عنوان زیرپروژه:

بررسی ابعاد و تفاوت های پیکره های برچسب داده ای و پیکره های خام در زبان فارسی

فهرست مطالب

1. مقدمه

پردازش زبان طبیعی عبارتست از استفاده از رایانه برای پردازش زبان گفتاری و نوشتاری. پردازش زبان طبیعی می‌تواند در سطوح مختلف از زبان صورت گیرد. این سطوح را می‌توان به شکل زیر تقسیم‌بندی نمود:

1. آواشناسی و صدا شناسی[۱] که به تشخیص آواها و صداها و بازشناسی گفتار می‌پردازد.
2. ریخت‌شناسی[۲] که به ساختارهای کلمات و ریشه‌یابی واژگان می‌پردازد.
3. نحو[۳] که به ارتباط کلمات به همدیگر و مباحث دستوری آن‌ها در گروه‌ها و جملات می‌پردازد.
4. معناشناسی[٤] که به ارتباطات معنایی کلمات می‌پردازد.
5. عمل‌گرایی[٥] که کاربردهای زبان برای رساندن یک مطلب به مخاطب یا مخاطبان، در حالت عملی و یا در نوشتار و گفتار طبیعی می‌پردازد.
6. مباحثه[٦] که به ارتباطات کلی یک زبان فرای یک یا چند جمله خاص می‌پردازد.

از کاربردهای اصلی پردازش زبان طبیعی می‌توان موارد زیر را ذکر کرد: 1) خلاصه‌سازی خودکار[۷]، 2) کمک به خواندن زبان‌های طبیعی دیگر[۸]، 3) کمک به نوشتن به زبان‌های طبیعی دیگر[۹]، 4) استخراج

---

[۱] Phonetics and Phonology

[۲] Morphology

[۳] Syntax

[٤] Semantics

[٥] Pragmatics

[٦] Discourse

[۷] Automatic Summarization

[۸] Foreign language reading aid

[۹] Foreign language writing aid

اطلاعات[۱]، 5) بازیابی اطلاعات[۲]، 6) ترجمه ماشینی[۳]، 7) تشخیص واحدهای اسمی[٤]، 8) تولید زبان طبیعی[٥]، 9) فهم زبان طبیعی[٦]، 10) نویسه‌خوان نوری[۷]، 11) تحلیل مرجع‌دارها[۸]، 12) سیستم سوال، پاسخ[۹]، 13) تشخیص گفتار[۱۰]، 14) مبدل متن به گفتار[۱۱]، 15) نظام‌های مکالمه گفتاری[۱۲]، 16) ساده‌سازی متن[۱۳]، 17) تایید متن[۱٤]

با توجه به کاربردهای فوق، برخی مسائل اساسی در پردازش زبان طبیعی را می‌توان به صورت زیر برشمرد:

1.قطعه‌بندی گفتار[۱٥]

2.قطعه‌بندی متن16

3.برچسب‌گذاری نقش کلمه1

---

[۱] Information Extraction

[۲] Information Retrieval

[۳] Machine Translation

[٤] Name Entity Recognition

[٥] Natural language generation

[٦] Natural language Understanding

[۷] Optical Character Recognition - OCR

[۸] Anaphora Reservation

[۹] Question Answering System

[۱۰] Speech Recognition

[۱۱] Text-to-Speech

[۱۲] Spoken Dialogue System

[۱۳] Text Simplification

[۱٤] Text Proofing

[۱٥] Speech Segmentation

[۱٦] Text Segmentation

4.ابهام‌زدائی از نقش کلمه2

5.ابهام‌زدائی نحوی3

6.هنجارسازی4

7.تشخیص اعمال گفتاری5

پردازش زبان طبیعی در حوزه متن یا خط و زبان به پردازش پیکره‌های متنی[۶] به عنوان نمایانگر زبان می‌پردازد. بنابر این مسائل اصلی در حوزه خط و زبان را می‌توان به موارد زیر کاهش داد:

1. قطعه‌بندی متن
2. برچسب‌گذاری نقش کلمه
3. ابهام‌زدائی از نقش کلمه
4. ابهام‌زدائی نحوی
5. هنجارسازی

