دانشکده مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد

بهبود ترجمه ماشینی آماری انگلیسی به فارسی با استفاده از اطلاعات زبان‌شناسی

تهیه و تنظیم:

رضا سعیدی

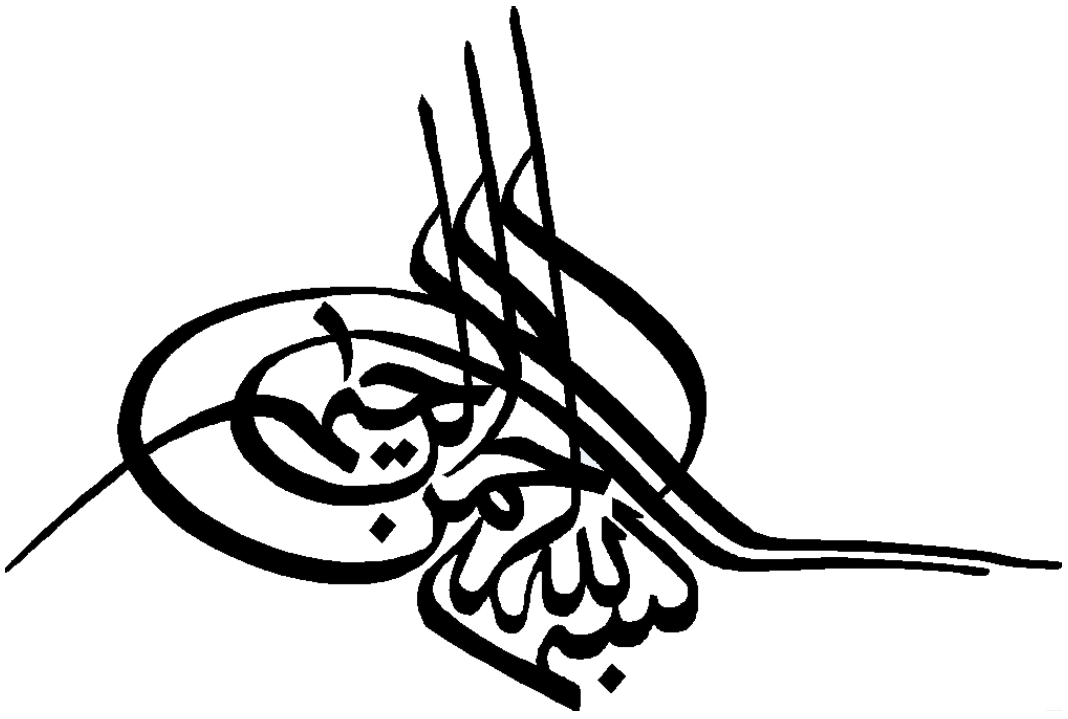
استاد راهنما:

دکتر محسن کاهانی

استاد مشاور:

دکتر نادر جهانگیری

تابستان 91


بسم الله الرحمن الرحيم

تعهدنامه

اینجانب رضا سعیدی دانشجوی دوره کارشناسی ارشد رشته کامپیوتر دانشکده مهندسی دانشگاه فردوسی مشهد نویسنده پایان‌نامه بهبود ترجمه ماشینی آماری انگلیسی به فارسی با استفاده از اطلاعات زبان‌شناسی تحت راهنمایی دکتر محسن کاهانی متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود و یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه فردوسی مشهد می باشد و مقالات مستخرج با نام "دانشگاه فردوسی مشهد" و یا "Ferdowsi University of Mashhad" به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند در مقالات مستخرج از رساله رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است، اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه فردوسی مشهد می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

تقدیم به

پدرم که عالمانه به من آموخت تا چگونه در عرصه زندگی، ایستادگی را تجربه نمایم

مادرم، دریای بی کران فداکاری و عشق، که وجودم برایش همه رنج بود و وجودش برایم همه مهر

و همسرم، اسطوره زندگیم، پناه خستگیم و امید بودنم

تقدیر و تشکر

سپاس گذار معلمی، هستم که اندیشیدن را به من آموخت، نه اندیشه ها را...

از جناب آقای دکتر محسن کاهانی که در طی این سالها صبورانه و دلسوزانه پیگیر و مشوق من بودند و راهنمایی ها و حمایت هایشان همواره کارگشای ایجانب بوده، کمال سپاس گذاری و تشکر را دارم. همچنین از جناب آقای دکتر نادر جهانگیری که همواره با حمایت ها و نظراتشان مشوق راه من بوده اند، متشکرم. همچنین از دوستان آزمایشگاه فناوری وب مخصوصا آقایان طوسی و بهمدی که صمیمانه یار و یاور من بودند کمال تشکر را دارم.

چکیده:

با گسترش روز افزون حجم داده‌ها و اطلاعات و همچنین گسترش تعاملات بین‌المللی برقراری ارتباط یکی از مهمترین جنبه‌های زندگی امروز است. مشکل عمده‌ای که در این زمینه وجود دارد عدم امکان برقراری ارتباط با استفاده از داده‌های به زبان دیگر است. از اینرو یکی از مهمترین مسائل زندگی امروز ارائه راه‌حلی خودکار جهت ترجمه از یک زبان به زبان دیگر است. ترجمه ماشینی یکی از راه‌هایی است که برای حل این مشکل ارائه شده است و به خاطر اهمیت آن، در سالیان اخیر توجه بسیار زیادی به آن شده است.

ترجمه ماشینی آماری به عنوان یکی از بهترین روش‌ها برای ترجمه از یک زبان به زبان دیگر شناخته می‌شود. برای زبان‌هایی که از لحاظ ساختار دارای شباهت زیادی به یکدیگر هستند خروجی این مترجم بسیار مناسب است. اما برای برخی جفت زبان‌ها مانند زبان‌های انگلیسی و فارسی تفاوت‌های ساختاری میان دو زبان و همچنین عدم وجود پیکره دوزبانه بزرگ باعث شده است که این روش برای ترجمه انگلیسی به فارسی ترجمه های مطلوبی را تولید نکند.

در این پایان‌نامه سعی شده است با کمک گرفتن از اطلاعات زبان‌شناسی، تا حد ممکن بر مشکلات این روش برای ترجمه انگلیسی به فارسی فائق آید. جهت انجام این کار ابتدا سعی در کاهش تفاوت ساختاری میان جملات انگلیسی و فارسی شده است. این عمل می‌تواند منجر به ایجاد مدل ترجمه بهتر شود. برای این منظور یکسری قوانین استخراج و بر روی جملات انگلیسی اعمال گردید. این تغییرات منجر به بهبود حدود 17 درصدی در معیار BLEU و حدود 21 درصدی در معیار NIST گردیده است. در ادامه نسبت به غنی‌سازی عبارات داخل پیکره با استفاده از برخی اطلاعات زبان‌شناسی از جمله برچسب‌های بخش‌های سخن[1] و ریشه کلمات اقدام شد. با این اطلاعات یک سیستم مترجم مبتنی بر فاکتور ایجاد گردید. خروجی این سیستم بهبود حدود 17 درصدی در معیار

[1] Part Of Speech

BLEU و حدود 25 درصدی در معیار NIST را نشان می‌دهد.

کلمات کلیدی:

ترجمه ماشینی، ترجمه ماشینی آماری، مدل زبانی، مدل ترجمه، بازآرایی نحوی، مدلسازی بر مبنای فاکتور

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل اول:

مقدمہ

1- مقدمه

رشد تعاملات بین‌المللی در زمینه‌های مختلف و وجود زبان‌های متفاوت در گوشه و کنار دنیا مشکلات زیادی برای افراد به منظور برقراری ارتباط با یکدیگر بوجود آورده است. از آنجا که نمی‌توان برای حل این مشکل آموزش زبان‌های مختلف را برای همه اجباری نمود و همچنین دسترسی به مترجم انسانی نیز در همه جا ممکن نیست؛ استفاده از کامپیوتر برای ترجمه به شدت احساس می‌شود. به این نوع مترجم اصطلاحاً مترجم ماشینی گفته می‌شود. درواقع اولین تلاش‌ها در این زمینه از سال 1940 آغاز گردید و تا به امروز پیشرفت‌های بسیار خوبی نیز به دست آمده است. اصولاً برای ایجاد یک مترجم ماشینی از دو رویکرد مبتنی بر قانون و مبتنی بر پیکره استفاده می‌شود. در رویکرد اول براساس زبان مبدا و مقصد یکسری قوانین نوشته شده و براساس آن عمل ترجمه صورت می‌گیرد که یکی از محدودیت‌های اصلی آن همین محدود بودن آن به زبان می‌باشد.

در رویکرد دوم براساس نمونه‌های قبلی و ترجمه‌های انسانی انجام شده به ترجمه متون جدید پرداخته می‌شود. در این رویکرد دیگر نیاز به قوانین برای ترجمه نیست و فقط نیازمند یک پیکره موازی و دوزبانه هستیم. یکی از روش‌های مهم در این رویکرد، روش ترجمه ماشینی آماری است که به خاطر عملکرد بسیار مناسب، در سالیان اخیر توجه زیادی به آن شده است. در این روش هدف استخراج مدل‌های لازم برای ترجمه از روی پیکره‌های موجود است. این مدل‌ها شامل مدل زبانی و مدل ترجمه هستند. در واقع با استفاده از این مدل‌ها دانش زبانی بصورت غیرمستقیم در مراحل ترجمه وارد می‌گردد. از اینرو می‌توانی به راحتی و بدون داشتن دانش زبانی زیاد از این مزیت استفاده نماییم.

نکته‌ای که در این روش بسیار مهم است دارا بودن یک پیکره موازی بزرگ است. در صورتی که بخواهیم دانش زبانی که بصورت ضمنی درون پیکره موازی وجود دارد استخراج و در قالب مدل‌های مطرح شده بیان نماییم، نیازمند یک پیکره موازی بسیار بزرگ هستیم. در واقع مشکل اصلی که در

این روش وجود دارد عدم دسترسی به این پیکره موازی است. برای ایجاد این پیکره ابتدا نیازمند منابع لازم برای ایجاد این پیکره و در ادامه روشی برای ترازبندی جمله به جمله این پیکره هستیم. هر دو این موارد جزو چالش‌های اساسی برای تولید پیکره‌های موازی هستند. از اینرو بایستی راهکارهایی بیابیم که بتوانیم با استفاده از پیکره‌های کوچکتر نیز خروجی بهتری ایجاد نماییم. برای این منظور بایستی برخی اطلاعات زبانی را صراحتاً درون پیکره وارد نماییم تا بخشی از مشکلات موجود بابت کوچک بودن پیکره مرتفع گردد.

برخلاف تحقیقات بسیار زیادی که در این زمینه بر روی زبان‌های مختلف در دنیا صورت گرفته است هنوز کار زیادی در زمینه‌ی زبان فارسی انجام نشده است. در واقع در حال حاضر بیشتر کارهای صورت گرفته در زبان فارسی ایجاد سیستم مترجم آماری صرف بوده است. نکته مهم درباره استفاده از اطلاعات زبانی در مترجم‌های ماشینی آماری این است که این اطلاعات زبانی بسته به زبان‌های مورد استفاده در سیستم مترجم متفاوت خواهد بود.

زبان فارسی یکی از زبان‌هایی است که دارای ساختار پیچیده‌ای است. در واقع زبان فارسی یک زبان تصریفی[1] بوده و کلمات بسته به زمان و شخص و ضمایر مربوطه با فرم‌های مختلف در جمله ظاهر می‌شوند. همین امر کار را بر روی این زبان با چالش مواجه می‌نماید. همچنین تفاوت ساختاری میان زبان فارسی و انگلیسی باعث بروز مشکل جهت ایجاد ترجمه با ترتیب و ترازبندی مناسب می‌گردد. همانطور که گفته شد، درصورتی که پیکره موازی ورودی به قدر کافی بزرگ باشد می‌توان بر این مشکلات فائق آمد اما ایجاد چنین پیکره‌ای بسیار مشکل است.

1-1- طرح پیشنهادی

در این پایان نامه هدف اصلی استفاده از برخی اطلاعات زبان‌شناسی جهت رفع برخی از مشکلات مطرح شده است به طوری که بتوان با استفاده از همین پیکره‌های کوچک نتایج را بهبود بخشید. برای

[1] Infelective

این منظور دو راهکار پیشنهاد شده است.

اولین راهکار کاهش تفاوت ساختاری بین جملات فارسی و انگلیسی است. یعنی سعی کنیم ساختار جملات دو زبان را از لحاظ نحو شبیه هم نماییم. این عمل باعث می‌شود که عباراتی که ترجمه یکدیگر هستند در توالی یکسانی در دو جمله همتراز انگلیسی و فارسی قرار گیرند. برای انجام اینکار لازم است تا ساختار دو زبان بررسی و یکسری قوانین در جهت تبدیل ساختاری استخراج گردد. این تبدیلات در سمت انگلیسی (زبان مبدا) صورت خواهد گرفت. با اعمال این تغییرات انتظار داریم که مدل ترجمه ایجاد شده به نسبت قبل بهبود یابد. این بهبود در نهایت منجر به بهبود نهایی مترجم ماشینی می‌گردد.

در راهکار بعدی استفاده از اطلاعات زبانی اضافه‌تر درون پیکره مطرح شده است. در حالت عادی درون پیکره فقط جملات انگلیسی (مبدا) و فارسی (مقصد) وجود دارند. همانطور که گفته شد بسیاری از کلمات و عبارات در زبان فارسی می‌توانند با ساخت و صرف‌های مختلف درون پیکره ظاهر شوند. در حالت عادی این صرف‌های مختلف از یک عبارت، متفاوت از یکدیگر در نظر گرفته می‌شوند و ارتباطی بین آن‌ها برقرار نمی‌گردد. از اینرو می‌توان اطلاعات زبانی دیگر نظیر ریشه[1] کلمات و برچسب‌های بخش‌های سخن[2] را نیز درون پیکره وارد نمود. این اطلاعات می‌توانند تا حدی مشکلاتی از این قبیل را رفع نموده و ما را در مراحل ایجاد مدل‌های لازم برای ترجمه یاری نمایند.

نتایج آزمایشات نشان دهنده بهبود روش‌های پیشنهادی در نتایج نهایی ترجمه است. ارزیابی بر روی نتایج در قالب دو معیار BLEU و NIST انجام شده است که نتایج حاصله افزایش هر دو معیار را نشان می‌دهد.

[1] Stem
[2] Part Of Speech

1 - 2 - مقالات مستخرج از پایان‌نامه

[1] **Reza Saeedi**, Mohsen Kahani, Seyed Ahmad Jakian Toosi, “Improving English to Persian Statistical Machine Translation Using Syntactic Reordering", 2012 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE’12) ,Hefei, China, Sep. 20-24, 2012.(**Accepted**)

[2] Seyed Ahmad Jakian Toosi, Mohsen Kahani, **Reza Saeedi**, “A Novel Method to Construct English-Persian Corpus”, 2th International eConference on Computer and Knowledge Engineering, Ferdowsi University of Mashhad, Mashhad, 2012.(**submitted**)

[3] **Reza Saeedi**, Mohsen Kahani, Ehsan Asgarian, “Multi Level Stemmer For Persian Language", sixth international symposium on communications (IST'2012).(**submitted**)

[4] **رضا سعیدی**، محسن کاهانی، سید احمد جکیان طوسی، "استفاده از بازآرایی نحوی جهت بهبود ترجمه ماشینی آماری انگلیسی فارسی", اولین کنفرانس بین المللی زبان فارسی, دانشگاه سمنان 2012.(**پذیرفته شده**)

[5] سید احمد جکیان طوسی، محسن کاهانی، **رضا سعیدی**، "معرفی یک مدل جدید برای ایجاد پیکره موازی انگلیسی- فارسی," هفدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران, دانشگاه صنعتی شریف، تهران, 2012.(**منتشر شده**)

[6] احمد استیری، محسن کاهانی، **رضا سعیدی**، "طراحی ابزار پارسر زبان فارسی," اولین کنفرانس بین المللی زبان فارسی, دانشگاه سمنان, 2012.(**پذیرفته شده**)

1 - 3 - ساختار پایان‌نامه

این نوشتار شامل چهار فصل دیگر به شرح زیر می‌باشد:

فصل دوم (مرور ادبیات): در این فصل در ابتدا مروری بر مشکلات ترجمه ماشینی خواهیم داشت و سپس در ادامه به روش‌های مختلف ترجمه ماشینی اشاره شده است. همچنین برخی تعاریف در زمینه پردازش زبان طبیعی در این فصل ارائه شده است. در انتها مروری بر کارهای مربوط صورت گرفته است.

فصل سوم (روش پیشنهادی): در این فصل دو روش برای ترکیب اطلاعات زبان شناسی با ترجمه آماری ارائه شده است. این روش‌ها بصورت اعمال پیش پردازشی بر روی پیکره اعمال می‌شوند تا در مرحله بعد وارد مرحله ایجاد مدل‌های مربوطه گردند.

فصل چهارم (پیاده سازی و ارزیابی) : در این فصل ابتدا مراحل و نحوه ساخت پیکره دوزبانه ارائه شده است. سپس نحوه پیاده‌سازی روش‌های ارائه شده بیان شده است و در انتها نتایج حاصل از ترجمه در قالب دو معیار BLEU و NIST ارائه شده است.

فصل پنجم (نتیجه گیری): در این فصل به نتیجه‌گیری روش‌های ارائه شده پرداخته شده و پیشنهادهایی برای کارهای آتی ارائه شده است.

فصل دوم:

مرور ادبیات

2- مرور ادبیات

2-1- تعاریف پایه زبان شناسی

در این بخش ابتدا تعاریف پایه و ابتدایی مورد نیاز توضیح داده شده است. در کاربردهای مختلف پردازش زبان طبیعی عموما از این تعاریف پایه استفاده می‌شود.

2-1-1- ریشه یابی

ریشه‌یابی به فرآیند کاهش دادن لغات به ریشه‌های آنها اطلاق می‌گردد[LOV68]. بنابراین "computer" و "compute"و "computing" به"compute" که ریشهٔ اصلی است کاهش می-یابند. نکته‌ای که مهم است این است که این ریشه می‌تواند در سطوح مختلف ارائه شود. مثلاً کلمه "دانشجویان" می‌تواند دارای سه ریشه "دانشجو"، "دانش" و "دان" باشد. در کاربردهای مختلف پردازش زبان طبیعی، هر یک از این سطوح مورد استفاده قرار می‌گیرند. همچنین این ریشه استخراج شده می‌تواند معنادار یا بی‌معنا باشد. در کاربردهایی نظیر ترجمه ماشینی از آنجا که قصد داریم ریشه استخراج شده معنای متفاوتی نسبت به حالت اصلی کلمه نداشته باشد، از بالاترین سطح ریشه که بامعنا نیز باشد استفاده می‌گردد. به این سطح ریشه اصلاحاً ریشه معنایی[1] گفته می‌شود. الگوریتم معروف پورتر[POR06] را برای ریشه‌یابی کلمات انگلیسی می‌توان نام برد. در[TAG05] ، [MIA06]، [SAN06] و [EST11] نیز چند ریشه‌یاب فارسی ارائه شده است.

2-1-2- برچسب زنی بخش‌های سخن (POS)

برچسب‌زنی بخش‌های سخن[2] عمل انتساب برچسب‌های نحوی به واژه‌ها و نشانه‌های تشکیل دهنده یک متن است به صورتی که این برچسب‌ها نشان دهنده نقش کلمات و نشانه‌ها در جمله

[1] Lemma
[2]- Part of Speech tagging

باشد[ELH09]. درصد بالایی از واژگان از نقطه نظر برچسب‌زنی نحوی دارای ابهام هستند زیرا کلمات در جایگاه‌های مختلف برچسب‌های نحوی متفاوتی دارند. بنابراین برچسب زنی نحوی، عمل ابهام زدایی از برچسب‌ها با توجه به زمینه (متن) مورد نظر است. برچسب‌زنی نحوی عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی از قبیل خلاصه‌سازی، خطایاب و تبدیل متن به گفتار می‌باشد.

در شکل 2-1 نمونه‌ای فرضی از یک مجموعه برچسب (Tagset) برای زبان فارسی معرفی شده است.

ADV	Adverb	قید
AJ	Adjective	صفت
CL	Classifier	شاخص
CONJ	Conjunction	حرف ربط
DET	Determiner	حرف تعریف
INT	Interjection	حرف صوت
N	Noun	اسم
NUM	Number	عدد
P	Preposition	حرف اضافه
POSTP	Postposition (را)	حرف اضافه پسین
PRO	Pronoun	ضمیر
PUNC	Punctuation	جدا کننده
RES	Residual:Arabic and Latin words, etc	متفرقه
V	Verb	فعل

شکل 2-1 نمونه‌های یک مجموعه برچسب‌زنی بخش‌های سخن

2-1-3- درخت تجزیه

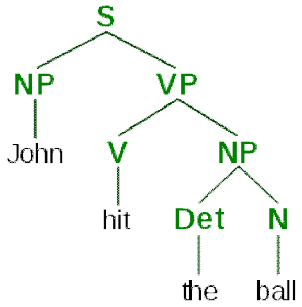
یک درخت تجزیه درختی است که نشان دهنده ساختار نحوی[1] یک رشته است. معمولاً این نمایش با توجه به دستور زبان‌های رسمی است. دریک درخت تجزیه، گره‌های داخلی، نقش‌های دستوری و برگ‌ها یا همان گره‌های خارجی، کلمات مربوط به آن نقش هستند[CHI07].

یکی از نیازمندیهای اساسی که در تولید برنامه‌ها نیاز است، این است که بایستی از لحاظ دستوری، درستی برنامه‌های نوشته شده تائید گردند. اغلب غیر ممکن است که یک برنامه نوشته شده با زبان‌های کامپیوتری را به شکل اولیه نوشته شده توسط برنامه‌نویس اجرا کنیم، زیرا اولاً برنامه نوشته

[1] syntax

شده به احتمال زیاد با خطاهای دستوری کاربر همراه است و ثانیاً اینکه فرم اولیه برنامه‌ها سطح بالا بوده و قابل تبدیل به کد، در همان شکل اولیه‌اش نیست. بنابراین برنامه بایستی در ابتدا به یک شکل بهتری تبدیل شود. بنابراین کامپایلرها سعی می‌کنند که برنامه نوشته شده را به یک فرم یکنوا و ساده‌تر که درخت تجزیه نام دارند، تبدیل کنند. درخت تجزیه با بحث کامپایلرها مرتبط است. برنامه‌ای که این چنین درخت هایی را تولید می‌کنند؛ تجزیه کننده[1] نامیده می‌شوند. درخت های تجزیه ممکن است برای جملات و عبارات زبان های روزمره مورد استفاده قرار بگیرند و یا در پردازش زبان های کامپیوتری بکار گرفته شوند.

در شکل 2-2، یک نمونه از درخت تجزیه نشان داده شده است. البته حالت‌های دیگری را نیز می‌توان متصور بود. درخت تجزیه یک ساختار کامل است که با جمله شروع می‌شود و به نقش‌های دستوری در برگ‌های انتهایی، پایان می‌یابد. البته این در صورتی درست است که عبارت مورد بررسی یک عبارت یا جمله از جملات روزمره باشد.



شکل 2-2 نمونه‌ای از یک درخت تجزیه

2-1-4- پیکره[2] تک زبانه

این پیکره متشکل از متون یک زبان خاص می‌باشد. از این پیکره برای کاربردهای خاص نظیر ترجمه ماشینی، خلاصه‌سازی و بسیاری از کاربردهای زبانی می‌توان استفاده نمود. در واقع با استفاده از چنین پیکره‌ای می‌توان الگوی زبانی زبان مربوطه را استخراج نمود. به عنوان مثال در ترجمه ماشینی که

[1] Parser
[2] corpus

موضوع اصلی این پایان‌نامه است، می‌توان مدلی به نام مدل زبانی[1] را از این پیکره اســتخراج نمـود. در بسیاری از تحقیقات در زمینه زبان فارسی برای یـک چنــین پیکـره‌ای از پیکـره همشـهری[ALE09] استفاده می‌گردد.

2-1-5- پیکره موازی

پیکره متنی موازی، در واقع یک منبع الکترونیکی ساخت یافته است که اطلاعات مورد نیاز جهت ترجمه، به وسیله ماشین‌های ترجمه آماری را فراهم می‌آورد. با اینکه می‌دانیم ماشین‌های ترجمه آماری، به قواعد زبانی ابداً بستگی ندارند اما وابستگی شدید آنها به پیکره های حجیم و موازی (دو یا چند زبانه) تا حد بسیاری تعیین کننده میزان قدرت ترجمه توسط این گونه از مترجم های ماشینی است[PIL11].

برای ساخت پیکره‌های موازی و البته دوزبانی، منابع مختلفی می‌تواند مورد استفاده قرار گیرد. مثلاً می‌توان از روزنامه‌ها، کتاب‌ها، وب سایت‌ها و یا صورت جلسات سازمانی و یا دولتی استفاده نمود. بدیهی است در این مورد، متونی می‌توانند حائز شرایط استفاده شدن، باشند که دارای ویژگیهای زیر باشند :

1. فاصله ساختاری میان متن اصلی و ترجمه متن کم باشد.

یعنی ترجمه انسانی نباید مفهومی یا معنایی باشد. بلکه تا حد امکان ترجمه لغت به لغت و وفادار به متن باشد.

2. میزان نویز در ترجمه حداقل باشد.

به این ترتیب که کلماتی که ترجمه آنها در جمله ترجمه شده، حذف شده است حداقل باشد. همچنین نتوان کلمه‌ای را در جمله ترجمه یافت که در جمله اصلی اصلاً وجود نداشته باشد.

[1] Language model

3. حجم جملات و ترجمه آنها قابل توجه و مناسب باشد.

