
بسم الله الرحمن الرحيم

دانشکده کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد

یک روش آماری برای ترکیب کلمات چندبخشی

آرزو ارجمندزاده

استاد راهنما:

دکتر مرتضی زاهدی

استاد مشاور:

مهندس مرضیه رحیمی

بهمن ۱۳۹۳

به مادرم، دریای بی کران فداکاری و عشق

و

به مهربان فرشتگانی که

لحظات ناب باور بودن، لذت و غرور دانستن، جسارت خواستن، عظمت رسیدن و تمام تجربه های یکتا و زیبای زندگیم را مدیون حضور سبز آنهاست

تقدیم به خانواده عزیزم

تشکر و قدردانی

نخستین سپاس و ستایش از آن خداوندی است که بنده کوچکش را در دریای بیکران اندیشه، قطره‌ای ساخت تا وسعت آن را از دریچه اندیشه‌های ناب استادانی بزرگ به تماشا نشیند. لذا اکنون که در سایه سار بنده نوازی هایش پایان نامه حاضر به انجام رسیده است، بر خود لازم می دانم تا مراتب سپاس را از بزرگوارانی به جا آورم که اگر دست یاریگرشان نبود، هرگز این پایان نامه به انجام نمی رسید.

از استاد فرزانه و فرهیخته جناب آقای دکتر مرتضی زاهدی که در راه کسب علم و معرفت مرا یاری نمودند و با دلگرمی ها و تشویق هایشان همواره راهنما و چراغ راه من بوده اند، سپاسگزار و متشکرم (هرچند تشکر کلمه گویایی در اینجا به نظر نمی رسد).

از اساتید خود جناب آقای دکتر علی اکبر پویان و جناب آقای دکتر حمید حسن پور که در دوره کارشناسی ارشد از محضر پر فیض تدریسشان، بهره ها برده ام، بسیار سپاسگزارم.

با سپاس از سرکار خانم مهندس مرضیه رحیمی به دلیل یاری ها و راهنمایی های بی چشمداشت ایشان که بسیاری از سختی ها را برایم آسانتر نمودند.

همچنین از خانواده عزیز، دلسوز و مهربانم که آرامش روحی و آسایش فکری را فراهم نمودند تا با حمایت های همه جانبه در محیطی مطلوب مراتب تحصیلی و نیز پایان نامه درسی را به نحو احسن به اتمام برسانم، سپاسگزاری می نمایم.

آرزو ارجمندزاده

بهمن ۱۳۹۳

تعهدنامه

اینجانب آرزو ارجمندزاده دانشجوی دوره‌ی کارشناسی ارشد رشته‌ی مهندسی کامپیوترگرایش هوش مصنوعی دانشکده‌ی مهندسی کامپیوتر و فناوری اطلاعات دانشگاه شاهرود نویسنده‌ی پایان‌نامه‌ی یک روش آماری برای ترکیب کلمات چندبخشی تحت راهنمایی دکتر مرتضی زاهدی متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام‌شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورداستفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه شاهرود» و یا « Shahrood University» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه‌ی مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه‌ی مراحل انجام این پایان‌نامه، در مواردی که به حوزه‌ی 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

- کلیه‌ی حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه شاهرود 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.

چکیده

زبان فارسی شامل کلماتی است که از چند بخش تشکیل شده‌اند و رسم‌الخط استاندارد این کلمات چندبخشی به این‌صورت هستند که باید میان تمام اجزاء آن نیم‌فاصله باشد تا به صورت یک جزء واحد شناسایی شوند. پیروی از این رسم‌الخط استاندارد فارسی در خوانا بودن متن فارسی نقش بسزایی دارد و خواننده می‌تواند درک درستی از مفهوم متن داشته باشد. علاوه برآن، در پردازش زبان طبیعی و ازجمله ترجمه ماشینی، مشخص‌بودن محدوده کلمات در کارایی سیستم تاثیر بسزایی دارد.

در این پایان‌نامه روش آماری جدیدی طبق اصول ترجمه ماشینی آماری برای ویرایش متن فارسی ارائه‌شده است. در این روش با کمک ترجمه ماشینی آماری فاصله بین کلمات چندبخشی با نیم‌فاصله جایگزین شده‌اند. اطلاعات زبان‌شناختی از پیکره موازی استخراج شده و سپس از این اطلاعات برای شناسایی و ویرایش کلمات چندبخشی استفاده شده است. در این روش برای آموزش به یک پیکره موازی با ساختار متن ویرایش‌نشده در یک سمت و متن ویرایش‌شده در سمت دیگر نیاز است، که در این پایان‌نامه ایجاد شده‌است.

نتایج حاصل نشان‌دهنده کارایی روش در شناسایی صحیح بیشتر کلمات چندبخشی و اصلاح فاصله میان اجزاء کلمات چندبخشی با نیم‌فاصله است.

کلمات کلیدی

کلمات چندبخشی فارسی، قوانین فاصله‌گذاری متن فارسی، ترجمه ماشینی آماری، پیکره موازی فارسی، ترکیب کلمات چندبخشی فارسی

فهرست مطالب

لیست اختصارات

MT	Machine Translation
SMT	Statistical Machine Translation
RBMT	Rule-based Machine Translation
LM	Language Model
EM	Expectation Maximization
HMM	Hidden Markov Model
CYK	Cocke–Younger–Kasami
CYK+	Cocke–Younger–Kasami +
SCFG	Synchronous Context-Free Grammar
PSCFG	Probabilistic Synchronous Context-Free Grammar
CF	Context-Free
CNF	Chomsky Normal Form
CFG	Context-Free Grammar
SAMT	Syntax Augmented Machine Translation
MERT	Minimmum Error Rate Training
nplcf	non-partially lexicalized context free grammars

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول: مقدمه

۱.۱ طرح مساله

کلمات چندبخشی ترکیبی از کلماتی هستند که دستور و معنای متفاوتی را با هریک از بخش‌های خود ایجاد می‌کنند [۱]. برای حفظ یکپارچگی این اجزاء باید از نیم‌فاصله میان آن‌ها استفاده کرد. بسیاری از نویسنده‌ها به‌صورت نادرست از فاصله به جای نیم‌فاصله استفاده می‌کنند [۲] که باعث بی‌ثباتی و نامفهومی متن می‌شود [۳]. هنگامی‌که مصدر و پیشوندها وپسوندها با فاصله ازهم می‌آیند، شناسایی محدوده کلمه با مشکل مواجه خواهد بود.

یک مشکل آن، آنالیز لغوی است که با مشکلات و چالش هایی در آنالیز محاسباتی[1] مواجه است که یکی از مشکلات اصلی آن نشانه گذاری[2] متن است. به دلیل آن‌که محدوده کلمه آن‌گونه که در متن آمده نیست،[۴] و اگر مصدر کلمه و پیشوندها وپسوندها با فاصله ازهم نوشته شده باشند، هرکدام به‌صورت واحدهای مجزا نشانه‌گذاری می‌شوند. که در این‌صورت پردازش بر روی متن، به عنوان مثال ترجمه ماشینی متن، را با مشکل مواجه خواهد کرد. بنابراین با پیش‌پردازشی بر روی متن ورودی می‌توان به کارایی بهتری دست یافت.

مشکل دیگر آن، خوانایی متن است. نویسنده یک متن را برای ارتباط با خواننده می‌نویسد. بنابراین یک صورت ناآشنا از کلمات باعث می‌شود که درک مطلب با مشکل مواجه شود [۵] و از طرفی ممکن است معنای کلمه دچار جابه‌جایی شود. به عنوان مثال اگر کلمه "نام‌آشنا" با فاصله نوشته شود به معنای اسم آشنا معنا شود[۳].

تاکیدبر خوانا بودن متن به این اشاره دارد که خواننده چقدر در خواندن و درک متن یا نوشته با موفقیت همراه است. خوانا بودن متن به ویژه در متونی اعم ازکتاب‌ها، نشریات، سخنرانی‌ها، اخبار و اسناد و مدارکی که بیشتر کاربرد دارند، اهمیت دارد.

[1] Computational analysis
[2] Tokenization

رسم‌الخط و شیوه ویرایش متن در خوانا بودن و قابل درک بودن متن تاثیر بسزایی دارد. مشخص بودن فاصله صحیح بین اجزاء یک کلمه و فاصله بین کلمات مستقل می‌تواند این دو مشکل را بهبود دهد. به این معنا که فاصله میان اجزاء کلمات چندبخشی باید کمتر از فاصله میان کلمات مستقل باشد. به چنین فاصله‌ای "نیم‌فاصله" گفته می‌شود. نیم‌فاصله در اتصال بخش‌های مختلف یک فعل، اسم یا صفت چندبخشی همانند "می‌خورم"، "حاصل‌ضرب"، "تجدیدنظرخواهی"، "پرده‌برداری"، "رحیم‌زاده"، "زمانی‌که" و "نیازمند" استفاده می‌شود.

زبان فارسی شامل پیشوندها و پسوندهایی است که هیچ گرامر خاصی ندارند [۶]. به صورت کلی بهترین حالت نوشتن کلمات چندبخشی این است که کلمات چندبخشی به صورت جدا و با نیم‌فاصله نوشته شوند، مگر آنکه کثرت استعمال نوع گسسته آن، صورت عجیبی داشته باشد. مثل کلمه "بهتر" که اگر به صورت "به‌تر" بنویسیم، عجیب است، هرچند که غلط نیست. جدول۱-۱ نمونه‌ای از کلماتی که به اشتباه به جای نیم‌فاصله، فاصله بین آن‌ها قرار گرفته است، آمده است.

جدول۱-۱.نمونه‌ای از کلمات چندبخشی

ویرایش نشده	ویرایش شده
در صورتی که	در صورتی‌که
می روند	می‌روند
در نظر گرفته شده است	درنظرگرفته‌شده‌است
محمد زاده	محمدزاده

به طور کلی ساختمان اسم‌های چندبخشی را می‌توان به دسته‌های زیر تقسیم کرد:

۱. قسمت معنی‌دار + قسمت معنی‌دار = اسم مرکب

نماز + خانه= نمازخانه

سپید + رود= سپیدرود

۲. قسمت معنی دار + قسمت معنی ساز=اسم مرکب

قلم +دان= قلم‌دان

دل + انگیز= دل‌انگیز

۳. قسمت معنی‌ساز به تنهایی معنی ندارد وقتی با کلمه ی دیگر ترکیب می شود، معنی جدید پیدا می‌کند. که شامل:

پیشوند = با +شعور ، باشعور

میانوند =سر+ تا +سر ، سرتاسر

پسوند = داد + گر ، دادگر

پسوندها تکواژ وابسته ایست که به پایان واژه‌ای می‌چسبد و معنای واژه را تغییر می‌دهند. مانند: "گون"، "آسا"،"مند" در کلماتی مثل "نهنگ‌آسا"، "برف‌آگین"، "سبزه‌زار"، "دیوسان"، "شیرسان"، "لاله‌گون"، "بهره‌مند" . پیشوندها نیز برخلاف پسوندها به ابتدای واژه می‌چسبد و معنای آن را تغییر می‌دهد. مانند "ابر"، "بی"، "پاد"، "تک"، "در"، "سر"، "فر"، "فرا"، "نا"، "وا" در کلماتی مثل "ابررسانا"، "بازآمدن"، "بی‌خرد"، "پادزهر"، "تک‌نگاری"، "درآمدن"، "سرآمدن"، "وارفتن"، "واکنش"، "فرارفتن".

ساختمان افعال چندبخشی نیز به صورت جزء غیرفعل (اسم، صفت، قید) به اضافه جزء فعلی است.

جزء غیرفعل اسم است؛ زمین (جزء غیرفعل) + خورد (جزء فعلی) = فعل مرکب: زمین خورد.

جزء غیرفعل صفت است؛ پاک (جزء غیرفعل) + کرد (جزء فعلی) = فعل مرکب: پاک کرد.

جزء غیرفعل قید است؛ بیرون (جزء غیرفعل) + آورد (جزء فعلی) = فعل مرکب: بیرون آورد.

۱.۲ مزیت روش‌های آماری

روش‌های آماری مزیت‌های بسیاری به سایر روش‌های غیرآماری دارد. اولین و اصلی‌ترین مزیت آن این است که، ترجمه بهتری نسبت به سایر روش‌ها دارد. به دلیل آنکه ارتباط بین کلمات، عبارات و ساختار گرامری معمولا به صورت ضعیف و مبهمی تعریف شده‌اند. توزیع احتمالاتی و تکنیک‌های آماری باعث شده‌اند که به این ضعف‌ها غلبه شود و از ساختارگرامری اطلاعاتی را بدست آوریم. روال آموزش در روش‌های آماری وابسته به داده‌های ورودی است و به منابع زبانی نیازی نیست [۷].

یک مدل آماری می‌تواند برروی مقدار زیادی داده اعمال شود و افزایش مقدار داده می‌تواند به مدل این اجازه را بدهد که اطلاعات بیشتری را از پدیده‌های زبان‌شناختی بدست آورد. بنابراین، افزایش مقدار داده‌های آموزشی می‌تواند به کیفیت بالای ترجمه منجر شود [۷].

از مزیت‌های دیگر روش‌های آماری این است که اتکا به خصوصیات خاصی از زبان مانند مدل‌های زبانی خاص جهت ترجمه و گرامر ندارد. بسیاری از خصوصیات مدل‌های ترجمه مستقل از زبان هستند و می‌تواند برای هر جفت زبان مبدا و مقصد با استفاده از پارامترهای مدل، تخمین‌زده شوند. به این دلیل سیستم‌های ترجمه ماشینی این توانایی را پیدا می‌کنند که با کمترین اصلاح و تغییر برای هر جفت از زبان‌های مبدا و مقصد ساخته شوند. جهت افزایش کیفیت ترجمه ماشینی، اطلاعات زبانی خاصی نیاز است: مدل‌های آماری ارائه شده‌اند تا اطلاعات زبانی را به‌کار گیرند. این اطلاعات زبانی شامل خصوصیات واژه‌ها، مرتب‌سازی دوباره[1] و مدل‌های گرامری است[۷] [۸].

بنابراین در این پایان‌نامه برروی روش‌های آماری تمرکز شده است.

[1] Reordering

۱.۳ تئوری اولیه ترجمه ماشینی آماری[1]

هدف ترجمه ماشینی، ترجمه جمله ورودی f از یک زبان به جمله خروجی e با همان معنی در زبان دیگر است. اولین گام در ترجمه ماشینی آماری (SMT)، تطبیق کلمه است که از مدل کانال نویزی[2] استفاده می‌کند. مدل اولیه سیستم ترجمه ماشینی براساس تطبیق کلمه به سه بخش تقسیم می‌شود[۹]:

۱. مدل زبانی[3] که احتمال $P(e)$ را به هر رشته مبدا اختصاص می‌دهد.

۲. یک مدل ترجمه[4] که احتمال $P(e|f)$ را برای هر جفت از رشته‌های مبدا و مقصد اختصاص می‌دهد.

۳. دیکدر : دیکدر جمله f را که قبلا ندیده است را می‌گیرد و سعی می‌کند که بیشترین احتمال $P(e|f)$ یا معادل آن $P(e)\ P(f|e)$ را داشته باشد.

ترجمه ماشینی در مورد پیداکردن جمله e با استفاده از ترجمه جمله داده شده f است. f و e برای جمله زبان فرانسه وانگلیس تعریف شده‌اند، به‌دلیل آن‌که این دوزبان در مقاله‌های مختلف موضوع بحث بوده‌اند. بنابراین در این پایان‌نامه که با زبان فارسی سروکارداریم بازهم به این تعریف‌گرهای تعیین‌شده وفادار مانده‌ایم.

ترجمه یک جمله ازیک زبان به زبان دیگر قطعی نیست. یک جمله می‌تواند با حفظ معنا به جمله‌های متفاوتی ترجمه شود. درحالی‌که، می‌تواند فقط به یک صورت تفسیر شود. به عبارت دیگر، یک جمله می‌تواند ترجمه‌های متفاوتی داشته باشد. به همین دلیل ترجمه ماشینی آماری هیچ جمله‌ای را برای ترجمه رد نمی‌کند. هرجمله یک کاندیداست. کاندیدی که بیشترین احتمال را دارد، به‌عنوان یک ترجمه خوب تعیین می‌شود[۱۰]. برای هرجفت جمله (e,f) یک احتمال $P(e|f)$ را محاسبه می‌کند، که در آن

[1] Statistical Machine Translation (SMT)
[2] Noisy channel
[3] Language Model (LM)
[4] Translation Model

e ترجمه f است. جمله‌ای انتخاب می‌شود که بیشترین احتمال ترجمه f را با درنظرگرفتن e دارد (معادله ۱-۱)[۱۰, ۱۱].

$$\hat{e} = argmax_e P(e|f) \qquad (۱-۱)$$

ترجمه ماشینی آماری اولیه، یک پیاده‌سازی ازمدل کانال نویزی است. دراولین گام تئوری Bayes در فرمول کانال تویزی اعمال می‌شود. به دلیل این‌که $P(f)$ برروی محاسبه $argmax$ تاثیری ندارد، می‌تواند نادیده گرفته شود. بنابراین معادله به صورت زیر می‌شود (معادله۲-۱) [۱۱] :

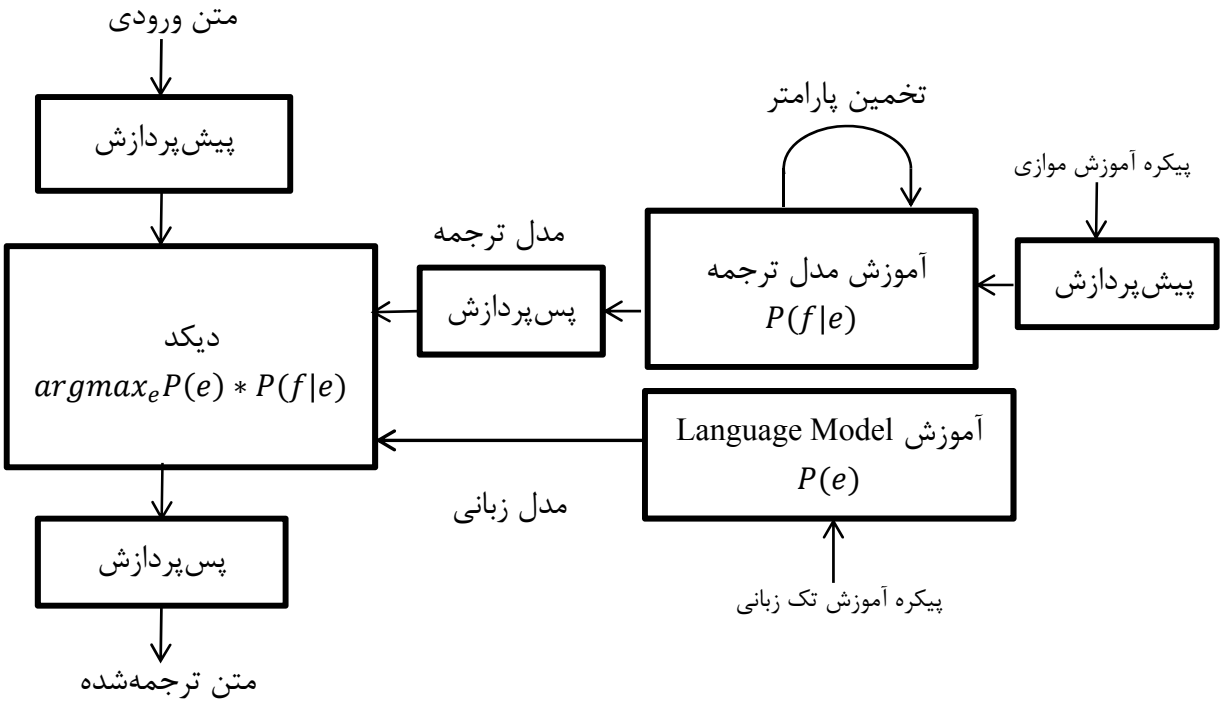
$$argmax_e P(e|f) = argmax_e P(e).P(f|e) \qquad (۱-۲)$$

بیشترین احتمال ترجمه عبارت f به e به صورت احتمال پیشین[۱] و احتمال f به شرط e ارائه شده است. در اولین نگاه این راه مفید به‌نظرمی‌رسد. اگرچه، $P(e)$ به ما این اجازه را می‌دهد که ترجمه‌های خوب را انتخاب کنیم. برای فهمیدن این موضوع به یادداشته باشید که $P(e|f)$ هیچگاه صفر نمی شود. هرمقدار e یک ترجمه‌ای از f است. اگر e نامفهوم باشد معنای آن این است که بسیاری ازفضای احتمالاتی به ترجمه‌ای داده شده است که جمله‌های آن ill-formed هستند. درحقیقت یک فضای قابل‌ملاحظه در فضای احتمالاتی است. احتمال $P(e)$ برای همین درنظرگرفته شده است. این احتمال " احتمال مدل زبانی"[۲] نامیده شده است [۱۰].