مسائل فوق تقریباً در تمامی کاربردهای پردازش زبان طبیعی در حوزه خط و زبان مطرح هستند و به عنوان مسائل زیرساختی خط و زبان می‌باید مورد بررسی قرار گیرند. زبان فارسی به عنوان یکی از زبان‌های طبیعی نیز از این قاعده مستثنی نیست. از طرفی همان طور که مطرح شد، در حوزه خط و

[۱] Part-of-Speech Tagging

[۲] Word Sense Disambiguation

[۳] Syntactic Disambiguation

[۴] Normalization

[۵] Speech acts

[۶] Text Corpus

زبان، پیکره‌های متنی به عنوان نمادی از زبان هستند که می‌باید با تحلیل آن‌ها به استخراج اجزا، قواعد و ساز و کار زبان پی برد.

دادگان مهم و زیرساختی حاصل از تحلیل پیکره‌های متنی در پردازش زبان طبیعی و حوزه خط و زبان را می‌توان به موارد زیر تقسیم نمود:

1. پیکره متنی
2. واژگان زبان (واژگان عمومی و تخصصی)
3. گنجواژه یا اصطلاح‌نامه[1]
4. الگوهای زبان
5. پیکره‌های برچسب داده‌ای
6. پیکره‌های تخصصی

از میان دادگان فوق، تمرکز این مستند بر روی پیکره‌های متنی و بویژه پیکره‌های برچسب داده‌ای است.

بنابراین، ابتدا به معرفی پیکره متنی و پیکره‌های برچسب داده‌ای پرداخته و سپس نقش پیکره‌های برچسب داده‌ای در دو مورد از کاربردهای پردازش زبان طبیعی (خلاصه‌سازی و ترجمه‌ی ماشینی) را برمی‌شماریم.

[1] Thesaurus

2. پیکره متنی

در علوم زبان شناسی و زبان شناسی رایانه‌ای، پیکره متنی، حجم بسیاری از متون ساخت‌یافته آن زبان است. در (Atkins and Clear, 1992) نیز تعریفی که برای پیکره ارائه شده است به این صورت است: حجم زیادی از داده‌های زبانی که براساس معیارهای مشخص برای هدف معینی جمع‌آوری و ذخیره شده‌اند به طوری که نمایندهٔ زبان یا گویش مورد مطالعه باشد. به طور کلی در طراحی و تهیه یک پیکره برچسب‌خورده یکی از مهمترین مسائلی که باید مورد توجه قرار گیرد مجموعه برچسب پیکره است که بر اساس هدف غایی پیکره و منظوری که از پیکره مد نظر است حاوی برچسب‌هایی خواهد شد که نیل به آن هدف و منظور را ممکن سازد.

پیکره‌های متنی به منظور تحلیل‌های آماری، صحت‌سنجی فرضیه‌ها و بررسی رخداد یا صحت قواعد زبانی در حوزه‌ای مشخص به کار می‌روند. پیکره‌های متنی به عنوان پایگاه دانش[1] اصلی در زبان‌شناسی رایانه‌ای خصوصا در حوزه خط و زبان شناخته می‌شوند. پیکره‌های می‌توانند تک زبانی[2] (شامل متون با یک زبان) و یا چند زبانی[3] (شامل متون با چندین زبان(باشند، که مورد اخیر معمولا برای مقایسه نظیر-به-نظیر ساختار دهی می‌شود و پیکره موازی منطبق[4] نامیده می‌شود.

به منظور مناسب‌سازی پیکره‌های متنی برای کاربرد مفیدتر در مطالعات زبان‌شناسی، پیکره‌ها می‌باید حاشیه‌نویسی[5] شوند. حاشیه‌نویسی عبارت است از تحلیل و افزودن برخی اطلاعات مانند اطلاعاتی در مورد نقش[6] و یا ریشه[7] کلمات موجود در متن به پیکره. پیکره‌های متنی اکثراً در سطح ساختاری