همانطور که قبلاً هم اشاره شد قدرت یک مترجم ماشینی به میزان حجم اطلاعات پیکره بستگی دارد، از این رو هر چه تعداد جفت جملات بیشتر باشد اعتبار ترجمه هم تا حد زیادی افزایش پیدا می‌کند.

اساساً یکی از مشکل‌ترین کارها در ابتدای عملیات تهیه پیکره‌های دو زبانی فارسی انگلیسی، پیدا کردن متونی است که در دو زبان، اطلاعات قابل انطباق با هم داشته باشند. مشکل دیگر دیجیتالی نبودن برخی از این منابع است که باید به طریقی تبدیل به فرمت مناسب برای مرحله ترازبندی شوند. با این حال به نظر می‌رسد منابع زیر بیش از همه قابل دسترس باشند.

قرآن، انجیل، تورات، خبرگزاری‌های دو زبانه، سایتهای دو زبانه، زیرنویس فیلم‌ها، ماشین‌های ترجمه از جمله Google translator، کتب انگلیسی در سفر، کتاب‌های ترجمه شده مرجع با رعایت حق کپی رایت و ویکی پدیا.

در [GAL93;HAR97;KAY93] الگوریتم‌هایی برای ترازبندی جملات ارائه شده است.

2 - 2 - ترجمه ماشینی[1]

ترجمه ماشینی روشی است که در آن از کامپیوتر به منظور ترجمه خودکار یک زبان به زبان دیگر استفاده می‌شود. تفاوت بین زبان‌ها و مخصوصاً ابهام ذاتی موجود در زبان‌ها این ترجمه را بسیار مشکل می‌سازد. روش‌های سنتی برای ترجمه ماشینی برروی اطلاعات زبان شناختی بشر و بصورت قوانین تبدیل متن از زبانی به زبان دیگر تکیه داشتند. با وجود وسعت زبان، این کار بسیار سخت و نیازمند دانش زیادی است. ترجمه ماشینی آماری یک رویکرد متفاوت بوده که بطور خودکار با استفاده از حجم زیادی از داده‌های آموزشی، به دانش گفته شده دست می‌یابد. این دانش که عموما بصورت

[1] Machine translation

احتمالاتی از قابلیتهای زبانی گوناگون می‌باشد، به منظور راهنمایی در فرایند ترجمه مورد استفاده قرار می‌گیرد.

ترجمه ماشینی درواقع مبحثی درگیر در علوم زبان شناسی، علوم کامپیوتر، هوش مصنوعی، تئوری ترجمه و آمار است. کار بر روی ترجمه ماشینی از سال 1940 آغاز گردید. اما ترجمه ماشینی آماری اولین بار در سال 1949 توسط شخصی بنام واران ویور[1] پیشنهاد شد[WEA55] اما فقط در دو دهه اخیر بصورت عملی، کار بر روی آن صورت گرفته است. این رویکرد با توجه به توسعه بسیار زیاد در تکنولوژی‌های کامپیوتری از لحاظ سرعت و ظرفیت ذخیره‌سازی و فراهم سازی حجم زیادی از داده‌های متنی بصورت عملی و شدنی در آمده است[RAM00].

در ترجمه ماشینی ویژگی‌هایی وجود دارد که نه فقط از نظر جاذبه و کشش علمی، بلکه، از دیدگاه اقتصادی و دیگر ضرورت‌ها و اقتضاهای عصر، انجام آن را کاملاً توجیه می‌کند. به عنوان مثال، در مقر سازمان ناتو در بروکسل و جامعه اروپا علیرغم آنکه حدود 1200 مترجم ورزیده به کار اشتغال دارند، در حال حاضر از ترجمه ماشینی نیز استفاده می‌شود. دلیل این امر سرعت و هزینه است. میزان کاری که مترجمی ورزیده در خلال چندین روز انجام می‌دهد، توسط کامپیوتر در عرض چند دقیقه انجام می‌شود. حتی اگر کیفیت و دقت ترجمه ماشینی کمتر از حاصل کار مترجم باشد، باز هم از جهات گوناگون اهمیت و ارزش خاص آن مشهود است.

فرآیند ترجمه به شرح زیر است[LOP08]:

1. رمزگشایی معنایی متن مبدا
2. کدگذاری دوباره این معنا در زبان مقصد

در پس این فرآیند به ظاهر آسان، عملیات شناختی پیچیده ای صورت می‌گیرد. به منظور رمزگشایی

[1] Warren Weaver

معنای متن مبدا، مترجم باید قابلیت تفسیر و تجزیه و تحلیل تمام ویژگی‌های متن را داشته باشد. یک فرآیند که احتیاج به دانش عمیقی از دستور زبان، جمله‌شناسی(نحو)، معناشناسی و اصطلاحات زبان مرجع دارد، به همان اندازه هم بایستی دانش مربوط به فرهنگ صحبت کنندگان آن زبان را نیز داشته باشد. چالشی که در ترجمه ماشینی وجود دارد این است که چگونه روشی را طراحی کنیم که یک کامپیوتر بتواند همانند یک انسان متنی را فهمیده و بتواند یک متن جدید در زبان مقصد بگونه‌ای بسازد که به نظر برسد توسط انسان نوشته شده است. این مساله را می‌توان به روش‌های مختلفی حل کرد.

در ادامه ابتدا مشکلاتی را که سر راه ترجمه یک متن از زبانی به زبان دیگر وجود دارد مورد بررسی قرار می‌دهیم.

2 - 3 - مشکلات موجود در ترجمه ماشینی

2 - 3 - 1 - ابهام[1]

کلمات و عبارات در یک زبان اغلب دارای چندین ترجمه و معادل در زبانهای دیگر هستند. مثلا کلمه bank در زبان انگلیسی هم به معنای بانک و هم به معنای ساحل دریاست. البته تعیین نوع ترجمه صحیح از روی محتوای متن امکان پذیر است.

همچنین هر زبان دارای یکسری اصطلاحات است که تشخیص آن از روی جمله بسیار مشکل است. بعنوان مثال جمله زیر را در نظر بگیرید.

India and Pakistan have broken the ice finally.

افعال ترکیبی نیز از دیگر ویژگی‌هایی است که امر ترجمه را مشکل می‌سازد. مثلاً فعل bring up را در سه جمله زیر در نظر بگیرید:

[1] ambiguity

They brought up the child in luxury.
They brought up the table to the first floor.
They brought up the issue in the house.

نوع دیگری از ابهام نیز وجود دارد که ابهام ساختاری نام دارد. جمله زیر را در نظر بگیرید.

Flying planes can be dangerous.

بر اساس اینکه آیا هواپیما یا حرفه پرواز (خلبانی) خطرناک است می‌توان ترجمه‌های متفاوتی ارائه کرد[RAM00].

2 - 3 - 2 - تفاوتهای ساختاری

این مسئله به قواعد زبانی موجود در یک زبان اشاره دارد. مثلاً ترتیب کلمات در یک جمله در زبان انگلیسی به ترتیب صورت فاعل، فعل و مفعول می‌باشد، بطوریکه این ترتیب در زبان فارسی بصورت فاعل، مفعول و فعل است. جدا از این ویژگی اساسی، زبانها در ترکیب برخی ساختارها و نحو دارای بایدها و نبایدهایی نیز هستند. در زمان ترجمه بایستی به این تفاوتها نیز توجه شود[DAV01]. مثلاً برای ترکیب موصوف و صفت در دو زبان فارسی و انگلیسی کاملاً برعکس یکدیگر عمل می‌شود مانند red car که در فارسی بصورت ماشین قرمز ترجمه می‌شود.

2 - 3 - 3 - تفاوتهای لغوی

برخی از کلمات در برخی از زبانها وجود دارند که معادلی در زبانهای دیگر ندارند و بایستی در هنگام ترجمه به این موارد نیز توجه شود[DAV01].

2 - 4 - روشهای موجود برای ترجمه ماشینی

در قسمت قبل دیدیم که زبانهای مختلف از لحاظ لغوی و ساختاری با یکدیگر متفاوتند. ترجمه ماشینی می‌تواند تا حد ممکن بعنوان یک فرایند برای کاهش این تفاوتها درنظر گرفته شود. این

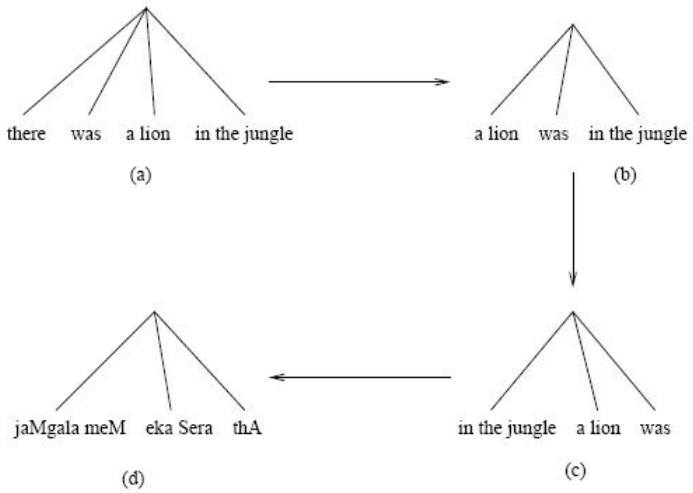
رویکرد می‌تواند منجر به شناخت مدل تبدیل[1] برای ترجمه ماشینی گردد[HUT92] و [JUR00].

2-4-1- رویکرد تبدیل

مدل انتقال شامل سه مرحله است: تحلیل، تبدیل و تولید. در مرحله تحلیل، جملات زبان منبع تجزیه[2] شده و ساختار و اجزای تشکیل دهنده جملات مشخص می‌گردند. در مرحله تبدیل، تبدیلات لازم بر روی درخت تجزیه زبان مبدا برای تبدیل به ساختار زبان مقصد صورت می‌گیرد و در نهایت در مرحله تولید، کلمات و توضیحات زمانی افعال، اعداد و غیره به زبان مقصد ترجمه می‌گردند. شکل 2-3 مراحل موجود در ترجمه جمله زیر را با روش تبدیل نشان می‌دهد.

There was a lion in the jungle.

در قسمت b نمایش درونی این جمله نشان داده شده که نزدیک به ساختار زبان انگلیسی، اما با حذف 'there' موجود در آن است. مرحله تبدیل این ساختار را مطابق با ترتیب کلمات هندی تبدیل می‌کند(قسمت c شکل). سرانجام در مرحله تولید، کلمات انگلیسی با کلمات هندی متناظر جایگزین می‌گردد.



شکل 2-3- ترجمه با استفاده از مدل تبدیل[RAM00]

[1] Transfer model
[2] parse

توجه داشته باشید که در مرحله تحلیل یک نمایش وابسته به زبان مبدا و در مرحله تولید یک ترجمه نهایی از یک نمایش وابسته به زبان مقصد تولید می‌گردند. بنابراین برای یک سیستم ترجمه ماشینی چند زبانه، برای هر طرف ترجمه به ازای هر جفت از زبان‌هایی که سیستم دربرمی‌گیرد، یک مولد تبدیل(مرحله دوم در این روش) جداگانه نیاز است. برای سیستمی‌که تمام ترکیبات ممکن از n زبان را در بر می‌گیرد، n مولفه برای تحلیل، n مولفه برای تولید و n×(n-1) مولفه برای تبدیل نیاز است.

اگر بتوانیم مرحله تبدیل را بگونه‌ای حذف کنیم بطوری که هر مولفه تحلیلی بتواند یک نمایش مستقل از زبان یکسان تولید کند و هر مولفه تولیدی بتواند از این نمایش، ترجمه را تولید نماید، توانسته‌ایم n×(n-1) سیستم ترجمه را فقط با ساختن n مولفه تحلیلی و n مولفه تولیدی ایجاد نماییم. این ایده دقیقاً ایده‌ایست که در پس رویکرد Interlingua [1] مورد استفاده قرار می‌گیرد.

2-4-2- رویکرد Interlingua

در این رویکرد ترجمه ماشینی طی دو مرحله صورت می‌پذیرد[HUT92]:

1- استخراج معنی جملات زبان مبدا به صورت یک فرم مستقل از زبان
2- تولید جملات زبان مقصد از روی معنای ایجاد شده در مرحله اول

فرم مستقل از زبانی که مورد استفاده قرار می‌گیرد Interlingua نامیده می‌شود. این فرم به ما اجازه می‌دهد تا یک نمایش رسمی[2] از هر جمله ایجاد کنیم، بطوریکه تمام جملاتی که دارای معنای یکسانی هستند بدون در نظر گرفتن زبان، بصورت یکسانی نمایش داده شوند.

نمایش interlingua معنایی کامل از هر جمله توسط مجموعه‌ای از مفاهیم و روابط جهانی بیان می‌دارد. به عنوان مثال یک نمایش ممکن برای جمله زیر در شکل 2-4 نشان داده شده است.

---

[1] یک زبان مصنوعی است که برای ترجمه ماشینی ایجاد شده است. اولین بار این زبان در سال 1951 براساس کلمات مشترک بین انگلیسی و زبانهای اروپایی بنا نهاده شد.

[2] Canonical representation

The child broke the toys with a hammer.

PREDICATE	BREAK
AGENT	CHILD number: SING
THEME	TOY number: PLUR
INSTRUMENT	HAMMER number: SING
TENSE	PAST

شکل 2-4- نمایش interlingual از جمله مثال بالا[HUT92]

مفاهیم و روابطی که مورد استفاده قرار می‌گیرند (که بعنوان آنتولوژی[1] شناخته می‌شوند) مهمترین جنبه در هر سیستم برمبنای interlingua هستند. آنتولوژی باید به اندازه کافی توانا باشد تا بتواند تمام جزییات و ظرایف معانی‌ای که می‌تواند با استفاده از هر زبانی بیان گردد بصورت زبان interlingua نمایش دهد.

در شکل 2-5 پیداست که وقتی به سمت بالا رفته و به سمت interlingua پیش می‌رویم، هزینه مولفه‌های تحلیل و تولید افزایش می‌یابد. در این روش بایستی از بین تجزیه‌های ممکن متفاوت برای جمله یکی را انتخاب کرده، مفاهیم جهانی که جملات به آنها اشاره می‌کنند را مشخص کنیم و رابطه بین مفاهیم مختلف تصریح شده در جمله را بیابیم. اما یک سیستم ترجمه برای یک جفت زبان خاص ممکن است تا این حد نیازمند تحلیل نباشد. این قضیه بسیاری از توسعه دهندگان را به استفاده از روشی سوق می‌دهد که رویکرد مستقیم[2] برای ترجمه ماشینی نامیده می‌شود.

2 - 4 - 3 - رویکرد مستقیم

جمله زیر را در نظر بگیرید:

We will demand peace in the country.

---

[1] Ontology
[2] Direct approach

برای ترجمه این جمله به فارسی لازم نیست تا نقش‌های شماتیک مفاهیم جهانی را مشخص کنیم. فقط بایستی تحلیل صرفی[1] انجام دهیم، اجزاء جمله را تعیین کنیم، آنها را براساس ترتیب اجزاء موجود در زبان فارسی (فاعل، مفعول و فعل) مرتب کنیم، کلمات را در یک دیکشنری انگلیسی به فارسی جستجو کنیم و بطور مقتضی زمان افعال را اعمال نماییم. ممکن است به نظر آید که در اینجا کار ساده‌تر نشده است بلکه کار بیشتری صورت گرفته است اما این اعمال، اعمالی هستند که معمولاً می‌توانند ساده‌تر و قابل اعتمادتر صورت بپذیرند. جدول 2-1 مراحل موجود برای ترجمه مستقیم جمله بالا را نشان می‌دهد.

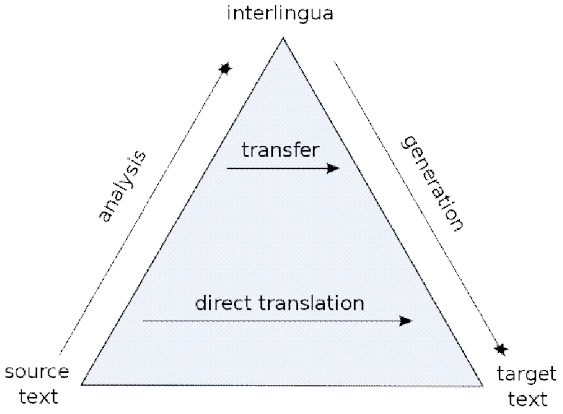
جدول 2-1- مثالی از ترجمه مستقیم

جمله انگلیسی	We will demand peace in the country
تحلیل ریخت شناسی	We demand FUTURE peace in country
تعیین اجزاء	<We> <demand FUTURE> <peace> <in country>
مرتب سازی مجدد	<We> <peace><in country> <demand FUTURE>
جستجو در دیکشنری انگلیسی به فارسی	ما صلح را در کشور خواستن **آینده**
اعمال زمان و ضمیر	ما صلح را در کشور خواستاریم.

تفاوت سیستمهای مستقیم ترجمه با سیستمهای تبدیل و interlingua در این است که آنها تحلیل ساختاری و معنایی[2] کمتری انجام می‌دهند[HUT92] [JUR00].

البته رویکرد مستقیم تا حدی موردی[3] است و نمی‌تواند بعنوان یک راه حل طولانی مدت برای ترجمه ماشینی مورد استفاده قرار گیرد. متدهای تبدیل و interlingua اگر چه دارای هزینه اولیه بیشتری هستند اما روش‌های مناسبتری به شمار می‌روند. در شکل 2-5 تفاوت بین روش‌های بیان شده نشان داده شده است.

[1] Morphological
[2] semantic
[3] ad hoc



شکل 2-5 مقایسه‌ای بین رویکردهای تبدیل، Interlingua و مستقیم در ترجمه[HUT92]

برای توضیحات و مثال‌های بیشتردر مورد سه رویکرد بیان شده به [HUT92] [JUR00] مراجعه گردد.

2 - 4 - 4 - رویکردهای برپایه پیکره[1]

همگی رویکردهایی که تا اینجا بررسی شده‌اند، از دانش زبان‌شناسی انسان برای حل مسئله ترجمه استفاده می‌کنند. اما برخی روش‌ها وجود دارند که از این دانش زبان‌شناسی صراحتاً استفاده نمی‌کنند. در این روش‌ها از یک پیکره موازی جهت انجام عمل ترجمه کمک گرفته می‌شود[HUT92]. در ادامه روشهایی که از این پیکره برای انجام عمل ترجمه بهره برده‌اند را بررسی می‌نماییم.

2 - 4 - 4 - 1 - ترجمه ماشینی برپایه نمونه[2]

این رویکرد برای ترجمه ماشینی، از تکنیکی استفاده می‌کند که انسان اغلب برای حل مسائل استفاده می‌کند[SOM99]. شکستن مسئله به مسائل کوچکتر، یافتن راه حل های گذشته برای حل زیرمسئله‌های مشابه، وفق دادن این راه حل‌ها با موقعیت های جدید و در نهایت ترکیب راه حل‌ها

[1] Corpus-based approach
[2] Example based

برای حل مسئله بزرگتر موجود.

برای اینکه ببینیم چطور ترجمه ماشینی از این روش استفاده می‌کند روال ترجمه جمله زیر را بررسی می‌کنیم.

He bought a huge house.

(1a) He bought a pen.

(1b) او یک خانه خرید.

(2a) All ministers have huge houses.

(2b) تمام وزرا خانه بزرگ دارند.

می‌کوشیم تا تکه‌های جمله را با استفاده از تطبیق آن‌ها با جملات داخل پیکره ترجمه نماییم. ابتدا عبارت He bought دقیقاً در پیکره یافت می‌شود (عبارت زیر خط دار در جمله 1a). عبارت huge house هم با جمله 2a تطبیق داده می‌شود با این تفاوت که جمله ما مفرد اما این جمله جمع است. بنابراین این دو تکه جمله را برداشته، تکه دوم را از لحاظ فعلی اصلاح می‌کنیم (از جمع به مفرد تبدیل می‌کنیم)، آنها را بگونه ای مناسب ترکیب می‌کنیم و در نهایت جمله زیر را بدست می‌آوریم:

"او یک خانه بزرگ خرید."

بنابراین روش ترجمه ماشینی براساس نمونه شامل سه مرحله است:

1- تطبیق تکه‌های جمله با پیکره موازی
2- وفق دادن تکه‌های تطبیق داده شده با زبان مقصد
3- ترکیب مجدد این تکه‌های ترجمه شده به گونه‌ای مناسب.

توجه داشته باشید که قبل از اینکه فرایند ترجمه در این روش آغاز گردد، جملات مترادف (جفت ترجمه‌ها) در پیکره موازی بایستی مشخص شوند. یک چنین پیکره‌ای اصطلاحاً یک پیکره تراز شده براساس جمله نامیده می‌شود. ترازبندی مناسب جملات برای تمام ترجمه‌های ماشینی از جمله ترجمه ماشینی آماری بسیار تعیین کننده خواهد بود.

2-4-4-2- ترجمه ماشینی آماری

مدلهای ترجمه ماشینی آماری از این منظر به مسئله نگاه می‌کند که هر جمله در زبان مقصد، با احتمالی، ترجمه‌ای از جمله زبان مبدا است. بهترین ترجمه، قطعاً جمله‌ای است که دارای بالاترین احتمال باشد. کلید حل مسأله در ترجمه ماشینی آماری عبارتست از تخمین احتمال یک ترجمه و در نهایت یافتن جمله ای که دارای بیشترین احتمال باشد. در ادامه به بررسی مفصل‌تری از روش ترجمه ماشینی آماری خواهیم پرداخت.

2-5- ترجمه ماشینی آماری

یک راه برای درک بهتر ترجمه ماشینی آماری این است که تصور کنید جمله f در زبان مبدا به تعدادی جمله e در زبان مقصد ترجمه شده است اما در هنگام انتقال از طریق یک کانال نویزدار دچار نویز شده است. اکنون فرض می‌کنیم که هر جمله در زبان مقصد با یک احتمال، ترجمه‌ای از زبان مبدا است و در نتیجه جمله‌ای را به عنوان ترجمه صحیح انتخاب می‌کنیم که دارای احتمال بیشتری باشد. این قضیه را بصورت رابطه (2-1) می‌توان بیان کرد[BRO90]:

$$e' = \arg\max_e p(e \mid f) \qquad (1\text{-}2)$$

اساساً $p(e \mid f)$ مبتنی بر دو فاکتور زیر است:

1- میزان احتمال بودن یک جمله در زبان E، این فاکتور بعنوان مدل زبانی[1] ($p(e)$) شناخته می‌شود.

2- روشی که جملات موجود در زبان E به جملات موجود در زبان F تبدیل می‌شوند. این فاکتور نیز بعنوان مدل ترجمه[1] ($p(f \mid e)$) شناخته می‌شود.

[1] language model

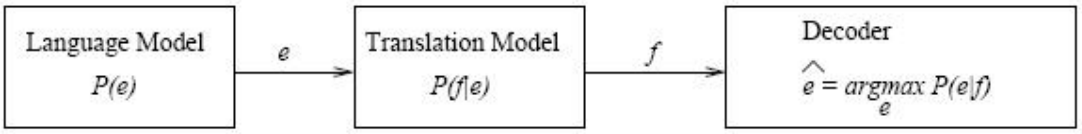
همانطور که در رابطه (2-2) نشان داده شده است این دو فاکتور از اعمال قانون بیز بر روی رابطه (2-1) حاصل می‌گردند.

$$e' = \arg\max_e \frac{p(e)p(f \mid e)}{p(f)} \qquad (2\text{-}2)$$

از آنجایی که مقدار f ثابت است آن را از بیشینه‌سازی حذف می‌کنیم. نتیجه نهایی در رابطه (2-3) نشان داده شده است.

$$e' = \arg\max_e p(e)p(f \mid e) \qquad (2\text{-}3)$$

این مدل برای ترجمه ماشینی در شکل 2-6 نشان داده شده است.



شکل 2-6- مدل ترجمه ماشینی آماری

مسئله‌ای که وجود دارد این است که $p(e \mid f)$ را مستقیماً با استفاده از رابطه (2-1) محاسبه نمی‌کنیم بلکه به جای آن $p(e \mid f)$ را با استفاده از قانون بیز به دو عبارت $p(e)$ و $p(f \mid e)$ تقسیم می‌کنیم. حاصل کار بسته به روشی است که مدل ترجمه ما عمل می‌کند که درواقع همان روشی است که با استفاده از آن $p(e \mid f)$ یا $p(f \mid e)$ را محاسبه می‌کنیم. عملاً مدلهای ترجمه زمانی که کلمات موجود در f عموماً ترجمه کلمات موجود در e باشند به $p(f \mid e)$ یا $p(e \mid f)$ احتمالات بالایی را نسبت می‌دهند. اما این مدل‌ها معمولاً به این نکته توجه نمی‌کنند که آیا ترتیب قرارگیری کلمات موجود در e کنار هم مناسب است یا نه. به عنوان مثال جمله زیر را در نظر بگیرید:

The weather is cold.

برای ترجمه این جمله به فارسی، مدل ترجمه احتمالات برابری به سه جمله زیر می‌دهد:

---

[1] translation model

هوا سرد است.

سرد هوا است.

است سرد هوا.

در اینجا اگر از رابطه (2-3) استفاده کنیم می‌توانیم به ترجمه درست دست یابیم زیرا برای مسئله $p(f|e)$ برای هر سه جمله دارای مقدار یکسانی است اما مدل زبانی در اینجا نقش مهمی بازی می‌کند و در واقع با دادن مقداری بیشتری به اولین ترجمه برای $p(e)$ نسبت به دو ترجمه دیگر، دو جمله نهایی را از لیست جواب‌های ما حذف می‌کند.