احتمال مدل زبانی می‌تواند به‌عنوان یک e که قراراست اتفاق بیافتد، درنظرگرفته شود. به‌طوری‌که، ترجمه نامفهوم با احتمال کمتری نسبت به یک ترجمه منسجم احتمال رخداد دارد. در جمله خوش‌فرم، $P(e)$ مقداربیشتری برای کلمه‌های بعدی نسبت به کلمه‌های قبلی دارد. احتمال $P(f|e)$ به نام احتمال مدل ترجمه[۳] نامیده می‌شود[۱۰].

[1] Prior probability
[2] Language model probability
[3] Translation model probability

احتمال مدل ترجمه، احتمالی است که جمله e را به‌عنوان ترجمه جمله f دارد. به‌صورت شهودی، حاصل‌ضرب $P(e)P(f|e)$ بزرگتر خواهدشد، اگر هم $P(e)$ و هم $P(f|e)$ زیاد شود. [۱۰]یک نمایش تصویری ازمطالب بالا است.این شکل۱-۱ اساس یک سیستم ترجمه ماشینی آماری را نشان می‌دهد.



شکل۱-۱. نمایش شماتیک ترجمه ماشینی آماری با جزئیات نشان می‌دهد. آموزش مدل زبانی با داده آموزشی (تک زبانی) جدا آموزش داده می شود.

[۱۲] مجموعه‌ای از ترجمه‌های ماشینی (MT) را براساس جانشینی کلمه با کلمه[1] و ترتیب‌دهی دوباره[2] را ارائه کرد. اما این روش شامل الگوریتم دیکد نبود. اگر دیکدکردن به چرخش‌های اطراف نودها در یک درخت دودویی محدود شده باشد، برای داشتن یک دیکد بهینه، می‌توان از الگوریتم high-polynomial

[1]Word-for-word
[2] reordering

استفاده کرد. در این‌صورت برای ترتیب‌دهی دوباره هر کلمه دلخواه، الگوریتم دیکد NP-Complete است [۱۳].

یک استراتژی قابل درک [۱۴]، [۱۵]، اعمال‌کردن زیرمجموعه بزرگی از دیکدهای احتمالی و انتخاب از میان آن‌ها است. البته احتمال این که یک ترجمه خوب را ازدست دهیم، وجود دارد. اگر دیکدر e' را برگرداند، اما امتیاز آن از e طبق $P(e|f) > P(e'|f)$ کمتر باشد، جست‌وجو خطا نامیده می‌شود. همان‌طورکه [۱۵] اظهارداشته است، مشکل می‌توان فهمید که جست‌وجو خطا اتفاق افتاده است یا نه. اما جست‌وجو خطا تنها راهی است که نشان دهیم که دیکد به صورت نیمه بهینه اتفاق افتاده است و یک امتیاز بالا ایجاد شده است [۹].

۱.۴ کارهای پیشین

تابحال تحقیقات بسیار محدودی بر روی ویرایش متن فارسی و به‌خصوص ویرایش نیم‌فاصله در کلمات چندبخشی انجام شده است. در ادامه به بررسی کارهای انجام شده، پرداخته شده است.

در مقاله [۳] جهت بهبود کیفیت فاصله‌گذاری در متن فارسی از رهیافت مدل زبانی برای اصلاح فاصله‌گذاری‌های نادرست استفاده شده است. این رهیافت متکی‌بر یک لغت‌نامه از کلمات نیم‌فاصله است. لغت‌نامه یک مدل نگاشت از فاصله به صورت نیم‌فاصله صحیح آن است. به این صورت شروع می‌شود که، یک جمله را می‌گیرد، سپس دیکد آن با استفاده از مدل زبانی به جست‌وجو جهت پیداکردن مسیری با بیشترین احتمال صورت می‌گیرد. در این مقاله از Peykare corpus [۱۶] و Persian dependencyand Treebank [۱۷] برای استخراج لغت‌نامه استفاده شده است. اشکالات روش ارائه‌شده در [۳] این است که، هیچ‌گونه اطلاعات زبان‌شناختی از متن استخراج نمی‌کند. بنابراین فقط می‌تواند کلماتی را که در لغت‌نامه

آمده‌اند را ویرایش کند. مشکل دیگر آن، استفاده از پیکره‌هایی است که در آن به‌درستی نیم‌فاصله رعایت نشده‌اند.

در مقاله [۲] متدهای مبتنی بر لغت‌نامه و مبتنی بر قانون را باهم جهت استانداردسازی متن فارسی، ترکیب کرده است. در گام آنالیز موفولوژی، مصدرها و پیشوندها و پسوندها با استفاده از صرف فعل و بعضی قوانین مورفولوژی اشتقاق در زبان فارسی، استخراج می‌شوند. این آنالیز با استفاده از برچسب‌های Part Of Speech (POS) برروی مصدرها و کل کلمه اعمال می‌شود تا بهترین مسیر را پیدا کند. این الگوریتم از اتوماتا حالت متناهی[1]، یک دیتابیس شامل کلمات و برچسب های POS، یک ساختار برای نگه‌داری مصدرها و برچسب‌های پیش‌بینی شده، دو لیست برای نگه‌داری پیشوندها و پسوندها، به‌طوری‌که از کلمه حذف شده باشد، و قوانینی برای آنالیز موفولوژی استفاده می‌کند. اشکال این روش نیز در مواجهه با کلماتی است که تابه‌حال ندیده است. به دلیل آنکه باید تمام قوانین زبان‌شناختی را برای سیستم تعریف کرده باشد.

مزیت استفاده از روش آماری این است که، احتیاج به هیچ‌گونه تعریف قوانین زبان‌شناختی ندارد و خود سیستم می‌تواند این قوانین را، حتی بدون درنظرگرفتن زبان مورد بررسی، استخراج کند. در [۳] از یک لغت‌نامه برای نگاشت کلمات به شکل ویرایش شده آن استفاده می‌کند. به‌طوری‌که اگر لغت دیده شده در زمان تست در آموزش دیده نشده باشد، نمی‌تواند آن را ویرایش کند. از طرفی [۲, ۳] هیچ‌گونه اطلاعات زبان‌شناختی ندارند، بنابراین برای کلماتی که با جایگزین کردن فاصله به جای نیم‌فاصله معنی آن‌ها تغییر می‌کند، نمی‌تواند به‌کارگرفته شوند.

۱.۵ رئوس مطالب

ساختار باقی پایان‌نامه به صورت زیر است:

[1] Finite State Automata(FSA)

فصل ۲ نگاهی بر روش تطبیق کلمه در ترجمه ماشینی آماری دارد. در این فصل به مدل‌های ترجمه‌ای برای محاسبه احتمالات کلمات تطبیق یافته و همچنین بهبود آن‌ها و نیز درباره چگونگی محاسبه احتمالات بیشینه در ترجمه پرداخته شده است.

فصل ۳ به معرفی روش آماری برای مدل کردن زبان می‌پردازد. این فصل، جزئیات محاسبه مدل زبانی و اشکالات آن و بهبود آن را مورد بحث و بررسی قرار داده است.

فصل ۴ به جزئیات مربوط به استخراج عبارات از یک متن موازی پرداخته است. در این فصل نحوه استخراج گرامر براساس عبارت و عبارت سلسله مراتبی با استفاده از کلمات تطبیق‌یافته ارائه شده است.

فصل ۵ روش‌های آموزش را در دودسته کلی مورد بحث قرار می‌دهد. روش‌هایی که برای آموزش براساس تطبیق‌های کلمه، عبارات و عبارات سلسله مراتبی مورد استفاده قرار گرفته‌اند.

فصل ۶ نگاهی به الگوریتم دیکد استفاده شده در این پایان‌نامه می‌پردازد. این الگوریتم دیکد از اطلاعات استخراج شده در زمان آموزش برای دیکد متن تست استفاده می‌کند.

فصل ۷ روش پیشنهادی، که مراحل آن در فصل‌های قبل مورد بحث قرارگرفت، را با استفاده از یک ابزار ترجمه ماشینی آماری پیاده‌سازی کرده است. نتایج حاصل با معیارهای مختلف مورد ارزیابی قرار گرفته است.

و در انتها در فصل ۸ به جمع‌بندی و نگاهی کلی به پایان‌نامه که شامل نیازها، روش و نتایج حاصل است، پرداخته شده است.

فصل دوم: تطبیق کلمه برای ترجمه ماشینی آماری

روش معمول برای ساخت سیستم ترجمه ماشینی آماری (SMT) به این‌صورت است که ابتدا یک مدل برای تطبیق‌دادن متن موازی[1] ایجاد کنیم. از این مدل برای هماهنگی کلمه‌های چندبخشی و نیز عبارات یک جمله در متن ویرایش‌نشده با متن ویرایش‌شده استفاده می‌شود. می‌توان به تطبیق به‌عنوان یک متغیر مخفی نگاه کرد و مدل‌ها را ساخت. معمولا از مدل‌ها استفاده می‌کنیم تا بهترین خروجی را به هر جمله در متن آموزش ارتباط دهیم. با استفاده از این اتصالات تطبیق‌ها، می‌توان عبارات ترجمه و قوانین عبارات را استخراج کرد [۷].

مدل‌های تطبیق برای تمام سیستم‌های ترجمه ماشینی استفاده می‌شود. این مدل‌ها برای سیستم‌های مبتنی بر عبارت[2] نیز برای استخراج هر جفت از عبارات از داده‌های آموزش به‌کار گرفته می‌شوند. به طوری‌که این مدل‌ها برای تولید و امتیازدهی فرض‌ها استفاده می‌شود. سیستم‌های مبتنی بر نحو[3] و سیستم‌های براساس عبارت سلسله مراتبی[4] [۱۸] نیز از مدل‌های ترجمه‌ایی که با متن‌های موازی مقداردهی می‌شوند، استفاده می‌کنند. تطبیق‌های صحیح میان متن‌های موازی می‌تواند منجر به ساخت سیستم‌های ترجمه‌ای بهتر شود. بنابراین، گسترش و آموزش مدل‌های تطبیق برروی کیفیت ترجمه تاثیر زیادی دارد[۷].

IBM MODEL ۲.۱

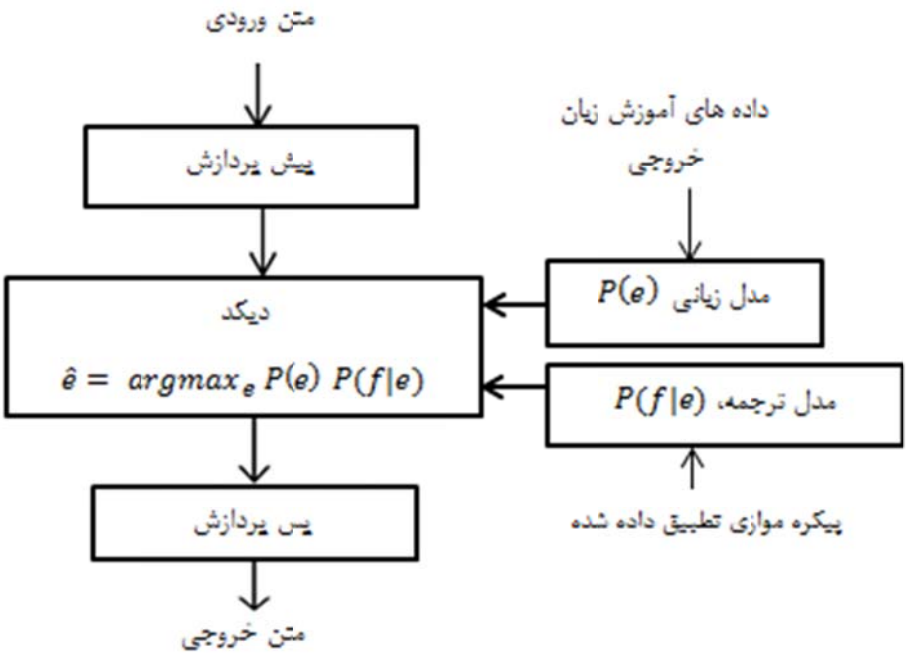
IBM MODEL [۲] مدل کانال مبدا را فرموله کرده است. به‌طوری‌که، فرض‌شده برای مدل‌کردن هدف e به‌عنوان خروجی f، با عبور از کانال نویزی به صورت P(f|e) تعریف شده است و می‌تواند جمله اولیه را از طریق f بازیابی کند. با استفاده از تئوری بیزین[5] می‌توان آن را به‌صورت

[1] Parallel corpora
[2] Phrase-based
[3] Syntax-based
[4] Hierarchical phrase-based
[5] Bayes' Theorem

معادله ۱-۲ نوشت:

$$\hat{e} = argmax_e P(e)\, P(f|e). \qquad (۲-۱)$$

که $P(e)$ مدل زبانی، $P(f|e)$ مدل ترجمه است. این فرایند به‌عنوان فرایند دیکد[1] شناخته می‌شود [۷]. مدل ترجمه با استفاده از متن موازی، آموزش داده می‌شود. متنی که قواعد مربوط به فاصله‌گذاری و قواعدی که در حفظ معنای کلمه‌های چند بخشی رعایت نشده است و نیز متن ویرایش‌شده معادل آن را که به‌طور دستی تهیه شده است را برای آموزش مدل ترجمه بکار گرفته شده است. به‌طور کلی، هرچه داده‌های آموزش بیشتر باشند، به نتیجه بهتری منجر خواهد شد. علاوه بر آن، یک متن ویرایش‌شده را نیز نیاز دارد تا مدل زبانی را آموزش دهد. مدل زبانی سیستم ترجمه ماشینی را برای تولید جمله‌های صحیح هدایت می‌کند [۷]. در شکل ۲-۱ [۷] سیستم مدل کانال نویزی و دیکدکردن نشان داده شده است.



شکل۲-۱. نمایش گرافیکی مدل کانال نویزی و دیکدکردن

[1] Decoding

در این فصل در مورد مدل‌های IBM MODEL 1 تا IBM MODEL 5 [۱۲] و نیز مدل تطبیق مخفی مارکوف[1] [۱۹, ۲۰] بحث شده است. تمام این مدل‌ها تجزیه‌های متفاوت احتمالاتی $P(f_1^J, a_1^J|e_1^J)$ را تولید می‌کنند. تطبیق a_1^J برای کلمات مبدایی که به هیچ‌یک ازکلمات مقصد تطبیق پیدا نمی‌کند، $a_j = 0$ است. تمام مدل‌ها شامل پارامتر لغوی $P(f|e)$ و یا پارامترهای اضافی که احتمال یک تطبیق را توصیف می‌کند، است.

دراینجا طرح کلی هریک ازشش مدل آورده شده است[۲۱]:

- تمام تطبیق‌ها در IBM-Model 1 احتمال یکسانی دارند.
- IBM-Model 2 مدل تطبیق مرتبه صفر[2] $P(a_j|j, I, J)$ استفاده می‌کند. به‌طوری‌که موقعیت‌های متفاوت آن‌ها ازهم مستقل هستند.
- مدل مخفی مارکوف از مدل مرتبه اول[3] $P(a_j|a_{j-1})$ به‌کار می‌گیرد. موقعیت تطبیق a_j به موقعیت تطبیق قبلی a_{j-1} وابسته است.
- در IBM Model 3 مدل تطبیق مرتبه صفر را به همراه مدل fertility اضافه شده داریم. به طوری که، تعداد لغات تطبیق داده شده به یک کلمه مقصد e را توصیف می‌کند.
- IBM Model 4 یک مدل تطبیق مرتبه اول $P(j|j')$ و مدل fertility $P(\emptyset|e)$ را داریم.
- IBM Model 3 و IBM Model 4 این ناکارامدی را دارند که، فضای احتمالاتی را برروی غیررشته به هدر می‌دهند. IBM Model 5 فرمول‌بندی جدیدی را از IBM Model 4 ارائه می‌دهند که غیرکارامدی آن را تصحیح می‌کند.

[1] Hidden Markov (HMM)
[2] Zero-order
[3] First-order

بنابراین تفاوت‌های اساسی این مدل‌ها (چه بامرتبه صفر چه با مرتبه یک) در مدل fertility است. چه کارا باشند یا نباشند.

مدل تطبیق مخفی مارکوف، IBM-Model 4 و IBM-Model 5 پارامترهای تطبیق را جهت دربرگرفتن وابستگی میان کلاس‌های کلمات در پیرامون موقعیت تطبیق، گسترش‌یافته‌اند. در مدل تطبیق مخفی مارکوف وابستگی کلمات یک کلاس $E = C(e_{a_{j-1}})$ را داریم. متقابلا چنین وابستگی از کلاس‌های کلمه از زبان مبدا به زبان مقصد در IBM-Model 4 و IBM-Model 5 [۱۲]را نیز وجود دارد. دسته‌بندی کلمات به تعداد کلاس‌های داده‌شده، بوسیله یک روال آموزش آماری گیری [۲۲] به‌صورت خودکار انجام می‌شود[۲۱].

روش آماری ترجمه ماشینی به این‌صورت است، جمله f_1^j زبان مبدا داده شده، هرجمله هدف e_1^i را به یک احتمال که منتج به یک ترجمه صحیح شود، اختصاص می‌یابد. برای راحتی پروسه ترجمه، این احتمال به صورت معادله۲-۲ بازنویسی شده است:

$$P(e_1^I|f_1^J) = \frac{1}{p(f_1^J)} . P(e_1^J). P(f_1^J|e_1^I) \qquad (۲-۲)$$

در آموزش پارامترهای مناسب برای واژه‌های بعدی، ازیک متن دو زبانه را استخراج می‌کند. دراین صورت، احتمال با مجموع تمام متغیرهای پنهان به نام تطبیق‌ها بیان می‌شود. با این پیش‌زمینه، در ادامه نگاهی به IBM Modelها می‌پردازیم.

IBM Model 1 ۲.۱.۱

IBM Model 1 یک مدل احتمالاتی است که فرض می‌کند که جمله مبدا با طول l_f به جمله مقصد e به صورت پیوسته زیر ترجمه شده است[۲۳]:

- طول m برای جمله e تولیدشده است.
- برای هرجمله هدف در موقعیت $j \in \{1, \dots, m\}$ داریم:

- کلمه تولید شده f_j در f (شامل کلمه تهی f_0)انتخاب شده است.
- کلمه هدف e_jدر موقعیت j توسط f_j تولید شده است.

IBM Model 1 به‌عنوان جمله بافرض این‌که تمام طول‌های f یک احتمال یکنواخت $\in$ دارد به این معنا که تمام انتخاب‌های کلمات تولیدشده ازجمله مبدا تقریبا احتمال یکسانی دارند و احتمال ترجمه کلمه زبان مقصد فقط وابسته به کلمه زبان مبدا است [۱۲]:

$$P(e|f) = \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} P(e_j|f_i) \qquad (۲-۳)$$

معادله ۳-۲ تخمین احتمال جمله هدف را به شرط جمله منبع، محاسبه می‌کند. ما به دنبال پیداکردن بیشترین احتمال تطبیق جمله مبدا به جمله هدف هستیم. به‌طوری‌که، با یک تطبیق برطبق مدل تولیدی فهمیده می‌شود که، کدام جمله هدف توسط کدام جمله مبدا تولیدشده است. IBM Model 1، مشابه مدل‌های تطبیق دیگر، باید هرجمله هدف توسط دقیقا یک جمله مبدا (با درنظرگرفتن جمله‌های تهی) تولیدشده باشد وتطبیق a با بردار$a_1, \dots, a_m$ نشان داده شود. a_j موقعیت جمله مبدایی است که f_j برطبق تطبیق آن را تولید می‌کند. IBM Model 1 بیشترین احتمال تطبیق $\hat{e}$ در e و f به صورت معادله ۴-۲ محاسبه می‌کند [۱۱, ۲۳]:

$$\hat{e} = argmax_e \prod_{j=1}^{m} P(e_j|f_{a(j)}) \qquad (۲-۴)$$

زمانی‌که IBM Model 1 را اعمال می‌کنیم، هیچ وابستگی میان a_jها نیست، به‌راحتی تطبیق با بیشترین احتمال برای هر j ، با انتخاب مقدار a_j، که منجر به بیشترین مقدار برای $P(e_j|f_{a(j)})$ است، را پیدا می‌کند.

