدگی محاسباتی به میزان زیادی به مقدار k(تعداد شروط هر قانون) و r (تعداد قوانین)وv(تعداد مقادیر صفات پیوسته) بستگی دارد.مقدار kوrوv از مجموعه داده ای به مجموعه داده ای دیگر بسیار متغیر است و به محتوای آن مجموعه داده بستگی دارد.

2-1-6-5- محاسن بکارگیری این روش در مقایسه با روشهای موجود

در این روش با توجه به اینکه حالت دستوری واژه مشخص می شود در ابهام زدایی از معنای واژه مانند روشهای معمولِ استفاده از پیکره که کل پیکره مورد جستجو قرار گیرد ، در این روش بعد از اینکه حالت دستوری واژه محرز شد و با توجه به تعریف مورچه های متخصص در این پروژه ، با توجه به حالت دستوری واژه تنها قسمتی از گراف مورد پیمایش قرار می گیرد ، از طرفی با توجه به مقدار دهی اولیه فرومون مسیر های موجود روی گراف که از روی پیکره زبان مقصد اسخراج می شود ، مورچه ها در پیدا کردن معنی واژه مبهم خیلی سریع تر می توانند به سمت معنی صحیح تر همگرا شوند.

نکته ی دیگر اینکه این روش مستقل از نوع زبان مقصد می باشد و به راحتی بر روی سایر زبانها نیز بعنوان زبان مقصد می تواند مورد استفاده قرار بگیرد.

7- نتیجه گیری

دراین پروژه با استفاده از پیکره زبان مقصد و بکارگیری الگوریتم کلونی مورچه ، به ابهام زدایی واژه های چند معنایی[1] زبان مبداء ، پرداخته شد.

1 Multi-meaning word

این رساله به بررسی یکی از مشکلات شاخص در رویه خودکار سازی تبدیل یک زبان نوشتاری به زبان دیگر، و ارائه راه حلی برای رفع آن می پردازد .

اصطلاح ترجمه ماشینی نامی استاندارد و رایج برای سیستم های کامپیوتری شده یا محاسباتی است که مسئول تولید ترجمه هایی از یک زبان طبیعی به زبان طبیعی دیگر با یا بدون کمک انسان باشند. در ماشین های ترجمه سعی میشود که ترجمه مناسبی برای رشته کلمات ورودی ارائه شود به گونه ای که ترجمه حاصل نه تنها معنای متن مبدأ را انتقال داده بلکه دارای سلاست کافی در زبان مقصد باشد.

ابهام واژگانی اشاره به حالتی دارد که یا یک واحد واژگانی متعلق به طبقات دستوری یا اجزاء کلام مختلف با معانی مختلف باشد و یا یک واحد واژگانی بیش از یک معنی داشته و معانی مختلف آن متعلق به یک طبقه دستوری باشند. هدف از ابهام زدایی معنی واژه ، رفع ابهام از حالت دوم است یعنی موردی که در آن معانی مختلف واژه متعلق به طبقه دستوری یکسانی باشند. ابهام زدایی معنی واژه اصطلاحی است که به استخراج معنی صحیح و مقتضی از واژه هایی که چند معنایی هستند اطلاق می شود و این کار غیر ممکنی نیست چرا که اگرچه واژه ها ممکن است معانی زیادی داشته باشند، طبیعتاً تنها یکی از آن معانی مناسب بافتی است که واژه مورد نظر در آن ظاهر می شود. هدف ابهام زدایی معنی واژه ، تعیین یکی از این معناها بر اساس بافت زبانی که این واژه در آن ظاهر شده می باشد. در بسیاری از سیستم هایی که کار ابهام زدایی معنی واژه را به عهده دارند سیستم برچسب گذاری اجزاء کلام وجود دارد که مقوله دستوری واژه ها را مشخص می نماید. مثلاً در مورد یک واژه خاص در صورتیکه سیستم برچسب گذاری اجزاء کلام تشخیص دهد که این واژه اسم است آنگاه سیستم ابهام زدایی معنی واژه ، کار خودرا شروع کرده و از میان معنای موجود، معانی اسمی این واژه را جدا کرده وسپس یکی را انتخاب می نماید.