---

[1] Knowledgebase

[2] Monolingual

[3] Multilingual

[4] Aligned Parallel Corpus

[5] Annotated

[6] Part-of-Speech - POS

[7] Lemma

حاشیه‌نویسی می‌شوند. پیکره‌هایی که کاملا تجزیه[1] و حاشیه‌گذاری شده باشند، پیکره‌های تجزیه شده یا بانک درختی[2] گفته می‌شود. عملا پیکره‌های کوچک (بین 1 تا 3 میلیون کلمه) امکان تجزیه کامل را دارند زیرا حصول اطمینان از این که حاشیه‌نویسی کاملا صحیح و سازگار است، عملی بسیار پیچیده، زمان‌گیر و پر هزینه خواهد بود. امکان حاشیه‌نویسی در سطوح ریخت‌شناسی، نحوی، معنایی و کاربردی معمولاً برای پیکره‌های کوچک امکان پذیر است.

از چالش‌های مطرح در خصوص پیکره‌های متنی می‌توان به این نکته اشاره کرد که چه پیکره‌ای با چه خصوصیاتی می‌تواند به بهترین و بیشترین شکل خصوصیات زبان را بیان سازد. با توجه به تغییرپذیری زبان در زمان، این نکته که چه متونی با چه ویژگی‌هایی می‌توانند پیکره‌ی مناسبی را شکل دهند اهمیت می‌یابد. این که چه سهمی از از پیکره از زبان معیار انتخاب شود، چه میزان آن از زبان فوق معیار باشد و چه بخشی از آن را زبان زیر معیار دربر گیرد از چالش‌های اساسی در انتخاب و ایجاد پیکره‌ها است. همچنین تعریف مناسب از زبان معیار، فوق و زیر معیار نیز در این امر تاثیر گذار است و می‌باید مورد توجه قرار گیرد.

در سه دهه‌ی گذشته، بیشتر کشورهای صنعتی به ایجاد بانک‌های زبانی خود پرداخته‌اند. این کشورها، ابتدا با گردآوری داده‌ها یا پیکره های متنی، و سپس با سازمان‌دهی آن‌ها در پایگاه‌های داده‌ها و بانک‌های اطلاعات زبانی، از شبکه‌های بین‌المللی داده های زبانی[3] بهره برداری می کنند .

بسیاری از پژوهشهای زبان‌شناختی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، تنها با استفاده از یک پیکره زبانی امکان‌پذیر است. بنابراین، پیکره متنی پایه‌ای‌ترین و یکی از مهم‌ترین دادگان ورودی در کاربردهای پردازش زبان طبیعی بشمار می‌رود. در واقع، پیکره‌های متنی، پایگاه دانش اصلی در زبان‌شناسی پیکره‌ای است. تحلیل و پردازش انواع مختلفی از پیکره‌ها، موضوع مقاله‌های بسیاری در حوزه‌ی زبان‌شناسی رایانشی و پردازش زبان طبیعی بوده است. ایجاد مدل‌های مارکف از پیکره‌های متنی، برای اهدافی همچون برچسب‌زنیِ نقش کلمه مورد استفاده قرار می‌گیرد. پیکره‌های متنی، کاربردهایی خارج از حوزه‌ی پردازش زبان طبیعی نیز دارند. بعنوان مثال می‌توان فهرست‌هایی از فراوانی کلمات را از پیکره‌ی

---

[1] Parse

[2] Treebank

[3] International Networks of Linguistic Data

متنـــــی اســـــتخراج کـــــرده و در آمـــــوزش زبـــــان مـــــورد اســـــتفاده قـــــرار داد.

3. پیکره‌های متنی برچسب داده‌ای و نقش آن در کاربردهای پردازش زبان طبیعی

پروژه‌های زیرساختی در حوزه خط و زبان فارسی را می‌توان پروژه‌هایی برای ایجاد دادگان زیرساختی در حوزه خط و زبان فارسی در نظر گرفت. از طرفی ابزارهایی نیز جهت حل مسائل زیرساختی مطرح شده در حوزه خط و زبان نیز نیاز هستند تا بتوان اقدام به تحلیل پیکره‌های متنی زبان جهت تولید دادگان زیرساختی نمود. بنابراین پروژه‌های زیرساختی خط و زبان فارسی تلفیقی از ابزارها و دادگان زیرساختی زبان خواهد بود. در این مستند، لزوم برچسب‌گذاری پیکره‌های متنی مورد بحث قرار می‌گیرد. بررسی ابزارهای برچسب‌گذاری و مقایسه‌ی میان آن‌ها، از حوزه‌ی مطالب این مستند خارج است.