پارامترهای IBM Model 1 برای یک جفت زبان به‌صورت معمول توسط الگوریتم بیشترین انتظار تخمین‌زده شده است. این پارامترها به‌عنوان داده‌های آموزشی یک متن موازی درنظرگرفته شده است.

آموزش به‌صورت معمول با تنظیم تمام توزیع احتمالاتی به‌صورت یکنواخت درتمام کلمات زبان هدف آغاز می‌شود.

IBM Model 2 ۲.۲.۲

IBM Model 2 یک مدل لغوی است و به این‌صورت کار می‌کند که، برای جمله f با طول n ، ابتدا طول جمله هدف m تولید می‌شود، سپس یک تطبیق $a =< a_1, a_2, \dots, a_m >$ ایجاد می‌شود. به این تاکید دارد، کدام کلمه مبدا (یا نشانه تهی) به کدام‌یک ازکلمه‌های مقصد می‌تواند ترجمه شود. در انتها m کلمه خروجی تولیدمی‌شود، به‌طوری‌که e_j فقط به $f_{a_{(j)}}$ وابسته است [۲۴, ۲۵].

بنابراین اولین گام آن ترجمه لغوی است، همان کاری که در IBM Model 1 انجام می‌گیرد که دوباره با استفاده از احتمال ترجمه $P(e|f)$ مدل می‌شود. گام دوم تطبیق است که، با استفاده از $P(i|j, l_e, l_f)$ محاسبه می‌شود [۱۱]. این دو گام به‌صورت معادله ۵-۲ باهم ترکیب می‌شوند [۱۱]:

$$P(e, a|f) = \epsilon \prod_{j=1}^{l_e} P(e_j|f_{a(j)})\ P(a(j)|j, l_e, l_f) \qquad (۲-۵)$$

۲.۲.۳ تطبیق مدل مخفی مارکوف

فرض ساده‌ای که در IBM Model 1 شده این است که، تمام تطبیق‌ها تقریبا احتمال مساوی دارند و فرض خوبی نیست. به دلیل اینکه، تطبیق‌ها تمایل دارندکه محلیت[1] خود را نگه دارند. به این معنی که، همسایگی کلمات دریک زبان تمایل دارند که به یک همسایگی ازکلمات در زبان‌دیگر تطبیق داده شوند. هدف تطبیق مدل مخفی مارکوف[2] [۲۶] این است که، خصوصیت محلیت را ضبط و استخراج می‌کند. مدل تطبیق می‌تواند ساختاری داشته باشد که کلیت همسایگی کلمات را ازدست‌ندهد (۶-۲)[۲۷]:

$$P(F, A|E) = P(J|e^I) \times \prod_{j=1}^{J} P(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^J) \qquad (۲-۶)$$

[1] Locality
[2] Hidden Markov Model (HMM)

$$P(F,A|E) = P(J|e_1^J) \times \prod_{j=1}^{J} P(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \times P(f_j|f_1^{j-1}, a_1^j, e_1^I)$$

با استفاده ازاین تجزیه، سه احتمال شامل: یک احتمال طول، یک احتمال مطابقت و یک احتمال لغوی به ترتیب درفرمول قراردارد. در مدل مخفی مارکوف یک وابستگی مرتبه اول[1] برای تطبیق‌های a_j فرض‌شده است. علاوه‌بر آن، احتمال لغوی فقط وابسته به کلمه با موقعیتa_j است. بنابراین، می‌توان فرمول ساده‌سازی شده زیر را استفاده کرد (معادله ۲-۷، معادله ۲-۸)[۲۷]:

$$P(a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) = P(a_j|a_{j-1}, I) \quad (۲-۷)$$

$$P(f_j|f_1^{j-1}, a_1^j, e_1^I) = P(f_j|e_{a_j}) \quad (۲-۸)$$

درنهایت بعدازآنکه طول مدل را به صورت ساده $P(J|e_1^I) = P(J|I)$ فرض شد، می‌توان همه را باهم جمع کرد و تجزیه براساس مدل مخفی مارکوف را به صورت زیر فرض کرد (معادله ۲-۹)[۲۷]:

$$P(F|E) = \sum_{a_1^J} \prod_{j=1}^{J} [P(a_j|a_{j-1}, I).P(f_j|e_{a_j})] \quad (۲-۹)$$

ایده مدل تطبیق مخفی مارکوف باعث شده است که، پارامترهای تطبیق مستقل از موقعیت‌های مطلق کلمه باشند. می‌توان فرض کرد که احتمالات تطبیق فقط وابسته به عرض پرش میان کلمات مطابق با گرفتن محلیت تطبیق‌ها است. بنابراین احتمال تطبیق $P(a_j|a_{j-1}, I)$ می‌تواند به صورت $P(i|i', I)$ نوشته شود، به‌طوری‌که احتمالات تطبیق جدید با استفاده ازیک تابع پرش محاسبه شده است (معادله ۲-۱۰) [۲۷]:

$$P(i|i', I) = \frac{c(i-i')}{\sum_{i''}^{I} c(i''-i')} \quad (۲-۱۰)$$

در فرمول اصلی مدل تطبیق مخفی مارکوف، هیچ کلمه تهی یا Null که کلمه مبدا هیچ مطابقت مستقیمی به کلمه مقصددارد، تولید نمی‌کند.

[1] First-order-dependency

IBM Model 3 ۲.۲.۴

IBM MODEL 3 نظریهfertility را جهت پوشش این حقیقت که کلمه‌های مبدا ممکن است به تعدادی از کلمه‌های مقصد مرتبط شوند، معرفی کرد. گاهی اوقات یک کلمه در سمت مبدا به یک کلمه در سمت مقصد مرتبط شود (fertility=1)، اما ممکن است یک کلمه زبان مبدا به n کلمه زبان مقصد مرتبط شود (fertility=n) و یا حتی هیچ کلمه مقصدی (fertility=0) به آن مرتبط نشود. fertility کلمه‌های مبدا با توزیع احتمال $n(\emptyset|f)$ مشخص می‌شوند. برای هرکلمه f زبان مبدا این توزیع احتمال به تعداد $\emptyset = 0,1,2, \ldots$ کلمات مقصدکه به آن ترجمه می‌شود، دلالت دارد. fertility صراحتا با قراردادن $\emptyset = 0$ کلمه مبدا را دور می‌اندازد. نشانه تهی[1] معرفی شد تا اجازه دهد کلماتی که هیچ منطبقی درسمت مبدا ندارد، در سمت مقصد درنظرگرفته شود. می‌توان مدل fertility مربوط به نشانه تهی را برای کلمات اضافه‌شده با توزیع احتمال شرطی $n(\emptyset|null)$ مدل کرد. به دلیل آنکه تعداد کلمات اضافه‌شده وابسته به طول جمله است، اضافه‌کردن تهی درگام جداگانه‌ای انجام می‌شود. این نتایج در چهار مرحله برای پردازش ترجمه برطبق IBM Model 3 به ترتیب گام fertility، گام اضافه‌کردن تهی، گام ترجمه لغوی و گام مرتب‌سازی دوباره صورت می‌گیرد[۲۵, ۲۸]:

fertility با استفاده ازتوزیع $n(\emptyset|f)$ مدل می‌شود (معادله ۱۱-۲).

$$\prod_{i=1}^{l_f} \emptyset_i!\ n(\emptyset_i|e_i) \qquad (۲-۱۱)$$

اضافه‌کردن تهی با احتمال‌های P_1 (یک نشانه تهی بعد از هر کلمه) و $P_0 = 1 - P_1$ (هیچ نشانه تهی نباشد) مدل می‌شود (معادله ۱۲-۲) .

$$P(\emptyset_0) = \binom{l_e - \emptyset_0}{\emptyset_0} P_1^{\emptyset_0} P_0^{l_e - 2\emptyset_0} \qquad (۲-۱۲)$$

ترجمه لغوی با استفاده از IBM Model 1 صورت می‌گیرد.

[1] Null

مرتب‌سازی دوباره همانند روش IBM Model 2 با توزیع احتمال $a(j|i,l_e,l_f))$ مدل می‌شود.

می‌توان گام fertility و اضافه‌کردن تهی را به صورت ریاضی به شکل معادله ۱۳-۲ ترکیب کرد[۲۸]:

$$\binom{l_e-\emptyset_0}{\emptyset_0} P_1^{\emptyset_0} P_0^{l_e-2\emptyset_0} \prod_{i=1}^{l_f} \emptyset_i!\ n(\emptyset_i|e_i) \qquad (۲-۱۳)$$

در انتها با ترکیب سه گام توضیح داده شده در یک فرمول داریم (معادله ۱۴-۲) :

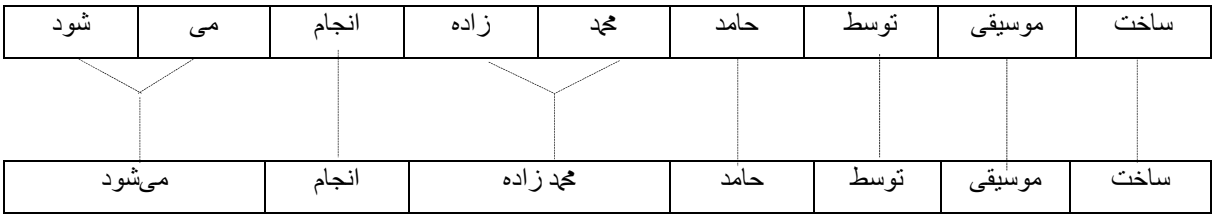
$$P(e|f) = \sum_a P(e,a|f) \qquad (۲-۱۴)$$

$$= \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \prod_{j=1}^{l_e} \binom{l_e-\emptyset_0}{\emptyset_0} P_1^{\emptyset_0} P_0^{l_e-2\emptyset_0} \prod_{i=1}^{l_f} \emptyset_0!\, n(\emptyset_i|e_i)$$

$$\times \prod_{j=1}^{l_e} t\left(e_j|f_{a(j)}\right) a(j|a(j),l_e,l_f)$$

IBM Model 3 هنوز یک روش قدرتمند در ترجمه کلمه‌به‌کلمه محسوب می‌شود. به‌طوری‌که جدول t (جدول ترجمه)، مرتب‌سازی دوباره، اضافه‌کردن کلمات تهی، حذف کلمات (کلمات با fertility مساوی صفر) و ترجمه (fertility>0) یک به چند برای کلمات محاسبه می‌شود.

بنابراین تطبیق کلمه یک فضا برای هر کلمه درنظر می‌گیرد. اگر دو کلمه مقصد به یک کلمه مبدا تطبیق داده‌شود، گفته می‌شود که کلمه مبدا fertility دو دارد. بر عکس، اگر کلمه مبدا هیچ معادلی در زبان مقصد نداشته باشد، fertility آن صفر است. شکل ۲-۲ پروسه تطبیق کلمه در متن موازی ویرایش‌نشده-ویرایش‌شده را نشان می‌دهد. به طوری که جمله ویرایش نشده به جمله ویرایش شده تبدیل می‌شود [۹].



شکل ۲-۲. تطبیق کلمه در پیکره موازی ویرایش‌شده-ویرایش‌نشده فارسی

در ابتدا کلمه ویرایش نشده یک fertility به آن نسبت داده می شود. این مقدار اختصاص داده شده با توجه به جدول $n(\emptyset \mid e_i)$ تعیین می‌شود. هر کلمه‌ای که fertility آن صفر است از جمله حذف می‌شود و کلمه‌هایی که مقدار fertility آن دو است را کپی کرده، تا فضا برای جایگزینی کلمه معادل داشته‌باشد. کلمه‌ای که fertility بزرگتر صفر داشته باشد را، کلمه fertility می‌گویند. اگر مقدار آن بزرگتر از یک باشد به آن بسیار fertility می‌گویند[۹]. در شکل ۲-۲ مقدار fertility برای کلمات "می" و "زاده" صفر است. سپس کلمه به کلمه جایگزینی صورت می گیرد (جایگزینی‌هایی که شامل تهی می‌شود، صورت می‌گیرد). این جایگزینی با استفاده از جدول $P(f_j \mid e_i)$ صورت می‌گیرد.

IBM Model 4 ۲.۲.۵

در داده‌های بزرگ اطلاعات کم و غیرواقعی از جابه‌جایی‌های کلمات داریم. این نقص در IBM Model 4 بهبودیافته است. کلماتی که در زبان مبدا در همسایگی هم بوده‌اند، در زبان مقصد نیز باید در همسایگی هم قرار داشته باشند. به عنوان مثال، ۲۲ امین ترجمه کلمه ورودی به ۲۲ امین کلمه خروجی وابسته به ۲۱ کلمه ماقبل آن است [۱۱] . بنابراین IBM Model 4 یک مدل "مرتب سازی نسبی[1]" ایجاد کرده است. این مدل به این اشاره دارد که جایگذاری کلمات ورودی نوعا براساس کلمات قبلی وروی وابسته است[۲۹].

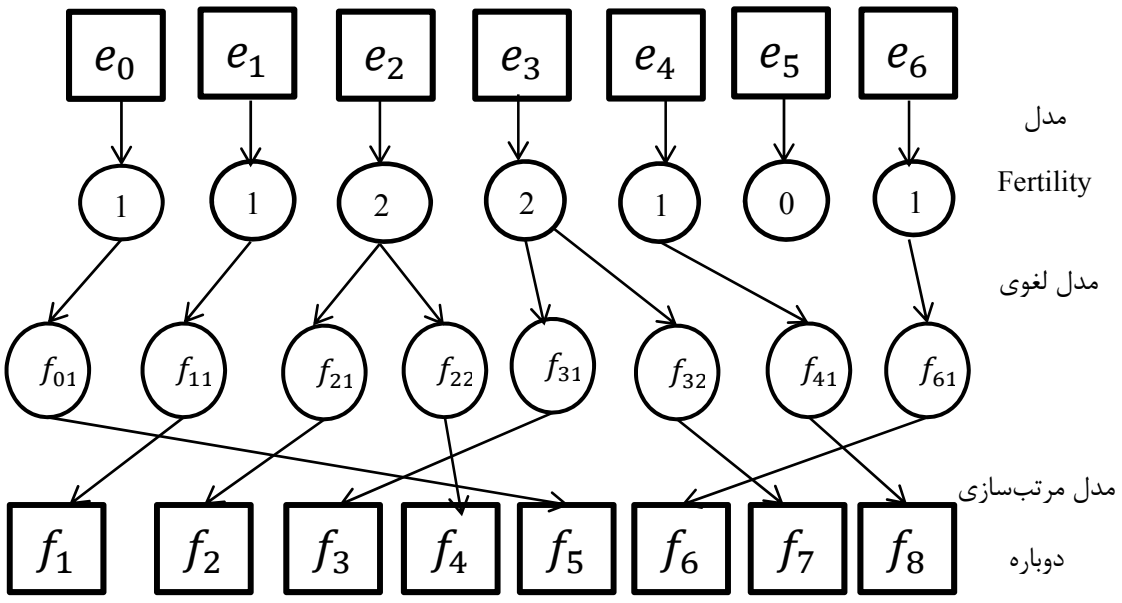
کلاس کلمات توسط IBM Model 4 معرفی شد. در آن کلمات یک زبان به تعدادی کلای گروه‌بندی می‌شوند و می‌توان احتمال شرطی روی این کلاس‌ها تعریف کرد[۱۱, ۲۹]. از این کلاس‌ها برای تعیین درست موقعیت‌های کلمات در ترجمه خروجی با استفاده از پردازش سمت دوم متن موازی که هم زبان با زبان خروجی است، انجام می‌شود.

[1] Relative distortion

IBM Model 5 ۲.۲.۶

برطبق IBM Model 3-4، این امکان وجوددارد‌که چندکلمه مقصد در یک موقعیت مبدا قراربگیرند. هیچ‌چیز از قرارگرفتن یک کلمه مقصد در یک موقعیت که درحال حاضر پرشده است، جلوگیری نمی‌کند. به‌عبارت دیگر، بسیاری ازتطبیق‌ها احتمالات مثبت دارند. IBM Model 5 این مشکل رارفع کرده است، به این معنا که این ناکارآمدی را حذف کرده است[۱۱, ۳۰]. این کار را با استفاده از دنبال‌کردن تعداد موقعیت‌های کلمه خالی و اجازه دادن فقط به این موقعیت‌ها برای جایگذاری،انجام می دهد. بنابراین مدل مرتب‌سازی براساس محل خالی است.

در شکل ۲-۳ [۱۱] از روند و پروسه تولیدکننده[1] مدل‌های IBM Model 1 تا IBM Model 5 نشان داده‌شده است.



شکل۲-۳ . پروسه تولیدکننده مدل‌های IBM Model 1 تا IBM Model 5

[1]Generative

فصل سوم: مدل آماری زبان[1]

[1] Statistical language model

در مدل آماری زبان از داده‌های آموزش برای تخمین آماری استفاده می‌شود. به دلیل آنکه زبان‌های متفاوتی با لغات زیادی مورد استفاده قرار می‌گیرد. تکنیک‌های تخمین آماری تعداد پارامترهای زیادی را باید تخمین بزنند، بنابراین نیاز به حجم داده بالایی دارند[۳۱]. بخش مهم سیستم ترجمه، مدل‌کردن زبان خروجی است. به طوری‌که، به ما اجازه می‌دهد که جملات مورد نظر به‌عنوان جملات گرامری درنظرگرفته شوند.

بیشتر تکنیک‌های آماری زبان از اطلاعات زبانی استفاده می‌کنند. متداول‌ترین مدل آماری زبان n-gram است. این مدل، زبان را مدل نمی‌کند، بلکه ترتیب نمادهای قراردادی را بدون هیچ ساختار پیچیده و یا فرض بیشتری مدل می‌کند. بنابراین می‌توان بدون داشتن اطلاعات خاصی از این مدل استفاده کرد.

به گفته Fred Jelinek باید هر زبان را به مدل زبان آن بازگرداند. متاسفانه تعداد انگشت‌شماری از تلاش‌ها تا به امروز در این زمینه توانسته ساختار زبانی را با استفاده از نظریه‌ها و دانش به مدل آماری زبان تبدیل کند و بسیاری از این تلاش‌ها به موفقیت نسبی دست یافته‌اند[۳۱].

تقریبا تمام مدل‌های زبان تا به امروز، احتمال یک جمله را به ضرب احتمالات شرطی تقسیم می‌کند (معادله ۱-۳) [۳۱].

$$P(s) = P(w_1\ w_2\ \dots\ w_n) = \prod_{i=1}^{n} P(w_i|h_i) \qquad (۳-۱)$$

همان‌طور که، گفته شد در سیستم ترجمه ماشینی نیاز به مدل زبانی دارد. مدل زبانی $P(e)$ با استفاده از متن‌های یک زبانه آموزش داده می‌شود. به طوری‌که، این امکان را می‌دهد که تا جایی‌که امکان دارد ترجمه درستی را ارائه دهد. به‌طور کلی، داده‌های آموزش زیاد منجر به ترجمه صحیح‌تری می‌شود [۷].

مدل می‌تواند برای ترجمه انواع مختلف داده‌ها از جمله اخبار وب سایت‌ها، تا خروجی سیستم‌های تشخیص صدا استفاده می‌شود. سیستم ترجمه ماشینی باید زبان با هر خصوصیتی را به خوبی ترجمه کند [۷].

۳.۱ مدل n-gram

فرض شده است که جملات از چپ به راست تولید شده‌اند و هرکدام به کلمه قبلی وابسته است. به عبارت دیگر، احتمال جمله قبلی $e = e_1, e_2, \dots, e_l = e_1^l$ به صورت معادله ۲-۳ محاسبه می‌شود [۷]:

$$P(e_1^l) = \prod_{i=1}^{l} P(e_i \mid e_1, e_2, \dots, e_{i-1}) \quad (۳-۲)$$

هر کلمه فقط به $n-1$ کلمه قبل خود وابسته است. این مدل زبانی به عنوان n-gram شناخته می‌شود. n-gram ابعاد تخمین مساله را با استفاده از مدل‌کردن زبان همانند مارکوف با مرتبه n-1، کاهش می‌دهد [۳۱]. احتمال جمله به صورت تقریبی به صورت معادله ۳-۳ محاسبه می‌شود[۷].:

$$P(e_1^l) \approx \prod_{i=1}^{l} P(e_i | e_{i-n+1}, \dots, e_{i-1}) \quad (۳-۳)$$

احتمالات n-gram $P(e_i \mid e_{i-n+1}^{i-1})$ با استفاده از متن بزرگی از زبان خروجی محاسبه می‌شود. تخمین بیشترین احتمال با استفاده از معادله ۴-۳ محاسبه می‌شود.

$$P(e_i | e_{i-n+1}) = \frac{c(e_{i-n+1}^{i})}{c(e_{i-n+1}^{i-1})} \quad (۳-۴)$$

تخمین احتمال بیشینه از پراکندگی داده‌ها رنج می‌برد. به عبارت دیگر، این مقدار زمانی‌که تعداد مثال‌های n-gram داده آموزش کم باشد، اشتباه می‌شود. مقدار احتمال صفر به n-gram یا کلماتی که در داده‌های آموزش نبوده‌اند، اختصاص داده می‌شود. درحالی‌که نمی‌خواهیم رخدادهایی با این مقدار وجود داشته باشند [۷]. متدهای هموارسازی[1] متفاوتی به‌کار گرفته شده‌اند تا تخمین‌های احتمالاتی بیشینه را جهت تولید توزیع‌های احتمالاتی صحیح، تنظیم کند.