ابهام زدایی معنی واژه یک امر ضروری در ترجمه ماشینی بوده و کاربرد زیادی درآن دارد چرا که سیستم ترجمه ماشینی بایستی قادر باشد تا مثلاً معنی واژه خاصی را در یک متن انگلیسی تشخیص داده و تعیین نماید که آن را به کدام کلمه معادل در فارسی ترجمه نماید.

در ماشین های ترجمه با بکارگیری شیوه های آماری[1] ، تحلیل دستوری[2] در سطح وابستگی ارتباطی و تحلیل معنایی[3] در سطح تشخیص جملات و اطلاعات لغوی ،همراه با مشخصه های دستوری در زبان مبداء ، وگرفتن بازخورد از کاربران ماشین میتوان با این ابهام مقابله کرد.

روشهای مختلفی برای ابهام زدایی از معنای واژه را مورد بررسی و مطالعه قرار دادیم و با توجه به نتایج بدست آمده ما معتقدیم که استفاده از پیکره زبان مقصد در امر ابهام زدایی واژه های چند معنایی زبان مبدأ کیفیت برونداد یک سیستم ترجمه ماشینی را تا حد قابل توجهی افزایش می دهد.

لذا بطور اجمالی کولونی مورچه و الگوریتم های بهینه سازی کولونی مطالعه شد و سپس با بکارگیری الگوریتم مذکور و پیکره زبان مقصد، روشی برای ابهام زدایی از معنی واژه ارائه شدکه با توجه به روشهای قبلی که بدین منظور استفاده شده است ،از دقت و صحت بالاتری برخوردار است . در مدل بیشترین احتمال همیشه گزینه ای با بیشترین احتمال به عنوان گزینه درست انتخاب می شود که دقت آن نسبت به روش موسوی و دلاور خلفی نیز پایین تر است . در روش موسوی و دلاورخلفی که از اعداد تصادفی[4] و آزمون نکویی برازش[5] استفاده نمودند، آنها با استفاده از اعداد تصادفی برای انتخاب بهترین معادل زبان مقصد برای یک واژه مبهم از زبان مبدأ در سیستم ترجمه ماشینی استفاده کرده و روش خود را در مورد پیکره زبانی نسبتاً بزرگی در حوضه روانشناسی در زبانهای انگلیسی و فارسی به ترتیب به عنوان زبانهای مبدأ و مقصد موردآزمایش قرار دادند. الگوریتم آنها روش توانست مشکل ابهام 604 واژه از 764 واژه مبهم یک پیکره زبانی انگلیسی را با استفاده از متن معادل آن در زبان فارسی حل نماید که بدین ترتیب صحت این مدل به 79% رسید (12). روش موسوی ودلاور خلفی نسبت به روش انتخاب بر اساس محتمل ترین حالت صورت می گیرد دقیق تر و علمی تر است

اما در این روش با بکارگیری الگوریتم کولونی مورچه و استفاده ازرابطه آماری AS بر روی یالی که مربوط با هم آیی واژه ها است نسب به روشهای قبلی از دقت بالاتری می تواندبرخورداراباشد و در نتیجه در کل باعث

[1] approaches Statistical-Based
[2] grammatical analysis
[3] semantic analysis
[4]- random numbers
[5]- goodness of fit test

ابهام زدایی از معنای واژه با دقت بالاتری می شود. در نهایت دیده شد که این روش با دقت 86% درصد کارایی قابل قبولی را از خودنشان می دهد.

مسیر کارهایی را که می توان بعنوان کار تحقیقاتی آینده در زمینه این پروژه معرفی نمود عبارتند از:

1- راهکارهایی برای مقدار دهی اولیه فرومون

2- راهکار هایی برای بروز رسانی بهینه تر فرومون

3- بررسی راهکارهایی برای تنظیم پویای تعداد مورچه های متخصص در هر مهارت. بعبارت دیگر تغییر تعداد عوامل متخصص در هر مهارت بر حسب نیاز کولونی

منابع و مراجع :

1. Frawley, W., Piatetsky-Shapiro, G., Matheus, C. J., “Knowledge Discovery in Databases: An Overview”, in Knowledge Discovery in Databases,G.Piatesky-Shapiro and W. Frawley, (Eds.), MIt Press, 1991

2. M. Dorigo and L. M. Gambardella, " Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem", IEEE Trans on Evolutionary Computation1(1), 53-66, 1997.