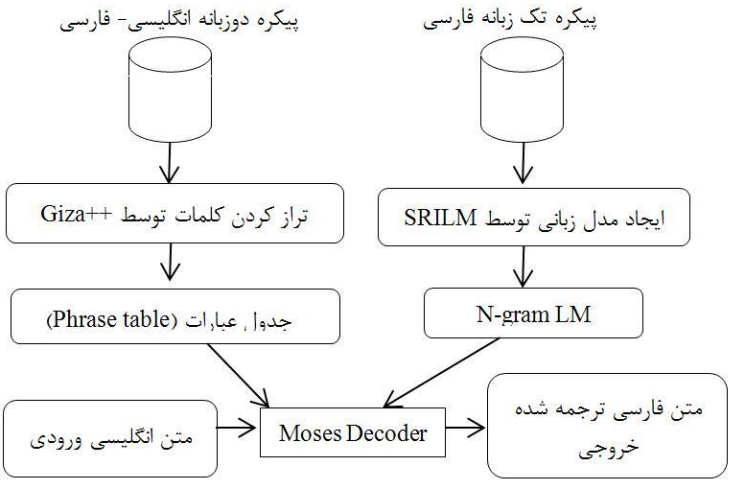
این مسئله منظر دیگری از مدل ترجمه ماشینی را نمایان می‌کند. بهترین ترجمه جمله‌ایست که به جمله اصلی وفادار باشد و جمله ای روان در زبان مقصد نیز باشد[JUR00]. برای مثال بالا سه جمله دارای میزان وفاداری یکسانی به جمله اصلی هستند اما بایستی جمله اول که میزان روانی آن در زبان فارسی بالاست به عنوان ترجمه انتخاب گردد.

براساس مسائل مطرح شده می‌توانیم از بین جملات کاندید برای ترجمه، جمله‌ای را به عنوان بهترین ترجمه بیابیم. اما نکته‌ای که مطرح است این است که این جملات بایستی چطور استخراج گردند. در واقع برای انجام این عمل نیازمند یک جستجوی اکشتافی برای استخراج جملات کاندید برای نرجمه هستیم. بنابراین ترجمه آماری نیازمند سه مولفه کلیدی زیر است:

1- مدل زبانی[1]
2- مدل ترجمه[2]
3- الگوریتم جستجو[3]

مراحل کلی ترجمه ماشینی آماری انگلیسی به فارسی در شکل 2-7- مراحل کار ترجمه ماشینی آماری‌نشان داده شده است.

[1] Language model
[2] Translation model
[3] Search algorithm



شکل 2-7- مراحل کار ترجمه ماشینی آماری

2 - 5 - 1 - مدل زبانی با استفاده از N-gramها

مدلسازی ترجمه انتساب یک احتمال به هر واحد متن است[KOE09]. در مبحث ترجمه ماشینی آماری، همانطور که در قسمت قبل توضیح داده شد واحدهای متن جمله هستند. براساس جمله e داده شده در این مدلسازی وظیفه محاسبه $p(e)$ است.

برای یک جمله شامل توالی کلمات $w_1\ w_2 \ldots w_n$ داریم:

$$P(e) = P(w_1\ w_2 \ldots w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1\ w_2)\ldots P(w_n \mid w_1\ w_2 \ldots w_{n-1})$$

(4-2)

مشکل اصلی در اینجا و در بسیاری از مدلهای زبانی، اسپارس[1] بودن داده هاست و اینکه چطور بتوان احتمالی مثل $P(w_n \mid w_1 w_2 \ldots w_{n-1})$ راحساب نمود. در هیچ پیکره‌ای نمی‌توان تمام توالی‌های ممکن از n کلمه را یافت. در واقع فقط کسر کوچکی از چنین جملاتی را می‌توان درون پیکره یافت. پس بایستی راهی وجود داشته باشد تا بگونه‌ای بتوان احتمال آمدن یک توالی را با کمک گرفتن از یک پیکره بررسی نمود. مدل N-gram یک راه برای انجام این کار فراهم می‌کند.

[1] sparse

2-5-1-1- تخمین N-gram

در یک مدل N-gram [JUR00]، احتمال وقوع یک کلمه با داشتن تمامی‌کلمات ماقبلش توسط محاسبه احتمال یک کلمه براساس N-1 کلمه ماقبلش محاسبه می‌گردد. با N=2 یک مدل bigram و با N=3 یک مدل trigram داریم.

N-gram ها دارای نتایج موفقیت آمیزی در تشخیص صدا[1]، بررسی املاء[2]، تگ زدن متن براساس گرامر[3] و سایر مکان‌هایی که نیازمند مدلسازی زبان هستند، بوده‌اند. این امر بخاطر سادگی این مدل در ایجاد و استفاده می‌باشد. مدلهای ممکن دیگری نیز وجود دارد. به عنوان مثال روشی که به هر ساختار جمله یک احتمالی را نسبت می‌دهد (بااستفاده از یک گرامر احتمالی). ایجاد چنین مدل‌هایی آسان نیست، اما مدل های N-gram می‌توانند با عمل بر روی یک پیکره ، براحتی ایجاد و مورد استفاده قرار گیرند.

احتمالات N-gram می‌تواند با استفاده یک روش مستقیم از روی یک پیکره محاسبه گردد. برای مثال احتمالات bigram می‌تواند توسط رابطه (2-5) محاسبه گردد.

$$P(w_n \mid w_{n-1}) = \frac{count(w_{n-1}\, w_n)}{\sum_w count(w_{n-1} w)} \qquad (5\text{-}2)$$

در این معادله $count(w_{n-1}w_n)$ نشان دهنده تعداد وقوع توالی $w_{n-1}w_n$ می‌باشد. مخرج کسر نیز نشان دهنده تعداد دفعاتی است که کلمه w_{n-1} قبل از کلمه دیگری ظاهر شده است که این درواقع همان تعداد وقوع کلمه w_{n-1} می‌باشد. پس رابطه (2-5) را می‌توانیم بصورت رابطه (2-6) بنویسیم.

$$P(w_n \mid w_{n-1}) = \frac{count(w_{n-1}\, w_n)}{count(w_{n-1})} \qquad (6\text{-}2)$$

مشکلی که در این روش وجود دارد صفر شدن رابطه (2-6) برای توالی کلماتی است که داخل پیکره

[1] Speech recognition
[2] Spell checking
[3] POS tagging

وجود ندارند. دلیل این مشکل به وجود آمده همان اسپارس بودن داده‌هاست که در ابتدای بحث N-gram ها به آن اشاره شد. ما نمی‌توانیم امیدوار باشیم که بتوانیم برای تمامیN-gramهای ممکن از روی هر پیکره محدودی، احتمالات قابل اعتمادی بیابیم. باید توجه داشت که وظیفه مدل زبانی این نیست که فقط به توالی‌هایی که در پیکره موجود وجود دارند احتمالات مناسبی را نسبت دهد بلکه این مدل بایستی زمانی که با یک توالی جدید مواجه شد نیز بتواند تخمین مناسبی برای آن بزند. برای حل این مشکل از روشی با نام هموارسازی[1] استفاده می‌شود.

2-5-1-2- هموارسازی

ساده ترین الگوریتم هموارسازی[KOE09] اضافه کردن عدد یک است. ایده اصلی در این روش این است که عدد یک را به تعداد تمام توالی های ممکن-آنهایی که در پیکره اتفاق افتاده اند به همراه آنهایی که اتفاق نیفتاده‌اند- اضافه نماییم و براساس آن احتمالات را محاسبه می‌کنیم.

مدل bigram هم، اکنون بر اساس این روش بصورت رابطه (2-7) خواهد بود:

$$P^*(w_n \mid w_{n-1}) = \frac{count(w_{n-1}\, w_n)+1}{count(w_{n-1})+V} \qquad (7\text{-}2)$$

بطوری که V تعداد نوع کلمه‌های موجود در فرهنگ لغت است.

رابطه (2-7) بدین معناست که فرض کرده‌ایم که هر bigram ممکن حداقل یکبار در پیکره آموزشی واقع شده باشد. بنابراین توالی‌ای که در پیکره واقع نشده باشد تعدادی برابر یک دارد و درنتیجه دارای احتمالی غیرصفر است. البته روش‌های دیگری نیز برای این کار وجود دارد[WIT91] که هدف همه آن‌ها حل مشکل انتساب احتمالات صفر، برای توالی کلماتی است که درون پیکره موجود نیستند.

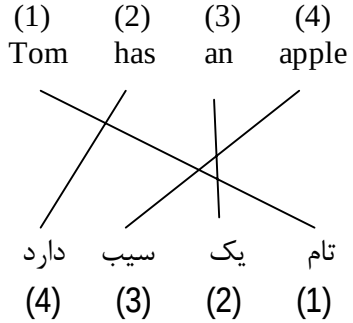
[1] smoothing

2-5-2- مدل ترجمه

همانطور که پیش از این توضیح داده شد وظیفه مدل ترجمه یافتن احتمال P(f|e) یعنی احتمال اینکه جمله f منبع، دارای ترجمه‌ای بصورت جمله e باشد[KOE09]. توجه داشته باشید که توسط مدل ترجمه P(f|e) محاسبه می‌شود و نه P(e|f). پیکره آموزشی برای مدل ترجمه، یک پیکره موازی ترازبندی شده براساس جملات زبانهای F(منبع) و E(مقصد) است.

واضح است که P(f|e) را نمی‌توانیم با شمارش جملات f و e موجود در پیکره موازی محاسبه نماییم. مجدداً مشکلی که با آن مواجه هستیم اسپارس بودن داده‌هاست. راه‌حلی که در ابتدا به نظر می‌رسد یافتن (یا تخمین) احتمال جمله ترجمه با استفاده از احتمالات ترجمه کلمات موجود در جملات است. احتمالات ترجمه کلمه را می‌توانیم به ترتیب از روی پیکره موازی محاسبه نماییم. اما یک مشکل کوچک وجود دارد. پیکره موازی فقط تراز شده جملات را به ما می‌دهد و به ما نمی‌گوید که چطور کلمات موجود در جملات تراز شده‌اند.

یک ترازبندی کلمه‌ای بین جملات موجود در پیکره به ما می‌گوید که دقیقاً هر کلمه موجود در جمله f به چه کلمه‌ای در e ترجمه شده است. شکل 2-8 یک مثال از یک چنین ترازبندی‌ای را نشان می‌دهد. این ترازبندی می‌تواند بصورت (1،۳،۲،۴) نیز نوشته شود تا موقعیت کلمات موجود در جمله فارسی را نسبت به کلمات موجود در جمله انگلیسی نشان دهد.



شکل 2-8 - یک مثال از ترازبندی

اما چطور می‌توان از روی یک پیکره آموزشی که فقط براساس جملات ترازبندی شده‌اند احتمالات ترازبندی کلمات را بیابیم. این مسئله با استفاده از الگوریتم بیشینه سازی امید[1] حل می‌شود.

2-5-2-1- بیشینه سازی امید

کلید اصلی الگوریتم EM این است که اگر بتوانیم تعداد دفعاتی که یک کلمه با کلمه دیگری در یک پیکره تراز شده است را بیابیم می‌توانیم احتمال ترجمه کلمه را به سادگی محاسبه نماییم[KOE09]. از طرف دیگر، اگر احتمالات ترجمه کلمات را نیز بدانیم می‌توانیم احتمال انواع ترازبندی‌ها را بیابیم.

ظاهراً در اینجا با مسئله مرغ و تخم مرغ روبرو هستیم. در هر حال اگر با یکسری احتمالات ترجمه کلمه یکسان، کار را شروع کنیم و احتمالات ترازبندی را محاسبه نماییم، سپس از این احتمالات ترازبندی برای بدست آوردن احتمالات ترجمه بهتر بهره برده و این کار را ادامه دهیم، می‌توانیم در نهایت به برخی مقادیر مناسب همگرا شویم. این فرایند تکرار شونده که الگوریتم بیشینه سازی امید (EM) نامیده می‌شود، به این خاطر همگرا می‌گردد که کلماتی که درواقع ترجمه یکدیگر هستند در پیکره ترازبندی شده بر اساس جمله، با همدیگر ظاهر می‌شوند یا بهتر است بگوییم در یک جمله ترازبندی شده ظاهر می‌گردند. مدلهای مختلفی برای مدلسازی ترجمه وجود دارد از جمله مدل‌های IBM1 تا IBM5.

2-5-2-2- مدل IBM 1

احتمال اینکه جمله f ترجمه جمله e باشد بصورت زیر نوشته می‌شود:

$$P(f \mid e) = \sum_{a \in \psi} P(f, a \mid e) \qquad (8\text{-}2)$$

در رابطه (2-8)، نماد Ψ نشان دهنده تمام ترازبندی‌های ممکن بین f و e و نماد a نیز نشان دهنده

[1] Expectation-Maximization

یک ترازبندی خاص بین f و e است. در واقع طرف راست رابطه (2-8) مجموع تمام راه‌هایی که f می‌تواند دارای ترجمه e باشد را نشان می‌دهد.

هدف از مدل ترجمه، بیشینه کردن P(f|e) روی تمام پیکره آموزشی است. به عبارت دیگر، احتمالات ترجمه کلمات را بگونه ای تنظیم کنیم که جفت ترجمه‌های موجود در پیکره آموزشی احتمالات بالایی دریافت کنند.

برای محاسبه احتمالات ترجمه کلمه، نیازمندیم تا بدانیم چندبار یک کلمه با کلمه دیگری تراز شده است. این تعداد را با شمارش هر جفت جمله تراز شده در پیکره آموزشی پیش بینی می‌کنیم. اما هر جفت جمله می‌تواند به طرق مختلف با هم تراز شود و هر ترازبندی اینچنینی دارای احتمالی است. بنابراین شمارش کلمات تراز شده‌ای که انجام می‌دهیم بصورت کسری خواهد بود و مجبوریم که این شمارش‌های کسری را روی هر ترازبندی ممکن جمع بزنیم. این کار نیازمند این است که احتمال یک ترازبندی خاص داده شده را از یک جفت ترجمه بیابیم.

بطور خلاصه اگر بخواهیم مدل IBM1 را توضیح دهیم بایستی با داشتن یک پیکره موازی از جملات تراز شده، بصورت زیر احتمالات ترجمه را تخمین بزنیم.

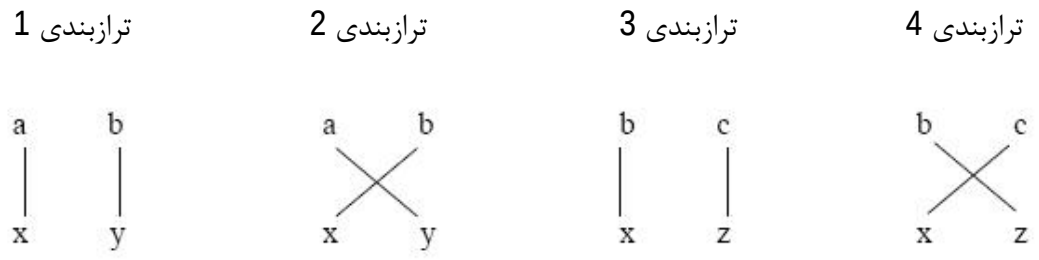
1- با یکسری مقادیر برای احتمالات ترجمه کار را شروع می‌کنیم.

2- تعداد (کسری) ترجمه‌های کلمه را محاسبه می‌کنیم.

3- این تعداد را جهت تخمین مجدد احتمالات ترجمه مورد استفاده قرار می‌دهیم.

4- دو مرحله قبل را تکرار می‌کنیم تا به یک مقدار همگرا شویم.

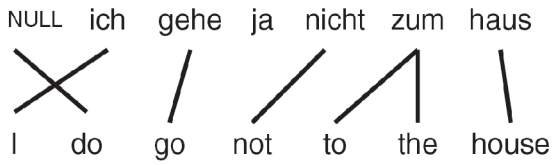
در [BRO93] ثابت شده است که با هر مقداردهی اولیه (به غیر از احتمالات صفر) برای احتمالات ترجمه، الگوریتم EM منجر به همگرا شدن به مقدار بیشینه می‌گردد.



شکل 2-9 - ترازبندی های ممکن برای کلمات در پیکره موازی

2 - 5 - 3 - روش‌های مختلف ترجمه ماشینی آماری

روش‌های مختلفی برای ترجمه ماشینی آماری ارائه شده است. ساده‌ترین و اولین روش ارائه شده ترجمه مبنتی بر کلمه[1] است[KOE09]. در ترجمه مبتنی بر کلمه، واحد اصلی ترجمه یک کلمه است. در یک ترجمه، نسبت طول توالی کلمات متن ترجمه باروری[2] نامیده می‌شود، که بیانگر این است که چه تعداد کلمه در ترجمه هر کلمه از زبان مبدا تولید می‌شود.



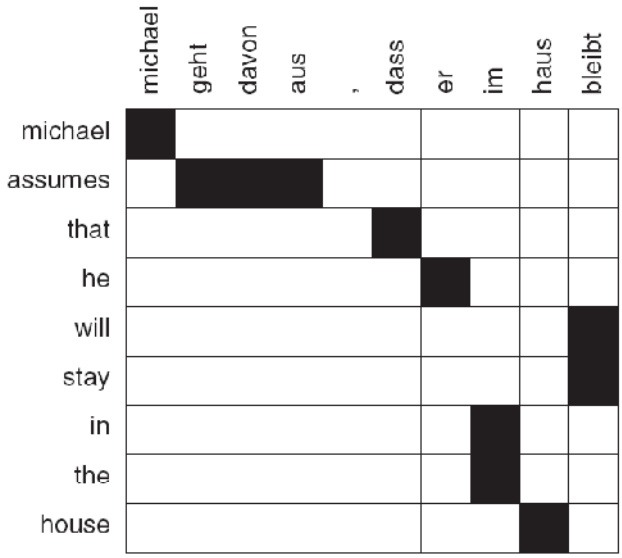
شکل 2-10 نمونه‌ای از ترازبندی کلمه به کلمه[KOE09]

ترجمه مبتنی بر کلمه معمولی، نمی‌تواند بین زبانهای با باروری متفاوت ترجمه مناسبی انجام دهد. ترجمه مبتنی بر کلمه می‌تواند بگونه‌ای توسعه داده شود که این مشکل را رفع نماید. در این توسعه می‌توان یک کلمه را به چندین کلمه نگاشت داد، اما نمی‌توان بصورت برعکس عمل نمود. ترجمه مبتنی بر کلمه امروزه بطور گسترده مورد استفاده قرار نمی‌گیرد؛ و ترجمه مبتنی بر عبارت[3] شایع‌تر است.

در ترجمه مبتنی بر عبارت[KOE03]، هدف این است که محدودیت های موجود در ترجمه مبتنی

[1] Word-based translation
[2] fertility
[3] Phrase-based translation

بر کلمه کاهش یابد. همانطور که در شکل 2-11 نشان داده شده است در این روش به جای ترجمه کلمات به دنبال ترجمه یک توالی از کلمات در قالب یک بلوک یا عبارت هستیم. این عبارات به طور معمول عبارات زبانی نیست بلکه عبارات با استفاده از روش های آماری از پیکره ایجاد می‌شوند. البته نشان داده شده است که محدود کردن عبارات به عبارات زبانی، باعث کاهش کیفیت ترجمه می‌شود.



شکل 2-11 ترازبندی مبتنی بر عبارت[KOE09]

مدل مبتنی بر عبارت، به نگاشت بخش‌های کوچک متن، بدون هر گونه استفاده صریح از اطلاعات زبانی نظیر اطلاعات صرفی، نحوی یا معنایی محدود است. چنین اطلاعات اضافی می‌توانند از طریق ترکیب آن‌ها با پیکره‌ها در مراحل پیش‌پردازش[1] و یا پس‌پردازش[2] نتایج کار را بهبود بخشند.

با این حال، ترکیب اطلاعات زبانی با مدل ترجمه به دو دلیل مطلوب است:

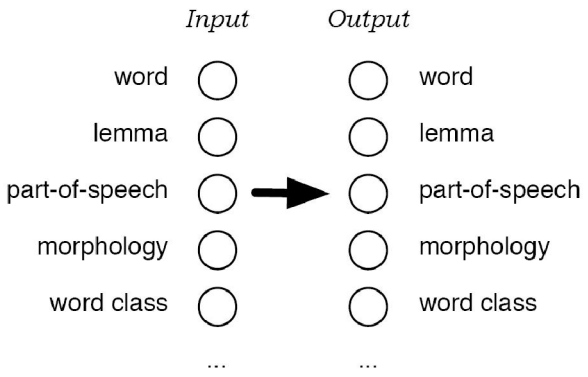
[1] Preprocessing
[2] Postprocessing

- مدل‌های ترجمه‌ای که بر روی نمایش‌های عمومی‌تر کلمات، مانند ریشه معنایی[1] به جای استفاده از اشکال اصلی[2] کلمات، عمل می‌کنند می‌توانند نتایج بهتری تولید نمایند و مشکل عدم وجود برخی حالت‌های کلمه در پیکره به دلیل محدودیت اندازه را رفع نمایند.
- بسیاری از جنبه‌های ترجمه را می‌توان در سطح صرفی، نحوی و یا معنایی بهتر توضیح داد. فراهم بودن این اطلاعات به مدل ترجمه اجازه می دهد تا الگوبرداری مستقیمی از این جنبه‌ها داشته باشد. به عنوان مثال بازآرایی در سطح جمله، عمدتا با اصول کلی نحوی، محدودیت‌های توافق شده محلی نشان داده شده در صرف و اینگونه مسائل مدیریت می‌شود.

بنابراین، رویکرد مبتنی بر عبارت به منظور استفاده از اطلاعات اضافی در فرایند ترجمه تحت عنوان ترجمه ماشینی برپایه فاکتور[3] توسعه داده شده است[KOE07]. این رویکرد جدید اجازه می‌دهد تا برچسب‌های اضافی در سطح کلمه مورد استفاده قرار گیرند. یک کلمه در این چارچوب ، یک توکن[4] تنها نیست، بلکه بصورت برداری از فاکتورهای مختلف جهت نمایش سطوح مختلف معنایی کلمه ارائه می‌شود.

این اطلاعات شامل برچسب‌هایی مانند فرم اصلی کلمه، ریشه معنایی، برچسب ادات سخن[5]، ویژگی های صرفی از قبیل جنسیت، شمارش و غیره هستند. شکل 2-12 زیر برخی از این فاکتورها را نشان می‌دهد.

[1] lemma
[2] surface forms
[3] Factored-based translation
[4] Token
[5] Part Of Speech



شکل 2-12 فاکتورهای مورد استفاده برای روی ترجمه برپایه فاکتور[KOE07]

2 - 6 - ارزیابی سیستم های ترجمه

در گذشته از معیار "کاربرد مفاهیم" برای ارزیابی سیستم های ترجمه ماشینی استفاده می‌شد. مثلاً آنهایی که از مفاهیم استفاده می‌کردند را با عبارت "خوب" و آنهایی را که از مفاهیم استفاده نمی‌کردند را با عنوان "بد" بر چسب گذاری می کردند . چون تقریبا هیچ سیستمی از مفاهیم استفاده نمی‌کرد لذا این معیار برای مدت طولانی دوام نیاورد. روشهای مختلفی برای تعیین ارزش ترجمه تولید شده، از راه ترجمه ماشینی وجود دارد .

به عنوان قدیمی ترین روش می توان از قضاوت انسانی استفاده کرد. با وجود این که این روش ممکن است بسیار وقت گیر و پرهزینه باشد در عین حال بهترین و قابل اعتماد ترین سیستم ارزش‌گذاری شناخته شده است. در کنار آن می‌توان از ابزارهای خودکاری همچون BLEU و NIST نیز نام برد .

بیشتر سیستم های ترجمه بر این اساس کار می کنند که میزان شباهت یک ترجمه تولید شده را به ترجمه مرجع اندازه گیری نمایند . ترجمه مرجع ترجمه‌ای است که از لحاظ انسان قابل قبول باشد . هر چه این فاصله کمتر باشد یعنی شباهت بیشتر بوده و ترجمه از کیفیت بالاتری برخوردار است .

2-6-1- سیستم ارزیابی BLEU[1]

در این معیار ارزیابی[PAP02]، میزان همبستگی بالا میان متن تولید شده توسط ماشین و متن ترجمه شده توسط انسان به صورت کمی، مورد بررسی قرار گرفته و درحقیقت به نوعی محور اساسی در این روش دقت ترجمه توسط ماشین قرار گرفته است.

ابتدا امتیازات برای تک‌تک قطعات متن محاسبه می‌شوند. هر قطعه را می‌توان یک عبارت در نظر گرفت. به این صورت که عبارت ترجمه شده توسط ماشین با عنوان عبارت کاندید با مجموعه‌ای از عبارات ترجمه شده توسط انسان موسوم به عبارات مرجع مقایسه می‌شوند، آنگاه میانگین این امتیازات در کل متن بدست می‌آید تا به این وسیله تخمین مناسبی از کیفیت را بدهد. نکته مهم در این روش این است که قابل فهم بودن و درستی ترجمه از حیث دستوری مورد قضاوت قرار نمی‌گیرد . مسئله دیگر اینست که اگر BLEU برای ارزیابی یک عبارت منفرد بکار گرفته شود نتایج بسیار بدی تولید می‌کند. در حالی که اگر برای یک متن با عبارات متنوع اعمال گردد نتایج خوبی تولید می‌شود.

کیفیت ترجمه با عددی بین 0 و 1 اندازه‌گیری می‌شود. این عدد نمایانگر میزان نزدیکی ترجمه به مجموعه‌ای از ترجمه‌های انسانی با کیفیت خوب است. هر چه این عدد به یک نزدیک‌تر باشد نشان دهنده آن است که مشابهت بیشتری میان ترجمه انسانی و ماشینی وجود داشته است.