[1] Smoothing

۳.۲ متد هموارسازی

متداول‌ترین روش در مدل زبانی، مدل‌های n-gram است. به‌طوری‌که، مستقیما برای هموارسازی استفاده می‌شود. این تکنیک تضمین می‌کند احتمالات بهتری را زمانی که داده‌های کافی برای تخمین احتمالاتی درست وجود نداشته باشد، استفاده کند. تکنیک‌های احتمالاتی زیادی براساس n-gram ارائه‌شده‌اند و بیشتر از آن هستند که بتوان آن‌ها را در اینجا ذکر کرد. روش‌هایی همانند Jelinek and Merce(1980)، Katz (1987)، Bell,Cleary and Witten(1990)، Ney,Essen and Kneser(1994)، Kneserand Ney (1995).

واژه هموارسازی تکنیک‌هایی برای تطبیق تخمین بیشترین احتمال را در فضای احتمالاتی برای تولید احتمالات صحیح‌تر توصیف می‌کند. هموارسازی از این حقیقت می‌آید که، این تکنیک‌ها باعث می‌شود که احتمالات بیشتر یکنواخت شوند. این یکنواختی بوسیله تطبیق احتمالات پایین مانند احتمالات صفر به مقدارهای بیشتر و احتمالات بالا به مقدار پایین‌تر صورت می‌گیرد. متدهای هموارسازی به طورکلی نه تنها از احتمالات صفر جلوگیری می‌کند، بلکه تلاش می‌کند تا صحت کلیت احتمال مدل را بهبود دهد. هر وقت که مقدار یک احتمال کم تخمین‌زده شده باشد، هموارسازی به صورت چشمگیری این مقدار را بهبود می‌دهد. بهترین و متداول‌ترین روش برای ارزیابی مدل زبانی، احتمالاتی است که مدل به داده‌های تست اختصاص می‌دهد یا معیار استخراج شده از cross-entropy و perplexity است [۳۲].

احتمال جمله $P(s)$ با استفاده از احتمالات $P(w_i \mid w_{i-n+1}^{i-1})$ که از هموارسازی n-gram بدست آورده است، به صورت معادله ۵-۳ محاسبه می‌شود [۳۲].

$$P(s) = \prod_{i=1}^{l+1} P(w_i \mid w_{i-n+1}^{i-1}) \qquad (۳-۵)$$

سپس به متن تست که از مجموعه جملات $(t_1, \dots, t_{l_T})$ ساخته شده است. می‌توان احتمال متن تست $P(T)$ را به عنوان ضرب احتمالات تمام جمله‌ها در این متن به حساب آورد (معادله ۶-۳).

$$P(T) = \prod_{i=1}^{l_T} P(t_i) \quad (۳-۶)$$

۳.۲.۱ متد هموارسازی Kneser-Ney

در این متد که در سال ۱۹۹۵ ارائه شد، گسترش‌یافته absolute discounting است. احتمال با مرتبه پایین با احتمال با مرتبه بالا ترکیب می‌شود و به صورت یک روش جدید ارائه می‌شود. در الگوریتم‌های قبلی به طور کلی، احتمال با مرتبه بالاتر به بیشترین مقدار توزیع احتمالاتی مرتبه پایین برده می‌شدند. زیرا زمانی‌که مقدار احتمالات مرتبه بالا کم و یا صفر باشد، احتمال با مرتبه پایین یک فاکتور قابل‌توجه می‌شود. بنابراین، این مقادیر باید بهینه شوند تا بتوانند در این مواقع به درستی اجرا شوند. در اینجا احتمالات متفاوتی را که توسط Kneser-Ney ارائه شده اند را مورد بررسی قرارداده شده است.

در این روش تنوع تاریخچه هر کلمه را را به عنوان متد هموارسازی درنظر گرفته است. به این منظور، با استفاده از معادله ۳-۷ تاریخچه برای یک کلمه را تعریف می‌کند[۳۲].

$$N_{1+}(\blacksquare w) = |\{w_i : c(w_i, w) > 0\}| \quad (۳-۷)$$

تخمین معمول بیشترین احتمال (Maximmum Liklihood) در مدل n-gram به صورت معادله ۳-۸ است.

$$P_{ML}(w) = \frac{c(w)}{\sum_i c(w_i)} \quad (۳-۸)$$

در هموارسازی Kneser-Ney مقدارهای خام را با مقدار تاریخچه آن برای یک کلمه جایگزین‌کرده است (معادله ۳-۹).

$$\mathrm{P_{KN}(w)} = \frac{\mathrm{N_{1+}(\blacksquare w)}}{\sum_{\mathrm{w_i}} \mathrm{N_{1+}(w_i w)}} \quad (۳-۹)$$

۳.۲.۲ مدل اصلاح شده هموارسازی Kneser-Ney

در این مدل درون‌یابی مدل‌های مرتبه بالا n-gram رابا استفاده از مدل‌های مرتبه پایین و متدهایی که فضای احتمالاتی مناسبی را به هرکدام از مرتبه‌ها اختصاص می‌هند و نیز چگونگی ساخت مولفه‌های مدل‌های n-gram، را در کنار هم قرار می‌دهد.

اول از همه ایده‌های درون‌یابی و روبه‌عقب را با هم ترکیب می‌کند و دوباره تابع درون‌یابی را به‌صورت معادله ۳-۱۰ تعریف کرده تا هر دو تابع در آن مشارکت داشته باشند [۳۲].

$$P_I(w_1 n\,|w_1,\dots,w_{n-1}) = \begin{cases} \alpha\,(w_n|w_1,\dots,w_{n-1}) & if\ c(w_1,\dots,w_n) \\ \gamma(w_1,\dots,w_{n-1})P_I(w_n|w_2,\dots,w_{n-1}) & otherwise \end{cases} \qquad (۳-۱۰)$$

برای n-gram در متن یک تابع α درنظرگرفته که مربوط به فضای احتمالاتی رزرو شده برای کلماتی که دیده‌نشده‌ای است که، بعد ازاین کلمه می‌آیند. این فرمول درون‌یابی رو به عقب[1] را به سمت تعریف اصلاح‌شده Kneser-Ney هدایت می‌کند.

۳.۲.۲.۱ فرمول α برای مدل n-gram مرتبه بالا

[۳۲]نسخه اصلاح‌شده هموارسازی Kneser-Ney را پیشنهاد کرد. به طوری که از متدی به نام absolute discounting برای کاهش فضای احتمالاتی برای رخدادهای دیده‌شده بهره می‌گیرد. برای $c(w_1,\dots,w_n) > 0$، absolute discounting بر تفریق یک مقدار ثابت D با $0 \le D \le 1$ از مقدارهای خام به مدل n-gram مرتبه بالا دلالت دارد (معادله ۳-۱۱)[۳۲] .

$$\alpha(w_n|w_1,\dots,w_{n-1}) = \frac{c(w_1,\dots,w_n)-D}{\sum_w c(w_1,\dots,w_{n-1},w)} \qquad (۳-۱۱)$$

[1] Interpolated back-off

[۳۲]یک مقدار ثابت برای D برای هر مقدار n-gram معرفی کردند، اما این مقدار باید وابسته به مدل n-gram باشد. مقدار بهینه آن به صورت **Error! Reference source not found.** ۳-۱۲ محاسبه می‌شود.

(۳-۱۲)

$$Y = \frac{N_1}{N_1 + 2N_2}$$

$$D_1 = 1 - 2Y\frac{N_2}{N_1}$$

$$D_2 = 2 - 3Y\frac{N_3}{N_2}$$

$$D_{3+} = 3 - 4Y\frac{N_4}{N_3}$$

مقدارهای N_c تعداد n-gram ها را با مقدارهای مشخص c را تعیین می‌کند. متناوبا، پارامتر D_i می‌تواند با استفاده از داده held-out بهینه شود.

۳.۲.۲.۲فرمول γ برای مدل n-gram مرتبه بالا

مقدار تابع γ فضای احتمال را برای رخدادهای دیده‌نشده درنظر می‌گیرد(معادله ۳-۱۳) .

$$\gamma(w_1, \dots, w_{n-1}) = \frac{\sum_{i\epsilon\{1,2,3+\}} D_i N_i(w_1,\dots,w_{n-1}\blacksquare)}{\sum_{w_n} c(w_1,\dots,w_n)} \qquad (۳-۱۳)$$

به‌طوری‌که N_i برای $i\epsilon\{1,2,3+\}$، براساس تعداد بسط تاریخچه یک کلمه $w_1, \dots, w_{n-1}$ به ترتیب ۱و۲و۳ و... . D_i را از هر n-gram با مقدار یک کم کرده است. بنابراین، D_1 بار n-gramها با مقدار یک برای γ استفاده می‌شود. این اتفاق برای سایر مقدارها نیز می‌افتد.

این فرمول بر روی اینکه تنوع کلمه‌های پیش‌بینی‌شده شمرده شود و فقط براساس فضای احتمالاتی discount بر روی فرکانس یک تاریخچه نباشد.

۳.۲.۲.۳ فرمول α برای مدل های n-gram مرتبه پایین

همان‌طورکه گفته‌شد n-gramها مرتبه‌پایین براساس توزیع احتمال تعداد تاریخچه‌های $N_{1+}(\blacksquare w)$ می‌توانند تخمین بهتری داشته باشند. به طوری‌که یک کلمه در آن ظاهر می‌شود، به‌جای آنکه تعداد تکرارهای خام آن درنظر گرفته شوند. بنابراین فرمول محاسبه α برای n-gram مرتبه پایین تغییر پیدا می‌کند (معادله ۳-۱۴) :

$$\alpha(w_n|w_1,\dots,w_{n-1}) = \frac{N_{1+}(\blacksquare w_1,\dots,w_n)-D}{\sum_w N_{1+}(\blacksquare w_1,\dots,w_{n-1},w)} \qquad (۳-۱۴)$$

D همانند قبل سه مقدار مختلف می‌گیرد.

۳.۲.۲.۲ فرمول γ برای مدل های n-gram مرتبه پایین

مدل n-gram مرتبه بالا فضای احتمال کنار گذاشته‌شده برای discount مقدارهای مشاهده شده با استفاده از تابع γ محاسبه می‌شود (معادله ۳-۱۵) [۳۳]:

$$\gamma(w_1,\dots,w_{n-1}) = \frac{\sum_{i\epsilon\{1,2,3+\}} D_i N_i(w_1,\dots,w_{n-1}\blacksquare)}{\sum_{w_n} c(w_1,\dots,w_n)} \qquad (۳-۱۵)$$

۳.۲.۲ درون‌یابی back-off

مدل‌های back-off معمولا براساس n-gramهای مرتبه بالا هستند و کلمه‌های پیش‌بینی‌شده و تاریخچه کلمات را با هم تطبیق می‌دهد. اگر دو n-gram متفاوت با یک تاریخچه یکبار در متن آموزش اتفاق بیافتد، به کلمات پیش‌بینی‌شده، احتمالات یکسانی اختصاص داده می‌شود. حتی اگر یک n-gram کم باشد ولی دیگری نماینده‌ای از یک اتفاق شایع باشد. اگر یک تاریخچه نادر باشد، بنابراین n-gram آن تاریخچه نیز باید کم باشد.

راه‌حل این مشکل این است که باید مقدارهای مدل‌های back-off با مرتبه پایین را درنظر گرفته شود، حتی اگر n-gram را دیده باشد. این کار را به سادگی با تطبیق تابع α با یک α_i درونیابی شده با اضافه کردن احتمال back-off انجام شده است (معادله ۱۶-۳).

$$\alpha_I(w_n|w_1,\dots,w_{n-1}) = \alpha(w_n|w_1,\dots,w_{n-1}) + \gamma(w_1,\dots,w_{n-1})P_I(w_n|w_1,\dots,w_{n-1}) \qquad (۳-۱۶)$$

مقدارهای تابع γنیز احتیاج به کاهش دارد [۳۲].

فصل چهارم: ترجمه براساس عبارت سلسله مراتبی

۴.۱ ترجمه براساس عبارت[1]

مدل‌های ترجمه اولیه فقط با تطبیق کلمه‌های میان دو کلمه سروکار دارد. به‌عبارت دیگر، واحدهای ترجمه، کلمه بودند. درحالی‌که بعضی توالی کلمه‌ها تمایل دارند که به‌صورت یک کلمه درنظرگرفته شوند. بنابراین این توالی از کلمات را باید در حالت مطلوب به صورت عبارت درنظرگرفته گرفت. به‌عبارت‌دیگر، کلمه‌های متوالی به‌صورت یک واحد درنظرگرفته می‌شوند [۷].

تمرکز بسیاری از تحقیقات اخیر ترجمه ماشینی آماری برروی مدل‌های ترجمه‌ای است که براساس رشته‌هایی از کلمه‌های به‌هم چسبیده در زبان مبدا و عبارات معادل با آن در زبان مقصد کار می‌کنند. ترجمه‌های براساس عبارت [۲۸, ۳۲, ۳۴-۳۶] موفق‌تر هستند و سیستم‌های ترجمه ماشینی بهتری را با استفاده از این روش می‌توان داشت. مولفه اصلی ترجمه براساس عبارت، تعیین مدل ترجمه با استفاده از متن موازی است که کلمه‌های آن تطبیق داده شده‌اند. جدول عبارت شامل عبارات زبان مبدا، ترجمه آن‌ها در زبان مقصد و احتمالات مرتبط با آن است. به‌طوری‌که، در زمان پیش‌پردازش و قبل از دیکدکردن متن تست، استخراج می‌شوند [۸].

ترجمه‌های براساس عبارت از این عبارات برای ترجمه استفاده می‌کنند. بنابراین گروهی از کلمه‌های متوالی در یک زبان می‌تواند به گروهی از کلمه‌های متوالی در زبان دیگر ترجمه شوند. این فرض باعث می‌شود که محتوای یک کلمه برروی ترجمه تاثیر بگذارد و تغییرات محلی در ترتیب کلمه‌ها میان زبان‌ها بتواند یادگرفته شود. علاوه‌بر آن، عبارات از متن واقعی استخراج شده‌اند. بنابراین کلمات، احتمال گرامری در درون خود عبارات دارند و سیستم می‌تواند ترجمه‌های بهتری را تولید کند [۷]. [۲۸] روش

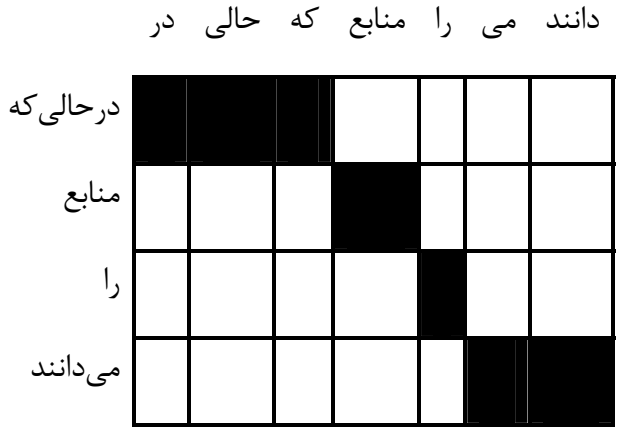
[1] Phrase-based

پیش‌پردازشی برروی داده‌های آموزش ارائه کرده است که به‌عنوان بخشی از ابزار متن باز Moses [۳۷] که براساس عبارت، ترجمه را انجام می‌دهد، اعمال شده است [۸].

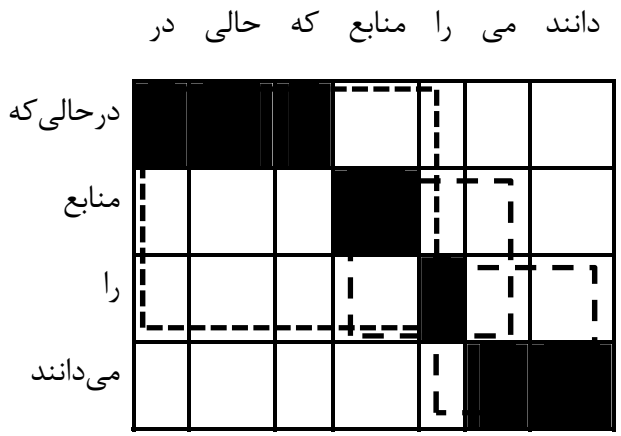
نحوه مدل‌کردن ترجمه براساس عبارت به‌این صورت است که ابتدا جمله ورودی به عبارات تقسیم می‌شود. به‌طوری‌که یک عبارت یک سلسله از کلماتی است که از نظر نحوی تغییری نکرده‌اند. نمونه‌ای از این تطبیق‌ها را در جدول ۴-۱ آمده است. عبارت‌ها سپس به‌صورت مجزا به زبان مقصد برطبق جدول ترجمه عبارات، ترجمه می‌شوند (شکل ۴-۱). جدول ترجمه عبارات با استفاده از تطبیق کلمه‌ها در متن‌های موازی تعیین می‌شوند [۷].

جدول ۴-۱ . نمونه‌ای از جدول تطبیق براساس عبارت

در حالی که منابع ↔ درحالی‌که منابع
را می دانند↔ را می‌دانند
در حالی که منابع را ↔ درحالی‌که منابع را
در حالی که منابع را می دانند ↔ درحالی‌که منابع را می‌دانند

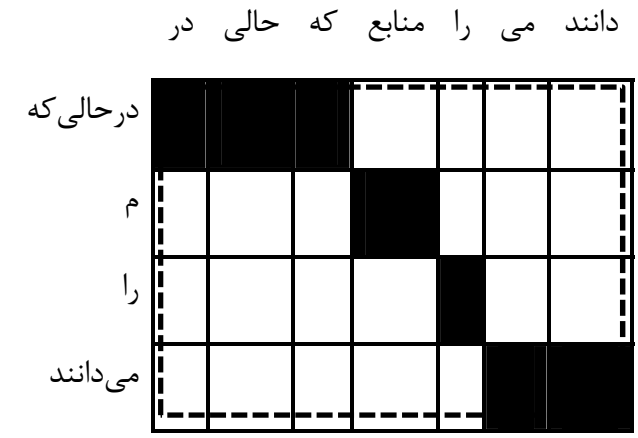


الف



ب

ب



ج

شکل ۴-۱. تطبیق براساس متن موازی ویرایش‌شده-ویرایش‌نشده فارسی. الف تا ج مراحل تطبیق عبارت را نشان می‌دهند.

۴.۲ نگاهی بر ترجمه براساس عبارت سلسله مراتبی

ترجمه براساس عبارت سلسله مراتبی[1] [۱۸] تعریف ترجمه براساس عبارت را گسترش داده است، به این‌صورت که فضای خالی میان کلمات یک عبارت را مدل کرده است. بنابراین برای مدل کردن فضای بین کلمات یک عبارت نیاز به استخراج گرامر مستقل از متن همگام[2] دارد. این گرامر از متن‌های موازی استخراج می‌شود و شامل خصوصیات مرتبط با متن هستند.

در ترجمه واقعی، استخراج گرامرها از متن‌های آموزشی بزرگ، بسیار بزرگتر از آن هستند که در حافظه جای گیرند. [۳۸] در این موارد، محاسبه ویژگی از لحاظ زمانی بسیار هزینه‌بر است و قوانین استخراج بسیار بزرگی جهت رخدادهای نسبی با احتمال ترجمه P(f|e) و احتمال ترجمه P(e|f) [۲۸] باید محاسبه شوند. بعد از گام‌های استخراج اگر داده‌های آموزش تغییر کنند، باید گرامرها دوباره استخراج شوند و این از لحاظ زمانی بسیار هزینه‌بر است. به‌خصوص زمانی‌که تمرکز برروی نشانه‌گذاری و قطعه‌بندی کلمه است.