3. M.dorigo and G.Di Caro,"The ant colony optimization meta-heuristic", In New Ideas in Optimization,D.Corne,M.Dorigo and F.Glover Eds.London,UK:McGraw Hill,pp.11-32,1999.

4. M.Dorigo,G.Di Caro and L. M. Gambardella, "Ant algorithms for discrete optimization",Artificial Life, vol.5,no.2,pp.137-172,1999.

5. Hutchins, W. J. and Somers, H, "An introduction to machine translation", London, Academic Press,1993.

6. Hutchins, W. J. , "Machine translation: half a century of research and use". Prepared for UNED Summer School at Avila, Spain. Web site: http://ourworld.compuserve.com/homepages/WJHutchins,2003

7. Frederking, R. and Nirenburg, S. "Tree heads are better than one". In: Proceeding of ANLP-94, Stuttgart, Germany. (1994)

8. Freivalds, J, " The technology of translation". Managemant Review Vol.8. No.7, pp.48-52, (1999).

9. Heyn, M. "Integrating machine translation into translation memory systems". In: European Assiciation for Machine Translation Workshop proceeding, ISSCO, Geneva. pp. 111-123, (1996).

10. Hutchins, J. "Translation technology and the translator". Proceeding of 11th Conference of the Institute of Translation and Interpreting. pp. 113-120,(1997).

11. Somers, H. " Computers and translation: a translator's guide". Amsterdam, John Benjamins, (2003).

12. Mosavi Miangah, T. and Delavar Khalafi "A. Word sense disambiguation using target language corpus in a machine translation system. Literary and Linguistic Computing", 20(2), pp. 237-249, (2005).

13. Mosavi Miangah, T. “Morphological analysis in the system of English-Persian machine translation”. Ph. D. Dissertation, Russian Academy of Sciences, Moscow. (2002).

14. Mosavi Miangah “Acom1075 Machine translation. Multilingual Computing”, Semester2 http://www.leeds.ac.uk/acom/teaching/mlc/handouts/lecture4_handout.pdf, (2003-04).

15. Seasly, J. “ Machine translation: a survey of approaches”, Technical Report, University of Michigan, http://www.personal.umich.edu/jseasly/home.html, (2003).

16. Robert E. Frederking ” Corpus-Based MT and Multi-Engine MT within Nespole”,Project IST-1999-11562 , WP5 - HLT development , 30th July 2002

17. Brown, P. F.; Stephen, A.; Della pietra, V.; Della Pietra, J. and Robert L. Marcer “The mathematics of statistical machine translation: parameter estimation. Computational Linguistics”, 19(2), pp. 263-311, (1993).

18. Somers, H, Review Article: “ Example-based machine translation. Machine Translation”, 14(2), pp. 113-157. (1999).

19. Turcato, D. and Popowich, F,What is example-based machine translation. In: “Proceedings of the workshop on Example-based Machine Translation, hosted by MT-Summit VIII”, Santiago de Compostela, Spain, September, 18-22, (2001).

20. J. Casillas, O.Cordon, F. Herrera, “ Learning Fuzzy Rules Using Ant Colony Optimization Algorithms”, Supported by CICYT, PB98-1319,1998.

21. Rayner, M. and Bouillon, P, “ Hybrid transfer in an English-French spoken language translator”, Proceeding of IA’95, Montpellier, France. http://www.cam.sri.com/tr, (1995).

22. Iomdin, L. and Streiter, O, “Learning from parallel corpora: Experiments in machine translation”. In: Proceeding of the Dialigue’99 International Seminar in Computational Linguistics and Application, 2, pp. 79-88, (1999).