برچسب‌گذاری پیکره‌های زبانی به طور کلی می‌تواند در چهار سطح زبانی انجام شود که عبارتند از (بی‌جن‌خان، 1381):

1- تعیین برچسب مقوله کلمه[1]

2- برچسب‌گذاری نحوی: که شامل پردازش جملات و به دست آوردن درخت نحوی آن‌ها است.

3- برچسب‌گذاری معنایی: که عبارت است از استخراج صورت منطقی جملات و به دست آوردن تعبیر معنایی آن‌ها. پیش‌نیاز این نوع برچسب‌دهی، انجام برچسب‌دهی مقوله کلمات است، به این تعبیر که پیش از آن‌که تعبیر معنایی کلمات و جملات یک متن به دست داده شود، تعیین مقوله کلمات آن‌ها ضروری است.

4- برچسب‌گذاری کاربردشناختی[1]: که عبارت است از تعیین روابطی که میان دو کلمه در یک متن وجود دارد. به عنوان مثال مشخص کردن ضمایر و مرجع آن‌ها در متن در حوزه برچسب‌دهی کاربردشناختی قرار می‌گیرد.

---

[1] Wordclass tagging

با توجه به سطوح زبانی در برچسب‌گذاری پیکره‌های زبانی، ابتدا باید یک مجموعه برچسب برای برچسب‌گذاری یک پیکره طراحی شود. تعیین برچسب مقولاتِ اصلی، اولین مرحله در تعیین برچسب‌های موجود در مجموعه برچسب است. بر این اساس در تهیه پیکره متنی زبان فارسی نیز باید مجموعه برچسب به گونه‌ای باشد که تمام مقوله‌های دستوری و انواع کلمات زبان فارسی را دربرگیرد و علاوه بر این مطابق با استانداردهای برچسب‌گذاری متون در زبان‌های دیگر مانند انگلیسی و زبان‌های اروپایی باشد.

برچسب‌هایی که برای یک پیکره در نظر گرفته می‌شود به طور کلی به سه دسته برچسب تقسیم می‌شوند که عبارتند از (Cloern, 1999):

1- برچسب‌های نحوی-ساختواژی[2]: که اصلی‌ترین برچسب‌ها هستند. این برچسب‌ها شامل مقوله‌های نحوی اصلی از جمله: فعل، اسم، صفت، قید و غیره هستند. اغلب کلماتی که در متون وجود دارند به یکی از این مقوله‌های اصلی تعلق دارند. اصلی‌ترین مقوله‌های نحوی که در اغلب پیکره‌های زبانی در نظر گرفته می‌شوند شامل مقوله اسم، صفت، حرف اضافه، حرف ربط، حرف تعریف، قید و عدد هستند.

2- برچسب‌های خاص[3]: که شامل کلماتی هستند که در طبقه مقوله‌های اصلی قرار نمی‌گیرند، اما تعیین برچسب آن‌ها در استخراج اطلاعات زبانی از پیکره حائز اهمیت است. ادات شرط، حرف ندا، تکواژ صفت‌ساز از این دسته‌اند.

3- برچسب‌های متفرقه[4]: نیز شامل کلماتی است که در طبقه مقوله‌های اصلی قرار نمی‌گیرند اما در متن وجود دارند و به عنوان کلمات مجزا استخراج شده‌اند. کلمات خارجی، نشانه‌ها و علائم ریاضی و نیز علائم اختصاری در این طبقه قرار می‌گیرند.

[1] Pragmatic tagging

[2] Morphosyntactic

[3] Unique tags

[4] Residual /Miscellaneous tags

جدای از این تقسیم‌بندی برچسب‌های مجموعه برچسب پیکره متنی زبان فارسی یا برچسب اصلی‌اند یا برچسب‌هایی هستند که به عنوان زیربخش‌های برچسب‌های اصلی استفاده می‌شوند. به عنوان مثال کلمه‌ای مانند "کتابهایم" در پیکره به صورت زیر نمایش داده شده است:

کتابهایم N N, COM, PL,1

ستون اول از چپ (N) برچسب اصلی کلمه، ستون دوم (N, COM, PL,1) برچسب سلسله‌مراتبی[1] و ستون آخر (کتابهایم) خود کلمه می‌باشد.