برای حل این مشکل، باید تمام قوانین زیرمجموعه‌های موجود استخراج شوند. به‌خصوص زمانی‌که مدل [۳۹] دنبال می‌شود و از آرایه پسوندی[3] برای استخراج قوانین جملات انتخاب شده زبان مبدا استفاده می‌شود. به‌طوری‌که خروجی‌ها در پهنه کوچکتری نسبت به داده‌های آموزش صورت می‌گیرد [۳۸, ۴۰]. [۴۱, ۴۲] نشان دادند که آرایه پسوندی به ترجمه براساس عبارت این اجازه را می‌دهد که گزینه‌های ترجمه را برای مدل ترجمه در زمان اجرا محاسبه کند. زیرمجموعه‌ای از پیشامدهای داده شده عبارات، برای محاسبه این احتمالات ترجمه استفاده شده‌اند. با دسترسی به متن هدف و داده‌های تطبیق کلمه، ترجمه‌های عبارات و پارامترهای مرتبط با مدل می‌تواند در زمان اجرا محاسبه شود. [۳۹] نشان داد که

[1] Hierarchical phrase-based
[2] Synchronous Context Free Grammar (SCFG)
[3] Suffix array

عبارت سلسله مراتبی می‌تواند از آرایه پسوندی در زمان اجرا استفاده کنند [۸]. آرایه‌های پسوندی ساختار داده فشرده‌ای از متن دارند که اجازه می‌دهد تطبیق الگو کارایی را در کل متن داشته باشد [۴۳].

[۴۱, ۴۲, ۴۴] مزیت استفاده از ساختار داده آرایه پسوندی برای پیاده‌سازی مدل‌های براساس عبارت در زمان اجرا اضافه شده است [۸]. [۴۴] با استفاده از آرایه پسوندی استخراج گرامر SCFG برای Hiero [۱۸] سریع‌تر شده است. Hiero [۱۸] مدل ترجمه SCFG خود را در گام‌های پیش‌پردازش مشابه آنچه که گفته شد، پیاده‌سازی کرده است [۸]. عبارات سلسله مراتبی پیچیده با استفاده از Query_Intersect function (برای مطالعه بیشتر به منبع [۸] مراجعه شود) ساخته می‌شود.

Moses [۳۷] یک دیکدر مبتنی بر عبارت است و با تخمین‌های تکرارهای مشابه از جمله‌های تطبیق داده شده در متن‌های موازی، دیکد را انجام می‌دهند.

دیکدر مبتنی بر عبارت می‌تواند به راحتی عبارات مورد نظر را در آرایه پسوندی پیدا کند، اما دیکدرهای سلسله مراتبی در مواجه با عبارات غیرپیوسته که با فاصله آمده‌اند، با مشکل مواجه است. در مواجه با چنین عباراتی یک الگوریتم افزایشی برای ساخت یک درخت خاص [۴۵] جهت ارائه گرامر ترجمه SCFG ارائه شده است [۴۴]. جمله مبدا با استفاده از این الگوریتم یک درخت پیشوندی[1] را با کمک ارتباطات پسوندی تمام عبارات پیوسته مبدا می‌سازد. از آرایه پسوندی مبدا برای جست‌وجو و پیداکردن ترجمه‌های پیوسته عبارات استفاده می‌کند. Joshua decoder [۴۶] یک ابزار متن باز براساس عبارت سلسله مراتبی است.

Joshua decoder [۴۶] تکنیک [۳۹] را پیاده‌سازی کرده است.این ابزار این امکان را می‌دهد که ترجمه جملات را با استفاده از متن‌های موازی، بدون اجباری جهت استخراج SCFG قبل از دیکد کردن انجام دهد. مزیت استفاده از این پیاده‌سازی این است که هر متن ورودی می‌تواند دیکد شود (برای live demos

[1] Prefix

یاreal world use مناسب است) و داده‌ها به فضای داده‌ای کمتری نسبت به جداول عبارات نیاز دارند و باعث می‌شود که مدت زمان اجرا پایین بیاید و دیکد و ترجمه‌های عبارات به سرعت اتفاق بیافتد [۸].

الگوریتم هایی که در Joshua decoder استفاده شده اند شامل الگوریتم‌های CKY chart-parsing ، ادغام مدل زبانی n-gram، حرص beam و Cube، الگوریتم استخراج K–بهترین، الگوریتم آموزش کمترین میزان خطا [38] و ماژول‌های مربوط به آن. [۴۷]. Joshua تمام الگوریتم‌های مورد نیاز SCFG [۱۸] را پیاده سازی کرده است [۳۸].

Chart-parsing برای محاسبه اولین بهترین و K–بهترین متن داده‌شده در الگوریتم CKY به‌کار گرفته می‌شود. به‌خصوص زمانی‌که الگوریتم دیکد یک چارت داشته باشد. آرایه‌های چارت از سلول تشکیل شده‌اند و هر سلول به نوبه خود فرایند تجزیه را با عناصر آن انجام می‌دهد. فرایند تجزیه به‌صورت تکراری با اعمال قوانین استنباط، آیتم‌های جدید را استدلال می‌کند. این آیتم‌ها به سلول چارت مناسبش اضافه می‌شود. آیتم‌ها شامل اشاره‌گری نیز هستند که به آیتم‌های قبلی اشاره می‌کنند و می‌توان با استفاده از آن‌ها K–بهترین را استخراج کرد[۳۸].

حرص Beam و Cube، هرس کردن سختگیرانه به ایجاد یک دیکدر جهت محاسبه SCFG برای کلمات متفاوت زبان هدف نیاز دارد. در Joshua دو روش هرس کردن Beam و Cube درنظرگرفته شده است[۳۸].

استخراج K-بهترین برای هر جمله زبان مبدا، الگوریتم char-parsing را برای ابرگراف به کار می‌گیرد. با استفاده از الگوریتم K-بهترین، K بهترین جواب را از ابرگراف استخراج می‌شود [۳۸].

ابزار Joshua استخراج گرامر آرایه پسوندی و آموزش براساس نرخ خطای کمینه[1] را پیاده سازی کرده است. این ابزار از تکنیک‌های پردازش موازی و توزیع شده برای داده های با حجم بالا

[1] Minimum Error Rate(MERT)

استفاده می کند [۳۸]. برای اطلاعات بیشتر در مورد پردازش موازی و توزیع شده به منبع [۴۸] مراجعه کنید. برای مطالعه بیشتر در مورد نحوه ساختار داده Joshua به منبع [۸] مراجعه کنید.

۴.۳ نگاهی بر گرامرهای مستقل از متن همگام

به‌دلیل آنکه گرامرهای مستقل از متن همگام[1] یک فرمول تئوری پایه است که اجازه می‌دهد نحو در ترجمه به‌کارگرفته شود، SCFG در ترجمه ماشینی استفاده شده است. در عمل تعداد بسیاربسیار زیادی ازترجمه‌های جایگزین و اشتقاق و احتمالات تعیین‌شده وجود دارد که به ما اجازه انتخاب بهترین ترجمه را برطبق مدل می‌دهند و جهت کاهش این فضای جست‌وجو، فقط قسمتی که بیشترین احتمال ترجمه را دارد، گسترش می‌یابد.

SCFG به صورت چندتایی زیر درنظرگرفته شده است:

$(N, S, T_\sigma, T_\tau, G)$

به‌طوری‌که N مجموعه‌ای از سمبل‌های غیرپایانه گرامر است، $S \epsilon N$ سمبل هدف، T_σ و T_τ به ترتیب سمبل پایانه‌ای منبع و مقصد کلمه‌ها است و G مجموعه‌ای از قوانین تولید گرامر است [۲۶]. هر قانون G به فرم زیر است:

$X \rightarrow < \alpha , \gamma, \sim >$

به‌طوری‌که $X \epsilon N$ یک سمبل غیرپایانه، α یک سلسله‌ای از سمبل‌های $N \cup T_\alpha$ است. γ یک سلسله از سمبل‌های $N \cup T_\alpha$ است. $\sim$ یک تطبیق یک‌به‌یک از سمبل‌های غیرپایانه α و γ است.

زبان SCFG مجموعه‌ای از جفت رشته‌های مرتب (f, e) که توسط مدل ترجمه محفوظ نگه داشته می‌شود. هر کاندید e بوسیله اعمال یک سلسله از قوانین تولید شده $(r_1 \dots r_n)$ ایجاد می‌شود [۲۶]. در منابع [۴۹-۵۲] و [۴, ۵۳-۵۶] در مورد PSCFG ها بحث شده است.

[1] Synchronous Context-Free Grammar(SCFG)

در Joshua از ابزار Thrax برای استخراج گرامر استفاده شده است. Thrax یک استخراج‌کننده گرامر مستقل از متن همگام است. Thrax هم گرامرهای سلسله مراتبی[18] و هم گرامرهای ترجمه ماشینی Syntax-augmented [57] را استخراج می‌کند. این استخراج‌کننده گرامر، برای بالابردن کارایی از سیستم توزیع‌شده استفاده می‌کند. برای این‌کار از Apache Hadoop استفاده کرده است. Joshua از Thax برای استخراج گرامرهای[۱۸] Hiero و[57]SAMT استفاده می کند. [۲۶]

هرگاه بخواهیم از این گرامرها در زمان تست و یا tune استفاده کنیم، به‌جای آنکه گرامر تمام داده‌های تطبیق داده‌شده متن آموزش را درنظر گرفته بگیرد، با استفاده از روش Kishore Papineni زیرمجموعه‌ای از داده‌های آموزش را که شامل جملات مفیدی هستند، را انتخاب می‌کند. روال کار انتخاب جملات مناسب به این‌صورت است که، برای متن‌های Tune و تست که برروی آن‌ها ترجمه صورت می‌گیرد، تمام n-gramها (تا ۱۰) را در نقشه ω جمع می‌کنند. مقدار اولیه این مقادیر صفر درنظرگرفته‌شده است. روال به ترتیب داده‌های آموزش ادامه پیدا می‌کند، برای هر جفت جمله مبدا به مقصد، اگر هر n-gram که در جمله مبدا قرار داد مقدار ω آن کمتر از k باشد، جمله انتخاب می‌شود. وقتی یک جمله انتخاب می‌شود، مقدار هر n-gram جملات مبدا آن در ω به مقدار رخدادهای آن درمتن مبدا افزایش می‌یابد[۳۸]. در Joshua از همین تکنیک برای انتخاب زیرمجموعه‌ای از داده‌های آموزش استفاده شده است.

۴.۴ ترجمه براساس نحو

مدل String-to-tree، ترجمه براساس نحو[1] [۵۸] را معرفی کردند. سیستم ترجمه مقدار احتمال $P(f|\varepsilon)$ را محاسبه می‌کند، به طوری‌که در آن ε درخت تجزیه‌کننده زبان هدف و f جمله ورودی است.

[1] Syntax-based

در این ترجمه به هر رشته زبان مقصد یک درخت تجزیه‌کننده نحوی[1] اختصاص داده می‌شود. [۵۹] یک دیکدر با اضافه‌کردن ترجمه عبارت به مدل نحوی ارائه کرده‌است.

مدل‌های String-to-tree در ترجمه براساس عبارت و ترجمه براساس عبارت سلسله مراتبی به صورت زیر تعریف می‌شوند. غیرپایانه‌ها درسیستم‌های ترجمه براساس عبارت و ترجمه براساس عبارت سلسله مراتبی با "X" نشان داده می‌شوند:

X --> [source] X_1 ||| [target] X_1

در مدل‌های نحوی، خروجی تجزیه‌گر جمله‌ها، به غیرپایانه‌ها این امکان را می‌دهد که برچسب زبان‌شناختی داشته باشند، مانند "ADJ" یا "NOUN ":

ADJ --> [source] ||| [target]

NOUN --> [source] ||| [target]

غیرپایانه‌ها در درخت CFG برچسب زبان‌شناختی می‌گیرند. این مدل‌ها با نام مدل‌های String-to-tree شناخته می‌شوند.

یک گرامر اضافه شده نیز وجود دارد که اگر هیچ قانونی جهت اعمال در زمان دیکد وجود نداشته باشد، می‌تواند به سادگی دو زیردرخت راباهم ترکیب کند. این قانون با نام قوانین glue معرفی شده‌اند. غیرپایانه‌ها در قوانین glue به‌صورت زیر نشان داده می‌شوند:

S --> <s> ||| <s>

S --> X_1 </s> ||| X_1 </s>

S--> X_1 X_2 ||| X_1X_2

گرامر glue فقط بر روی تست و Tune اجرا می‌شود.

[1] Syntax parser

۴.۴.۱ Hiero و SAMT

گرامر [18] Hiero یک نوع SCFG با فقط یک سمبل غیرپایانه است که به صورت معمول با X برچسب‌گذاری می‌شود. گرامر [18]Hiero به‌صورتی‌که در ادامه آمده است، می‌تواند از متن‌های موازی که کلمات آن در جفت جملات تطبیق داده شده‌اند، استخراج شود [۲۶]:

اگر (f_i^j, e_k^l) یک زیرعبارت از یک جفت جمله باشد می‌گوییم با جفت تطبیق داده شده سازگار است، اگر هیچ‌کدام از کلمات در f_i^j به کلمه‌ای خارج از e_k^l تطبیق داده نشده باشد و برعکس. زیرعبارت سازگار ممکن است به‌عنوان یک قانون SCFG استخراج شده باشد. به‌علاوه، اگر یک زیرعبارت سازگار شامل یک زیرعبارت سازگار دیگر در خودش باشد، یک قانون سلسله مراتبی از آن استخراج می‌شود. این کار با جایگذاری قسمت کوچکتر با یک غیرپایانه انجام می‌شود [۲۶].

سیستم Hiero با استفاده از SCFGها برای ترجمه ماشینی آماری توسط Chiang [۴۰] معرفی شده است. این سیستم روش ترجمه براساس عبارت را در ترجمه ماشینی آماری گسترش داده است. این کار برروی عباراتی که شامل فاصله هستند، گسترش یافته است. Chiang به توصیف چگونگی استخراج مستقیم عبارت سلسله مراتبی با گسترش متدهای استانداردی [۲۸, ۶۰-۶۲] که جفت عبارات را تطبیق می‌دهد، پرداخته است [۴۷].

به جای استفاده از تمام قدرت فرمول SCFGها، سیستم Hiero از یک گرامر ساده استفاده می‌کند. در این سیستم با یک سمبل غیرپایانه X مدل‌های مبتنی بر عبارت را جهت اینکه بین کلماتشان فاصله باشد، گسترش داده است [۴۷]. سیستم Hiero به‌صورت تکنیکی یک رهیافت مبتنی‌بر گرامر در ترجمه ماشینی است. اما هیچ‌گونه اطلاعات نحوی در گرامر خود به همراه ندارد.

گرامر SAMT(Syntax Augmented Machine Translation)[63][۵۷] شبیه به گرامر Hiero است. به‌جز اینکه مجموعه سمبل غیرپایانه بزرگتر است و برچسب‌های آن از یک درخت تجزیه در منبع و مقصد

استخراج شده‌اند [۴۷, ۶۴]. SAMT [۵۷] شامل گرامرهای glue و گرامرهای سلسله مراتبی است [۲۶].در این سیستم به جای آنکه به تمام جفت عبارت استخراج شده برچسب"X"زده شود، همان کاری که Hiero انجام می‌دهد، از برچسب‌های زبان‌شناختی درخت تجزیه استفاده می‌شود.

فصل پنجم: رهیافت‌های آموزش برای ترجمه ماشینی آماری

۵.۱ مدل تولیدکننده[1]

۵.۱.۱ الگوریتم انتظار بیشینه

آموزش تطبیق کلمه‌به‌کلمه با استفاده ازالگوریتم انتظار بیشینه[2] صورت می‌گیرد. الگوریتم EM به صورت زیر کار می‌کند [۱۱, ۳۰]:

۱-مدل را با استفاده ازتوزیع یکنواخت مقداردهی اولیه می‌کند. به این معنا که، هرکلمه ورودی f در متن موازی، فرض شده است که با احتمال یکسانی به هریک از لغات زبان مقصد e ترجمه می‌شود.

۲-مدل به داده‌ها اعمال می‌شود (گام Expectation). به این معنا که، توزیع احتمال فعلی جهت تطبیق کلمات اعمال می‌شود.

۳-آموزش مدل با استفاده ازداده (گام Maximization). به این معنا که، شمارشی راکه مدل اعمال شده به داده‌ها جهت تخمین یک مدل اصلاح‌شده انجام داده است را، جمع می‌کند. یعنی یک مدل شامل فرکانس واقعی ازتطبیق‌های متفاوت است.

۴-گام ۲تا۳تازمانی که همگرا شود ادامه می یابد.

جزئیات ریاضی الگوریتم EM در [۱۱] آمده است.

آموزش تمام مدل‌های تطبیق کلمه با الگوریتم EM با استفاده از متن‌های آموزش موازی $(f^{(s)}, e^{(s)}); \mathrm{s} = 1, \ldots, \mathrm{S}$ انجام می‌شود (معادله۱-۵) :

$$P(e|f;e,f) = \frac{\sum_{(e,f)} c(e|f;e,f)}{\sum_e \sum_{(e,f)} c(e|f;e,f)} \qquad (۵-۱)$$

به همین‌صورت، احتمالات تطبیق وfertility می‌توانند تخمین زده شوند. مدل‌های IBM Model، IBM Model 2 و HMM فرم ریاضیاتی ساده‌ای دارند. بنابراین الگوریتم EM می‌تواند به صورت دقیق اجرا

[1] Generative
[2] Expectation Maximization (EM)

شود، به این معنا که درگام Expectation این امکان وجود دارد که تمام تطبیق‌ها به صورت کارائی درنظرگرفته شوند. برای HMM این کار را با استفاده از الگوریتم Baum-Welch [۶۵] انجام می‌شود[۲۱]. به‌دلیل آنکه هیچ راه کارائی برای اجتناب از مجموع تمام تطبیق‌ها در الگوریتم EM وجود ندارد، شمارش فقط برروی یک مجموعه ازتطبیق‌های قابل اعتماد انجام می‌شود. برای IBM Model 3 ، IBM Model 4 و IBM Model 5 شمارش فقط برروی تعدادی ازتطبیق‌های خوب صورت می‌گیرد. برای آنکه آموزش سریع صورت بگیرد فقط باید بخش کوچکی از تطبیق‌ها را درنظربگیریم. می‌توان سه راه ممکن[۲۱] برای بدست آوردن زیرمجموعه‌ای ازاین تطبیق‌ها را درنظر گرفت:

ساده‌ترین راه این است که با استفاده ازآموزش viterbi می‌توان بهترین تطبیق راپیداکرد. محاسبات تطبیق Viterbi به خودی‌خود بسیار زمانبر است و از متدی که در [۱۲] توصیف شده است، استفاده می‌کند.

در [۶۶] پیشنهاد شده است که از تطبیق‌های همسایگی (به‌عنوان مثال، تطبیق‌های متفاوتی با یک جابه‌جائی یا حرکت) بهترین تطبیق قابل دسترس، استفاده شود.

در [۱۲] یک مجموعه بزرگتر از تطبیق‌ها استفاده شده است. این مجموعه شامل تطبیق‌های"pegged" است. در این مجموعه، مدل‌های متفاوت برروی داده‌های یکسان متوالی جهت آموزش به‌کار گرفته شده‌اند. مدل‌های متفاوت با جانشینی برروی داده‌های یکسان آموزش داده شده‌اند. به‌طوری که، مقدارهای پارامتر نهایی یک مدل ساده‌تر به‌عنوان نقطه شروع برای یک مدل پیچیده‌تر نگه‌داشته می‌شود. استفاده از HMM به جای IBM Model 2 کیفیت راه انداز طی IBM Model 4 وIBM Model 5 به‌صورت چشمگیری افزایش می‌دهد.