23. Carl, M., Iomdin, L., Pease, C. and Streiter, O. “Toward a dynamic linkage of example-based and rule-based machine translation. In: Hybrid Approaches to Machine

Translation",Streiter, O. Carl, M. and Haller, J. (Eds). IAI Web-based working paper. http://www.iai.uni-sb.de, (2000).

24. Carl, M. and Hansen, S, "Linking translation memories with example-based machine translation", In: Hybrid Approaches to Machine Translation.Streiter, O. Carl, M. and Haller, J. (Eds). IAI Web-based working paper. http://www.iai.uni-sb.de,2000

25. http://www.trajome.org

26. Justeson, J. S. and Katz, S. M., " Principled disambiguation: discriminating adjective senses with modified nouns", Computational Linguistics, 21(1), pp. 1-28, (1995).

27. http://www.persian-language.org/

28. Yarowsky, D , "Word sense disambiguation using statistical models of Roget's categories trained on large corpora",Proceeding of 15th International Conference on Computational Linguistics, pp. 454-460, (1992).

29. M.Dorigo,A. Colorni and V.Maniezzo, "The Ant System: optimization by a colony of cooperating agents",IEEE Trasactions on Systems, Man, and Cyberneticspart B, vol.26,no.1,pp.29-41,1996.

30. Bruce, R. and Wiebe, J, "Word sense disambiguation using decomposable models. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, (1994).

31. Ng, H. T. and Lee, H. B, "Integrating multiple knowledge sources to disambiguate word sense: an example-based approach", Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp. 40-47, (1996).

32. Dagan, I. And Itai, "A. Word sense disambiguation using a second language monolingual corpus. Computational Linguistics, 20(4), pp. 563-596, (1994).

33. Hyun Ah Lee, Jong C. Park and Gil Chang Kim," Lexical Selection with a Target Language Monolingual Corpus and an MRD", Korean Advanced Institute of Science and Technology,Dept. of Computer Science

34. E. Suzuki, M.Gotoh and Y. Choki, "Bloomy Decision Tree for Multi-Objective Classification " ,©Springer-Verlag Berlin Heidelberg,2001

35. U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "from data mining to knowledge discovery: an overview",In: Advances in Knowledge Discovery & Data Mining, U.m. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy(Eds) Cambridge, MA:AAA/MIT, pp.1-34,1996.

36. P.L. Carbone, "Data Mining or Knowledge Discovery in Databases", An Overview© MITRE Corporation, 1997.

37. G. Di Caro, "An Introduction to swarm Intelligence Issues",IDSIA, USI/SUPSI,Lugano(CH),2002.

38. M.Wooldridge and N. R. Jennings, " Intelligent agents: Theory and practice", The Knowledge Engineering Review, 10(2): 115-152, 1995.

39. U. Fayyad, G. P. Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, 37-54, Fall 1996.

40. E.Bonabeau, M. Dorigo and G. Theraulaz, Swarm Intelligence: "From Natural to Artificial Systems". New York, NY: Oxford University Press, 1999.

41. R.S. Parpinelli, H.S.Lopes and A.A.Freitas, "Data Mining with an Ant Colony Optomization Algorithm",IEEE Trans on Evolutionary Computation, special issue on Ant Colony Algorithms,6(4), 2002.

42. R.S. Parpinelli, H.S.Lopes and A.A.Freitas, "An Ant Colony Algorithm for Classification Rule Discovery", Data Mining: a Heuristic Approach, 191-208, Idea Group Publishing ,London ,2002

43. D. A. Holway, A. V. Suarez, "Colony-structure variation and inter-specific competitive ability in the invasive Argentine ant",Oecologia(2004),Received: 28 May 2003 / Accepted: 17 September 2003 / Published online: 18 October 2003 @Springer Verlag 2003.

44. Hedden, T. D. "Machine translation: a brief introduction". http://www.hedden.org/intro-mt.html, (1992-2003).

45. Zh.I. Reznikova,"Government and nepotism in social insects:New dimension provided by an experimental approach",c Euroasian Entomological, Journal, Volume 2,Number 1,2003.