برچسب‌گذاری ادات سخن[2] عمل انتساب برچسب‌های واژگانی به کلمات و نشانه‌های تشکیل‌دهنده یک متن است، به صورتی که این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جلمه باشند. درصد بالایی از کلمات از نقطه نظر برچسب واژگانی دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های واژگانی متفاوت دارند. بنابراین برچسب‌گذاری واژگانی عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینه[1] مورد نظر است. برچسب‌گذاری واژگانی عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی از قبیل ترجمه ماشینی، خطایاب و تبدیل متن به گفتار می‌باشد.

برچسب‌گذاری پیکره‌ها به عنوان یکی از مهم‌ترین حاشیه‌نویسی‌ها برای پیکره‌ها مطرح است. برچسب‌گذاری کاربردهای فراوانی در استخراج الگوهای اجزای واژگانی کلام در زبان دارند، به عبارت دیگر دستورات و قوانین زبان را می‌توان از طریق برچسب‌گذاری به مدلی آماری تبدیل نمود که قابل استفاده در زبان‌شناسی رایانه‌ای هستند. با استفاده از چنین مدل‌هایی می‌توان از کلمات و عبارات رفع ابهام کرد، عبارات نادرست دستوری را تشخیص داد، همچنین می‌توان از آن به عنوان مقدمه‌ای جهت استفاده از لایه‌های معنایی زبان یاد کرد.

در ادامه‌ی این بخش، نقش و ارتباط پیکره‌های برچسب داده‌ای در برخی از کاربردهای پردازش زبان طبیعی را مورد بررسی قرار خواهیم داد.

---

[1] Hierarchical tag

[2] Part of Speech - POS

3-1.خلاصه‌سازی

خلاصه‌سازی عبارت است از نمایش فشرده و دقیق متن ورودی به نحوی که متن خروجی مفاهیم مهم متن ورودی را در برداشته باشد. همگان بر این باورند که فرآیند خلاصه‌سازی خودکار متون باید بر پایه فهم کامل متن باشد و فرآیندی که انسان در این مورد طی می‌نماید را به نحوی تقلید نماید. به طور کلی خلاصه‌سازی یکی از زیر مجموعه مشکلات پیچیده پردازش زبان طبیعی است که هنوز بصورت کامل حل نشده است.

هدف خلاصه‌سازی خودکار آن است که مراحلی که توسط انسان انجام می‌شود، شناسایی شده و در خلاصه‌سازی خودکار انجام شود. بدین‌صورت که تمام متن خوانده و فهمیده شود و سپس خلاصه تولید شود. فهمیدن متن شامل تشخیص قسمت‌های مهم و غیر مهم متن است. مرحله تبدیل متن ورودی به متن خلاصه شامل مشخص کردن کلمات کلیدی، مفهوم اصلی، کلمه‌های مهم و جمله‌های مهم می‌باشد.

در خلاصه‌سازی باید در نظر داشت که: 1) به چه صورت یک توصیف نرمال از متن ورودی می‌توانیم داشته باشیم. 2) چگونه مهمترین بخش متن را تشخیص دهیم 3) چگونه متن را تجزیه کرده و متن خلاصه را ایجاد نماییم. وجود یک سیستم تطابق دهنده که متن ورودی انتخابی را به پایگاه داده از قبل ایجاد شده مرتبط نماید ضروری به نظر می‌رسد.

خلاصه‌سازی از دید ماشین، شامل درک زبان طبیعی است. درک زبان طبیعی به اطلاعاتی همچون «دانش صوت‌شناسی»، «دانش مورفولوژی»، «دانش نحوی» و «دانش عملی» و «دانش سخن» نیازمند است.