این الگوریتم از یک قالب اصلاح شده سنجش دقت برای مقایسه میان عبارات کاندید و عبارات مرجع استفاده می‌کند. در فرم ساده سنجش دقت، نسبت تعداد کلماتی که بین دو ترجمه ماشینی و مرجع مشترک هستند به کل کلمات به عنوان کمیتی برای کیفیت، ارائه می‌شود.

[1] Bilingual Evaluation Understudy

دراین روش هر چه تعداد واژه های موجود در کلمه مرکب n گرمی بیشتر باشد، دقت ارزیابی متن نیز بالا می‌رود.

2 - 6 - 2 - سیستم ارزیابی NIST

روش ارزیابی NIST، که توسط اداره استاندارد آمریکا پیشنهاد شده است دارای الگوریتمی مشابه الگوریتم BLEU است، با این تفاوت که در مدل BLEU برای هر کلمه مرکب n گرمی، وزن یکسانی فرض شده است در حالی که در اینجا کلمات مرکب چند گرمی بر اساس تعداد تکرارشان در متن وزن دهی می‌شوند[MAR10]. به عنوان مثال ترکیب "In the" که به طور متناوب در یک متن تکرار می‌شود دارای ارزش کمتر و ترکیب "Intresting calculation" که وقوع آن کمتر بوده از وزن برخوردار می‌شود.

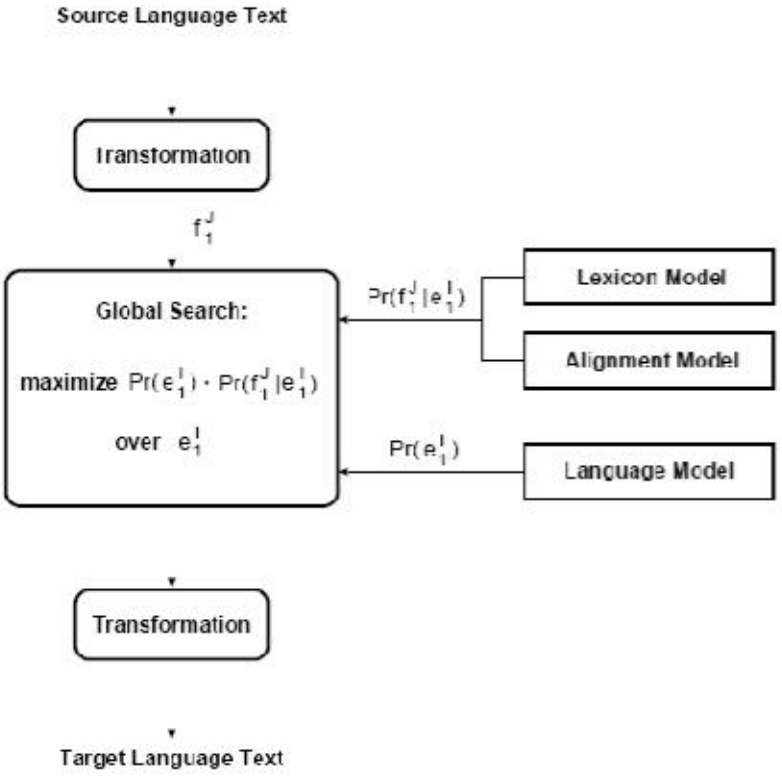
2 - 7 - پیشینه تحقیق

زمانی که صحبت از ترجمه آماری صرف می‌گردد، اکثراً تفاوتی بین جفت زبان‌های مورد استفاده در بیان نحوه عملکرد نیست و نحوه کار تفاوتی نخواهد داشت و تنها نتایج خروجی بسته به میزان شباهت بین دو زبان متفاوت خواهد بود. اما زمانی که بخواهیم از اطلاعات زبان‌شناسی در زمینه ترجمه ماشینی استفاده نماییم، آن‌گاه جفت زبان‌های مورد بررسی مهم و روش کار بسته به این دو زبان خواهد بود. دلیل این امر نیز تفاوت‌های ذاتی، نحوی و ساختاری بین زبان‌های مختلف است. در این روش‌ها الگوریتم‌ها و قوانین بسته به نوع زبان متفاوت خواهد بود. به همین منظور در ادامه برخی از روش‌های ارائه شده برای بهبود ترجمه ماشینی آماری که در آن‌ها از اطلاعات زبان‌شناسی بهره گرفته شده است مرور می‌گردد. البته بایستی به این مورد اشاره شود که از آنجایی که تفاوت و گوناگونی زبان‌ها بسیار زیاد است فقط بخشی از این تحقیقات در این گزارش ارائه شده است. در ابتدا برخی از روش‌های استفاده شده بر روی زبان‌های غیر فارسی بیان شده است و در انتها تحقیقات انجام

شده در زمینه ترجمه ماشینی آماری انگلیسی به فارسی معرفی شده است.

2 - 7 - 1 - استفاده از اطلاعات زبان‌شناسی جهت بهبود ترجمه ماشینی آماری

Niessen و Ney به این مورد اشاره کرده‌اند که پیکره‌های دوزبانه ما هیچ‌گاه نمی‌توانند آنقدر بزرگ باشند که تمام ویژگی‌ها و حالات زبانی را در خود جای دهند؛ از اینرو بایستی اطلاعات زبان‌شناسی نیز به کمک ترجمه آماری آمده و کاستی‌های بیان شده را بهبود بخشد. هرچند این اطلاعات بطور غیرمستقیم و از روی پیکره استخراج می‌گردد اما بایستی با برخی روش‌ها، این اطلاعات را صراحتاً درون مراحل ترجمه دخالت دهیم تا خروجی کار بهبود یابد. آن‌ها اثر ویژگی‌های نحوی و لغوی را در ترجمه آلمانی- انگلیسی و بر روی دو پیکره متفاوت مورد بررسی قرار دادند. شکل 2-13 نمایی از معماری کلی ارائه شده در این مقاله نشان داده شده است.



شکل 2-13 معماری کلی سیستم مترجم آماری ارائه شده[NIE00]

آنها در این تحقیق تغییرات زیر را در پیکره‌شان ایجاد نمودند[NIE00]:

- جداسازی پیشوندهای فعلی: برخی از افعال در زبان آلمانی دارای یکسری پیشوند متصل به قسمت اصلی فعل هستند. این افعال بصورت "پیشوند|قسمت اصلی" در پیکره از هم جدا گردیدند.
- جداسازی کلمات مرکب: برخی کلمات مرکب در زبان آلمانی وجود دارند که زمان ترجمه بصورت یک کلمه واحد در نظر گرفته می‌شوند و دیگر در زمان ترازبندی و ایجاد مدل ترجمه اجزای آن نمی‌توانند تراز شده و برای آنها نیز ترجمه ایجاد شود. برای حل این مشکل این گروه، کلمات مرکب را به اجزای تشیل دهنده‌شان تجزیه نمودند.
- برچسب‌زنی کلمات با برچسب‌های بخش‌های سخن[1]: برای رفع ابهام در زمان ترجمه می‌توان از برچسب‌های بخش‌های سخن استفاده نمود. برای همین منظور کلمات پیکره آموزشی و آزمایشی با این برچسب‌ها غنی گردیدند.
- حل مشکل کلمات ناشناخته: یکی از بزرگترین مشکلات در روش‌های مبتنی بر پیکره عدم پوشش تمامی کلمات ممکن است. از اینرو بایستی برای حل این مشکل فکری اندیشید. برای همین منظور مولفین سعی کردند با تبدیل کلمات به فرم کلی‌شان این مشکل را حل کنند. برای مثال می‌توان تاثیر شخص را بر روی فعل از بین برد و فعل را بصورت زمان گذشته مورد استفاده قرار داد. همچنین می‌توان کلمات مرکب را همانطور که قبلاً اشاره شد تجزیه نمود تا کلمات تشکیل دهنده‌شان ایجاد گردند.

نتایج حاصل از اعمال موارد فوق برحسب میزان خطا بر روی دو پیکره در جدول 2-2 و جدول 2-3 نشان داده شده است.

[1] POS Tags

جدول 2-2 نتایج بر روی پیکره Verbmobil [NIE00]

پیش‌پردازش	SSER[%]
سیستم پایه	20.3
جداسازی پیشوندهای فعلی	19.4
جداسازی کلمات مرکب	20.3
برچسب‌زنی POS	19.7
برچسب‌زنی POS+ترکیب کلمات	19.5
برچسب‌زنی POS+ترکیب کلمات+ جداسازی پیشوندهای فعلی	18

جدول 2-3 نتایج بر روی پیکره EuTrans [NIE00]

پیش‌پردازش	SSER[%]
سیستم پایه	57.4
جداسازی کلمات مرکب	52.9
جداسازی کلمات مرکب+ حل مشکل کلمات ناشناخته	51.8
جداسازی کلمات مرکب + جداسازی پیشوندهای فعلی	50.3

Niessen و Ney در کار بعدی خود [NIE04] به بحث درباره یک لغت نامه[1] سلسله مراتبی پرداخته‌اند. در مقاله آن‌ها اشاره شده است که در برخی از زبان ها مثل زبان آلمانی، کلمات می‌توانند براساس زمان و شخص جمله به شکل‌های متفاوتی در جمله ظاهر شوند. این مساله در زبانی مثل زبان انگلیسی به ندرت روی می‌دهد. همین مساله باعث می‌شود که ترازبندی ترجمه‌ایی که بایستی انجام شود دچار مشکل گردد. برای همین به ارائه لغت نامه سلسله مراتبی پرداخته است تا بتواند برای کلماتی که ریشه یکسان اما شکل‌های متفاوتی دارند، ترازبندی مناسبی به دست آورد. برای انجام این کار علاوه بر حالت اصلی کلمات در داخل پیکره، به اطلاعات اضافه‌ای همچون فرم پایه (ریشه[2]) کلمات و برچسب‌های لغوی و نحوی مثل برچسب‌های بخش‌های سخن، زمان و شخص نیازمندیم.

[1] lexicon
[2] stem

نتایج آزمایشات آن‌ها نشان می‌دهد که با ترکیب اطلاعات نحوی و لغوی حتی با یک پیکره کوچک هم می‌توان به نتایج مطلوبی رسید.

جدول 2-4 نتایج حاصل از استفاده از لغتنامه سلسله مراتبی[NIE04]

BLEU	
31.6	سیستم پایه
36.5	لغتنامه سلسله مراتبی

در یکی دیگر از تحقیقات انجام شده در این زمینه، Lee به بررسی ترجمه آماری عربی به انگلیسی پرداخته است. در [LEE04] اشاره شده است که به خاطر کاربرد پسوندها در عربی، متناوباً مشاهده می‌گردد که یک کلمه عربی به چند کلمه انگلیسی تراز می‌گردد. ایشان الگوریتمی را توسعه دادند که در آن واژهایی که در حالت نهایی بایستی حذف یا ترکیب شوند را مشخص می‌کند. در این روش کلمات موجود در پیکره به شکل "پیشوندها+ریشه+پسوندها" شکسته می‌شوند. سپس براساس الگوریتم ارائه شده عباراتی که بایستی حذف یا ترکیب شوند مشخص می‌گردند. نتایج حاصل از این تغییرات در قالب معیار BLEU در جدول 2-5 نشان داده شده است.

جدول 2-5 تاثیر تغییرات صرفی بر روی پیکره در نتیجه ترجمه[LEE04]

اندازه پیکره	**سیستم پایه**	**تغییرات صرفی**
3.5K	17	24
35K	24	29
350K	32	36
3.3M	36	39

در [GOL05] مولفین آزمایشات متنوعی بر روی ترجمه ماشینی آماری چک– انگلیسی انجام داده‌اند. زبان چک یکی از زبان‌های تصریفی است که کلمات در آن بسته به نوع تصریفشان می‌توانند به حالت‌های مختلف ظاهر شوند. مولفین در این مقاله به بررسی حالت‌های مختلف کلمات پرداخته و فهمیدند که برخی برچسب‌های واژی زمانی که بصورت کلمات جدا در نظرمی‌شود مفیدترند؛ در حالی

که برخی دیگر از همین برچسب‌ها زمانی که مستقیماً به کلمه‌ای که به آن وابسته‌اند، متصل می‌شوند مفیدتر خواهد بود. در این مقاله نشان داده شده است با اعمال این قوانین بر روی پیکره و اصلاح ورودی‌های زبان چک، می‌توان ترجمه ماشینی آماری را بهبود داد.

در [POP06] به بررسی ترجمه ماشینی با استفاده از پیکره‌های با حجم کم پرداخته و برخی روش‌ها برای بهبود نتایج را معرفی کرده است. جفت زبان‌های مورد بررسی در این مقاله اسپانیایی–انگلیسی و صربستانی–انگلیسی است. در این مقاله ابتدا به بررسی این موضوع پرداخته شده است که اگر ترتیب کلمات بین زبان‌های مبدا و مقصد که از لحاظ ساختار زبان با هم متفاوتند را تغییر داده و به هم شبیه نماییم، در نتیجه ترجمه، بهبود حاصل خواهد شد. برای همین ترتیب کلمات زبان مبدا را به فرم زبان مقصد تبدیل کرده است. برای این منظور ابتدا به بررسی ساختار دو زبان پرداخته و بیان شده است که صفات در زبان اسپانیایی برخلاف زبان انگلیسی، بعد از اسم مربوطه قرار می‌گیرد. به همین منظور، جایگاه صفات موجود در پیکره به طور مناسب تغییر داده شده است. در این مقاله همچنین بیان شده است که در زبان اسپانیایی صفت ها براساس نوع کاربرد به چهار حالت متفاوت ظاهر می‌شوند که همین امر عمل تراز بندی عبارات و ایجاد مدل ترجمه را دچار چالش می‌نماید، زیرا یک نوع صفت در سمت انگلیسی بایستی به چهار نوع صفت در سمت اسپانیایی تراز گردد. برای همین تمامی‌حالت‌های مختلف صفات موجود در پیکره مبدا را به حالت پایه تبدیل کرده است. ضمناً یک دیکشنری عمومی انگلیسی-اسپانیایی نیز برای کاهش اثر کلماتی که در داخل پیکره وجود ندارند به پیکره اضافه شده است. خروجی کار برای ترجمه اسپانیایی به انگلیسی در جدول 2-6 و برای ترجمه انگلیسی به اسپانیایی در جدول 2-7 نشان داده شده است. بهبود حاصله از تغییرات اعمال شده با حجم‌های مختلف پیکره، در این جداول قابل مشاهده است.

جدول 2-6 نتایج ترجمه برای ترجمه اسپانیایی به انگلیسی[POP06]

اسپانیایی به انگلیسی		WER	PER	BLEU
دیکشنری	سیستم پایه	60.4	49.3	19.4
	+ بازآرایی صفات	59.4	47.4	20.1
	+ حالت پایه صفات	56.4	46.8	23.8
1K	سیستم پایه	52.4	40.7	30
	+دیکشنری	48	36.5	36
	+ بازآرایی صفات	45	35.3	39.8
	+ حالت پایه صفات	44.5	34.8	40.9
13K	سیستم پایه	41.8	30.7	43.2
	+دیکشنری	40.6	29.6	46.3
	+ بازآرایی صفات	38.5	29.2	48.9
	+ حالت پایه صفات	38.3	29	49.6
1.3M	سیستم پایه	34.5	25.5	54.7
	+ بازآرایی صفات	33.5	25.2	56.4

جدول 2-7 نتایج ترجمه برای ترجمه انگلیسی به اسپانیایی[POP06]

انگلیسی به اسپانیایی		WER	PER	BLEU
دیکشنری	سیستم پایه	67.6	55.9	14.1
	+ بازآرایی صفات	66.3	55.2	15.7
	+ حالت پایه صفات	65.7	54.5	16.5
1K	سیستم پایه	60.1	47.4	23.9
	+دیکشنری	56	43.2	28.3
	+ بازآرایی صفات	54	42	30.5
	+ حالت پایه صفات	53.9	42	30.6
13K	سیستم پایه	49.6	37.4	36.2
	+دیکشنری	48.6	36.3	37.2
	+ بازآرایی صفات	47.4	36	38.6
	+ حالت پایه صفات	47.3	35.7	39.1
1.3M	سیستم پایه	39.7	30.6	47.8
	+ بازآرایی صفات	39.6	30.5	48.3

در ادامه این مقاله ترجمه صربستانی به انگلیسی و بالعکس مورد بررسی قرار گرفته است. زبان صربستانی نیز همانند زبان اسپانیایی یک زبان تصریفی است و کلمات براساس شخص و زمان مورد استفاده در جمله حالت‌های مختلفی به خود می‌گیرند. به همین منظور ابتدا تمامی کلمات به حالت پایه‌شان تبدیل می‌شوند زیرا تصریف مورد استفاده در کلمات در حین ترجمه به انگلیسی کاربرد خاصی ندارند. اما این قضیه برای افعال صادق نیست و زمان و شخص افعال در حین ترجمه به انگلیسی مورد استفاده قرار می‌گیرند. به همین منظور و برای حفظ این اطلاعات از برچسب‌های بخش‌های سخن[1] استفاده شده است. در نهایت افعال در قالب حالت پایه فعل و یک برچسب بخش‌های سخن مشخص می‌شوند. همچنین برای تفکیک افعال مثبت و منفی نیز از یک پیشوند استفاده می‌گردد. به این صورت که برای افعالی که حالت منفی دارند یک پیشوند به ابتدای آن‌ها به نشانه منفی بودن، متصل می‌گردد. در مورد ترجمه انگلیسی به صربستانی فقط به این نکته اشاره شده است که حروف تعریف مانند the یکی از پرکاربردترین دسته از کلمات در انگلیسی هستند اما در طرف مقابل در زبان صربستانی اصلاً حرف تعریف نداریم. از اینرو این دسته از کلمات از پیکره انگلیسی حذف می‌گردند. نتایج حاصل از اعمال مطالب اشاره شده در ترجمه صربستانی به انگلیسی در جدول 2-8 و در ترجمه انگلیسی به صربستانی در جدول 2-9 نشان داده شده است.

جدول 2-8 نتایج ترجمه برای ترجمه صربستانی به انگلیسی[POP06]

صربستانی به انگلیسی		WER	PER	BLEU
0.2K	سیستم پایه	65.5	60.8	8.3
	+عبارات	65	59.8	10.3
	+ ریشه کلمات	59.2	54.8	13.9
	+POS افعال+حالت منفی	60	52.6	14.8
2.6K	سیستم پایه	44.5	37.9	32.1
	+ ریشه کلمات	42.9	37.4	35.4
	+POS افعال+حالت منفی	41.9	34.7	34.6

[1] POS Tags

جدول 2-9 نتایج ترجمه برای ترجمه انگلیسی به صربستانی[POP06]

BLEU	PER	WER	انگلیسی به صربستانی	
6.8	68.4	73.4	سیستم پایه	0.2K
9.3	67.5	71.9	+عبارات	
9.4	62.2	66.7	+ حذف حروف تعریف	
23.1	45.8	51.8	سیستم پایه	2.6K
24.6	44.6	50.4	+ حذف حروف تعریف	

در [MUC10] هدف یافتن اطلاعات لغوی و نحوی زبان مبدا است بطوری که باعث بهبود ترجمه شود. در این مقاله به این نکته اشاره شده است که ترجمه ماشینی آماری در زبان‌هایی که تصریفی هستند به تنهایی نمی‌تواند موفق باشد. در اینگونه زبان‌ها و در مراحل ترجمه ماشینی آماری، بسیاری از کلمات که دارای ریشه یکسانی بوده اما به خاطر تصریف زبانی دارای شکل ظاهری متفاوتی هستند، مستقل در نظر گرفته می‌شوند؛ در اینگونه موارد درصورتی که ریشه و مصدر آنها را بیابیم، می‌توانیم چالش‌های موجود در زمان ترازبندی را کاهش داده و در نتیجه ترجمه بهتری داشته باشیم. اما نکته‌ای که مهم است این است که برای رساندن عبارت به ریشه خود بایستی برخی از اطلاعات حذف گشته و برخی تبدیلات بر روی آن‌ها صورت گیرد. در این مواقع اگر حذفیات و تبدیلات به صورت مناسب صورت نگیرد و برخی اطلاعات لازم برای ترجمه حذف شوند ترجمه مناسبی نخواهیم داشت. ایده اصلی مقاله کاهش اطلاعات زبان‌شناسی لازم برای هنگام ترجمه است. یعنی هدف یافتن اطلاعات لغوی و نحوی در زبان مبدا و مقصد با استفاده از یک الگوریتم بهینه سازی عمومی است، بطوری که فقط اطلاعات لازم برای زمان ترازبندی حفظ گردد و سایر اطلاعات حذف شوند. نتایج آزمایشات انجام شده در این زمینه و بر روی ترجمه اسلونی – انگلیسی در جدول 2-10 نشان داده شده است. در جدول نشان داده شده عبارت MSD[1] برچسب‌های تصریف شده کلمات است.

[1] morpho-syntactic descriptions

جدول 2-10 نتایج آزمایشات [MUC10]

واحد مدلسازی	انواع	BLEU	تعداد MSD
کلمات	63575	34.44	-
ریشه معنایی	44916	34.21	-
ریشه معنایی + POS	47800	34.23	12
ریشه معنایی + MSD کامل	83237	33.83	1024
ریشه معنایی + MSD کاهش یافته	53565	35.77	105

در [ELK10] مولفین به بهبود ترجمه انگلیسی – ترکی پرداخته‌اند. ابتدا به این مطلب اشاره شده است که زبان ترکی زبان بسیار پیچیده‌تری نسبت به انگلیسی است و همین امر عمل ترجمه بین این دو زبان را دچار چالش می‌نماید. به مسائل زیر در این مقاله اشاره شده است:

- بررسی اثر نمایش‌های مختلف صرفی کلمات در هر دو سمت انگلیسی و ترکی بر ترجمه ماشینی آماری.
- کمک به ترازبندی با انجام بازآرایی محلی کلمات در سمت انگلیسی و ایجاد شباهت ساختاری نحوی با زبان ترکی.
- کاهش اثر پیکره دو زبانه کوچک با به کار بردن ریشه کلمات در داخل پیکره
- اصلاح کلمات خروجی ترکی از لحاظ ریخت‌شناسی

در این مقاله مولفین براساس مولفه‌های زبان‌شناسی و نحوی، سطوح مختلفی برای نمایش اطلاعات پیکره ایجاد نموده‌اند. همچنین برای ترتیب بهتر کلمات خروجی یک ترتیب متناسب با جملات ترکیه‌ای بر روی جملات انگلیسی انجام دادند. در مقاله آن‌ها نشان داده شده است که با غنی‌سازی پیکره و ترکیب اطلاعات زبان‌شناسی با ترجمه آماری می‌توان بهبود بسیار زیادی در ترجمه ترکی به انگلیسی ایجاد نمود. نتایج حاصله در جدول 2-11 نشان داده شده است.

جدول 2-11 نتایج ترجمه برای تمامی تاثیرات صرفی بر روی مراحل ترجمه[ELK10]

	مراحل	BLEU
0	سیتم پایه مبتنی بر کلمات	15.23
1	کدگذاری با پارمترهای اصلاح شده	20.16
2	کدگذاری با مدل زبانی برپایه صرف	21.9
3	(2) + اضافه کردن محتوای آموزشی	22.2
4	(3) + امتیازدهی مجدد با مدل زبانی برپایه کلمات	22.81
5	(4) + بازآرایی	23.69
6	(5) + افزایش داده	24.83
7	(6) + تصحیح کلمات ناقص	24.9
8	(7) + تصحیح کلمات OOV[1]	25.17

در [WAN07] به منظور بهبود ماشین ترجمه آماری از چینی به انگلیسی از بازآرای نحوی استفاده شده است. در این مقاله مولفین ابتدا به بررسی زبان‌های چینی و انگلیسی و تفاوت‌های موجود بین آن‌ها پرداخته‌اند. سپس براساس تفاوت‌های استخراج شده یکسری قوانین براساس درخت تجزیه زبان چینی برای کاهش فاصله ساختاری زبان چینی به انگلیسی استخراج نموده‌اند. استخراج این قوانین در گروه‌های مختلف نحوی زبان چینی صورت گرفته است. در انتها با اعمال قوانین بازآرایی استخراج شده، نتایج حاصل از ترجمه ماشینی آماری چینی به انگلیسی با قبل از اعمال این قوانین مقایسه شده است که نشان دهنده بهبود در نتایج آزمایشات می‌باشد. نتایج حاصل از آزمایشات این مقاله در جدول 2-12 نشان داده شده است.

[1] out-of-vocabulary

جدول 2-12 نتایج حاصل از بازآرایی نحوی بر روی ترجمه ماشینی آماری چینی به انگلیسی[WAN07]

Gain	BLEU	
-	31.57	سیستم پایه
+1.14	32.71	قواعد VP
+0.66	32.23	قواعد NP
+0.02	31.59	قواعد LC
+1.29	32.86	تمام قواعد

خلاصه‌ای از سیستم‌های مترجم آماری معرفی شده در جدول 2-13 نشان داده شده است.