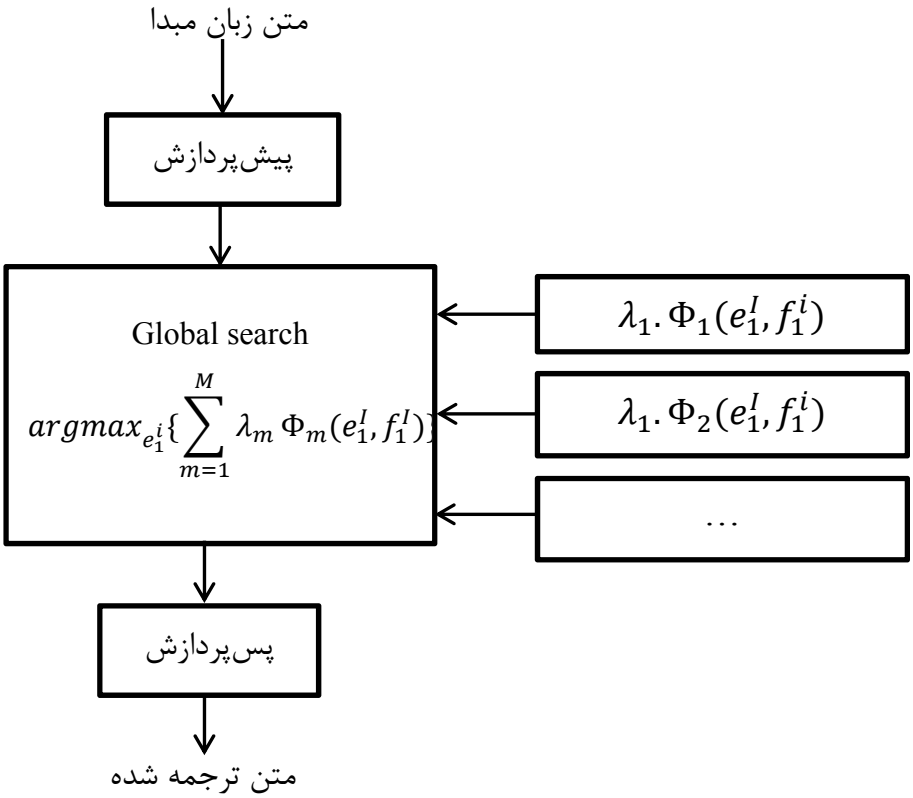
۵.۲ یادگیری discriminative

در این بخش، می‌خواهیم کارائی ترجمه را بهینه کنیم. ما از متدهای آموزش ماشینی جهت تمایز قائل‌شدن میان ترجمه‌های خوب و بد استفاده می‌کنیم و سپس مدل خود را با آن تطبیق می‌دهیم تا یک ترجمه خوب را به سایر ترجمه‌ها الویت دهد.

روال سیستم ترجمه ماشینی به این‌صورت است که به جست‌وجو ترجمه کاندید با بیشترین امتیازی که مولفه‌های متفاوت مدل به آن اختصاص داده‌اند، می‌گردد و این کاندیداها را به‌عنوان ترجمه فرضی بازمی‌گرداند. لزومی به احتمالاتی بودن هرکدام از این مدل‌ها نیست و در عوض باید مطابق با ویژگی‌های زوج.(ترجمه کاندید، جمله ورودی) باشند [۶۷].

یک ترجمه ممکن برای یک جمله با نام "ترجمه کاندید" شناخته می‌شود و ویژگی‌هایی باخود به همراه داشته باشد. این ویژگی‌ها از یک سری خصوصیات استخراج می‌شوند و وزن‌های اختصاص داده شده به این ویژگی‌ها به درجه اهمیت آن‌ها اشاره دارد. وظیفه ترجمه ماشینی این است که آموزش این وزن‌ها را به خوبی انجام دهد. زمانی‌که مدل را به یک جمله ورودی در زبان مقصد اعمال کنیم، مقدارهای ویژگی ترجمه کاندید برطبق مدل وزن‌دهی شده‌اند و ترکیب می‌شوند تا یک مقدار کلی را تعیین کنند. بیشترین امتیاز کاندید، بهترین ترجمه برطبق مدل است. که به این مدل یادگیری Discriminative ویا Tunning [۱۱] گفته می‌شود.

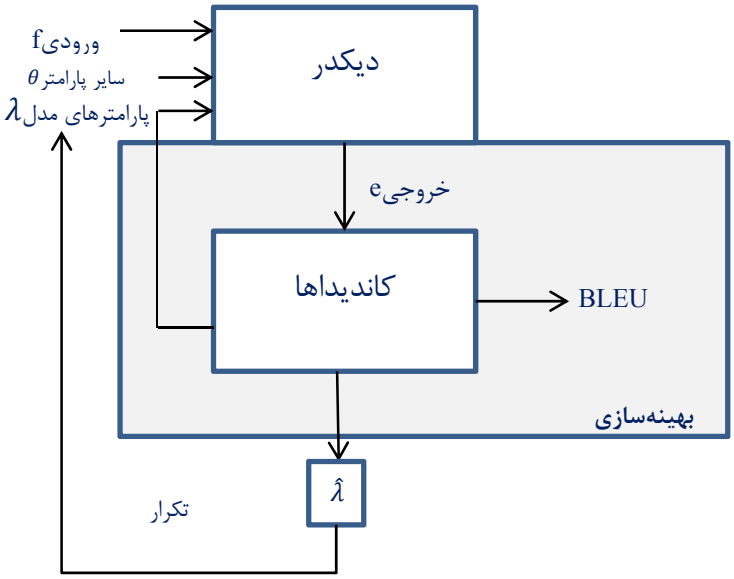
پیرو [۶۸]، این مدل از رهیافت کانال نویزی فاصله گرفته و از مدل کلی‌تر Log-linear استفاده کرده است. مولفه‌های سیستم ترجمه ماشینی به روش Log-linear باهم ترکیب می‌شوند. مدل Log-linear با مدل کانال نویزی در این تفاوت دارد که می‌تواند براساس یک تعداد مشخص از ویژگی‌ها ترجمه را امتیازدهی کند. در این‌صورت می‌تواند به عنوان یک مدل کلی‌تر توصیف شود. در شکل۵-۱ [۶۹] نمایش شماتیکی از مدل Log-linear آمده است.



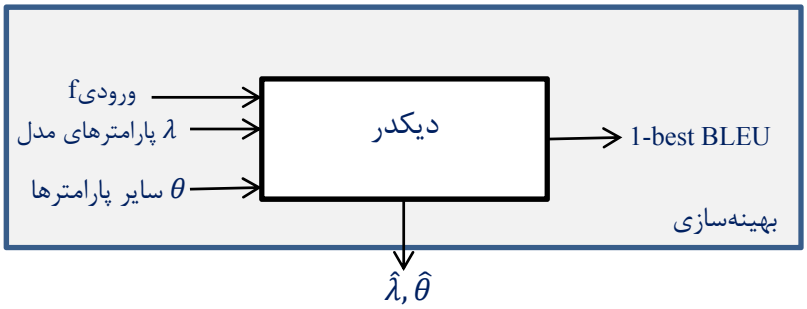
شکل۵-۱. نمایش شماتیکی از سیستم Log-linear

۵.۲.۱ مدل Log-linear (فاز MERT)

روند مدلLog-linear به این‌صورت است که باید یک وزن به هریک ازخصوصیات اختصاص دهیم [۳۰]. این مدل یک شاهد تجربی برای تاثیر تنظیم وزن‌ها در محاسبه معیار ارزیابی ترجمه ماشینی است. این فرایند به عنوان فاز MERT (Minimum Error Rate Training) در آموزش‌های موازی یک سیستم ترجمه ماشینی ، شناخته می‌شود. [70] روند الگوریتم MERT را در مقایسه با جست‌وجو مستقیم به ترتیب در شکل ۵-۲ و شکل ۵-۳ نشان داده شده است.



شکل ۵-۲ . نمایش کلی سیستم MERT



شکل ۵-۳ . نمایش شماتیک سیستم جست‌وجو مستقیم برای پیداکردن بهترین ترجمه

به‌صورت ریاضی، از متد Log-linear در سیستم ترجمه استفاده شده است. به دنبال خروجی (e) که بیشترین احتمال را داشته باشد. به طوری‌که، مجموع تابع‌های خصوصیات $\Phi_m(e,f)$ بیشینه شود (معادله ۲-۵) [۷۱] :

$$\hat{e} = \arg\max_e P(e|f) \qquad (۵-۲)$$

$$= \arg\max_e \sum_{m=1}^{M} \lambda_m \Phi_m(e,f)$$

جمله داده شده در زبان مبدا f باشد و سیستم ترجمه ماشینی آن را به e ترجمه کند که بهترین ترجمه در زبان هدف است. این کار با انتخاب جمله هدف با بیشترین احتمال شرطی برروی جمله مبدا انجام می‌دهد.

احتمال پسین $P(e|f)$ با استفاده از مدل Log-linear می‌تواند مدل شود. این مدل یک جفت جمله (e,f) را به یک بردار ویژگی $\Phi(e,f) = \{\Phi_1(e,f), \dots, \Phi_M(e,f)\}$.مرتبط می‌کند و سپس یک امتیاز را به آن اختصاص می‌دهد (معادله ۳-۵) [۶۷]:

$$s_\Lambda(e,f) \stackrel{\text{def}}{=} \Lambda.\Phi(e,f) = \sum_{m=1}^{M} \lambda_m \Phi_m(e,f) \qquad (۵-۳)$$

به طوری‌که $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_M\}$ یک بردار وزن برای خصوصیات M است. حالا احتمال پسین به صورت معادله ۴-۵ تعریف می‌شود [۶۷]:

$$\mathrm{P(e|f)} \stackrel{\text{def}}{=} \frac{\exp(\mathrm{s}_\Lambda.\Phi(\mathrm{e,f}))}{\sum_{\acute{\mathrm{e}}} \exp(\mathrm{s}_\Lambda.\Phi(\acute{\mathrm{e}},\mathrm{f}))} \qquad (۵-۴)$$

و بنابراین ترجمه ماشینی ، بهترین ترجمه را انتخاب می‌کند [۶۷]:

$$\hat{\mathrm{e}} = \mathrm{argmax_e P(e|f)} = \mathrm{argmax_e}\ \frac{\exp(\mathrm{s}_\Lambda.\Phi(\mathrm{e,f}))}{\sum_{\acute{\mathrm{e}}} \exp(\mathrm{s}_\Lambda.\Phi(\acute{\mathrm{e}},\mathrm{f}))} \qquad (۵-۵)$$

$$= \mathrm{argmax_e}\ \ \mathrm{s}_\Lambda(\mathrm{e,f})$$

۵.۲.۱.۱ ویژگی های استخراج شده از گرامرهای مستقل از متن همگام برای آموزش به روش کمترین میزان خطا

آموزش به روش کمترین میزان خطا[1] براساس SCFG وزن داده شده، کار می‌کند. ساختار ابتدایی SCFG به این‌صورت است که قوانین را به همراه جفت تطبیق‌های سمت راست آن‌ها بازنویسی می‌کند:

$X \rightarrow < \gamma, \alpha, \beta >$

[1] Minimmum Error Rate Trainning (MERT)

به‌طوری‌که X یک غیرپایانه، γ و α هر دو رشته‌ای از پایانه‌ها و غیرپایانه‌ها هستند و $\sim$ یک تطابق یک-به-یک میان رخدادهای غیرپایانه‌ای در γ و رخدادهای غیرپایانه‌ای α است. بازنویسی قوانین با جفت اتصالات سمبل‌های Start آغاز می‌شود. در هر گام، دو غیرپایانه با اندیس یکسان با استفاده از دو مولفه یک قانون نوشته می‌شوند. به‌طوری‌که، به هیچ‌کدام از سمبل‌هایی که در حال حاضر وجود دارند، متصل نشده‌اند. این مدل در فرم معادله ۶-۵ نشان داده شده است:

$$w(X \rightarrow < \gamma, \alpha >) = \prod_i \Phi_i (X \rightarrow < \gamma, \alpha >)^{\lambda_i} \qquad (۵-۶)$$

به‌طوری‌که Φ_i ویژگی‌هایی است که در قوانین تعریف شده‌اند. γ سلسله‌ای از رشته‌های غیرپایانه و پایانه مبدا است، α سلسله‌ای از رشته‌ای از غیرپایانه‌ها و پایانه‌های هدف است و $\sim$ تطبیق یک‌به‌یک برای غیرپایانه‌های دیده‌شده در γ و α است.

هفت ویژگی به‌کارگرفته‌شده در ابزار آموزش استفاده شده در این پایان‌نامه، Z-MERT [۶۷]، در ادامه مورد بحث قرار گرفته است [۴۰, ۷۲].

- قانون مدل ترجمه$P(\gamma|\alpha)$ و $P(\alpha|\gamma)$ ، که دومی در رهیافت کانال نویزی وجود نداشته است اما قبلا توسط [۶۸] به‌کار گرفته‌شده است.
- وزن‌های لغوی $P_w(\gamma|\alpha)$ و $P_w(\alpha|\gamma)$ [۲۸, ۷۳]. به‌طوری‌که، چقدر ترجمه کلمه‌های f به کلمه‌های e خوب است.
- مدل زبانی مربوط به $P_{LM}(T)$ n-gram
- جریمه کلمه
- Word's Arity

۵.۲.۱.۲ روش برآورد پارامتر براساس متد Och

در این قسمت به بررسی نحوه تنظیم بردار وزن Λ پرداخته شده است. [۳۳] با استفاده از متریک‌های ارزیابی نشان داده است که با انتخاب مناسب بردار وزن، کارایی سیستم برروی بیشتر مجموعه‌های آموزشی، افزایش می‌یابد. سطح خطا در این روش smooth نمی‌شود. به این معنا که تکنیک‌های بهینه‌سازی براساس گرادیان نمی‌تواند مورد استفاده قرارگیرد. همچنین یک جست‌وجو grid با تکرار بهینه‌سازهای خطی گزینه خوبی نیست. به دلیل آنکه تابع جهت ارزیابی در نقطه داده‌شده $p\epsilon\Re^M$ هزینه‌بر است. به‌طوری‌که به امتیازدهی دوباره مجموعه کاندید برای هر جمله جهت پیدا کردن ترجمه اول-بهترین در نقطه p نیاز خواهد داشت. Och روشی بسیار کارا برای بهینه‌سازی پیشنهاد کرد.

فرض شده است بهینه‌سازی خطی در بعد dام انجام می‌شود. به این معنا که یک بردار وزن $\Lambda = \{\lambda_1, \dots, \lambda_d, \dots, \lambda_M\}$ داده شده و بدنبال یک بردار وزن بهینه است. به‌طوری‌که، dامین بعد بهینه شود و سایر ابعاد ثابت نگه‌داشته شوند. یک جمله هدف f را درنظر بگیرید و مجموعه کاندید برای f به‌صورت $\{e_1, \dots, e_k\}$ باشد. همان‌طورکه در **Error! Reference source not found.** ۵-۵ گفته شده، اولین بهترین کاندید از Λ، ترجمه‌ای است که بیشترین $s_\Lambda(e_k, f)$ را داشته باشد. به‌طوری‌که معادله ۳-۵ به صورت $\Phi(e,f) = \{\Phi_1(e,f), \dots, \Phi_M(e,f)\}$ تعریف می‌شود، $\sum_{m=1}^{M} \lambda_m \Phi_m(e_k, f)$ بردار ویژگی است. می‌توان آن را به صورت مجموع نوشت (معادله۷-۵) [۶۷]:

$$s_\Lambda(e_k, f) = \lambda_d \Phi_d(e_k, f) + \sum_{m \neq d} \lambda_m \Phi_m(e, f) \qquad (۵-۷)$$

دومین عبارت با درنظرگرفتن λ_d ثابت است و بنابراین $\Phi_d(e_k, f)$ ثابت است. اگر آن دو مقدار به‌صورت $offest_\Lambda(e_k)$ و $slop(e_k)$ دوباره نامگذاری شوند، معادله به‌صورت معادله ۸-۵ می‌شود[۶۷]:

$$s_\Lambda(e_k, f) = slop(e_k)\lambda_d + offest_\Lambda(e_k) \qquad (۵-۸)$$

این معادله یک خط است و بنابراین زمانی‌که λ_d را تغییر دهیم، امتیاز یک کاندید به صورت خطی تغییرمی‌کند. به این معنا که، اگر امتیاز را برای هر ترجمه کاندید نسبت به λ_d رسم شود، این کاندید یک خط را نشان می‌دهد. اگر خطوط را برای تمام کاندیدها رسم کنیم، قسمت بالائی خطوط بهترین کاندیدا برای هر مقداری از λ_d است [۶۷].

غیرهمواربودن سطح تابع خطا اختیاری نیست، اما به‌صورت قطعه‌ای در طول بعد λ_d خطی است. به این دلیل که خطا براساس ترجمه اولین-بهترین محاسبه می‌شود و یک تغییرکوچک در λ_d معمولا بالاترین کاندید را تغییر نمی‌دهد.

اگر بتوان این نقاط تلاقی را برای هر جمله تعیین کرد و سپس تمام آن‌ها را به عنوان یک مجموعه نقاط تلاقی باهم ادغام کرد، یک مجموعه کلی ازمقدارهای اساسی بعد λ_d داریم که با هر مقدار ترجمه اولین-بهترین برای جمله مبدا تغییر می‌کند.

به این معنا که اگر احتمال خطا را برای یک مقدار λ_d درست قبل از مقدارهای اساسی محاسبه کنیم، احتمالات برای مقدار λ_d درست بعد از هر مقدار اساسی می‌تواند به راحتی محاسبه شود: به راحتی احتمالات اولیه را برطبق تغییراتی که نقاط تلاقی کاندیداها نشان می‌دهند، تطبیق می‌دهد. در این روش نیازی به امتیازدهی دوباره کاندیداها نداریم و با پیمایش بعد λ_d با درنظرگرفتن فقط نقاط تلاقی می‌توان مقدار بهینه آن را پیداکرد. پیداکردن این مقدارهای اساسی یعنی پیداکردن نقاط تلاقی خطوط نمایش کاندیداها.

یک بخش دیگر از مساله باقی می‌ماند. فرض شده است که به مجموعه ترجمه‌های کاندید برای هر جمله ورودی، در دسترس است. چگونه این مجموعه به دست مارسیده است؟ می‌توان تلاش کرد تا تمام کاندیداهای ممکن برای پوشش تمام فضای جست‌وجو دسترسی داشته باشیم، اما این امکان وجودندارد و به‌هرحال این کار بسیار هزینه‌بر است. بنابراین نیاز به تخمین مجموعه کاندید داریم. برای این کار نیاز به

۳۰۰ کاندید عالی برطبق مقداراولیه بردار وزن داریم. اما این مجموعه تقریبا متمرکز است و معرف خوبی از فضای جست‌وجو نیست. درعوض، میان بردار وزن و تولید مجموعه کاندیداهای عالی، هر بار که ادغام مجموعه کاندید با مجموعه موجود انجام می‌شود، بهینه‌سازی صورت می‌گیرد. این پروسه تا زمانی‌که همگرا شود تکرار می‌شود و به کیفیت مجموعه کاندید تاکید دارد نه سایز آن [۶۷].

ماژول MERT [۳۳] برای ابزار Joshua پارمترهای وزن را بهبود می‌هد تا کارایی ترجمه را برروی مجموعه Tune با استفاده از یک الگوریتم ارزیابی خودکار همانند BLEU بیشتر کند. بهینه‌سازی شامل یک سری از بهینه‌سازهای خطی است که در طول ابعاد متناظر با پارامترها به کار گرفته شده است. جست‌وجو در میان یک بعد از متد کارا ارائه شده در [۳۳] استفاده می‌کند [۳۸].

هر تکرار در پیاده‌سازی MERT شامل بروزرسانی چندین وزن است که با استفاده از انتخاب حریصانه بعدی است که بیشترین منفعت را داشته است. هر بار که وزن‌ها بروزرسانی می‌شوند، مقدارهای قبلی نگه داشته می‌شوند [۳۸]. برای مطالعه بیشتر در مورد MERT به منبع [۶۷] مراجعه کنید.

فصل ششم: دیکد کردن

یک الگوریتم دیکد مناسب برای داشتن سیستم ترجمه ماشینی آماری[1] خوب، حیاتی است. کار دیکدر پیداکردن ترجمه‌ای است که بیشترین شباهت را با پارامترهایی که در زمان آموزش دیده است (ونیز فرمولی که این پارامترها را باهم ترکیب کرده است) داشته است. دیکدر به کار گرفته شده در این پایان‌نامه، با استفاده از تجزیه‌گر CYK+ و جست‌وجو دسته جمعی به دنبال بهترین ترجمه برای جمله ورودی f طبق معادله ۱-۶ است.

$$\hat{e} = e(\text{argmax}_{f(D)=f} P(D)) \qquad (۶-۱)$$

$P(D)$ مجموع وزن‌هایی است که به هر جمله ترجمه شده در زمان آموزش اختصاص داده شده است. تا زمانی‌که فضای ترجمه‌ها بزرگ باشد، الگوریتم‌های دیکد خاص فقط قادر خواهند بود که قسمتی از آن را جست‌وجوکنند. بنابراین ریسک ازدست دادن جواب‌های خوب وجود دارد [۹] .