برخلاف زبان انگلیسی که در آن هم حروف و هم لغات کاملا متمایز از یکدیگرند، در زبان فارسی پیوستگی میان برخی علائم با لغات وجود دارد و علاوه بر آن تنوع نگارش در کلمات نیز موجود می‌باشد. ریشه یابی فعل که یکی از مراحل مهم پیش پردازش متن برای خلاصه‌سازی می‌باشد، در زبان فارسی چالش‌های خاص خود را دارد. به عنوان مثال در یک لغت به هم پیوسته هم بن فعل، شناسه، علامت زمان فعل و حتی شناسه‌های مفعولی می‌توان داشت که کار پردازش لغات را پیچیده‌تر می‌نماید به طوری که نمی‌توان از دانش، تجربه و نرم‌افزارهای موجود در این زمینه استفاده نمود و تولید نرم‌افزاری که قادر به حل تمامی این پیچیدگی‌ها باشد، فرایندی زمان‌بر و مستلزم تلاش فراوان می‌باشد.

مسائل و خصوصیات نگارشی زبان فارسی، امر پیش‌پردازش خودکار خلاصه‌سازی متون را دچار چالش می‌نماید. این چالش‌های شامل چالش‌های نگارشی و چالش‌های ریشه‌یابی است. با در دست داشتن پیکره‌ی برچسب داده‌ای، برخی از چالش‌های نگارشی و ریشه‌یابی را می‌توان مرتفع نمود.

خلاصه‌سازی متن از سه گام اصلی تشکیل شده است. مرحله پیش پردازش متن، مرحله پردازش متن و تولید خلاصه. مرحله پیش پردازش شامل حذف اطلاعات اضافی و غیر مهم و ریشه یابی می‌باشد. به عبارت دیگر، پیش پردازش متن در این بخش انجام می‌شود. بعد از پیش پردازش، بخش مهم متن باقی می‌ماند. بخش اصلی خلاصه‌سازی همان تفسیر متن و در حقیقت مرحله پردازش متن می‌باشد. این مرحله شامل پیدا کردن و امتیازدهی به کلمه‌های مهم می‌باشد. در مرحله آخر، خلاصه متن بر اساس جمله‌ها و امتیازدهی مربوطه ایجاد خواهد شد.

اهمیت تحلیل متن برای هر کاربردی در پردازش زبان طبیعی کاملا روشن است اما اهمیت درک متن برای هر کاربردی متفاوت است. مثلا درک سوال در یک سیستم پرسش و پاسخ برای سیستم بسیار مهم‌تر از تحلیل و درک متن در یک سیستم تشخیص صحبت می‌باشد. به هر حال مهمترین قسمت در هر سیستم کاربردی پردازش زبان طبیعی تحلیل متن ورودی خواهد بود.سطوح مختلف تحلیل متن که بیانگر هدف تحلیل می‌باشند را می توان به صورت ذیل تقسیم کرد:

- تصحیح متن: تبدیل جمله به فرم استاندارد.
- تجزیه جمله
- استخراج اطلاعات ویژه: هدف در این حالت استخراج اطلاعات خاص می‌باشد
- درک معنی متن: تبدیل هر جمله به یک فرم انتزاعی که بیانگر معنای جمله باشد. هدف از تحلیل در این مورد پی بردن به معنای جمله و یا جملات است.
- یافتن موضوع متن: تفسیر هر قسمت یا بخش از متن برای یافتن مفهوم، موضوع یا عنوانی که متن در توضیح آن آمده است.
- ارتباط متون: یافتن ارتباط‌های منطقی معنایی و ظاهری بین جملات مختلف بر اساس تشابه یا تفاوت بین جملات یا معنای آن‌ها.

لازم به ذکر است حسب پردازشی که روی متن انجام خواهد پذیرفت ممکن است یک و یا چندین مورد از موارد بالا در مرحله پیش‌پردازش متن انجام نشود.

بخشی از تحلیل متن استخراج اطلاعات است. این استخراج اطلاعات می‌تواند با استفاده از برچسب‌زنی کلمات متن باشد. برچسب‌زنی کمک قابل توجهی به تجزیه صحیح جملات و تصمیم‌گیری در مورد کلماتی که در شرایط مختلف معانی متفاوتی دارند، خواهد نمود. با تعیین دقیق نوع برچسب‌های مورد استفاده در پیکره‌های برچسب داده‌ای، می‌توان مرحله‌ی استخراج اطلاعات را تسهیل کرد.