46. دکتر علی محمدی ، ابوذر جمال نیا ،" الگوریتم های لانه مورچه و کاربرد آن در نگهداری پیشگیرانه"، اولین همایش ملی مدیریت صنعتی،1386

47. رضا شمسایی،"الگوریتم کشف دانش چند کلاسه با استفاده از الگوریتم بهینه سازی کولونی مورچه"، پایان نامه کارشناسی ارشد،دانشگاه علم و صنعت،صفحات 68-60،1383

48. بهناز شب رو، ماشین های ترجمه مبتنی بر پیکره،سمینار کارشناسی ارشد،دانشگاه علم وصنعت،صفحات 15-10،پاییز 1386

49. طیبه موسوی میانگاه ،" *درآمدی برترجمه ماشینی*"، انتشارات یلداقلم، تهران،فصل سوم،صفحات 93-85،1386

پیوست

قسمتی ازبرونداد پیکره برای چندواژه مبهم به شکل جفت های خصیصه – ارزش

ردیف	واژه مبهم	معنی واژه مبهم	واژه همبسته	معنی واژه همبسته	میزان همبستگی دوواژه درپیکره (AS)
1	**power**	اختیار	**play**	بازی	.0012
2	**power**	اختیار	**play**	نمایش	7.14E-05
3	**power**	اختیار	**play**	تفریح	6.7E-06

4	**power**	قدرت	**play**	بازی	.0001
5	**power**	قدرت	**play**	نمایش	.0026
6	**power**	قدرت	**play**	تفریح	.00018
7	**power**	حق	**play**	بازی	.0014
8	**power**	حق	**play**	نمایش	.0006
9	**power**	حق	**play**	تفریح	.00012
10	**wordy**	پرگو	**conflict**	مشاجره	NaN
11	**wordy**	پرلغت	**conflict**	مشاجره	NaN
12	**wordy**	لغت دار	**conflict**	مشاجره	NaN
13	**wordy**	لفظی	**conflict**	مشاجره	0.0006
14	**wordy**	پرگو	**conflict**	تضاد	6.77E-06
15	**wordy**	پرلغت	**conflict**	تضاد	1.22E-06
16	**wordy**	لغت دار	**conflict**	تضاد	1.44E07
17	**wordy**	لفظی	**conflict**	تضاد	8.88E-05
18	**wordy**	پرگو	**conflict**	درگیری	3.7E-7
19	**wordy**	پرلغت	**conflict**	درگیری	NaN
20	**wordy**	لغت دار	**conflict**	درگیری	1.93E-07
21	**wordy**	لفظی	**conflict**	درگیری	0.0001
22	**wordy**	پرگو	**conflict**	ستیزه	NaN

ردیف	واژه مبهم	معنی واژه مبهم	واژه همبسته	معنی واژه همبسته	میزان همبستگی دوواژه درپیکره (AS)

23	**wordy**	پرلغت	**conflict**	ستیزه	NaN
24	**wordy**	لغت دار	**conflict**	ستیزه	1.43E-07
25	**wordy**	لفظی	**conflict**	ستیزه	1.2E-05
26	armed	مسلح	conflict	مشاجره	5.77E-05
27	armed	مسلحانه	conflict	مشاجره	3.70E-6
28	armed	مجهز	conflict	مشاجره	.00010
29	armed	مسلح	conflict	تضاد	2.05E-5
30	armed	مسلحانه	conflict	تضاد	2.41E-5
31	armed	مجهز	conflict	تضاد	0.004
32	armed	مسلح	conflict	درگیری	.016
33	armed	مسلحانه	conflict	درگیری	4.95E-6
34	armed	مجهز	conflict	درگیری	0.00012
35	armed	مسلح	conflict	ستیزه	0.00045
36	armed	مسلحانه	conflict	ستیزه	2.22E-5
37	armed	مجهز	conflict	ستیزه	5.77E-05
38	political	سیاسی	figure	نماد	0.00037
39	political	سیاسی	figure	عدد	5.33E-07
40	political	سیاسی	figure	شخصیت	0.0044
41	political	سیاسی	figure	نشانه	2.57E-05
42	figure	نماد	peace	آرامش	0.00098
43	figure	عدد	peace	آرامش	NaN
44	figure	شخصیت	peace	آرامش	0.00017