پروسه دیکد کردن که یک نوع تجزیه‌کردن است، از یک الگوریتم برنامه‌نویسی پویا Cocke-Kasami-Younger(CKY) استفاده می‌کند. این الگوریتم بهترین استخراج را با استفاده از قوانین گرامری احتمالاتی پیدا می‌کند [۴۷]. برای هر قانون اگر سمت هدف فقط با یک درخت تجزیه پوشش داده شود، برچسب سازگارش را با سمبل غیرپایانه به آن قانون، اختصاص می‌دهیم. به‌عبارت دیگر، با فرم گسترش‌یافته آن به‌صورت $C_1 + C_2$، C_1/C_2 یا $C_2 \setminus C_1$ بر این تاکید می‌کند که سمت هدف دو همسایگی سازگار دارد و نشان‌دهنده این است که هر هدف از دو همسایگی تشکیل شده است. به ترتیب، C_1 که C_2 در همسایگی راست آن است و یا C_1 که C_2 در همسایگی چپ آن است. مزیت اصلی استخراج SCFGها این است که دارای اطلاعات زبان‌شناختی هستند. به همین دلیل فقط نیاز به تجزیه‌کردن فقط دریک سمت متن موازی دارند.

[1] Statistical Machine Translation (SMT)

۶.۱ الگوریتم تعمیم‌یافته CYK

الگوریتم CYK+ نوعی ازالگوریتم CYK است که به نام CYK+ نامیده شده است. به طوری که می‌تواند یک کلاس ازگرامرهای CF را که بسیار بزرگتر از کلاس گرامرهایCNF که توسط الگوریتمCYK استاندارد پیاده‌سازی شده است را تجزیه کند. این کلاس ازگرامرها که با نام nplcf (Non- Partially Lexicalized CF grammers) شناخته می‌شوند. درحالی‌که، کارائی الگوریتم تجزیه راحفظ می‌کند، اجازه می‌دهد که گرامرهای ساده‌تری (بامحدودیت کمتری نسبت به گرامرهای CNF) نوشته شود. بعلاوه، الگوریتم تجزیه با گرامرهای احتمالاتی سروکار دارد و قادراست که احتمالN-بهترین را با تفسیر نحوی برای هر بخش از جمله ورودی، حتی برای کلمات دیده‌نشده، تولیدکند. در انتها، CYK+ قادراست که قالب‌های مرکب را نیز درنظربگیرد (با ترکیب قوانین به‌صورت $X \rightarrow W1\ W2 \dots \dots \dots \dots \dots . Wn$).

الگوریتم CYK+ یک الگوریتم پائین به بالا است که برای گرامرهای مستقل‌ازمتن احتمالاتی استفاده می‌شوند. این الگوریتم قادراست [۷۴]:

۱. با تفسیرهای چندگانه جملاتی که شامل کلمات مرکب هستند، سروکار داشته باشد.

۲. استخراج تجزیه‌های N –محتمل‌ترین در زمان $O(n^3)$ و به‌صورت همزمان تمام تجزیه‌های هر قسمت از دنباله ورودی را به همراه احتمالشان، محاسبه کند.

۳. در مواجه به لغاتی که خارج ازدیکشنری هستند، مناسب است.

صراحتا استخراج تمام درخت‌های تجزیه نسبت داده‌شده به یک جمله ورودی داده شده یا هر زیررشته یک جمله ورودی را ممکن می‌سازد. هزینه این کار بسته به پیچیدگی گرامردارد، اما حتی درمواردی که

تعداد درخت‌های تجزیه به‌صورت نمایی هستند، چارت با استفاده از الگوریتمCYK+ از فضای محاسباتی با پیچیدگی $O(n^2)$ استفاده می‌کند.

این الگوریتم ازجدول CYK به‌عنوان یک چارت استفاده می‌کند. به‌طوری‌که، شامل آیتم‌های کلی مشابه است. این الگوریتم کاملا مشابه الگوریتم‌های پائین به بالا است [۵۱، ۷۵، ۷۶]. به صورت دقیق‌تر، براساس طرح‌واره تجزیه، می‌توان به این الگوریتم به‌عنوان یک کاهنده تجزیه به یک کلاس خاص از گرامر نگاه کرد (نسبت به گرامرهای CNF[1] که به طور معمول توسط CYK درنظر گرفته می‌شوند، محدودیت کمتری دارد).

همان‌طور که گفته شد، کلاس گرامری که توسط CYK+ پذیرفته می‌شود یک زیرکلاس CFG است. این کلاسnplCFG و شامل "non partially lexicalized" است. به این معناست که پایانه‌ها می‌توانند فقط در قوانینی که به‌صورت $X \rightarrow w_1 \dots w_n$ هستند، اعمال شوند. به طوری‌که X یک غیرپایانه و w_cها پایانه‌ها هستند. به‌صورت عملی این قوانین لغوی درگرامرها نوشته نشده‌اند، اما بخشی از لغت هستند. به دلیل آنکه پردازش "قوانین لغوی" فقط به گام مقداردهی اولیه محدود می‌شود، محدودیت کلاس nplCFG در روند الگوریتم تغییری ایجاد نمی‌کند. با این حال این الگوریتم به راحتی با هر گرامر CFG می‌تواند گسترش پیدا کند. به این الگوریتم می‌توان به عنوان اشتقاقی ازالگوریتم های پائین به بالا (با آیتم‌های عمومی‌شده و بدون پیش‌بینی) و یا می‌توان به عنوان اشتقاقی ازالگوریتم CYK (اما بدون اجرای گرامر باینری دینامیک) درنظرگرفت. همچنین می‌توان به عنوان گسترش‌یافته الگوریتم که در [۷۶] ارائه‌شده است، در نظر گرفت. ساختارداده پایه‌ای که در الگوریتم استفاده شده است، یک ماتریس با $\frac{n(n+1)}{2}$ سلول است. به‌طوری‌که n سایز رشته ورودی $w_1 \dots w_n$ است که باید تجزیه شود [۷۴].

[1] Chomsky Normal Form(CNF)

الگوریتم +CYK مشابه تمام الگوریتم‌های CYK، نه تنها تعیین می‌کند که رشته از لحاظ نحوی درست است یا نه، بلکه یک ارائه فشرده‌ای از تمام تجزیه‌ها برای تمام زیررشته‌های، رشته ورودی را تولید می‌کند. علاوه‌بر آن، استخراج تمام درختان تجزیه ممکن مربوط به یک جمله با این الگوریتم ساده است. مشکل افتادن در حلقه زمان تجزیه رشته ورودی این‌طور است که، یک گرامر در یک دور می‌افتد (پروسه استخراج ممکن است که درحلقه بی‌نهایت گیرکند). که این مشکل با اجتناب از عبور از یک لینک برای بیشتر از یک‌بار حل‌شده است. این محدودیت باعث می‌شود که کوچکترین درخت نحوی رادر زمان چرخه ایجاد شود.

با استفاده از طرح تجزیه و نشانه‌های[۷۷] الگوریتم +CYK به صورت سیستم تجزیه $< \mathcal{L}, \mathrm{H}, \mathrm{D} >$ برای هر گرامر nplcfg گسترش‌یافته است. به طوری‌که:

$\mathcal{L}$ دامنه است:

$$\{[A \rightarrow \alpha \blacksquare \beta, i, j] | A \rightarrow \alpha\beta \epsilon P, i \geq 1, j \geq 1\}$$

H مجموعه فرضیاتی به صورت زیراست:

$$\{[w, i, j] |\, w = w_j \ldots w_{j+i-1}, 1 \leq j \leq n\,, 1 \leq i \leq n - j + 1\}$$

D مجموعه استنتاجی ساخته شده ازاجتماع است:

$$\mathrm{D}^{\mathrm{Init}} = \{[\mathrm{w}, \mathrm{i}, \mathrm{j}] \vdash \mathrm{A}[\mathrm{A} \longrightarrow \mathrm{w}\blacksquare, \mathrm{i}, \mathrm{j}] |\, \mathrm{A} \rightarrow \mathrm{w}\epsilon\mathrm{P}\}$$

دلیل این‌که چرا محدود به کلاس گرامر nplcfg شده است، این است که، این الگوریتم این اجازه را می‌دهد که تعداد گام‌های اجرایی را کاهش دهد و مطابق با نوع گرامری که برروی آن‌ها کار شده است، عمل کند. کاهش به این حقیقت اشاره دارد که، رفتار غیرپایانه‌ها می‌تواند محدود به گام اولیه شود. بعلاوه، این گام اولیه می‌تواند (باپیچیدگی زمانی $O(n^2)$) به فرم کلمات مرکب تعمیم یابد.

۶.۳روند الگوریتم +CYK در مواجه با احتمالات

محاسبه احتمال بیشینه برای هر درخت تجزیه تقریبا ساده است. به دلیل اینکه احتمال غیرپایانه‌ها در SCFG نتیجه حاصل‌ضرب احتمالات شرطی قوانینی است که، توسط حاصل‌ضرب احتمالات غیرپایانه‌های سمت راست یک قانون تولید شده‌اند. به دلیل اینکه الگوریتم‌های پائین به بالا همیشه ابتدا غیرپایانه‌های سمت راست را می‌سازند و به دلیل اینکه حاصل‌ضرب بیشینه شامل اجزاء بیشینه است، الگوریتم قادر خواهد بود که احتمال یک غیرپایانه جدید را با یک سربار اضافه در طول فاز تجزیه محاسبه کند.

الگوریتم +CYK محدود به گرامرهای CNF نیست. مزیت دیگر آن علاوه‌بر متغیرهای احتمالاتی شناخته‌شده درCYK [۵۲, ۷۵] این است که، درطول فازتجزیه، N تجزیه با بیشترین احتمال (با نام N-بهترین تجزیه) را می‌تواند محاسبه کند. درانتهای فاز تجزیه، الگوریتم N-بهترین تجزیه را برای جمله ورودی (همچنین برای هریک از زیررشته‌ها) تولید می‌کند.

درمواردی که روش N-بهترین از کارایی لازم برخوردار نیست، می‌توان به‌صورت بازگشتی تمام تجزیه‌های جمله را استخراج کرد و احتمالات آن‌ها را در طول استخراج با روش بالا به پائین و حاصل‌ضرب احتمالات اجزای سازنده آن، محاسبه کرد[۷۴].

تنها مساله باقی‌مانده‌ای که در اینجا باید در مورد آن بحث کرد، محاسبه احتمالات در چرخه گرامرها است. این کار همانند [۷۸]، به‌صورت پیش‌محاسبه انجام می‌شود. این پیش‌محاسبه یک‌بار برای تمام گرامرهای داده‌شده به این‌صورت محاسبه می‌شود که، یک ماتریس از احتمالات چرخه‌های گرامر ایجاد می‌شود (برای مطالعه بیشتربه منبع[۷۸] مراجعه کنید). این ماتریس به عنوان یک جدول Look up زمانی‌که الگوریتم، یک چرخه را شناسایی کرد به آن مراجعه می‌کند (همانندآیتم‌هایی که دوباره دیده شده‌اند).

برای کاربردهای عملیNLP، الگوریتم بالا باید برای لغات خارج از دیکشنری نیز قابل تعمیم باشد. در این روش، روند لغات دیده‌نشده در چارچوب احتمالاتی جاداده شده است. این قسمتی از تعریف SCFGها است که برای هر غیرپایانه (A) در گرامر، احتمالات شرطی برای تولید A باید محدودیت احتمالاتی به صورت زیر باشد (معادله۲-۶) :

$$\sum_{\alpha} P(A \to \alpha) = 1 \qquad (۶-۲)$$

برای پیاده‌سازی کلمات دیده‌نشده این محدودیت را آرام می‌کنیم (معادله ۳-۶):

$$\sum_{\alpha} P(A \to \alpha) \leq 1 \qquad (۶-۳)$$

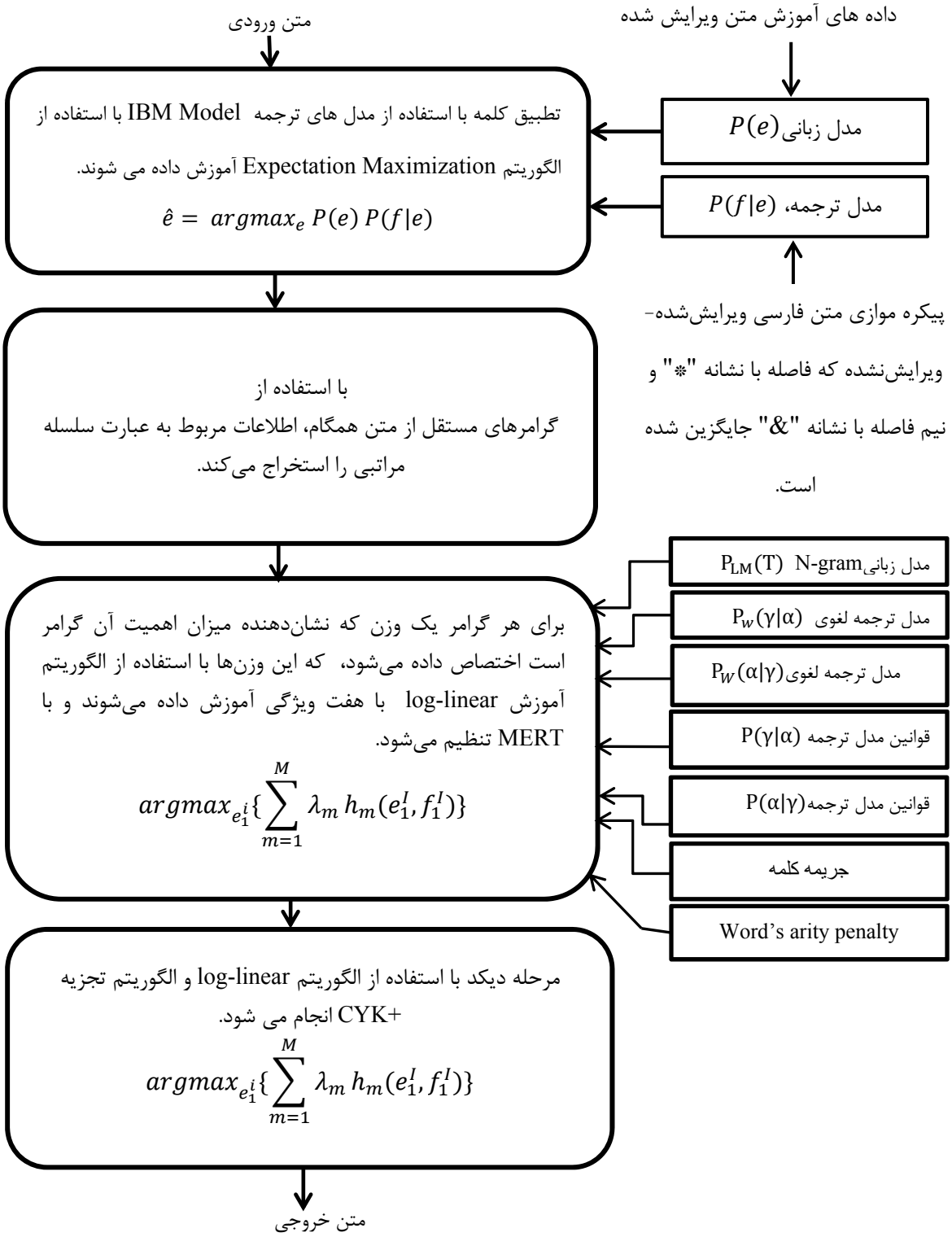
و یک "غیرپایانه باز[1]"برای هر غیرپایانه تعریف می‌شود، به طوری‌که نامساوی بالا تاکید دارد. برای چنین غیرپایانه‌هایی کمیت $P_{\underline{u}}(A) = 1 - \sum_{\alpha} P(A \to \alpha)$ معرفی می‌شود. این کمیت می‌تواند به این‌صورت تفسیر شودکه، احتمالی که کلمه دیده‌نشده ($\underline{u}$) می‌تواند با قانون(جا افتاده) $A \to \underline{u}$ تولید شود. کلمه دیده‌نشده یک پایانه جدید است (که درحال حاضر در گرامر پایانه‌ها وجود ندارد). به طوری‌که، به هرکلمه دیده‌نشده نگاشت پیدا می‌کند. هر وقت که به یک کلمه دیده‌نشده در جمله ورودی مواجه شود، تمام قوانین $A \to \underline{u}$، A یک"غیرپایانه باز" است، توسط الگوریتم همانند هر قانون پایانه استاندارد درنظر گرفته می‌شود[۷۴].

[1] Open non-terminal

فصل هفتم: روش پیشنهادی و بررسی نتایج و ارزیابی

۷.۱ روش پیشنهادی

در فصل های گذشته هر یک از مراحل روش ارائه شده مورد بررسی قرار گرفته است. در این بخش نگاهی کلی به هریک از این مراحل و نیز ابزارهای استفاده شده برای آن‌ها پرداخته می‌شود. شکل ۷-۱ نمایش گرافیکی از روند الگوریتم پیشنهادی برای اصلاح متن فارسی آمده است. ابزار استفاده شده برای تطبیق کلمه به کلمه ابزار GIZA++، ابزار استفاده شده برای استخراج گرامر با استفاده از تطبیق کلمه به کلمه Thrax، ابزار استفاده شده برای آموزش این گرامرها Z-MERT است و در انتها ابزار Joshua 5.0 [۷۹] برای دیکد به کار گرفته شده است. مجموعه‌های تست اجرا شده و خروجی‌های آن مورد بحث قرار گرفته است.



شکل ۷-۱ . نمایش گرافیکی از روند روش پیشنهادی

۷.۱.۱ گام اولیه: پیکره آموزش و تست

روش ارائه شده در این پایان‌نامه توسط یک پیکره موازی فارسی ویرایش‌نشده-ویرایش‌شده آموزش داده‌شده است. پیکره‌های موازی در دسترس فارسی، اغلب فارسی-انگلیسی هستند همانند پیکره شیراز[۸۰] که به‌عنوان اولین پیکره موازی انگلیسی-فارسی معرفی‌شده، پیکره Tehran English-Persian Corpus(TEP)، MULTEXT-EAST framework، Parallel English-Persian new corpus (PEN) و یا این پیکره‌ها جهت برچسب‌گذاری کلمات تولید شده‌اند همانند "بی‌جن‌خان"[۸۱]. بازم به ذکر است که در هیچ‌کدام قاعده نیم‌فاصله رعایت نشده است. پیکره دیگر فارسی، که موازی نیست، پیکره همشهری است که در دانشگاه تهران ایجاد شده و شامل مجموعه مقالات وب‌سایت همشهری است. همان‌طورکه در شکل ۷-۲ دیده می‌شود در این پیکره نیز قاعده نیم‌فاصله رعایت نشده است.

بنابراین یک پیکره موازی فارسی ویرایش‌نشده-ویرایش‌شده در این پایان‌نامه ایجاد شده است. این پیکره شامل مقاله‌های سایت‌های خبری در حوزه هنری و فرهنگی، ورزشی، علمی، پزشکی و سایر زمینه‌ها است که از وب‌سایت‌های خبری جمع‌آوری شده است.

ویژگی دیگر پیکره ایجاد شده این است که، برای آنکه تطبیق کلمه در ترجمه ماشینی آماری صورت گیرد، یک طرف متن ویرایش‌نشده و سمت دیگر متن ویرایش شده قرار دارد. در این‌صورت، زمانی که در تست یک کلمه چندبخشی دیده شود، می‌تواند با استفاده از اطلاعات و گرامرهای استخراج‌شده کلمه چندبخشی مناسب را جایگزین کرد.

برای آنکه بتوان قوانین مربوط به فاصله و نیم‌فاصله را استخراج کرد به گونه‌ای که بتوان گرامر مربوط به ساختار نیم‌فاصله کلمات چندبخشی و فاصله میان کلمات مستقل را استخراج کرده، باید یک نشانه[1] برای

[1] Token

فاصله و یک نشانه برای نیم‌فاصله درنظرگرفت. نشانه "*" برای فاصله و نشانه "&" برای نیم‌فاصله تعریف شده است. نمونه‌ای از پیکره ایجاد شده را در شکل ۷-۳ آمده است.