تعیین کلمات کلیدی و استخراج واژه‌های موجود در متن از جمله دیگر کارهایی است که در بخش استخراج اطلاعات در تحلیل متن به کار می‌آید

3-2. ترجمه ماشینی

ترجمه ماشینی عبارت است از تبدیل یک متن از زبان مبدأ به زبان مقصد با استفاده از یک نرم‌افزار ترجمه کننده؛ چه این عمل با کمک انسان باشد چه بدون کمک انسان. نرم افزارهای مترجم، در بهترین حالت، عمل ترجمه را با دقتی در حدود 70 درصد انجام می‌دهند. برای به دست آوردن نتیجه بهتر، لازم است قبل و بعد از ترجمه، مقداری ویرایش روی متن انجام شود.

ترجمه ماشینی شاهد 3 نسل از ابزارها و روش‌ها بوده است. صرفنظر از دوره‌ی اول، امروزه روش‌های مورد استفاده در دوره‌ی دوم و سوم کاربرد بیشتری دارند.

رویکردهای مبتنی بر آمار و مبتنی بر مثال، از سال 1989 به بعد مورد استفاده قرار گرفته‌اند. ویژگی این رویکردها، قید-محوری و توجه بیشتر آن‌ها به جنبه لغوی در سیستم‌های قاعده-محور[1] و سیستم‌های تک‌زبانه بود.

دوره بعد از سال 1990 به بعد بوده و نسل سوم ماشینهای ترجمه در این زمان بوجود آمدند که در این نسل ماشینهای ترجمه از ترکیب رویکرد قاعده-محور نسل دوم، با روشهای متکی بر پیکره[2] بوجود آمده بودند. ویژگی‌های این نسل شامل تحلیل دستوری در سطح وابستگی ارتباطی و تحلیل معنایی در سطح

[1] Rule-based

[2] Corpus-based

تشخیص جملات و اطلاعات لغوی از فرهنگ لغات بدست آمده که با مشخصه دستوری در زبان مبدا نگهداری شده، است. متد های مبتنی بر مثال و آماری نیز به کار گرفته شدند، همچنین دانش جمع‌آوری شده محدود به حوزه مشخصی گردیده و از بازخورد کاربران ماشین هم در جمع آوری آماری اطلاعات استفاده شد.

استفاده از تکنیک‌های مبتنی بر پیکره منجر به تلاش برای ترجمه‌ی متون پیشرفته‌تر شد. چرا که این رویکردها تفاوت‌های گونه‌شناسی زبان[۱]، بازشناسیِ عبارات، ترجمه‌ی اصطلاحات و هم‌چنین جداسازی نابهنجاری در متون را بهتر پوشش می‌دهد.

با دردست داشتن داده‌های کافی، این تکنیک‌های ترجمه‌ی ماشینی قادر به ترجمه‌ی تقریبیِ متن از زبان مبدأ به زبان مقصد هستند. در حقیقت، مشکل اصلی در این تکنیک‌ها جمع‌آوری مقدار کافی از متن‌هایی با نوعِ خاص است. در روش‌های آماری پیکره‌های چندزبانه‌ی عظیمی مورد نیاز است، در حالی‌که در روش‌های مبتنی بر نحو چنین نیست. اما در عین حال، در توسعه‌ی روش‌های مبتنی بر نحو به زبان‌شناس‌های متبحر نیاز است تا بتوان گرامر زبان را بدرستی و با دقت بالایی طراحی کرد. مشکل دیگر روش‌های مبتنی بر نحو، دقیق نبودن گرامرهای فعلی برای برخی از زبان‌ها است. زبان فارسی از جمله‌ی این زبان‌ها است.

بنابر توضیحات فوق، امروزه 2 رویکرد مختلف در فن‌آوری‌های ترجمه ماشینی را می‌توان شناسایی کرد: ترجمه‌ی مبتنی بر قاعده[۲]، و ترجمه‌ی مبتنی بر پیکره[۳].

رویکردهای مبتنی بر پیکره، از داده‌های موجود در پیکره‌های متنی استفاده می‌کنند.