45	figure	نشانه	peace	آرامش	9.98E-05
46	figure	نماد	peace	صلح	.0011

ردیف	واژه مبهم	معنی واژه مبهم	واژه همبسته	معنی واژه همبسته	میزان همبستگی دوواژه درپیکره (AS)
47	figure	عدد	peace	صلح	0.00049
48	figure	شخصیت	peace	صلح	0.00048
49	figure	نشانه	peace	صلح	0.00019
50	figure	نماد	peace	دوستی	3.95E06
51	figure	عدد	peace	دوستی	7.44E-06
52	figure	شخصیت	peace	دوستی	2.87-6
53	figure	نشانه	peace	دوستی	3.09E-05
54	civil	داخلی	code	رمز	3.71E-05
55	civil	کشوری	code	رمز	2.44E-05
56	civil	غیرنظامی	code	رمز	0.00014
57	civil	مدنی	code	رمز	NaN
58	civil	داخلی	code	قانون	0.0019
59	civil	کشوری	code	قانون	3.15E-05
60	civil	غیرنظامی	code	قانون	0.00012
61	civil	مدنی	code	قانون	.00076
62	civil	داخلی	code	علامت	3.14E-05
63	civil	کشوری	code	علامت	0.0002

64	civil	غیرنظامی	code	علامت	9.76E-06
65	civil	مدنی	code	علامت	6.54E-05
66	civil	داخلی	war	جنگ	.094
67	civil	کشوری	war	جنگ	0.0057
68	civil	غیرنظامی	war	جنگ	0.0014
69	civil	مدنی	war	جنگ	0.0044
70	closed	بسته	session	جلسه	2.84E-05
71	closed	سرّی	session	جلسه	6.44E-05

ردیف	واژه مبهم	معنی واژه مبهم	واژه همبسته	معنی واژه همبسته	میزان همبستگی دوواژه درپیکره (AS)
72	closed	خاتمه یافته	session	جلسه	1.67E-06
73	closed	غیرعلنی	session	جلسه	0.0076
74	closed	بسته	session	قسمت	1.15E-05
75	closed	سرّی	session	قسمت	2.80E-5
76	closed	خاتمه یافته	session	قسمت	7.48E-07
77	closed	غیرعلنی	session	قسمت	NaN
78	closed	بسته	session	دوره	NaN
79	closed	سرّی	session	دوره	4.55E-05
80	closed	خاتمه یافته	session	دوره	0.00016
81	closed	غیرعلنی	session	دوره	NaN
82	academic	علمی	exchange	بورس	9.65E-05

83	academic	فرهنگی	exchange	بورس	9.53E-05
84	academic	دانشگاهی	exchange	بورس	2.72E-05
85	academic	علمی	exchange	ارز	3.31E-05
86	academic	فرهنگی	exchange	ارز	NaN
87	academic	دانشگاهی	exchange	ارز	5.7E-07
88	academic	علمی	exchange	مبادله	3.42E-05
89	academic	فرهنگی	exchange	مبادله	0.0016
90	academic	دانشگاهی	exchange	مبادله	0.00087
91	exchange	بورس	rate	سرعت	2.05E-07
92	exchange	ارز	rate	سرعت	.00011
93	exchange	مبادله	rate	سرعت	.00011
94	exchange	بورس	rate	میزان	.00016
95	exchange	ارز	rate	میزان	.00099
ردیف	واژه مبهم	معنی واژه مبهم	واژه همبسته	معنی واژه همبسته	میزان همبستگی دوواژه درپیکره (AS)
۹۶	exchange	مبادله	rate	میزان	5.79E05
97	exchange	بورس	rate	نرخ	.00071
98	exchange	ارز	rate	نرخ	.38
99	exchange	مبادله	rate	نرخ	.071
100	exchange	بورس	rate	درجه	.00028
101	exchange	ارز	rate	درجه	9.67E-05
102	exchange	مبادله	rate	درجه	.00042