جدول 2-13 خلاصه اطلاعات سیستم‌های مترجم آماری معرفی شده

خروجی بهبود یافته	خروجی پایه	معیار ارزیابی	حجم پیکره	اعمال صورت گرفته	جفت زبان ماشین ترجمه	
18	20.3	SSER[%]	k45	برچسب‌زنی POS، ترکیب کلمات، جداسازی پیشوندهای فعلی	آلمانی به انگلیسی	[NIE00]
39	36	BLEU	3.3 M	تغییرات صرفی بر روی پیکره	عربی به انگلیسی	[LEE04]
49.6	43.2	BLEU	k13	استفاه از دیکشنری، بازآرایی صفات، تغییر صفات به حالت پایه	اسپانیایی به انگلیسی	[POP06]
39.1	36.2	BLEU	k13	استفاه از دیکشنری، بازآرایی صفات، تغییر صفات به حالت پایه	انگلیسی به اسپانیایی	[POP06]
35.4	32.1	BLEU	k2.6	استفاده از ریشه کلمات،POS افعال، مشخص کردن حالت منفی	صربستانی به انگلیسی	[POP06]
24.6	23.1	BLEU	k2.6	حذف حروف تعریف	انگلیسی به صربستانی	[POP06]
35.77	34.44	BLEU	k98	ریشه معنایی، MSD کاهش یافته	اسلونی به انگلیسی	[MUC10]

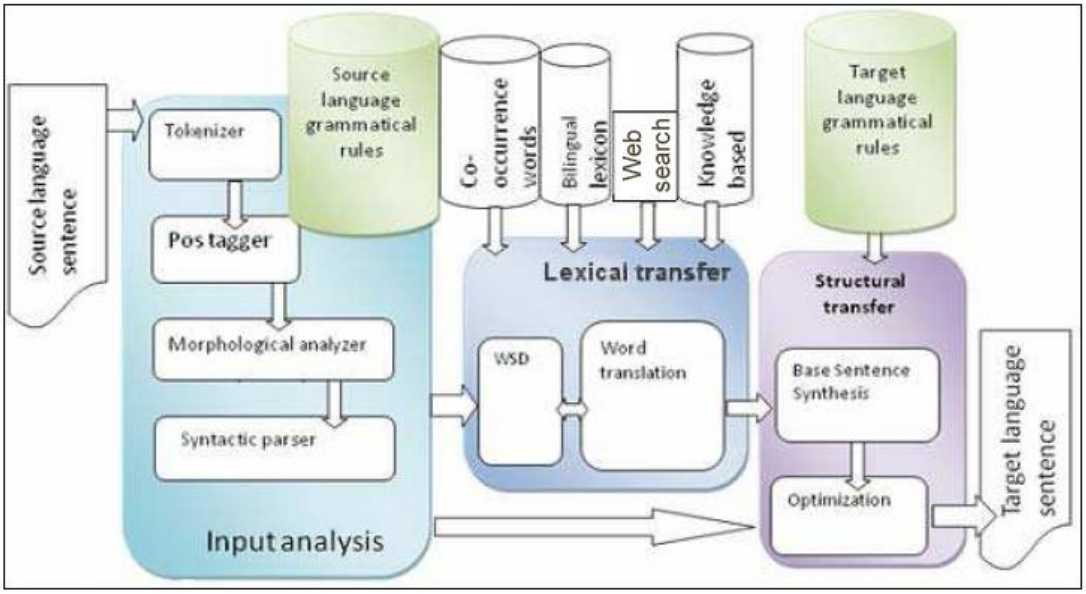
[ELK10]	انگلیسی به ترکی	کدگذاری با مدل زبانی برپایه صرف، اضافه کردن محتوای آموزشی، بازآرایی، افزایش داده، تصحیح کلمات ناقص، تصحیح کلمات OOV	k44	BLEU	15.23	24.9
[WAN07]	چینی به انگلیسی	اعمال بازآرایی نحوی	k637	BLEU	31.57	32.86

2 - 7 - 2 - ترجمه ماشینی آماری انگلیسی به فارسی

ترجمه‌هایی که برای زبان فارسی صورت گرفته است اکثراً با استفاده از روش‌های قانونمند است. اولین کار آکادمیک بر روی زبان فارسی در حدود اواخر دهه نود در دانشگاه شیراز صورت گرفت و گام نخست را در این زمینه را این دانشگاه برداشت[AMT00]. اما این شروع ادامه ای نداشت و همان کار نیز خروجی دقیقی نداشت تا بتوان نظر درستی در مورد آن داد.

اما بعد از آن چند گروه بر روی ترجمه ماشینی فعالیت‌هایی را انجام داده‌اند. در [SAE09] فعالیت‌های انجام گرفته اکثراً بر اساس قوانین زبان‌شناسی بوده و بیشتر کار براساس قوانین نحوی است؛ اما در مرحله دوم زمانی که نمی‌توان براساس قوانین موجود در پایگاه دانش، ابهام ترجمه کلمات را رفع نمود از یک پیکره استفاده می‌شود تا براساس هم وقوعی و تکرار کلمات، ابهامات موجود را رفع نماید. معماری کلی سیستم پیاده‌سازی شده در این مقاله در شکل 2-14 نشان داده شده است که مرحله رفع ابهام مرحله WSD[1] است.

[1] Word Sense Disambiguation



شکل 2-14 معماری کلی سیستم PEnT2 [SAE09]

تلاش صورت گرفته بعدی توسط محقق و صراف‌زاده در دانشگاه مسی استرالیا می‌باشد. در [MOH09] کاملاً به روشهای آماری برای ترجمه انگلیسی به فارسی پرداخته شده است. در واقع می‌توان گفت این اولین سیستم مترجم آماری انگلیسی به فارسی ارائه شده است.[1] ابتدا یک پیکره دوزبانه براساس اخبار BBC و یک پیکره تک زبانه براساس پیکره همشهری[ALE09] ایجاد شده است و هدف اصلی ترجمه از انگلیسی به فارسی بوده است. لازم به ذکر است که پیکره تک زبانه براساس همشهری موجود است اما پیکره دو زبانه را خود این گروه تراز و ایجاد نموده است. در انتها براساس پیکره‌های ایجاد شده و با استفاده از ابزار موجود در زمینه ترجمه ماشینی آماری یک سیستم مترجم آماری ایجاد شده و خروجی‌های آن استخراج گردیده است.

این گروه در ادامه کار و در [MOH09] به بررسی میزان حجم پیکره‌ها و همچنین رابطه بین حجم پیکره دو زبانه و تک زبانه و اثر آن‌ها بر ترجمه، پرداخته شده است. در این مقاله خروجی مترجم آماری با تغییرات اعمال شده بر اندازه پیکره تک زبانه و موازی استخراج شده است. خروجی‌های این

[1] البته شرکت گوگل اولین سیستم مترجم آماری انگلیسی به فارسی و بالعکس را ارائه کرده است و منظور از اولین سیستم اشاره شده در متن، اولین سیستمی می‌باشد که مقاله‌ای در مورد آن ارائه گردیده است.

مقاله در جدول 2-14و جدول 2-15 و جدول 2-16 نشان داده شده است.

جدول 2-14 نتایج بدست آمده با استفاده از پیکره موازی با اندازه 817 جمله[MOH09]

اندازه پیکره = 817 جمله			
7005	1066	864	اندازه پیکره تک زبانه (مدل ترجمه)
0.0805	0.0920	0.1061	BLEU
1.6721	1.6838	1.8218	NIST

جدول 2-15 نتایج بدست آمده با استفاده از پیکره موازی با اندازه 1011 جمله[MOH09]

اندازه پیکره = 1011 جمله			
7005	1066	864	اندازه پیکره تک زبانه (مدل ترجمه)
0.0888	0.0986	0.0882	BLEU
1.5512	1.5301	1.5338	NIST

جدول 2-16 نتایج بدست آمده با استفاده از پیکره موازی با اندازه 2343 جمله[MOH09]

اندازه پیکره = 2343 جمله			
7005	1066	864	اندازه پیکره تک زبانه (مدل ترجمه)
0.1148	0.1127	0.0806	BLEU
1.7554	1.6961	1.7364	NIST

در انتها براساس نتایج حاصل از ترجمه با اندازه‌های مختلف پیکره‌های تک زبانه و موازی بیان شده است که با افزایش پیکره موازی کیفیت ترجمه بهبود خواهد یافت اما اندازه پیکره تک زبانه بایستی متناسب با اندازه پیکره موازی باشد.

این گروه در کار بعدی[MOH11] خود به افزایش حجم پیکره پرداخته‌اند که اندازه پیکره‌های ایجاد شده در جدول 2-17 نشان داده شده است.

جدول 2-17 اندازه پیکره‌های توسعه یافته[MOH11]

پیکره‌ها	نوع	جملات انگلیسی	کلمات انگلیسی	جملات فارسی	کلمات فارسی
سسیستم 1	اخبار	9351	208961	9351	226759
سسیستم 2	اخبار	18277	334440	18277	362326
سسیستم 3	اخبار	27773	437781	27773	472679
سسیستم 4	اخبار	37560	506972	37560	548038
سسیستم 5	اخبار	46759	708801	46759	776154
TEP	زیرنویس فیلم	612086	3920549	612086	3810734
NSPEC	زیرنویس فیلم-اخبار	618039	5370426	618039	5137925

همچنین آن‌ها به بررسی نتایج در اندازه‌ها و دامنه‌های مختلف پرداخته‌اند و دلیل ایجاد پنج سیستم متفاوت هم همین امر است. در ادامه با استفاده از دو پیکره تک زبانه(مدل زبانی)، خروجی سیستم را استخراج نموده‌اند که نتایج حاصل در جدول 2-18 و جدول 2-19 نشان داده شده است.

جدول 2-18 خروجی سیستم مترجم PEEN-SMT [MOH11]

مدل زبانی = Europarl v4		
سیستم	BLEU	NIST
سیستم 1	0.1208	2.5952
سیستم 2	0.1277	2.5592
سیستم 3	0.2005	3.5310
سیستم 4	0.2415	3.2908
سیستم 5	0.2576	3.1892
TEP	0.0414	2.1196
NSPEC	0.1796	3.1622

جدول 2-19 خروجی سیستم مترجم PEEN-SMT [MOH11]

مدل زبانی = News-Commentary		
سیستم	BLEU	NIST
سیستم 1	0.1318	2.8344
سیستم 2	0.2656	3.2458
سیستم 3	0.2910	3.4425
سیستم 4	0.3056	3.7057
سیستم 5	0.3332	3.8085
TEP	0.0621	2.2952
NSPEC	0.1975	2.9907

کار بعدی در زمینه ترجمه ماشینی توسط پیله‌ور و فیلی صورت گرفت است[PIL10]، که این کار هم کاملاً بر اساس روش‌های آماری است. این گروه نیز ابتدا یک پیکره دو زبانه بر اساس زیرنویس فیلم، سه کتاب در زمینه کامپیوتر و همچنین استفاده از پیکره موازی موجود در سایت KDE4، ایجاد کردند[PIL11] و برای پیکره تک زبانه هم از پیکره همشهری[ALE09] استفاده نمودند. بر اساس اطلاعات منتشر شده تا کنون این پیکره بزرگترین پیکره موجود در زمینه ترجمه می‌باشد. البته اکثر کلمات پیکره شامل زیرنویش فیلم و بصورت محاوره‌ای است. با پیکره موازی ایجاد شده این گروه توانستند برای ترجمه زیرنویس فیلم‌ها به دقت بهتری نسبت به مترجم گوگل دست یابند. نتایج حاصل از سیستم ترجمه ایجاد شده توسط این گروه با دو مجموعه تست متفاوت با مترجم گوگل و یک مترجم ساده که فقط کار جایگزینی براساس دیکشنری را انجام می‌دهد مقایسه شده است. نتایج حاصل از این مقایسه در جدول 2-20 نشان داده شده است.

جدول 2-20 مقایسه سیستم PersianSMT با گوگل و یک مترجم ساده[PIL10]

سیستم‌های ترجمه	مجموعه تست Subtitle	مجموعه تست EGIU
	BLEU	BLEU
PersianSMT	25.46	11.22
مترجم Google	0.96	5.35
مترجم ساده	کمتر از 0.5	کمتر از 0.5

در انتها بایستی اشاره شود که هنوز یک پیکره استاندارد برای ترجمه ماشینی انگلیسی به فارسی وجود ندارد؛ از اینرو در هر یک از کارهای مطرح شده مؤلفین خود به ایجاد پیکره موازی اقدام نموده‌اند. همچنین مجموعه مورد نظر برای تست هم در هر یک از مقالات معرفی شده متفاوت است. همین امر باعث شده است که نتوان مقایسه‌ای بین آن‌ها انجام داد. البته تنها مقایسه‌ای که شاید بتوان انجام داد مربوط به کیفیت پیکره‌های مورد استفاده می‌باشد که خارج از حوزه این پایان‌نامه است.

نتایج سیستم‌های مترجم آماری ایجاد شده برای ترجمه نگلیسی به فارسی در جدول 2-21 خلاصه شده است.

جدول 2-21 خلاصه اطلاعات سیستم‌های مترجم آماری انگلیسی به فارسی

	جفت زبان ماشین ترجمه	اعمال صورت گرفته	حجم پیکره	BLEU
[MOH09]	انگلیسی به فارسی	ایجاد پیکره و ایجاد سیستم مترجم آماری مبتنی بر عبارت	k2	11.48
[MOH11]	انگلیسی به فارسی	افزایش پیکره	k46	25.76
[PIL10]	انگلیسی به فارسی	ایجاد پیکره و ایجاد سیستم مترجم آماری مبتنی بر عبارت	M5 توکن	25.46

2 - 8 - چالش‌های موجود

همان طور که گفته شد برای انجام ترجمه ماشینی آماری دو مدل ایجاد می‌شود. مدل زبانی که از روی پیکره تک زبانه مقصد ایجاد می‌گردد و مدل ترجمه که از روی پیکره دوزبانه تراز شده مبدا و مقصد ایجاد می‌گردد. ایجاد مدل زبانی بسیار ساده بوده و مشکل خاصی در مورد آن وجود ندارد اما مدل ترجمه مخصوصاً برای ترجمه انگلیسی به فارسی یا بالعکس دارای مشکلات زیر است:

2 - 8 - 1 - نبود پیکره دوزبانه مناسب

هنوز پیکره مناسبی برای ایجاد مدل ترجمه برای زبان فارسی موجود نیست. این پیکره بایستی آن قدر بزرگ باشد که بتواند تمام ترکیبات و لغات ممکن زبانی را به شکل مناسب در خود جای داده باشد. به دلیل نبود منابع لازم و کافی زبانی ترجمه شده مناسب، ایجاد چنین پیکره‌ای بسیار مشکل است.

2 - 8 - 2 - محدودیت در مورد کار با پیکره‌های بزرگ

اگر هم بتوان پیکره مناسب و بسیار بزرگی ایجاد نماییم استفاده از آن دارای محدودیت‌های خواهد بود از جمله فضا و زمان زیاد مورد نیاز برای انجام عمل ترجمه که در بسیاری از مواقع مشکلات و محدودیت‌های عدیده‌ای را به وجود می‌آورد.

2 - 8 - 3 - تفاوت ساختاری بسیار زیاد بین زبان فارسی و انگلیسی

در مرحله ایجاد مدل ترجمه از روی جملات تراز شده بایستی به کلمات یا عبارات تراز شده برسیم. با توجه به تفاوت بسیار زیاد ساختاری بین زبان‌های انگلیسی و فارسی، استخراج چنین ترازهایی مشکل و درنهایت نامناسب خواهد بود که باعث ترجمه نامناسب نهایی خواهد شد. ضمناً زبان فارسی از جمله

زبان‌هایی است که اصطلاحاً تصریفی[1] نامیده می‌شود بطوری که کلمات آن به اشکال مختلف در جمله ظاهر می‌شوند. مثلاً زمان و شخص باعث ایجاد اَشکال مختلف فعلی می‌گردد که این مساله برای زبانی مثل زبان انگلیسی برقرار نمی‌باشد.

براساس آزمایشات انجام گرفته بر روی یک پیکره کوچک و با توجه به مشکلات مطرح شده مشخص شد که ترجمه آماری صرف برای انگلیسی – فارسی چندان مناسب عمل نمی‌کند و بایستی برخی مسائل دیگر از جمله ساختار زبان‌های مبدا و مقصد نیز وارد مراحل ترجمه گردد.

2 - 9 - خلاصه فصل

روش های سنتی ترجمه ماشینی نیازمند حجم زیادی از دانش زبانشناسی برای تولید قوانین برای ترجمه هستند. ترجمه ماشینی آماری یک راه خودکار برای یافتن رابطه بین قابلیت های دو زبان مبدا و مقصد با استفاده از ایجاد مدل ترجمه و مدل زبانی از یک پیکره موازی را ارائه می‌نماید.

ترجمه ماشینی آماری کاملاً بی نیاز از دانش زبانشناسی است اما دارای نتایج قابل رقابتی نسبت به سایر روش‌ها بوده است. با توجه به این که دانش زبانشناسی در برخی موقعیت‌ها توانایی حل مسئله را ندارد این روش می‌تواند این کاستی‌ها را رفع نماید. محققین معتقدند که استفاده از دانش زبانشناسی می‌تواند به کمک ترجمه ماشینی آماری آمده و باعث حصول نتایج مطلوبتری گردد. این اطلاعات ممکن است به فرم قوانین صرفی، قوانین مربوط به ترتیب کلمات، اصطلاحات کاربردی، کلمات شناخته شده و معروف و از این قبیل داده‌ها باشد. برای ترجمه بین انگلیسی و فارسی به دلیل تفاوت‌های ساختاری و لغوی زیاد، این اطلاعات زبان‌شناسی می‌تواند بسیار موثر واقع شود.

یکی از مشکلات بسیار مهم در زمینه کار با ترجمه آماری ایجاد پیکره موازی است. مخصوصاً برای پیکره موازی بین زبان فارسی و انگلیسی زیرا برای اینکه این روش جواب مناسبی داشته باشد بایستی

[1] reflective

یک پیکره موازی بزرگ برای استفاده موجود باشد. مسئله بعدی بحث ترازبندی بین جملات و به تبع آن ترازبندی بین کلمات موجود در پیکره موازی می‌باشد.

ترجمه ماشینی آماری برای بسیاری از زبانها بصورت مطلوب عمل کرده و حاصل آن را می توان در سایت گوگل دید اما برای برخی جفت زبان‌ها مانند انگلیسی و فارسی هنوز جواب مناسبی حاصل نشده است که به نظر مشکل از نبود پیکره موازی مناسب است و ترازبندی مناسب بین جملات و پیکره‌ها باشد. امید است با ترکیب برخی اطلاعات زبان‌شناسی بتوان تا حدی بر این مشکلات فایق آمد.

فصل سوم:

روش پیشنهادی

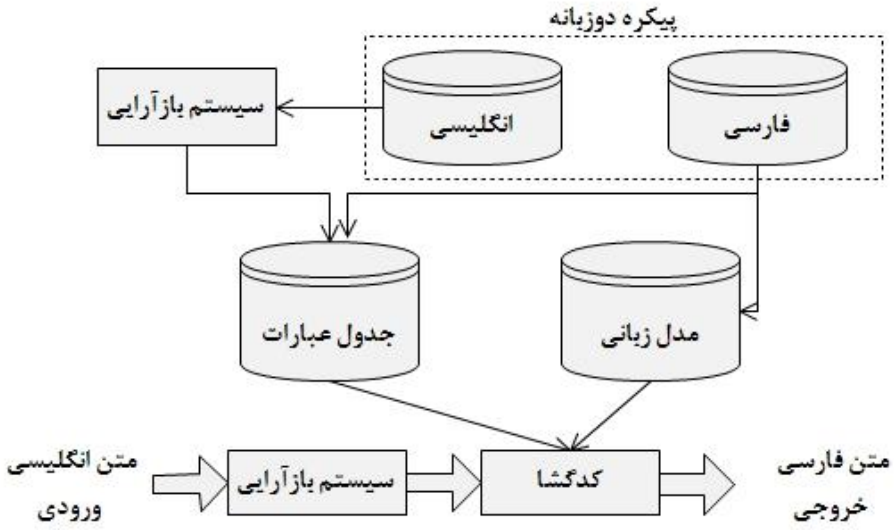
3 - روش پیشنهادی

در این فصل به بررسی برخی از راه‌کارهای مورد استفاده در جهت بهبود ترجمه ماشینی آماری انگلیسی به فارسی پرداخته شده است. در این راه‌کارها، اطلاعات زبان‌شناسی صراحتاً در مراحل ترجمه درگیر شده‌اند. ابتدا سعی شده است با استفاده از بازآرایی نحوی[1] عمل ترازبندی و در نتیجه نتایج ترجمه را بهبود دهیم. در ادامه مدل مبتنی بر فاکتور بر روی سیستم ترجمه آماری انگلیسی به فارسی آزمایش شده است. برای ایجاد این مدل یک ریشه‌یاب نیاز است که روال طراحی و ایجاد آن نیز توضیح داده شده است.

3 - 1 - بازآرایی نحوی جملات

در روش پیشنهادی برای بهبود خروجی سیستم مترجم آماری یک رویکرد بازآرایی نحوی، برای ترجمه انگلیسی به فارسی با استفاده از درخت تجزیه ارائه شده است. درابتدا برای انجام این روش، با بررسی گرامر زبان فارسی و انگلیسی، برخی قوانین تبدیلی، در جهت کاهش تفاوت ساختاری دو زبان انگلیسی و فارسی، استخراج گردیده است. در مرحله بعد این قوانین بصورت یک عمل پیش پردازشی و فقط در سمت زبان انگلیسی، بر روی هر دو داده آموزشی و آزمایشی اعمال می‌گردد. با انجام این عمل شباهت جملات موجود در پیکره دوزبانه از لحاظ ساختار نزدیک به هم می‌شوند و همین امر عمل ترازبندی عبارات ترجمه را بهتر و در نتیجه مدل ترجمه را بهبود می‌بخشد. در شکل 3-1 معماری کلی سیستم ارائه شده نشان داده شده است.

[1] Syntactic Reordering



شکل 3-1 مراحل کار سیستم پیشنهادی برای بازآرایی نحوی

3-1-1- قوانین بازآرایی

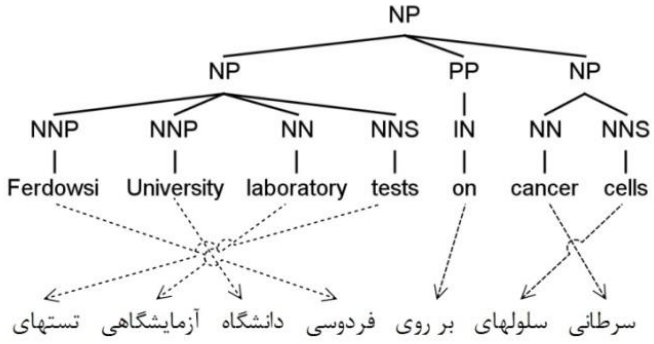
زبان فارسی و انگلیسی از لحاظ ساختار نحوی تفاوتهای بسیار زیادی با یکدیگر دارند. بصورت کلی، زبان فارسی را یک زبان SOV[1] و زبان انگلیسی را یک زبان SVO[2] می‌دانند. البته این نکته یکی از تفاوتهای موجود میان این دو زبان را به شکلی ساده بیان می‌کند با این حال وجوه تمایز پیچیده‌تری نیز وجود دارند که نمی‌توان آنها را به سادگی قاعده فوق نشان داد. از این رو لازم است با بررسی انواع عباراتی که درون جمله روی می‌دهند قواعد بیشتری را استخراج نمود. در ضمن بایستی به این نکته نیز توجه شود که زبان فارسی در بسیاری از ترکیبات دارای ترتیب خاصی نیست و اصطلاحاً دارای ترتیب آزاد[3] است و کلمات می‌توانند با ترکیبات متفاوت در عبارات قرار گیرند. البته همین مساله، کار را برای استخراج قوانین بسیار مشکل می‌نماید.

[1] Subject Object Verb
[2] Subject Verb Object
[3] Free order

در روش ارائه شده در این پایان‌نامه برای استخراج قوانین بازآرایی جهت اعمال بر روی پیکره موازی و مجموعه آزمایشی از برچسب های دستوری درخت تجزیه استفاده شده است. همچنین کلیه قوانین بازآرایی در دو گروه اسمی و فعلی به طور مجزا مورد بررسی قرار گرفته است.

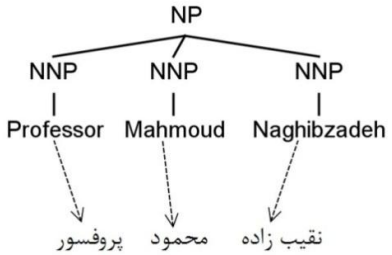
3-1-1-1- عبارات اسمی[1]

بررسی‌های دستوری زبان فارسی در مورد عبارات اسمی، نشان می دهند که ترتیب قرار گرفتن چند اسم در یک عبارت اسمی دقیقا برعکس قرار گرفتن عبارت اسمی معادل در زبان انگلیسی است. شکل 3-2 نمونه‌ای از این عبارات را نشان می‌دهد.



شکل 3-2 نمونه‌ای از عبارت اسمی بازآرایی شده شامل چند اسم

البته استثنائاتی نیز برای این قانون وجود دارد. به عنوان مثال در شکل 3-3 ترتیب اسامی افراد در ترجمه انگلیسی تغییر نمی‌کند.

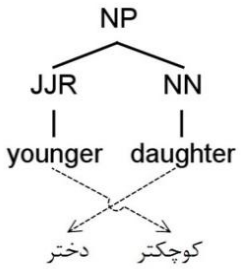


شکل 3-3 نمونه‌ای از عبارت اسمی مربوط به نام افراد

[1] Noun phrases

پس در صورتی که اجزای عبارت اسمی مربوط به اسامی افراد نباشند می‌توان با معکوس نمودن ترتیب هریک از آنها عمل بازآرایی را انجام داد.

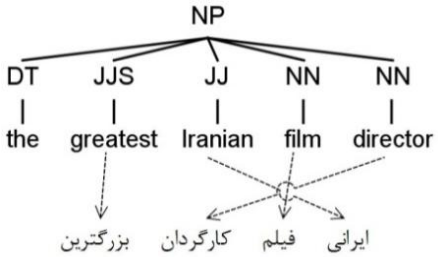
تفاوت دیگر میان زبان فارسی و انگلیسی در ترتیب قرارگرفتن موصوف و صفت مشاهده می‌شود. در زبان فارسی ابتدا موصوف و بعد از آن صفت بیان می‌گردد. این در حالی است که در زبان انگلیسی صفت قبل از موصوف می‌آید. شکل 3-4 این تفاوت را با یک مثال نشان می‌دهد.