رویکرد اول این نوع از ماشین های ترجمه، مبتنی بر مثال نام دارند که در آنها در پیکره یا مجموعه اطلاعات زبانی به دنبال مشابه و نمونه ای که قبلا آمده بوده و اینکه چگونه ترجمه شده بوده می‌گردد.

---

[۱] Linguistic typology

[۲] Rule-Based MT

[۳] Corpus-Based MT

رویکرد آماری، زبان مبدا را بدون استفاده از قوانین به زبان مقصد نگاشت می‌کند و از مدل زبانی و اطلاعاتی که در پیکره وجود دارد استفاده می‌کند. به‌طور کل دانش زبانی کمی دراین روش بکار گرفته می‌شود و از احتمال رخ دادن ترکیبات جملات استفاده می نماید.

مراحل یک ترجمه به صورت کلی عبارت است از :

1. تفسیر پیکره زبان مقصد به صورت غیربرخط
2. تفسیر جملات زبان مبدأ
3. بهره برداری از پیکره زبان مقصد برای ارائه بهترین ترجمه
4. ترکیب و ساخت ترجمه خروجی

در مرحله‌ی تفسیر پیکره زبان مقصد، عبارت‌ها اندیس‌گذاری شده و طبقه‌بندی می‌شود. این عمل باعث می‌شود که جستجوها برای یافتن بهترین تطبیق، با سرعت بیشتری انجام شود.

تفسیر جملات زبان مقصد درست قبل از اینکه به الگوریتم تطبیق داده شوند انجام میگیرد. بدین ترتیب اطلاعات زبانی لازم در اختیار این الگوریتم قرار می‌گیرد. ابتدا، برچسب‌گذاری انجام شده و توسط ابزارهای بخش‌کننده[1] به قسمت‌های متنهاهی بخش‌بندی شده می‌شود. سپس به قسمت‌های برچسب‌دار، مجموعه لغات دوزبانه اعمال شده و ابتدا ترجمه روی ترکیبات کلمه‌ای و سپس روی تک‌تک کلمات صورت می‌گیرد. خروجی این مرحله، کلیه‌ی ترجمه‌های ممکن به زبان مقصد هستند. بعد از این مرحله، از این منابع برای ساخت مجموعه دو زبانی لغات استفاده شده و پس از بررسی سنخیت و دقت آن با نیاز سیستم، همگون می‌گردد.

در برخی از روش‌ها نیز پیشنهاد شده است که می‌توان با استفاده از یک پیکره‌ی متنی موازی و توصیف گرامری مجموعه‌ی کوچکی از جملات پیکره (بصورت دستی)، گرامر یک زبان را یاد گرفت.

از چالش‌های ترجمه‌ی ماشینی در روشهای ترکیبی (مبتنی بر قوانین و آماری)، می‌توان به «فراهم کردن دادگان آن» و ابزار تجزیه‌کننده اشاره کرد. کلیه‌ی موارد زیرشاخه‌ی فراهم کردن دادگان، در روش‌های و نیز برخی از موارد زیرشاخه‌ی ابزار تجزیه‌کننده، در مراحل تهیه‌ی پیکره های برچسب داده‌ای (یک زبانی و چندزبانی) موردپوشش قرار می‌گیرد.

[1] Chunker

4. نتیجه‌گیری

در بخش پیش، ارتباط و نیازمندی برخی کاربردهای پردازش زبان طبیعی به پیکره‌ی متنی را برشمردیم. همان‌گونه که آمد، پیکره‌های برچسب داده‌ای، حاوی اطلاعات ارزشمندی هستند که تقریباً در تمامی کاربردهای پردازش زبان طبیعی می‌تواند مورد استفاده قرار گیرد. در این مستند تنها به دو کاربرد از کاربردهای پردازش زبان طبیعی پرداختیم، لکن امتیازات پیکره‌های برچسب داده‌ای، در دیگر کاربردها نیز مشاهده می‌شود. لازم بذکر است که برتری پیکره‌های برچسب داده‌ای به پیکره‌های خام، منحصر به زبان خاصی نیست و به تمامی زبان‌ها قابل تعمیم است.