شکل 3-4 مثالی از ترتیب قرار گرفتن موصوف و صفت درون یک گروه اسمی

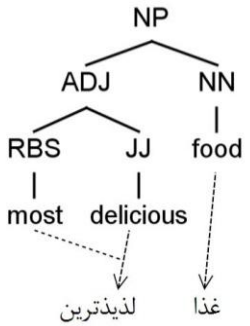
از این رو قانون دیگر برای بازآرایی چنین بیان می شود؛ در مواقعی که در یک گروه اسمی، یک صفت یا گروه صفتی قبل از یک اسم یا گروه اسمی‌قرار گیرد، صفت می‌بایست به بعد از آن منتقل شود.

البته باز هم در اینجا استثنائی وجود دارد. به عنوان مثال در مورد صفتهای برتر که با برچسب JJS مشخص می‌شوند این قانون صدق نمی‌کند و عمل تغییر مکان نبایستی صورت بگیرد. شکل 3-5 نشان می‌دهد که قاعده فوق در این حالت صدق نمی‌کند.



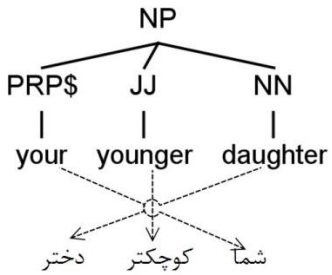
شکل 3-5 مثالی از ترتیب قرار گرفتن ترکیب صفت برتر و اسم در یک گروه اسمی

همچنین بر اساس شکل 3-6 هر گاه در عبارت انگلیسی قبل از صفت، قیدی بـا برچسـب RBS قـرار گیرد؛ صفت تبدیل به صفت برتر می شود و نیازی به جابجایی آن نیست.



شکل 3-6 مثالی از تبدیل یک صفت ساده به یک صفت برتر با استفاده از قید با برچسب RBS

در یک گروه اسمی‌قیدهایی که با برچسب RPR$ نیز مشخص می‌شوند بایستی بـه بعـد از اسـامی و صفتهای بعد خود منتقل گردند. شکل 3-7 نمونه ای از این عبارات را نشان می‌دهد.

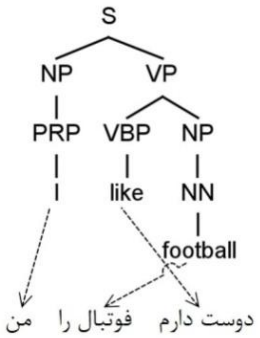


شکل 3-7 مثالی از یک ضمیر ملکی که بایستی به بعد از اسامی و صفتهای بعد خود منتقل شود.

3-1-1-2- عبارات فعلی[1]

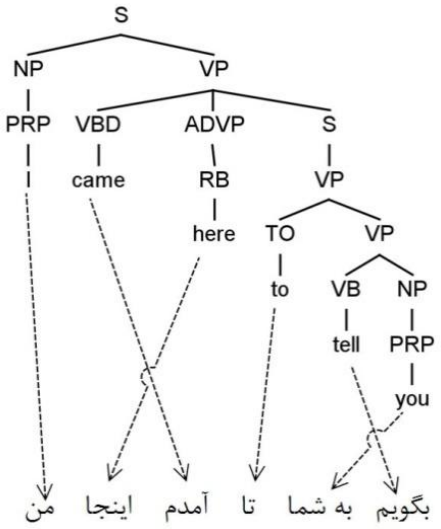
یکی از تفاوتهای اصلی زبان‌های انگلیسی و فارسی مربوط به جایگاه فعل در جمـلات اسـت. اصـولاً در زبان فارسی افعال در انتهای جملات قرار می‌گیرند. همانطور که در شکل 3-8 دیـده مـی شـود بـرای بازآرایی، ابتدا افعال در عبارت انگلیسی جستجو شده، سپس به انتهای عبارت فعلی منتقل می‌شود.

[1] Verb phrases



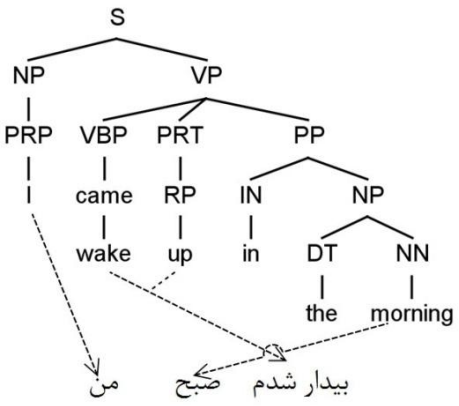
شکل 3-8 تغییر جایگاه فعل در عبارت فعلی

همچنین هرگاه درون عبارت فعلی، برچسب شروع جمله (این برچسب‌ها با S شروع می‌شوند) مشاهده گردد، یعنی حالتی که یک جمله درون جمله دیگر واقع شود می‌بایست فعل به قبل از این برچسب منتقل گردد. در شکل 3-9 این مطلب نشان داده شده است.



شکل 3-9 مثالی از قاعده بازآرایی در مورد جایگاه فعل در جملات تو درتو

برخی افعال نیز در انگلیسی وجود دارند (مانند wake up) که بیش از یک قسمت دارند. در این مواقع بایستی توجه شود که هر دو قسمت فعل با هم جابجا گردد. در شکل 3-10 نمونه‌ای از این وضعیت نشان داده شده است.



شکل 3-10 مثالی از قانون بازآرایی در مورد افعال چند قسمتی

قوانین مستخرج در این پایان نامه تمامی قوانین موجود نیست، برخی از قوانین نیز وجود دارند که با بررسی‌های انجام شده تعداد وقوع آن‌ها بسیار کم و در نتیجه تاثیر زیادی در نزدیک نمودن شباهت ساختار انگلیسی به فارسی نداشتند، از اینرو از این قوانین صرف نظر شد.

3 - 2 - مدل بر مبنای فاکتور[1]

همانطور که در قمست مرور ادبیات اشاره شد در این مدلسازی برخی اطلاعات زبانی به پیکره اضافه می‌گردد تا صراحتاً اطلاعات زبان‌شناسی در مراحل ایجاد مدل‌های مورد نیاز برای ترجمه ماشینی آماری مورد استفاده قرار گیرد، برای همین منظور ابتدا بایستی فاکتورهای مورد نیاز زبانی برای هر زبان معلوم و استخراج گردد، به همین منظور برای هر دو زبان فاکتورهای زیر استفاده گردید:

- خود کلمه[2]
- ریشه معنایی[3] کلمه
- برچسب بخش‌های سخن[4]

[1] Factored based model
[2] Surface form
[3] Lemma
[4] POS tag

در ابتدا برای استخراج فاکتورهای اشاره شده بایستی ابزاری فراهم آورده شود. از آنجایی که این ابزار برای زبان انگلیسی موجود هستند مشکلی در سمت زبان انگلیسی وجود ندارد. اما ابزار پردازش متن برای زبان فارسی یا وجود ندارند یا اگر هم وجود دارند در دسترس عموم نیستند. در نتیجه درصدد ایجاد این ابزار برآمدیم. در ادامه پس از ایجاد ابزار مورد نیاز بایستی با استفاده از فاکتورهای استخراجی، مدل مترجم آماری مبتنی بر فاکتور ایجاد گردد. توضیحات بیشتر در مورد پیاده‌سازی این روش در فصل بعد ذکر شده است.

در ادامه روال ایجاد هر یک ابزار جهت تولید فاکتورهای مورد نیز از عبارات توضیح داده شده است.

3-2-1- ریشه‌یاب فارسی

برای استخراج ریشه کلمات نیازمند یک ابزار ریشه‌یاب هستیم. تعدادی ریشه‌یاب برای زبان فارسی طراحی و ایجاد شده است[TAG05] [TAS02] [MIA06] [SAN06] [EST11]. اما این روش‌ها، سعی در حذف وندهای کلمات، بدون درنظر گرفتن معنای کلمه و جایگاه آن در جمله دارند. مشکل دیگر این روش‌ها در نظر نگرفتن کاربرد عملیات ریشه‌یابی در حوزه زبان‌شناسی است. به عنوان مثال اغلب روش‌های بررسی شده سعی در حذف بیشترین وندهای ممکن دارند که برای کاربرد بازیابی اطلاعات مناسب هستند. ولی این امر، باعث از بین رفتن معنای کلمات و ایجاد مشکل در کاربردهایی از قبیل ترجمه ماشینی، خلاصه‌سازی و سایر تحلیل‌های معنایی زبان می‌شود. در نتیجه در روش پیشنهاد شده در این پایان‌نامه سعی در برطرف نمودن این مشکلات با در نظر گرفتن وندها استثنای کلمات، قواعد حذف وندها و مشخص نمودن سطح ریشه‌یابی مورد نظر هستیم.

در روش مطرح شده ابتدا نوع عبارات براساس فعل و غیرفعل بودن تشخیص داده می‌شوند. سپس بر این اساس عمل ریشه‌یابی با روالی جداگانه برای هر کدام انجام می‌گیرد.

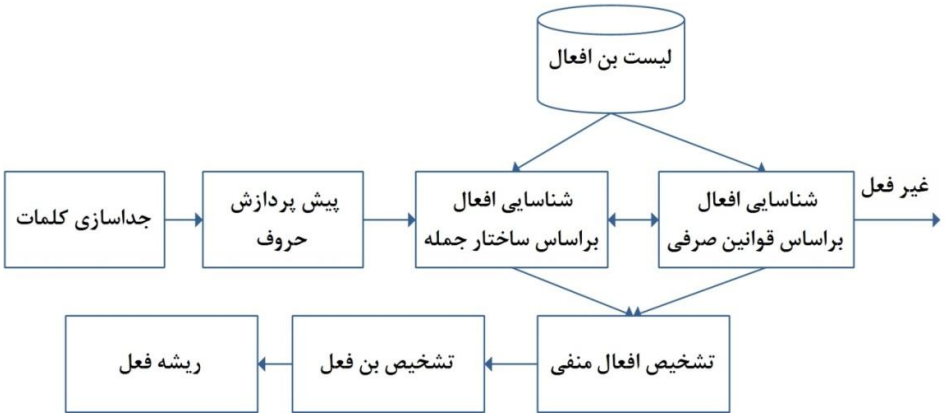
3-2-1-1- ریشه یابی عبارات فعلی

در این حالت ابتدا فعل را به دو صورت منفی یا مثبت تفکیک می‌نماییم، سپس در صورت منفی بودن فعل پسوند منفی‌ساز فعل از آن حذف می‌گردد. در مرحله بعد ساختار فعل با یکی از انواع ساختارهای فعل در زبان فارسی تطبیق داده می‌شود. پس از تطبیق و یافتن زمان مناسب فعل، براساس نوع فعل به ترتیب پسوندها و پیشوندهای موجود در فعل حذف می‌شوند. نکته‌ای که بایستی به آن دقت شود این است که در صورتی که فعلی مانند "خواب دیده بود" بصورت مرکب و چند قسمتی باشد فقط قسمت تصریف شده آن (در این مثال "دیده بود") ریشه‌یابی می‌گردد و سایر قسمت‌های قبلی بصورت مستقیم به همراه بن فعل به عنوان ریشه معرفی می‌گردد.

در انتها پس از بدست آوردن ریشه فعل، دو زمان حال و آینده در قالب بن‌های گذشته و حال ارائه می‌گردند. برای این منظور در فرهنگ لغتی حدود 400 بن فعل حال به همراه معادل گذشته‌شان ذخیره شده است. پس از یافتن و استخراج قسمت اصلی فعل آنرا داخل این فرهنگ لغت جستجو نموده و بن‌های متناسب استخراج می‌گردند. همچنین اطلاعات جانبی در مورد زمان و شخص فعل نیز در ادامه فعل درج می‌گردد. نکته مهمی که در این قسمت به آن دقت شده است این است که برخی کلمات که دارای بن فعلی هستند براساس شرایط و بافت جمله، ممکن است جایگاه فعل خود را از دست دهند و ریشه‌یابی برای آن‌ها بایستی بصورت دیگری انجام گیرد. برای این منظور لیستی از حروف اضافه استخراج شده است که زمانی که قبل از کلمه مورد نظر که شامل بن فعلی است قرار گیرند حالت کلمه را از حالت فعل به اسم تغییر می‌دهند. به عنوان مثال عبارت "به گفته" باعث می‌شود که کلمه "گفته" به اسم تبدیل شود و ریشه آن به صورت "گفت+ه" مشخص می‌گردد. جدول 3-1 حروف اضافه مورد استفاده در این موارد را نشان می‌دهد. مراحل کار در شکل 3-11 نشان داده شده است.

جدول 3-1 فهرست حروف اضافه تغییر دهنده نقش فعل به اسم

از	به	در	با	بر	درباره	برای	



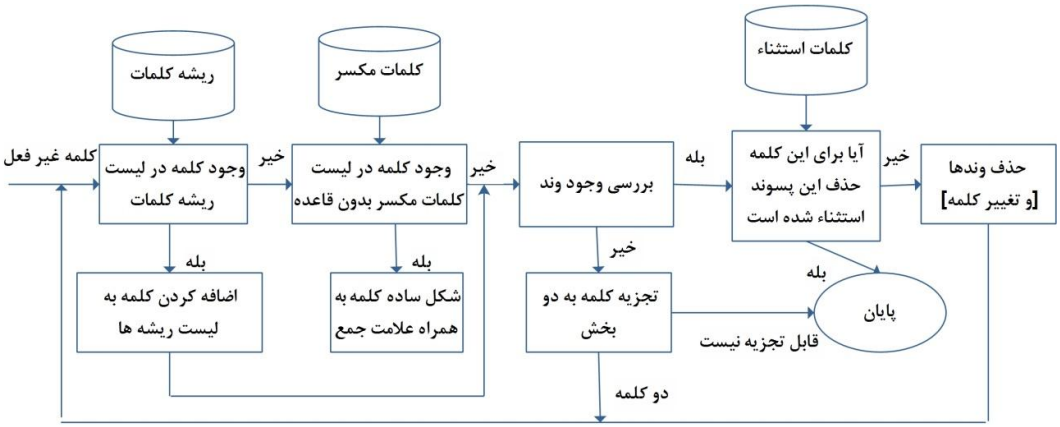
شکل 3-11 مراحل ریشه‌یابی افعال

3-2-1-2- ریشه یابی عبارات غیرفعل

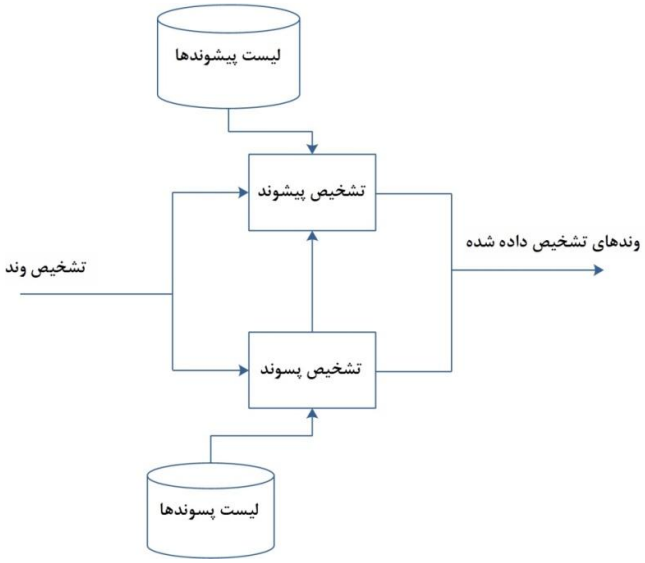
در این مرحله ابتدا کل کلمه درون فرهنگ لغت شامل ریشه کلمات جستجو می‌گردد. اگر کلمه وجود داشته باشد، خود کلمه به عنوان ریشه معرفی می‌گردد. در غیر اینصورت کلمه درون فرهنگ لغت مربوط به کلمات جمع مکسر و بدون قاعده جستجو می‌گردد. اگر کلمه در این فرهنگ لغت یافت شود، معادل مفرد آن به همراه یک نشانگر مبنی بر جمع مکسر بودن آن، برگردانده می‌شود. سپس با استفاده از قوانین مربوطه تمامی وندهای ممکن موجود در کلمه یافت می‌شود. سپس از کوچکترین پسوند شروع کرده و پسوندها حذف می‌شوند. در هر مرحله و پس از حذف پسوند مجدداً کلمه درون فرهنگ لغت مربوط به کلمات ریشه جستجو می‌شود. درصورتی که کلمه یافت شود می‌تواند به عنوان ریشه معرفی گردد. اما در این رویکرد ممکن است مشکلاتی به وجود آید. مثلاً در کلمه "ماشین" پسوند "ین" شناخته شده و ریشه "ماش" نیز در فرهنگ لغت موجود است اما این ریشه برای کلمه "ماشین" صحیح نمی‌باشد. از آنجایی که استخراج قانون برای چنین شرایطی سخت و در بعضی مواقع غیر ممکن است، از اینرو از یک فرهنگ لغت که شامل استثنائات مربوطه به حذف وندها است، بهره

گرفته شده است. پس از اینکه پسوند تشخیص داده شده حذف و ریشه محتمل استخراج گردید، آن ریشه درون فرهنگ لغت استثنائات جستجو شده و در صورتی که این ترکیب(ریشه+پسوند) جزو استثنائات نباشد؛ کلمه مربوطه به عنوان ریشه برگردانده می‌شود. در غیر اینصورت کار ریشه‌یابی خاتمه می‌یابد. در هنگام حذف پسوند در صورتی ه کلمه باقیمانده درون فرهنگ لغت کلمات ریشه نباشد، مجدداً کلمه درون فرهنگ لغت مربوط به جمع‌های مکسر جستجو می‌شود و در صورت یافت شدن، حالت مفرد کلمه به عنوان ریشه برگردانده می‌شود. این جستجو برای شرایطی صورت می‌گیرد که کلمه مورد نظر برای ریشه‌یابی شامل یک کلمه جمع مکسر به همراه پسوند متصل به آن باشد. به عنوان مثال می‌توان کلمه "اساتیدی←اساتید+ی"یا "اساتیدشان←اساتید+شان" را نام برد. در صورتی که باز هم کلمه در فرهنگ لغات کلمات جمع مکسر یافت نشود مجدداً عملیات حذف پسوند، با حذف پسوند بعدی ادامه می‌یابد. همین مراحل برای پیشوندها نیز صورت می‌گیرد و پیشوندها نیز از کلمه مربوطه حذف می‌گردند تا ریشه‌های نهایی به دست آید. همچنین ممکن است دو کلمه در متن به هم متصل شده باشند و برنامه نتواند طبق روال فوق آن‌ها را ریشه‌یابی نماید. به منظور رفع این مشکل، برای جداسازی کلمات به هم چسبیده در متن، هر کلمه که در فرهنگ ریشه کلمات و جمع‌های مکسر یافت نشود به دو بخش بزرگتر از دو حرف شکسته شده و درصورتی که هر دو بخش در فرهنگ ریشه کلمات وجود داشته باشد، عمل جداسازی انجام می‌شود. می‌توان کلمه "آبمعدنی" را به عنوان مثالی برای این شرایط بیان نمود که به دو کلمه "آب" و "معدنی" شکسته می‌شود. این عملیات بصورت تکراری ادامه می‌یابد تا در نهایت ریشه نهایی استخراج گردد. مراحل کار در شکل 12-3 و شکل 13-3 نشان داده شده است. با تکرار این عملیات، تمامی سطوح مختلف برای ریشه‌یابی اعمال شده و ابزار این قابلیت را دارد که بتوان سطوح معنایی مختلف ریشه را از عبارات استخراج نمود. به عنوان مثال برای کلمه "خلاصه‌سازی" سه ریشه "خلاصه‌سازی"، "خلاصه‌ساز" و "خلاصه" به عنوان خروجی بازگردانده می‌شود که بسته به نوع استفاده می‌توان از هر کدام از آنها استفاده نمود.

به عنوان مثال دیگر برای کلمه "دانشجویان" نیز سه ریشه "دانشجو"، "دانش" و "دان" استخراج می‌گردد.



شکل 3-12 مراحل ریشه‌یابی کلمات غیرفعلی



شکل 3-13 مرحله تشخیص وندها

نکته‌ای در اینجا حائز اهمیت است این است که در کارهایی نظیر ترجمه از بین ریشه‌های یافت شده بایستی بالاترین سطح و در واقع بزرگترین ریشه استخراج و مورد استفاده قرار گیرد.

3 - 2 - 2 - برچسب زنی ادات سخن

مقالات متعددی در زمینه تولید برچسب ادات سخن وجـود دارد[AZIM08;MEG04;SHA08] امـا هیچ کدام به ابزاری که در اختیار عموم قرار داده شوند تبدیل نشده است. البته یک پلاگین برای ابـزار Gate توسط آزمایشگاه فناوری وب دانشگاه فردوسی مشهد ایجاد شده است که به دلیل سرعت پـایین این ابزار درصدد برآمدیم تا یک ابزار برچسب زنی آماری ساده جهت انجام کار ایجـاد نمـاییم. از اینـرو ابتدا با استفاده از پیکره برچسب خورده بی‌جن‌خان[ORO06] برچسب‌های بخش‌های سـخن کلمـات استخراج گردید. برای کلماتی که ابهام داشتند یک فرض ساده در نظر گرفته و بیشترین بسـامد مـورد نظر، به عنوان برچسب کلمه انتخاب گردید.

3 - 3 - خلاصه فصل

در این فصل روش‌های پیشنهادی برای بهبود ترجه ماشینی آماری انگلیسی به فارسی مطرح شد. ابتدا قوانین مربوط به بازآرایی نحوی برای کاهش فاصله ساختاری میان زبان انگلیسی و فارسی بیـان شـده است. این کاهش فاصله ساختاری می‌تواند موجب بهبود خروجی مترجم آماری گردد. در بخش بعـدی این فصل استفاده از فاکتورهای زبانی جهت بهبود مترجم ماشینی مطرح شده اسـت. بـرای اسـتخراج فاکتورهای زبانی بیان شده نیازمند یک ریشه‌یاب و برچسب‌زن ادات سخن هستیم. مراحل طراحی این ابزار نیز در این فصل توضیح داده شده است.

فصل چهارم:

پیاده سازی و ارزیابی

4 - پیاده سازی و ارزیابی

4 - 1 - مقدمه

در این فصل جزییات مربوط به پیاده‌سازی دو مدل پیشنهاد شده در فصل قبل و نتایج حاصل از پیاده‌سازی شرح داده شده است. ابتدا تاثیر بازآرایی نحوی در سمت زبان مبدا بر روی ترجمه آماری انگلیسی به فارسی بررسی شده و نتایج بیان گردیده است. سپس مدل ترجمه ماشینی مبتنی بر فاکتور برای ترجمه انگلیسی به فارسی ایجاد و نتایج مشخص شده است. در این بین توضیحاتی نیز درباره ابزار ایجاد شده مورد نیاز برای این مدل بیان شده است.

همانطور که قبلاً اشاره شده است برای ایجاد سیستم مترجم آماری نیاز به پیکره تک زبانه برای ایجاد مدل زبانی و یک پیکره دو زبانه برای ایجاد مدل ترجمه هستیم. به همین منظور برای پیکره تک زبانه از متون همشهری[ALE09] و برای پیکره دو زبانه از نسخه رایگان پیکره دانشگاه تهران موسوم به TEP استفاده نمودیم. پیکره TEP شامل 10000 جفت عبارت هم‌تراز بوده که به صورت خودکار بر اساس زیرنویس‌های انگلیسی فارسی فیلم های سینمایی تهیه شده است. اما پس از پیاده‌سازی قوانین بازآرایی و اعمال آن بر روی پیکره مشاهده نمودیم که نتایج ترجمه نسبت به قبل از اعمال این قوانین دچار افت شده است. نتایج این آزمایش در جدول 4-1 نشان داده شده است.

جدول 4-1 تاثیر بازآرایی نحوی بر روی پیکره‌ای شامل جملات زیرنویس فیلم

مجموعه تست	BLEU قبل از بازآرایی	BLEU بعد از بازآرایی
تست 1	16.73	14.28
تست 2	10.57	10.5

البته پیش از اعمال این قوانین این نکته مشهود به نظر می‌رسید. دلیل آن هم این است که پیکره TEP اکثراً شامل جملات مربوط به زیر نویس فیلم است. از آنجایی که زیرنویس فیلم نزدیک به زبان گفتار است و در زبان گفتار فارسی اکثراً از قواعد نحوی فارسی پیروی نمی‌شود، این پیکره

نمی‌توانست برای آزمایش روش پیشنهادی این پایان‌نامه مورد استفاده قرار گیرد. از اینرو و به خاطر عدم وجود پیکره دو زبانه مناسب، درصدد برآمدیم تا ابتدا یک پیکره دو زبانه ایجاد نماییم. مراحل ساخت پیکره در ادامه توضیح داده شده است.

4-2- تولید پیکره آموزشیFEP

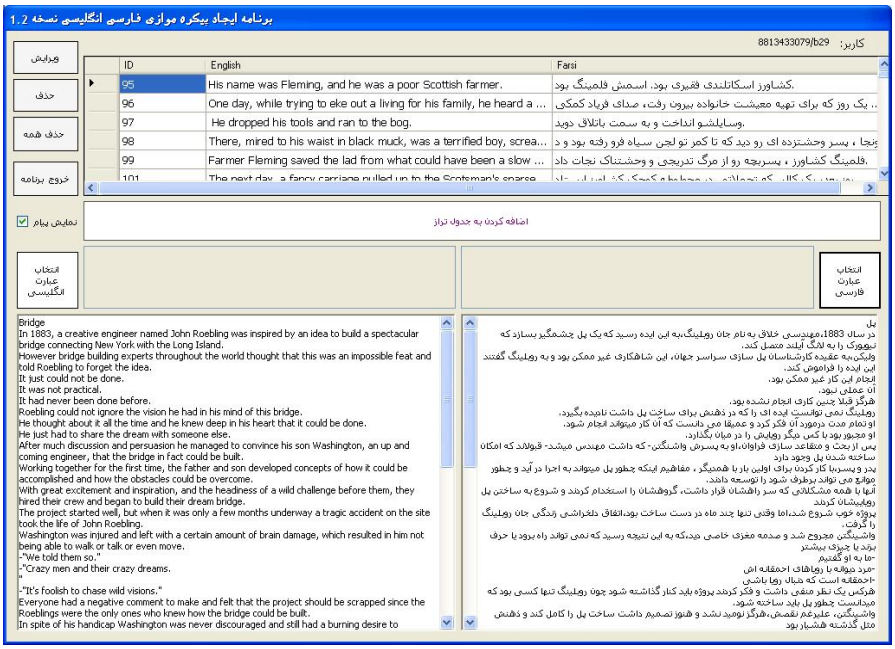
برای ایجاد پیکره ابتدا نیازمند منابع ترجمه شده بودیم. به همین منظور از منابع مختلف در جهت ایجاد این پیکره بهره جستیم. جدول 4-2 دامنه متون داخل پیکره و درصد آن‌ها در داخل پیکره را نشان می‌دهد.

جدول 4-2 توزیع مطالب جمع آوری شده جهت استفاده در پیکره دوزبانه

منبع	درصد
داستان کوتاه	8.5
شعر	1.5
طنز	0.2
مذهبی و کتب دینی	8
تغذیه	1
علوم تربیتی	2.5
آشپزی	2.5
اخبار علمی و اطلاعات عمومی	1.5
کامپیوتر	15
اخبار و بیانیه‌های رهبری و سازمان ملل	8
پزشکی	2.5
مطالب متفرقه	0.5
مقالات علمی	40
جملات عمومی زبان	10

پیکره FEP پیکره‌ای است که در طول ایـن پایـان‌نامـه ایجـاد و گسـترش یافـت. عملیـات مربـوط بـه ترازبندی عبارات با همکاری دانشجویان آزمایشگاه فناوری وب دانشگاه فردوسـی صـورت پـذیرفت. بـه این ترتیب که ابتدا نرم افزار مناسب برای انجام این عملیات با واسط کاربری مناسـب تولیدگردیـد. در این سامانه برای هر ترازکننده انسانی یک حساب کاربری و به هر حساب کاربری تعدادی مـتن جهـت ترازبندی نسبت داده شد.

در شکل 4-1 نمایی از این نرم افزار، قابل مشاهده است. نتایج جفت عبارات هر نرم افزار در یک فایـل Access 2007 ذخیره شده و سپس در یک سیستم الحاق‌گر دیگر این فایلها بـه مخـزن اصـلی پیکـره اضافه می شوند.



شکل 4-1 نمایی از صفحه اصلی نرم افزار ایجاد پیکره

در نهایت با تولید پیکره آموزشی FEP به روش انسانی داده‌های مورد نیـاز بـرای ارزیـابی سیسـتم در اختیار قرار گرفت. در جدول 4-3 حجم جملات موجود در پیکره ایجاد شده نشان داده شده است.

جدول 4-3 اندازه پیکره دوزبانه FEP

	فارسی	انگلیسی
جملات	11314	11539
کلمات	154958	145527

4 - 3 - حذف حروف و علائم زائد

به این دلیل که منابع تولید پیکره از بخشهای مختلف گردآوری شده بودند لازم بود پیش از بکارگیری پیکره آموزشی یک سری عملیات پیش پردازشی بر روی داده های تراز شده صورت بگیرد. به عنوان نمونه وجود اعراب بر روی برخی کلمات یا تفاوت در استفاده از برخی حروف فارسی عربی در مراحل بعدی می‌تواند سیستم را با خطاهایی روبرو کند. در نهایت این حروف می‌بایست حذف و یا با معادل فارسی خود جایگزین می‌شدند. علاوه بر این‌ها برخی علائم وجود دارند که می‌بایست از متن پاکسازی شوند. در جدول 4-4 تعدادی از این جایگزینی‌ها نشان داده شده است.

جدول 4-4 جایگزینی‌های انجام شده بر روی پیکره

بعد از اصلاح	قبل از اصلاح
ی	ي
ک	ك
ک	ك
ی	یٔ
ا	أ
ا	إ
و	ؤ
ه	ۀ
وو	ئو
وو	ؤو
یی	ئی
یی	ؤی
ه	یی
فاصله	نیم فاصله

4 - 4 - ایجاد سیستم مترجم آماری

برای ایجاد سیستم مترجم آماری نیاز به ابزاری برای ایجاد مدل‌های زبانی و ترجمه داریم. ابزار مورد استفاده در این پایان‌نامه در ادامه مختصراً توضیح داده شده است.

SRILM[STO02] یک ابزارایجاد و بکارگیری مدل‌های زبانی آماری است. این مدل‌ها در کاربردهایی نظیر تشخیص گفتار[1]، برچسب‌زنی و قطعه‌بندی آماری[2] و ترجمه ماشینی مورد استفاده قرار می‌گیرند. این ابزار در آزمایشگاه تکنولوژی و تحقیقاتی گفتار SRI توسعه یافته است. شروع کار این نرم افزار از سال 1995 بوده است و از آن زمان تا کنون نسخه‌های متعددی از این ابزار توسعه یافته است. نسخه مورد استفاده ابزار SRILM در این پایان‌نامه مبتنی بر لینوکس بوده و برای نصب آن نیازمند کامپایل و نصب بسته‌های جانبی می‌باشد. نحوه ارتباط با این ابزار نیز از طریق ترمینال لینوکس می‌باشد.

GIZA++[OCH03] نیز ابزاری جهت ایجاد مدل ترجمه است. این ابزار توسعه‌ای بر برنامه GIZA (قسمتی از ابزار ترجمه ماشینی آماری EGYPT) بوده است که توسط تیم ترجمه ماشینی آماری از سال 1999 تا کنون توسعه یافته است. این ابزار تحت لینوکس بوده و جهت نصب آن نیازمنده بسته‌های نرم‌افزاری زیادی هستیم. برقراری ارتباط و کار با این نرم افزار نیز از طریق ترمینال لینوکس است.

Moses[KOE07] یک موتور ترجمه ماشینی آماری رایگان است که امکان ایجاد مدل‌های لازم برای ترجمه را با استفاده از مجموعه‌ای از پیکره‌های موازی و تک زبانه متنی فراهم می‌کند. این ابزار نیز تحت لینوکس بوده و برای نصب آن از طریق ترمینال لینوکس اقدام می‌گردد. پیش از نصب این ابزار نیاز به نصب ابزارهای SRLM و GIZA++ بر روی سیستم و معرفی آن‌ها در هنگام نصب به این ابزار هستیم. اجرای نهایی عمل ترجمه با استفاده از این ابزار صورت می‌پذیرد. کدهای تمامی ابزار بیان

---

[1] speech recognition
[2] statistical tagging and segmentation

شده مبتنی بر زبان‌های Perl و C++ هستند.

در این پایان‌نامه با استفاده از ابزار بیان شده ابتدا یک سیستم مترجم آماری انگلیسی به فارسی پایه به منظور مقایسه نتایج حاصل از پیشنهادات مطرح شده ایجاد شده است. برای این منظور، ترجمه ماشینی مبتنی بر عبارت[1] را به کار گرفته و از ابزار متن‌باز Moses [KOE07] برای ایجاد این سیستم استفاده نمودیم. برای ایجاد مدل زبانی و N-gramها در این سیستم، از SRILM استفاده شده است[STO02]. همچنین برای ایجاد مدل ترجمه و جدول عبارات نیز از Giza++ [OCH03] استفاده گردید. قبل از شروع به ایجاد مدل‌ها و سیستم ترجمه ابتدا جملات داخل پیکره توکن شدند و تمامی حروف در پیکره انگلیسی به حروف کوچک تبدیل گشتند. در انتها با استفاده ابزار بیان شده مدل‌های زبانی و ترجمه و در نهایت سیستم مترجم مبتنی بر عبارت از روی پیکره‌های موازی و تک زبانه ایجاد گردید.

4 - 5 - مترجم آماری انگلیسی به فارسی بر مبنای بازآرایی نحوی

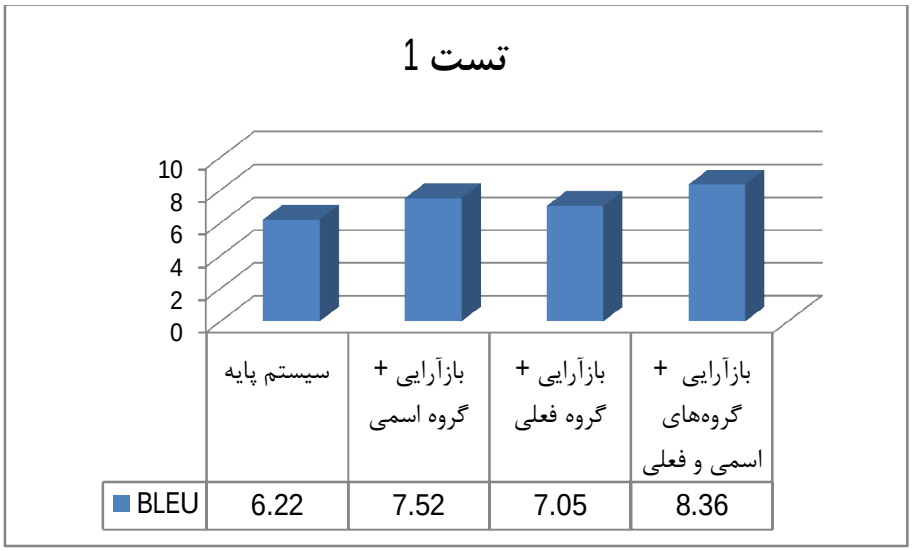
پس از ایجاد پیکره و در مرحله بعد، کلیه عبارات انگلیسی پیکره به وسیله پارسر استنفورد[MAR08] تجزیه گردید. پس از آنکه برچسب های مورد نظر برای عبارت انگلیسی بدست آمد به کمک قوانین بازآرایی، ترتیب واژگان بر اساس ساختار دستوری جملات در زبان فارسی مرتب شد. به عبارت دیگر با این کار فاصله موجود میان ساختار دو زبان کاهش یافت. سپس با استفاده از پیکره انگلیسی جدید ایجاد شده دوباره مدل ترجمه ایجاد گردید.

برای آزمایش سیستم ترجمه ماشینی، از دو مجموعه تست که هر کدام شامل 500 جمله می‌باشد بهره گرفته شد. این دو مجموعه تست از دامنه‌های کاملاً متفاوت انتخاب شده است که دامنه مجموعه تست 1 فاصله بیشتری با جملات درون پیکره دو زبانه دارد. خروجی سیستم بر روی مجموعه تست 1 با استفاده از معیارهای BLEU و NIST در جدول4-5، شکل4-2 و شکل4-3 نشان داده شده است.

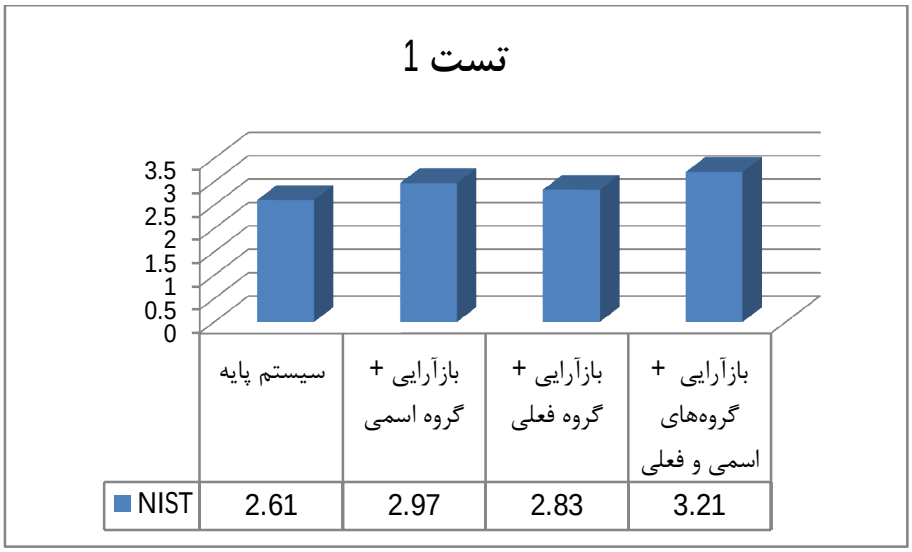
[1] Phrase-based Statistical Machine Translation

جدول4-5 نتایج حاصل از اعمال تغییرات بازآرایی بر روی مجموعه تست 1

	انگلیسی ← فارسی	BLEU	NIST
تست 1	سیستم پایه	6.22	2.61
	+بازآرایی گروه اسمی	7.52	2.97
	+بازآرایی گروه فعلی	7.05	2.83
	+ بازآرایی گروه‌های اسمی و فعلی	8.36	3.21



شکل4-2 خروجی BLEU حاصل از بازآرایی نحوی مجموعه تست 1

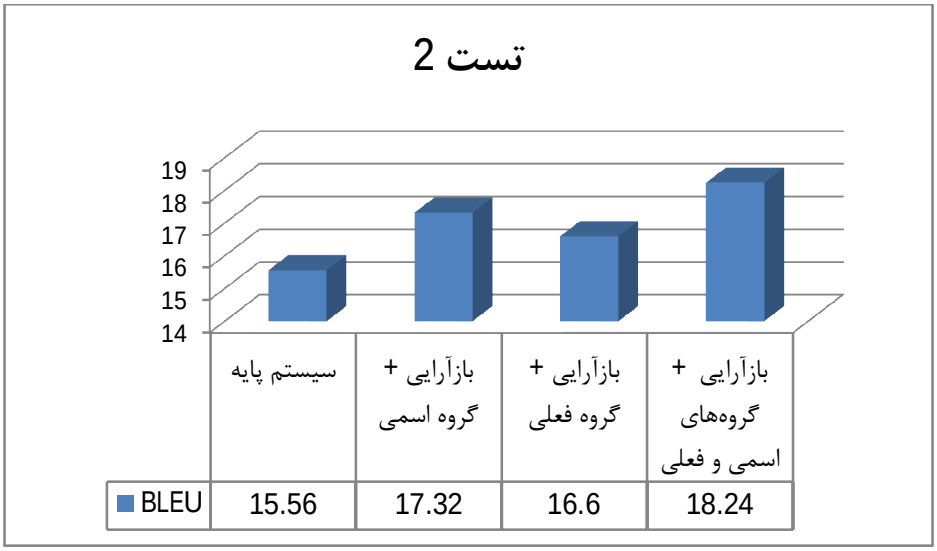


شکل4-3 خروجی NIST حاصل از بازآرایی نحوی مجموعه تست 1

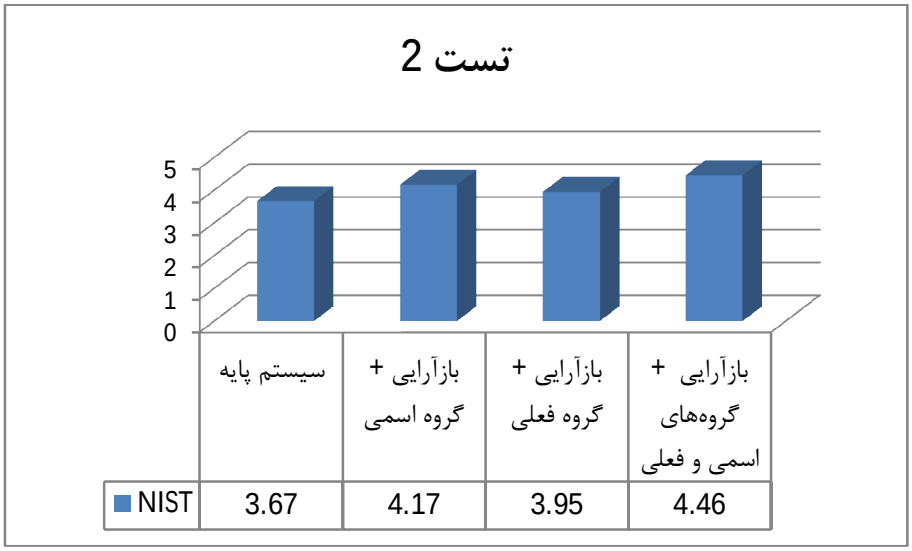
همچنین خروجی بازآرایی نحوی بر روی مجموعه تست 2 نیز در قالب جدول4-6، شکل4-4 و شکل4-5 نشان داده شده است.

جدول4-6 نتایج حاصل از اعمال تغییرات بازآرایی بر روی مجموعه تست 2

	انگلیسی ← فارسی	BLEU	NIST
تست 2	سیستم پایه	15.56	3.67
	+بازآرایی گروه اسمی	17.32	4.17
	+بازآرایی گروه فعلی	16.6	3.95
	+ بازآرایی گروه‌های اسمی و فعلی	18.24	4.46



شکل4-4 خروجی BLEU حاصل از بازآرایی نحوی مجموعه تست 2



شکل4-5 خروجی BLEU حاصل از بازآرایی نحوی مجموعه تست 2

4 - 6 - ایجاد ترجمه ماشینی آماری مبتنی بر فاکتور

ابتدا براساس توضیحاتی که در فصل قبل بیان گردید، ابزار مورد نیاز در سمت فارسی برای ساخت این مدل ایجاد گردید.

در الگوریتم پیاده‌سازی شده برای این ابزار، از هر دو رویکرد استفاده از فرهنگ لغت و قوانین نحوی استفاده شده است. برای اجرای مراحل الگوریتم از پنج فرهنگ لغت استفاده شده است.

1. فرهنگ لغت برای نگه داشتن تمامی ریشه‌های مربوط به کلمات غیر فعلی (شامل اسامی، صفتها، قیدها و ...).
2. فرهنگ لغت برای نگهداری کلمات جمع مکسر و حالتهای جمع بدون قاعده به همراه حالت مفرد این کلمات
3. فرهنگ لغت شامل بن‌های گذشته و حال افعال زبان فارسی
4. فرهنگ شامل تمامی وند(پیشوند و پسوند)های زبان فارسی
5. فرهنگ لغت برای کلمات استثناء

برای ایجاد فرهنگ لغت مربوط به بن‌های گذشته و حال حدود 6000 فعل زبـان فارسـی از مجموعـه دادگان زبان فارسی[RAS11]، مورد بررسی قرار گرفت و در نهایت حـدود 400 بـن گذشـته و حـال افعال استخراج گردید. به عنوان مثال دو فعل "خواب دیدن" و "دیدن" یک فعل در نظر گرفتـه شـده است. سپس بن‌های گذشته و حال آن به شکل "دید" و "بین" در فرهنگ لغت ذخیره می‌گردد. دلیل این عمل این است که تنها بن فعل در جمله تغییر حالت می‌دهد و سایر قسـمت‌هـای فعـل[1] بـه ازای زمان و شخص‌های مختلف، تغییر شکل نمی‌دهند.

در الگوریتم ریشه‌یابی، ابتدا بایستی نوع عبارت (فعل یا غیرفعل)، تشخیص داده شود. برای انجـام ایـن عمل از دو روش استفاده شده است. در روش اول ابتدا از روی افعال استخراج شده از مجموعه دادگـان تمامی تصریف‌های مختلف افعال درون یک فرهنگ لغت ایجاد می‌شوند و سپس با جستجوی کلمه در این مجموعه نوع عبارت تشخیص داده می‌شود.

در روش دوم شناسایی صحیح انواع فعل پیشوندی، ساده و مرکـب در زمـان‌هـای مختلـف در مـتن بـا بهره‌گیری از قواعد نحوی گرامر زبان فارسی، موقعیت لغات(بافت جمله) و تحلیل ریخت‌شناسی لغـات (مطالعه ساختار لغات) مورد بررسی قرار می‌گیرد. برای شناسایی افعال موجود در متن، احتمال ظهـور فعل در موقعیت‌هایی خاص نظیر لغات قبل از جداکننده‌های جمـلات از قبیـل علامـت‌هـای "."، "؛"، "!"، "؟"، "?"، ":" و "خط جدید"، بررسی می‌گردد. همچنین گاهی حروف ربط می‌توانند بـه عنـوان آغاز کننده جملات باشند مثل "که"، "اساسا"، "البته"، "تا"، "اما"، "اگر"، "ولی"، "زیرا"، "سـپس"، "همچنین"، "و"و "یا" که فعل می‌تواند قبل از آنها بیاید. احتمال فعل بـودن لغـات در موقعیـت‌هـای ذکر شده با بررسی مطابقت با انواع ساختار افعال فارسی توسط عبـارات مـنظم بررسـی مـی‌گـردد. در صورتی که لغات مورد ارزیابی با هر کدام از ساختارهای فعل توسط عبارات منظم، تطبیق پیـدا کننـد،

[1] مانند فعلیار و حرف اضافه فعلی

گروه فعلی را تشکیل خواهند داد. سایر عباراتی که در این قالب نمی‌گنجند نیز بصورت غیرفعل شناسایی می‌گردند.

در روش پیشنهادی ترکیبی از دو روش فوق استفاده شده است. به این صورت که ابتدا با استفاده از روش اول عبارات کاندید برای فعل یافت شده و سپس با توجه به قوانین مطرح شده براساس جایگاه افعال در جمله، برچسب فعل عبارت مورد نظر تایید یا رد می‌گردد. دلیل ترکیب این دو روش، وجود برخی از کلمات با ساختار فعل هستند که می‌توانند نقش فعل یا غیرفعل داشته باشند. مثال‌هایی از این افعال در جدول 4-7 نشان داده شده است.

جدول 4-7 مثال‌هایی از کلماتی که می‌توانند هم حالت فعلی و هم حالت غیرفعلی داشته باشند

عبارت	حالت فعلی	حالت غیرفعلی
متن	فعل امر دوم شخص مفرد از بن تنیدن	مثلاً متن نامه
بازی	فعل مضارع اخباری دوم شخص مفرد	مثلاً بازی بچه‌ها
دوم	فعل مضارع اخباری اول شخص مفرد	مثلاً جایگاه دوم مسابقات

در مرحله بعد برای هر کدام از عبارات فعل و غیر فعل بصورت جداگانه و براساس قوانین خاص خود عمل می‌شود.

مطابق با الگوریتمی که برای هر عبارت فعل و غیرفعل در قسمت روش پیشنهادی بیان شد، ابزار مربوط به عمل ریشه‌یابی فارسی ایجاد گردید. نمایی از ابزار ریشه‌یابی ایجاد شده در شکل 4-6 نشان داده شده است.

شکل 4-6 نمایی از ریشه‌یاب چند سطحی ایجاد شده

البته این ریشه‌یاب چند سطح ریشه را استخراج می‌نماید که برای عملیات ترجمه نیازمند بالاترین سطح آن هستیم. دقت خروجی ابزار برای ریشه‌یابی در بالاترین سطح بر روی سه داده تست از مجموعه همشهری[ALE09][1] در جدول ۴-۸ نشان داده شده است.

جدول 4-8 دقت ریشه‌یاب ایجاد شده در ایجاد ریشه در بالاترین سطح

مجموعه تست	مجموع کلمات	تعداد کلمات نیازمند ریشه‌یابی	تعداد ریشه‌یابی صحیح	دقت
تست 1	723	102	89	0.87
تست2	1463	253	231	0.91
تست 3	3354	441	407	0.92

[1] پیکره استاندارد همشهری به منظور پردازش زبان‌های طبیعی و بازیابی اطلاعات در سال 2009 در دانشگاه تهران تهیه شد.

برای ارزیابی پایین‌ترین سطح ریشه‌یابی در ریشه‌یاب ارائه شده، از پیکره همشهری ویرایش دوم استفاده شده است[ALE09][1] . این پیکره شامل 318,000 متن از روزنامه همشهری بین سال‌های 1996 تا 2007 به همراه برچسب موضوع هر متن می‌باشد. در این آزمایش بر روی قسمتی پیکره همشهری عمل خوشه‌بندی بدون استفاده از ریشه‌یابی و به همراه آن صورت گرفته است. نتایج در جدول 4-9 به همراه مقایسه ریشه‌یاب ایجاد شده با ریشه‌یاب PerStem[JAD10][2] نشان داده شده است. دلیل مقایسه با این ریشه‌یاب این بود که فقط خروجی این ریشه‌یاب برای عملیات مقایسه در دسترس بود[3].

جدول 4-9 تاثیر ریشه یابی با استفاده از روش پیشنهادی در انجام عمل خوشه‌بندی

نام مجموعه داده	HAM2(96/06/22)			HAM2(07/01/01-07/01/14)		
تعداد اسناد	79			1065		
تعداد کلاس	7			8		
ریشه‌یابی	بدون ریشه‌یابی	PerStem	روش پیشنهادی	بدون ریشه‌یابی	PerStem	روش پیشنهادی
تعداد کلمات	5191	4202	2598	22184	19275	5852
دقت	%48.10	%48.11	%49.40	%33.90	%33.15	%36.81

در ادامه با استفاده از ابزارهای ایجاد شده بالاترین سطح ریشه و برچسب بخش‌های سخن عبارات موجود در پیکره استخراج گردید. در پیکره انگلیسی نیز برای استخراج ریشه عبارات از ابزار morph adorner[4] مربوط به دانشگاه نورس‌وسترن[5] و برای استخراج برچسب‌های بخش‌های سخن از ابزار Senna[TOU03] استفاده شده است. این ابزار جزو بادقت‌ترین ابزار موجود برای زبان انگلیسی هستند.

[1] http://ece.ut.ac.ir/dbrg/hamshahri/
[2] http://sourceforge.net/projects/perstem/
[3] http://sourceforge.net/projects/perstem/
[4] http://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf
[5] Northwestern

نمونه‌ای از اطلاعات داخل پیکره بعد از غنی‌سازی توسط فاکتورهای بیان شده در شکل 4-7 نشان داده شده است.

فردی|فرد|N_SING شام|شام|N_SING منزل|منزل|N_SING دوستان|دوست|N_PL قدیمی‌اش|قدیمی|ADJ_SIM دعوت شد|دعوت شد|V_PA .

Someone|Someone|NN was|be|VBD invited|invite|VBN to|to|TO the|the|DT home|home|NN of|of|IN some|som|DT old|old|JJ friends|friend|NNS for|for|IN dinner|dinner|NN .

اینها|این|PRO آهنگهای|آهنگ|N_PL خوبی|خوب|ADJ_SIM هستند|است|V_PRS .

These|this|DT are|be|VBP good|good|JJ songs|song|NNS .

شکل 4-7 نمونه‌ای از جملات غنی شده داخل پیکره

در مدل مبتنی بر عبارت هدف ترجمه تکه‌های متن مبدا به تکه‌های متن مقصد است. در واقع احتمالات برای ترجمه عبارت بصورت مستقیم از روی پیکره با استفاده از تکه‌های پیکره استخراج می‌گردد. اما در مدل مبتنی بر فاکتور مورد استفاده مراحل ترجمه به مراحل زیر شکسته می‌شود.

1- ترجمه ریشه ورودی به ریشه خروجی
2- ترجمه برچسب‌های بخش‌های سخن
3- تولید ترجمه عبارت خروجی با استفاده از ترجمه‌های دو مرحله قبل

برای درک بهتر مطلب، این مراحل با ذکر یک مثال توضیح داده می‌شود. فرض کنید به دنبال ترجمه کلمه houses در سمت فارسی می‌گردیم. برای این کلمه فاکتورهای زیر در پیکره وجود است.

Houses(خود کلمه)|house(ریشه کلمه)|NNS(برچسب بخش‌های سخن کلمه)

ابتدا ریشه کلمه ترجمه می‌گردد. در واقع نگاشت ریشه صورت می‌گیرد.

House → خانه, ساختمان, منزل, پله

سپس برچسب بخش‌های سخن نگاشت می‌یابد.

NNS → N_PL, N_SING

و در انتها ترجمه مورد نظر تولید می‌گردد.

خانه‌ها → N_PL | خانه

خانه → N_SING | خانه

ساختمان‌ها → N_PL | ساختمان

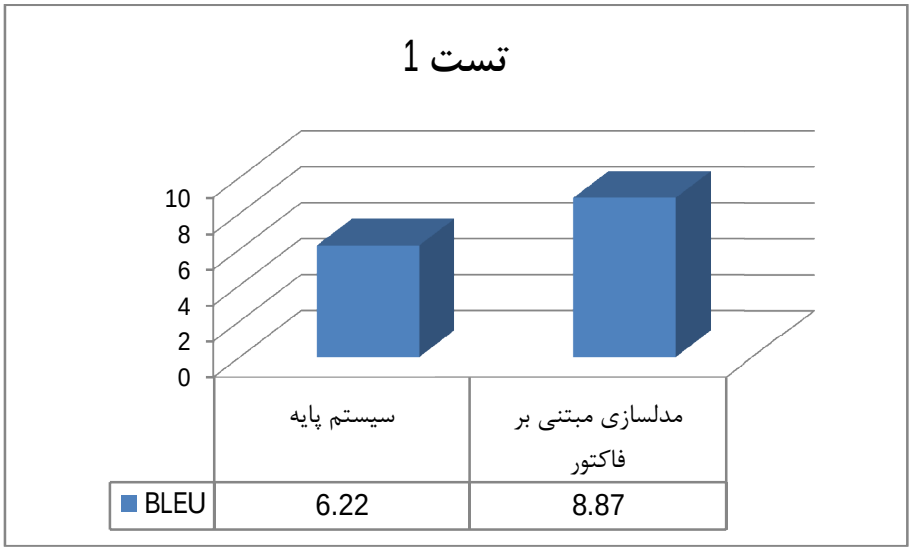
...

همانطور که در این مثال دیده می‌شود حتی اگر حالتی از یک کلمه در داخل پیکره نباشد باز هم می‌توان بر اساس حالت‌های دیگر کلمه ترجمه انجام داد اما در مدل مبتنی بر عبارت اگر کلمه‌ای داخل پیکره نباشد مدل ترجمه ایجاد شده قادر به ترجمه آن کلمه نخواهد بود.

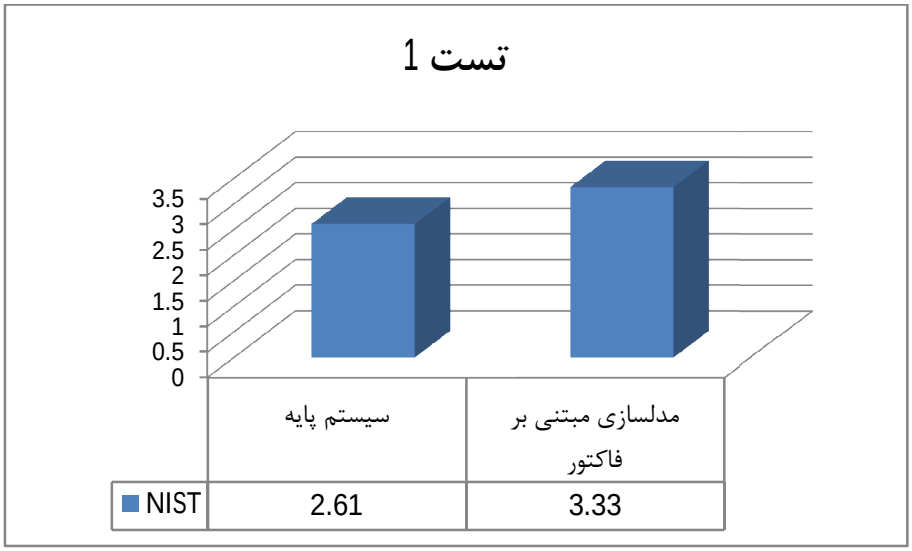
برای ایجاد مدل زبانی از SRILM [STO02]، برای ایجاد مدل ترجمه از GIZA++ [OCH03] و با استفاده از ابزار Moses به ایجاد سیستم پرداختیم. خروجی‌های بدست آمده و مقایسه آن‌ها با مدل ترجمه مبتنی بر عبارت بر روی مجموعه تست 1، در قالب جدول 4-10، شکل 4-8 و شکل 4-9 نشان داده شده است.

جدول 4-10 نتایج حاصل از اعمال فاکتورهای اضافی بر روی مجموعه تست 1

انگلیسی ← فارسی		BLEU	NIST
تست 1	سیستم پایه	6.22	2.61
	مدلسازی مبتنی بر فاکتور	8.87	3.33



شکل 4-8 خروجی BLEU حاصل از اعمال فاکتورهای اضافی بر روی مجموعه تست 1

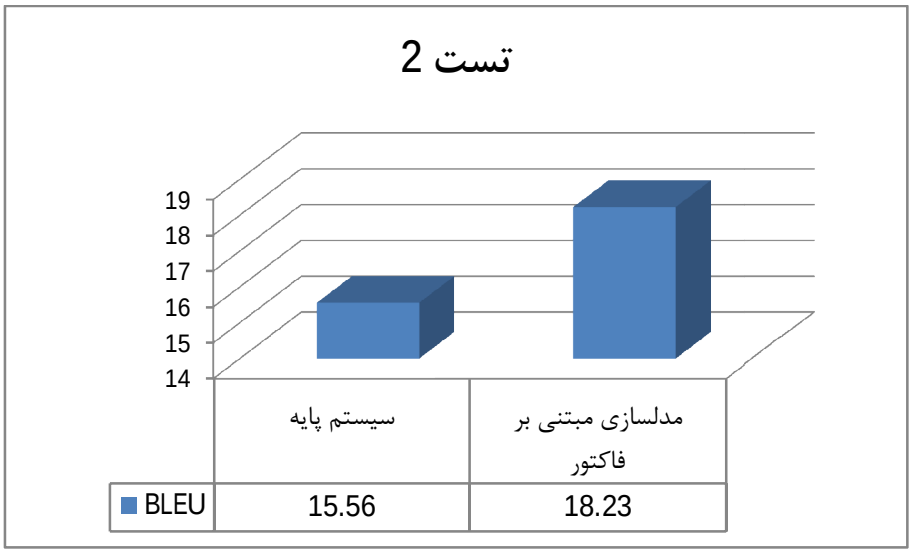


شکل 4-9 خروجی NIST حاصل از اعمال فاکتورهای اضافی بر روی مجموعه تست 1

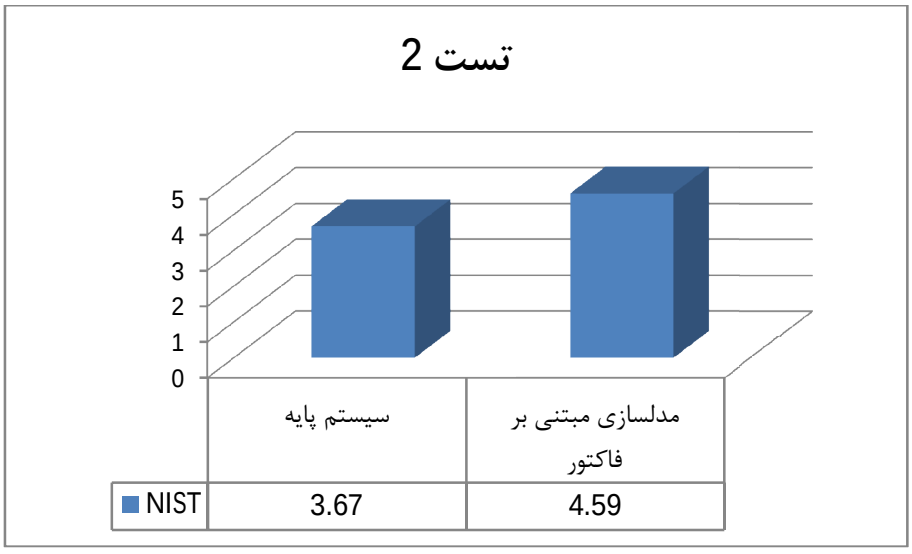
همچنین مقایسه نتایج مربوط به مجموعه تست 2 نیز در جـدول 4-11، شـکل 4-10 و شـکل 4-11 نشان داده شده است.

جدول 4-11 نتایج حاصل از اعمال فاکتورهای اضافی بر روی مجموعه تست 2

انگلیسی ← فارسی		BLEU	NIST
تست 2	سیستم پایه	15.56	3.67
	مدلسازی مبتنی بر فاکتور	18.23	4.59



شکل 4-10 خروجی BLEU حاصل از اعمال فاکتورهای اضافی بر روی مجموعه تست 2



شکل 4-11 خروجی NIST حاصل از اعمال فاکتورهای اضافی بر روی مجموعه تست 2

4 - 7 - خلاصه فصل

در این فصل مراحل ایجاد سیستم مترجم آماری انگلیسی به فارسی توضیح داده شد. ابتدا درباره ایجاد پیکره موازی توضیحاتی مطرح گردید. در ادامه روال ساخت سیستم مترجم آماری انگلیسی به فارسی مبتنی برعبارت به عنوان سیستم پایه بیان شده است. تمامی مقایسات با این سیستم پایه صورت گرفته است. در ادامه فصل پیاده‌سازی روش‌های پیشنهادی مطرح شده در فصل قبل بیان شده است. نتایج پیاده‌سازی هر روش بر روی دو مجموعه تست با نتایج سیستم مترجم آماری انگلیسی فارسی مبتنی بر عبارت مقایسه گردیده است. نتایج نشان دهنده بهبود سیستم به ازای روش‌های مطرح شده در هر دو معیار BLEU و NIST می‌باشد.

فصل پنجم:

نتیجه گیری و

کارهای آتی

5- نتیجه‌گیری و کارهای آتی

5-1- نتیجه‌گیری

در این پایان‌نامه ابتدا یک سیستم مترجم آماری انگلیسی به فارسی مبتنی بر عبارت ارائه شده است. از آنجایی که یکی از معضلات موجود در ساخت چنین سیستمی نبود پیکره بزرگ و مناسب است، دو روش که در آن‌ها استفاده صریح از اطلاعات زبان‌شناسی مطرح شده است، جهت بهبود خروجی مترجم ماشینی آماری انگلیسی به فارسی پیشنهاد شده است. در روش اول به کاهش فاصله ساختاری بین زبان انگلیسی و فارسی پرداخته شده است. این عمل باعث ترازبندی بهتر عبارات معادل در دو زبان و در نتیجه ایجاد مدل ترجمه بهتر می‌گردد. خروجی سیستم بر روی پیکره‌ای شامل حدود یازده هزار جمله تراز شده تحت دو معیار BLEU و NIST و برای دو مجموعه تست متفاوت استخراج گردیده است. مقایسه نتایج نشان دهنده بهبود روش پیشنهادی در نتایج نهایی ترجمه می‌باشد.

در ادامه به بررسی تاثیر فاکتورهای زبانی در مراحل ترجمه پرداخته شده است. برای این منظور اثر اضافه نمودن فاکتورهای ریشه کلمات و برچسب‌های بخش‌های سخن درون پیکره موازی بررسی شده است. در این مدل ترجمه مبتنی بر فاکتور پیاده‌سازی گردیده است. در این روش ریشه کلمات و برچسب‌های بخش‌های سخن به کمک ترازبندی آمده و منجر به بهبود مدل ترجمه می‌گردند. همچنین با استفاده از ریشه کلمات موجود در پیکره تا حدی مشکل ترجمه برخی کلمات خارج از پیکره نیز مرتفع می‌گردد. خروجی حاصل از اعمال این روش باز هم منجر به بهبود معیارهای BLEU و NIST نسبت به سیستم پایه گردیده است.

5-2- کارهای آتی

یکی از مسائل مهم در زمینه روش‌های بهبود سیستم‌های مترجم ماشینی آماری، بررسی میزان بهبود

سیستم به ازای اندازه‌های مختلف پیکره است. از آنجایی که حجم پیکره در این پایان‌نامه به قدری نبود که بتوانیم این آزمایشات را انجام دهیم می‌توان به عنوان کارهای آتی به ایجاد یک پیکره نسبتاً بزرگ پرداخت و اثر روش‌های پیشنهاد شده را بر روی اندازه‌های متفاوت آزمایش نمود. البته بدیهی است که اگر اندازه پیکره خیلی بزرگ باشد میزان اثر این روش‌ها کاهش خواهد یافت.

یکی دیگر از نکاتی که مهم به نظر می‌رسد بررسی سایر فاکتورهای زبانی از جمله جنسیت[1]، تعداد[2]، نقش‌های معنایی کلمات و سایر تحلیل‌های صرفی موجود است. همچنین با بالا بردن دقت ابزار تولید شده در این پایان‌نامه می‌توان در جهت بهبود نتایج نهایی در ترجمه گام برداشت.

نکته دیگری که می‌توان روی آن تمرکز نمود ارائه ابزار تجزیه‌گر برای زبان فارسی است. با انجام این کار می‌توان عمل بازآرایی را برای زبان فارسی انجام داد و سیستم مترجم آماری فارسی به انگلیسی را بهبود بخشید. از آنجایی که زبان انگلیسی از لحاظ ترتیب اجزاء قانونمندتر از زبان فارسی است این عمل می‌تواند بهبود مناسبی در نتایج نهایی ترجمه ایجاد نماید.

[1] gender
[2] count

مراجع

6- مراجع

[ALE09] AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems, 22*(5), 382-387.

[AMT00] Amtrup, J. W., Rad, H. M., Megerdoomian, K., & Zajac, R. (2000). Persian-English machine translation: An overview of the Shiraz project. *Memoranda in Computer and Cognitive Science MCCS-00-319, NMSU, CRL*.

[AZIM08] Azimizadeh, A., Arab, M. M., & Quchani, S. R. (2008). *Persian part of speech tagger based on Hidden Markov Model.* Paper presented at the 9th JADT.

[BRO90] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., . . . Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics, 16*(2), 79-85.

[BRO93] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics, 19*(2), 263-311.

[CHI07] Chiswell, I., & Hodges, W. (2007). *Mathematical logic*: Oxford University Press, USA.

[DAV01] Dave, S., Parikh, J., & Bhattacharyya, P. (2001). Interlingua-based English–Hindi Machine Translation and Language Divergence. *Machine translation, 16*(4), 251-304.

[ELK10] El-Kahlout, I. D., & Oflazer, K. (2010). Exploiting Morphology and Local Word Reordering in English-to-Turkish Phrase-Based Statistical Machine Translation. *Audio, Speech, and Language Processing, IEEE Transactions on, 18*(6), 1313-1322.

[ELH09] Elhadj, Y. O. M. (2009). Statistical Part-of-Speech Tagger for Traditional Arabic Texts. *Journal of Computer Science, 5*(11), 794-800.

[EST11] Estahbanati, S., Javidan, R., & Nikkhah, M. (2011). A New Multi-Phase Algorithm for Stemming in Farsi Language Based on Morphology. *International Journal of Computer Theory and Engineering, 3*.

[GAL93] Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics, 19*(1), 75-102.

[GOL05] Goldwater, S., & McClosky, D. (2005). *Improving statistical MT through morphological analysis*. Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada.

[HAR97] Haruno, M., & Yamazaki, T. (1997). High-performance bilingual text alignment using statistical and dictionary information. *Natural Language Engineering, 3*(01), 1-14.

[HUT92] Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation*: Academic Press New York.

[JAD10] Jadidinejad, A., Mahmoudi, F., & Dehdari, J. (2010). Evaluation of perstem: a simple and efficient stemming algorithm for Persian. *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 98-101.

[JUR00] Jurafsky, D., Martin, J. H., Kehler, A., Vander Linden, K., & Ward, N. (2000).

Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (Vol. 2): Prentice Hall New Jersey.

[KAY93] Kay, M., & Röscheisen, M. (1993). Text-translation alignment. *Computational linguistics, 19*(1), 121-142.

[KOE09] Koehn, P. (2009). *Statistical machine translation*: Cambridge University Press.

[KOE07] Koehn, P., & Hoang, H. (2007). *Factored translation models.* Paper presented at the EMNLP.

[KOE07] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . Herbst, E. (2007). *Moses: open source toolkit for statistical machine translation*. Paper presented at the Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic.

[KOE03] Koehn, P., Och, F. J., & Marcu, D. (2003). *Statistical phrase-based translation.* Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, Edmonton, Canada.

[LEE04] Lee, Y.-S. (2004). *Morphological analysis for statistical machine translation.* Paper presented at the Proceedings of HLT-NAACL 2004: Short Papers, Boston, Massachusetts.

[LOP08] Lopez, A. (2008). Statistical machine translation. *ACM Comput. Surv., 40*(3), 1-49.

[LOV68] Lovins, J. B., & LAB., M. I. O. T. C. E. S. (1968). *Development of a stemming algorithm*: MIT Information Processing Group, Electronic Systems Laboratory.

[MAR08] Marneffe, M.-C. d., & Manning, C. D. (2008). *The Stanford typed dependencies representation*. Paper presented at the Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, Manchester, United Kingdom.

[MAR10] Martin, A., & Greenberg, C. (2010). *The 2009 NIST language recognition evaluation*. Paper presented at the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic.

[MUC10] maučec, M. S., & Brest, J. (2010). Reduction of Morpho-Syntactic Features in Statistical Machine Translation of Highly Inflective Language. *Informatica, 21*(1), 95-116.

[MEG04] Megerdoomian, K. (2004). *Developing a Persian part of speech tagger*. Paper presented at the 1st Workshop on Persian Language and Computer.

[MIA06] Miangah, T. M. (2006). Automatic lemmatization of Persian words*. *Journal of Quantitative Linguistics, 13*(01), 1-15.

[MOH09] Mohaghegh, M., & Sarrafzadeh, A. (2009). *An analysis of the effect of training data variation in English-Persian statistical machine translation.* Paper presented at the Proceedings of the 6th international conference on Innovations in information technology, AI-Ain, United Arab Emirates.

[MOH11] Mohaghegh, M., & Sarrafzadeh, A. (2011, 25-27 April 2011). *An overview of the challenges and progress in PeEn-SMT: First large scale Persian-English SMT system.* Paper presented at the Innovations in Information Technology (IIT), 2011 International Conference on.

[MOH09] Mohaghegh M., & Sarrafzadeh A. (2009). *The first English-Persian statistical machine translation.* Paper presented at the New Zealand Postgraduate Conference.

[NIE00] Nießen, S., & Ney, H. (2000). *Improving SMT quality with morpho-syntactic analysis*. Paper presented at the Proceedings of the 18th conference on Computational linguistics - Volume 2, Saarbrücken, Germany.

[NIE04] Nießen, S., & Ney, H. (2004). Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Comput. Linguist., 30*(2), 181-204. doi: 10.1162/089120104323093285

[OCH03] Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist., 29*(1), 19-51. doi: 10.1162/089120103321337421

[ORO06] Oroumchian, F., Tasharofi, S., Amiri, H., Hojjat, H., & Raja, F. (2006). Creating a feasible corpus for Persian POS tagging. *Department of Electrical and Computer Engineering, University of Tehran.*

[PAP02] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a method for automatic evaluation of machine translation*. Paper presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania.

[PIL10] Pilevar M.T., & Faili, H. (2010). *PersianSMT: A First Attempt to English-Persian Statistical Machine Translation.* Paper presented at the 10th International Conference on statistical analysis of textual data.

[PIL11] Pilevar, M. T., Faili, H., & Pilevar, A. H. (2011). *TEP: Tehran English-Persian parallel corpus*. Paper presented at the Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II, Tokyo, Japan.

[POP06] Popovi, M., & Ney, H. (2006). *Statistical Machine Translation with a Small Amount of Bilingual Training Data*. Paper presented at the 5th LREC SALTMIL Workshop on Minority Languages.

[POR06] Porter, M. F. (2006). An algorithm for suffix stripping. *Program: electronic library and information systems, 40*(3), 211-218.

[RAM00] Ramanathan, A. (2000). Statistical Machine Translation (D. o. C. S. a. E. I. I. o. Technology, Trans.).

[RAS11] Rasooli, M. S., Moloodi, A., Kouhestani, M., & Minaei-Bidgoli, B. (2011). *A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank.* Paper presented at the 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics.

[SAE09] Saedi, C., & Shamsfard, M. (2009). *Translating Persian documents into English using knowledge based WSD*. Paper presented at the Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on.

[SAN06] Sani, G. G., & Hesami, R. (2006). *A Stemming Algorithm for Farsi Language*. Paper presented at the Proceeding of 11 International CSI Computer Conference (CSICC'2006).

[SHA08] Shamsfard, M., & Fadaee, H. (2008). *A hybrid morphology-based pos tagger for Persian.* Paper presented at the LREC.

[SOM99] Somers, H. (1999). Review article: Example-based machine translation. *Machine translation, 14*(2), 113-157.

[STO02] Stolcke, A. (2002). *SRILM-an extensible language modeling toolkit*. Paper presented at the Proceedings International Conference on Spoken Language Processing.

[TAG05] Taghva, K., Beckley, R., & Sadeh, M. (2005). *A stemming algorithm for the farsi language*. Paper presented at the International Conference on Information Technology: Coding and Computing (ITCC'05).

[TAS02] Tashakori, M., Meybodi, M., & Oroumchian, F. (2002). Bon: The persian stemmer. *EurAsia-ICT 2002: Information and Communication Technology*, 487-494.

[TOU03] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network*. Paper presented at the Association for Computational Linguistics.

[WAN07] Wang, C., Collins, M., & Koehn, P. (2007). *Chinese syntactic reordering for statistical machine translation*. Paper presented at the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague.

[WEA55] Weaver, W. (1955). Translation. *Machine translation of languages, 14*, 15-23.

[WIT91] Witten, I. H., & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *Information Theory, IEEE Transactions on, 37*(4), 1085-1094.

Abstract*:*

With the increasing volume of data and information as well as expanding international interactions, communication is one of the important aspects of today's life. A major problem That there is in here is the impossibility of communicating with the data in another language. Hence one of life's most important issues is to provide an automated solution for the translation from one language to another. Machine translation is one way that is presented to solve this problem and because of its importance, it has a lot of attention in recent years. Statistical machine translation is one of the best methods to translate from one language to another language. For languages that are very similar to each other, the output of this translator is very good. But for some pairs of languages such as English and Persian, structural differences between the two languages and also the lack of a large bilingual corpus is made that this technique does not produce the desired language translation of English to Farsi translations.

This thesis is trying to overcome the problems of this method to translate English into Persian partly, with the help of linguistic information. To do this, first try to reduce the structural difference between sentences in English and Persian. This can lead to a better translation model. For this purpose a series of rules extracted and on the English sentences were applied. These changes lead to 17% improvement in BLEU score and 21% improvement in NIST score. Then enrichment within the body is done using some linguistic information, including part of speech tags and the word stems. With this information a factor-based translation system was created. Output system 17 percent improvement in the BLEU score and 25% improvement in NIST score.

Keywords:

Machine translation,Statistical machine translation, Language model, Translation model, Syntactic reordering, Factor based model

Department of Computer Engineering

Ferdowsi University of Mashhad

M.Sc. Thesis

Improving English – Persian statistical machine translation with linguistics information

By:

Reza Saeedi

Supervisor:

Dr. Mohsen Kahani

Advisor:

Dr. Nader Jahangiri

July 2012