<TEXT>
<IMAGE>/1385/851017/news/009333.jpg</IMAGE>
<![CDATA[
گروه ادب و هنر- همایش بین المللی نمایش و دین طی مراسمی در خانه هنرمندان ایران گشایش یافت.
به گزارش خبرنگار همشهری، در نخستین روز همایش کریستوفر ایتس از کانادا درباره نقش مفاهیم دین بنیاد در خلق درام، فرهاد مهندس پور درباره روایت حماسی و روایت دین ایرانی، رحمت امیتی درباره بررسی تاثیرات متقابل نمایش های مذهبی و متاثر معاصر ایران بر یکدیگر و سمولز از انگلیس درباره نمایش و نیایش سخنرانی کردند.
همچنین در بخش دوم نخستین روز همایش پیتر چلکوفسکی از آمریکا درباره تعذیه و جهانشیر یاراحمدی درباره روضه خوانی به مثابه یک نظام نمایشی مقاله ارائه دادند.
دیروز در بخش پایانی همایش ابوالفضل حرمی درباره نمایش دینی و نقش تماشاگر، کلاوز هوفمان از آلمان درباره ابعاد محتوی و دینی تئاتر معاصر اروپا و فرح یگانه با عنوان بررسی تعزیه طبق مدل رویداد تئاتری مقاله خود را ارائه دادند.
دبیر نخستین سمیتار بین المللی نمایش و دین این سمیتار را سرآغازی بر فعالیت های جدید تئاتر ایران در عرصه بین الملل می داند و می گوید: در میان تمام نظرات محققان در خصوص ریشه های پیدایش هنر، نظریه آنیتی- مذهبی بیش از همه مورد توجه بوده و جوهره این نظریه بر مبنای این مفهوم استوار شده که بشر از آغازین روزهای تجربه زندگی در این کره خاکی در پی برقراری ارتباط با خالق خود بوده است؛ ارتباطی که از یکسو نیازهای روحانی و معنوی را برآورده می کند و از سوی دیگر تضمینی بر امنیت و زندگی جسمانی او بوده است. در این راستا هنر نمایش از آن جهت سرآمد دیگر هنرهاست که نخستین وسیله موثر ارتباطی انسان با معبود خود بوده و به تدریج به زیور خلاقیت و

شکل ۷-۲ . نمونه‌ای از پیکره همشهری

این پیکره با سایزهای مختلف ۱۰۰۰۰، ۲۰۰۰۰، ۳۰۰۰۰، ۴۰۰۰۰ و ۵۰۰۰۰ کلمه‌ای ایجاد شده است و در هربار که سایز پیکره آموزش افزایش می‌یابد، آموزش وتست دوباره بر روی آن انجام می‌شود. اجزاء سازنده پیکره که شامل تعداد کلمات و تعداد جملات است در جدول۷-۱ آمده است.

پیکره مدل زبانی نیز ایجاد شده است. این پیکره باید به اندازه کافی بزرگ باشد تا بتوانیم نتایج قابل قبولی را در خروجی ترجمه داشته باشیم. پیکره مدل زبانی که برای این سیستم ایجاد شده شامل ۱۱۷۲۷ کلمه است، که برای تمام مراحل آموزش یکسان درنظر گرفته شده است.

به*گوشش*خورد:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می*کنیم،*ولی*اگر*کسی*بی*نیازی*بورزد*و*دست*حاجت*پیش*مخلوقی*دراز*نکند،*خداوند*او*را*بی*نیاز*می*کند.*آن*روز*چیزی*نگفت.*و*به*خانه*خویش*برگشت.*باز*با*هیولای*مهیب*فقر*که*هم*چنان*بر*خانه*اش*سایه*افکنده*بود*روبرو*شد،*ناچار*روز*دیگر*به*همان*نیت*به*مجلس*رسول*اکرم*حاضر*شد،*آن*روز*هم*همان*جمله*را*از*رسول*اکرم*شنید:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می*کنیم،*ولی*اگر*کسی*بی*نیازی*بورزد،*خداوند*او*را*بی*نیاز*می*کند.این*دفعه*نیز*بدون*این*که*حاجت*خود*را*بگوید،*به*خانه*خویش*برگشت،*چون*خود*را*هم*چنان*در*چنگال*فقر*ضعیف*و*بیچاره*و*ناتوان*می*دید،*برای*سومین*بار*به*همان*نیت*به*مجلس*رسول*اکرم*رفت،*باز*هم*لب*های*رسول*اکرم*به*حرکت*آمد*و*با*همان*آهنگ*که*به*دل*قوت*و*به*روح*اطمینان*می*بخشید*همان*جمله*را*تکرار*کرد.*

به*گوشش*خورد:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می*کنیم،*ولی*اگر*کسی*بی*نیازی*بورزد*و*دست*حاجت*پیش*مخلوقی*دراز*نکند،*خداوند*او*را*بی*نیاز*می*کند.*آن*روز*چیزی*نگفت.*و*به*خانه*خویش*برگشت.*باز*با*هیولای*مهیب*فقر*که*هم*چنان*بر*خانه*اش*سایه*افکنده*بود*روبرو*شد،*ناچار*روز*دیگر*به*همان*نیت*به*مجلس*رسول*اکرم*حاضر*شد،*آن*روز*هم*همان*جمله*را*از*رسول*اکرم*شنید:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می*کنیم،*ولی*اگر*کسی*بی*نیازی*بورزد،*خداوند*او*را*بی*نیاز*می*کند.این*دفعه*نیز*بدون*این*که*حاجت*خود*را*بگوید،*به*خانه*خویش*برگشت،*چون*خود*را*هم*چنان*در*چنگال*فقر*ضعیف*و*بیچاره*و*ناتوان*می*دید،*برای*سومین*بار*به*همان*نیت*به*مجلس*رسول*اکرم*رفت،*باز*هم*لب*های*رسول*اکرم*به*حرکت*آمد*و*با*همان*آهنگ*که*به*دل*قوت*و*به*روح*اطمینان*می*بخشید*همان*جمله*را*تکرار*کرد.*

الف- متن ویرایش نشده سمت زبان مبدا

به*گوشش*خورد:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می&کنیم،*ولی*اگر*کسی*بی&نیازی*بورزد*و*دست*حاجت*پیش*مخلوقی*دراز*نکند،*خداوند*او*را*بی&نیاز*می&کند.*آن*روز*چیزی*نگفت.*و*به*خانه*خویش*برگشت.*باز*با*هیولای*مهیب*فقر*که*هم&چنان*بر*خانه&اش*سایه*افکنده*بود*روبرو*شد،*ناچار*روز*دیگر*به*همان*نیت*به*مجلس*رسول&اکرم*حاضر*شد،*آن*روز*هم*همان*جمله*را*از*رسول&اکرم*شنید:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می&کنیم،*ولی*اگر*کسی*بی&نیازی*بورزد،*خداوند*او*را*بی&نیاز*می&کند.این*دفعه*نیز*بدون*این&که*حاجت*خود*را*بگوید،*به*خانه*خویش*برگشت،*چون*خود*را*هم&چنان*در*چنگال*فقر*ضعیف*و*بیچاره*و*ناتوان*می&دید،*برای*سومین*بار*به*همان*نیت*به*مجلس*رسول&اکرم*رفت،*باز*هم*لب&های*رسول*اکرم*به*حرکت*آمد*و*با*همان*آهنگ*که*به*دل*قوت*و*به*روح*اطمینان*می&بخشید*همان*جمله*را*تکرار*کرد.*

به*گوشش*خورد:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می&کنیم،*ولی*اگر*کسی*بی&نیازی*بورزد*و*دست*حاجت*پیش*مخلوقی*دراز*نکند،*خداوند*او*را*بی&نیاز*می&کند.*آن*روز*چیزی*نگفت.*و*به*خانه*خویش*برگشت.*باز*با*هیولای*مهیب*فقر*که*هم&چنان*بر*خانه&اش*سایه*افکنده*بود*روبرو*شد،*ناچار*روز*دیگر*به*همان*نیت*به*مجلس*رسول&اکرم*حاضر*شد،*آن*روز*هم*همان*جمله*را*از*رسول&اکرم*شنید:*«هر کس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می&کنیم،*ولی*اگر*کسی*بی&نیازی*بورزد،*خداوند*او*را*بی&نیاز*می&کند.این*دفعه*نیز*بدون*این&که*حاجت*خود*را*بگوید،*به*خانه*خویش*برگشت،*چون*خود*را*هم&چنان*در*چنگال*فقر*ضعیف*و*بیچاره*و*ناتوان*می&دید،*برای*سومین*بار*به*همان*نیت*به*مجلس*رسول&اکرم*رفت،*باز*هم*لب&های*رسول*اکرم*به*حرکت*آمد*و*با*همان*آهنگ*که*به*دل*قوت*و*به*روح*اطمینان*می&بخشید*همان*جمله*را*تکرار*کرد.*

ب. متن ویرایش‌شده سمت زبان مقصد

شکل ۷-۳ . نمونه‌ای از پیکره فارسی ایجاد شده ویرایش‌نشده-ویرایش‌شده

متن تست نیز به همین ترتیب پیکره آموزش، ایجاد شده است. نکته‌ای که در ایجاد متن تست درنظرگرفته شده، این است که، سیستم متن تست را نباید در پیکره آموزش دیده باشد. متن تست در چهار دسته فرهنگی و هنری، ورزشی، علمی و پزشکی ایجاد شده است. از ویرایش شده متن تست برای ارزیابی ترجمه استفاده شده است.

جدول ۷-۱. اجزاء پیکره موازی فارسی ایجاد شده جهت آموزش

Train	۱	۲	۳	۴	۵
تعداد کلمات	۱۰۲۸۱	۲۰۳۰۹	۳۰۹۳۹	۴۱۶۷۱	۵۲۹۶۰
تعداد جملات	۹۸۴	۱۸۷۰	۲۸۵۶	۳۶۹۰	۴۴۳۴

۷.۲ بحث و بررسی نتایج حاصل و ارزیابی روش

۷.۲.۱ بررسی خروجی حاصل

همان‌طور که گفته شد، از ابزار Joshua برای ترجمه استفاده شده است. Joshua [۷۹]یک ابزار متن باز شناخته شده‌ای است که براساس رهیافت مبتنی بر عبارت سلسله مراتبی کار می‌کند. Joshua پیاده‌سازی دوباره‌ای از سیستم ترجمه ماشینی Hiero است که با اضافه‌کردن گرامرهایی همانند SAMT گسترش‌یافته است. Joshua به زبان Java نوشته شده است. این ابزار الگوریتم‌هایی همچون مدل زبانی n-gram، chart-parsing، استخراج K-بهترین ترجمه واحد و الگوریتم‌های هرس Beam وCube به‌کار گرفته است. این سیستم به دلیل استفاده از محاسبات موازی، می‌تواند برای سیستم‌های با مقیاس بزرگ استفاده شود. تولید خروجی برای متن تست پنج مرحله دارد [۸۲] :

۱. آموزش مدل زبانی با استفاده از ابزار SRILM [۸۳].

۲. استخراج گرامرها برای متن تست. این گام به‌خودی‌خود شامل زیرمراحلی همانند آماده‌سازی پیکره موازی، تطبیق کلمه با استفاده از ابزار GIZA++[84]، استخراج گرامر با استفاده از فراساختار آرایه پسوندی که با ابزار Thrax صورت می‌گیرد.

۳. پیداکردن وزن‌های بهینه برای ترکیب مدل‌های متفاوت و تابع‌های ویژگی با استفاده از MERT.

۴. در آخر، اجرای دیکدر جهت تولید K-بهترین ترجمه برای هر جمله در متن تست.

تست در چهار دسته متن فرهنگی و هنری، ورزشی، علمی و پزشکی تقسیم شده است. در این متن‌ها کلماتی وجود دارند که مختص به همان زمینه است، بنابراین می‌توان با متن‌های مختلف گستره‌ای از واژه‌ها را داشت. در جدول ۷-۲ خصوصیات متن‌های تست آمده است.

جدول ۷-۲ . متن تست ایجاد شده برای ارزیابی سیستم

تست	فرهنگی و هنری	ورزشی	علمی	پزشکی
تعداد کلمات	۳۶۴۱	۱۵۶۳	۱۴۸۰	۲۴۹۴
تعداد جملات	۱۲۳	۴۲	۸۳	۸۶

برای تنظیم پارامترها در زمان آموزش پیکره موازی ایجاد شده است که ساختار آن همانند ساختار پیکره آموزش ایجاد شده است. این پیکره شامل ۱۲۰۶۶ کلمه و ۴۴۹ جمله است. این کار به دلیل آنکه داده‌ها بدون هیچ نویزی باشند، صورت می‌گیرد.

در جدول ۷-۳ بخشی از نتیجه بهترین ترجمه برروی داده تست پزشکی که با استفاده ازپیکره آموزش ۴۰۰۰ کلمه‌ای آموزش دیده، آمده است. همان‌طور که دیده می‌شود، این روش توانسته کلمه‌های چندبخشی همانند "هیچ‌وجه"، "نمی‌تابد"، "می‌تواند"، "چروک‌ها"، "درنظرگرفته‌شده‌است"، "تدارک‌دیده‌شود"، "پیشگیری‌کننده‌اند"، "ضعیف‌تر" را به درستی شناسایی و ویرایش کند.

نتیجه بسیار چشمگیری که از مشاهده خروجی دیکد دیده می‌شود این است که برای شناسایی و ویرایش کلمات چندبخشی در زمان تست، لزومی به وجود عین کلمات چندبخشی برای شناسایی آن در زمان آموزش وجود ندارد. فقط با داشتن تعداد کافی از کلمات چندبخشی با ساختار مشابه می‌توان آن را ویرایش کرد. به عنوان مثال، در داده آموزش کلمه‌ای که با "ها" جمع بسته می‌شود به تعداد کافی در زمان آموزش دیده شده است و ساختار مربوط به "ها" و نیم‌فاصله‌ای که در آموزش با نشانه " &" در زمان آموزش وزن‌دهی شده، استخراج کرده است.

به ساختار یا گرامر، وزنی با استفاده از پیکره آموزش، مدل زبانی و پیکره Tune اختصاص داده‌شده است. وزن‌های تعیین شده توسط الگوریتم Log-linear با الگوریتم آموزش MERT همان‌طور که گفته شد، تنظیم می‌شود. این وزن‌ها به هریک از ویژگی‌های Log-linear اختصاص داده می‌شود، نشان‌دهنده درجه اهمیت آن است.

با استفاده از ساختار یا گرامر استخراج‌شده می‌تواند هر کلمه چندبخشی که با ساختار مشابه با آن را در زمان تست آمده است را شناسایی و ویرایش کند.

اما کلمات چندبخشی همانند " زیر چشم "، " تا حدی "، " بر طرف " را نتوانسته شناسایی کند. به دلیل آنکه تعداد تکرارهای این کلمات در زمان آموزش به اندازه کافی نبوده است. همان‌طور که گفته شد، می‌توان با بالابردن فرکانس کلمات چندبخشی با ساختار مشابه در پیکره آموزش این نقص را برطرف کرد.

از طرفی در زبان فارسی کلمات چندبخشی وجود دارند که هر جزء آن یک کلمه معنی دار است مانند "به‌ویژه". اجزاء کلمه چندبخشی "به‌ویژه" می توانند به صورت مستقل و یا به عنوان جزئی از یک کلمه چندبخشی در متن آمده باشند. این روش با استفاده از اطلاعات زبان‌شناختی که از متن استخراج می‌کند، می‌تواند به درستی فاصله‌گذاری میان کلمات چندبخشی را که از اجزاء معنی‌دار تشکیل شده‌اند را ویرایش کند.

جدول ۷-۳. خروجی بهترین ترجمه داده پزشکی آموزش‌دیده با داده آموزش ۴۰۰۰ کلمه‌ای

خروجی تست	ورودی تست
این نوع کرم‌ها را باید به پوست زیر چشم زد . برای پلک بالا نیازی به استفاده از کرم نیست زیرا آفتاب به هیچ‌وجه به پوست آن نمی‌تابد. اگر فردی به دلیل رعایت نکردن اصول مراقبتی، دچار چروک دور چشم شد ، آیا استفاده از کرم دور چشم می‌تواند باعث بهبود و برطرف‌شدن چروک‌ها شود یا باید به فکر روش دیگری بود ؟ برای اصلاح نسبی چروک‌ها و تیرگی زیر چشم می‌توان از نوعی لیزر استفاده کرد. توکسین بوتولینیوم a نیز می‌تواند تا حدی چروک‌های قسمت خارجی پلک را برطرف کند. در صورتی‌که چروک‌ها بیش از حد و باعث افتادگی پلک‌شده‌باشند ، با کمک جراحی قابل درمان هستند. کرم‌های ضد چروک فقط پیشگیری کننده‌اند و خاصیت درمانی ندارند. کرم‌های دور چشم قوی یا انواع ضعیف‌تر که برای زیر ۴۰ سالگی درنظرگرفته‌شده‌است ؟ حسینی در پایان یادآور شد با توجه به جلساتی که با ریاست دامپزشکی داشته‌ایم مقرر شد با افزایش نظارت ها برنامه‌هایی تدارک‌دیده‌شود.	این نوع کرم ها را باید به پوست زیر چشم زد. برای پلک بالا نیازی به استفاده از کرم نیست زیرا آفتاب به هیچ وجه به پوست آن نمی تابد. اگر فردی به دلیل رعایت نکردن اصول مراقبتی ، دچار چروک دور چشم شد ، آیا استفاده از کرم دور چشم می تواند باعث بهبود و بر طرف شدن چروک ها شود یا باید به فکر روش دیگری بود؟ برای اصلاح نسبی چروک ها و تیرگی زیر چشم می توان از نوعی لیزر استفاده کرد. توکسین بوتولینیوم a نیز می تواند تا حدی چروک های قسمت خارجی پلک را بر طرف کند. در صورتی که چروک‌ها بیش از حد و باعث افتادگی پلک شده باشند ، با کمک جراحی قابل درمان هستند. کرم‌های ضد چروک فقط پیشگیری‌کننده اند و خاصیت درمانی ندارند. کرم های دور چشم قوی یا انواع ضعیف تر که برای زیر ۴۰ سالگی در نظر گرفته شده است ؟ حسینی در پایان یادآور شد با توجه به جلساتی که با ریاست دامپزشکی داشته ایم مقرر شد با افزایش نظارت‌ها برنامه هایی تدارک دیده شود.

۷.۲.۲ ارزیابی با استفاده از معیار BLEU[1]

جهت ارزیابی در مورد کیفیت ترجمه می‌توان از ارزیابی انسان استفاده کرد. اما این ارزیابی بسیار هزینه‌بر است. به دلیل آنکه ارزیابی‌های انسان ممکن است ماه‌ها به طول انجامد تا تمام شود و نیاز به افرادی داردکه فقط یک‌بار در مورد ترجمه نظر بدهند. بنابراین باید از معیارهای ارزیابی اتومات برای بررسی کیفیت ترجمه استفاده کرد.

[1] Bilingual Evaluation Understudy (BLEU)

یک روش استفاده از معیار BLEU است. روشی است که ترجمه ماشینی را به‌صورت اتومات سریع،ارزان و بدون وابستگی به زبان و نیز بسیار به ارزیابی انسان نزدیک است و هزینه حاشیه‌ای کمی در هر بار اجرا دارد،ارزیابی می‌کند.

کار اصلی برنامه BLEU [70] پیاده‌سازی مقایسه N-gramهای متن ترجمه شده با N-gram ترجمه رفرنس قرارداده شده و شمارش تعداد تطبیق‌ها است. این مطابقت‌ها وابسته به موقعیت نیستند و آن‌هایی که تطبیق بیشتری دارند، کاندیداهای بهتری برای ترجمه هستند.

اساس این متریک یک معیار دقت است. برای محاسبه دقت، تعداد کلمات ترجمه کاندید (unigram) را همان‌گونه که در هر متن منبع و مرجع اتفاق افتاده را می‌شمارد و سپس به کل کلمات ترجمه کاندید تقسیم می‌کند.

مشکل ازاین جا نشات می‌گیرد که یک کلمه مرجع بعداز آن‌که با کلمه کاندید مطابقت داده شد کنار گذاشته نمی‌شود. با درنظرگرفتن این مشکل یک روشی به نام "دقت unigram اصلاح شده" فرموله شده است. برای محاسبه آن، ابتدا ماکزیمم تعداد رخدادهای یک کلمه در هریک ازترجمه‌های رفرنس محاسبه‌شده، تعداد کل هر کلمه کاندید و ماکزیمم تعداد آن کلمه در متن رفرنس باهم جمع می‌شوند و به تعداد کلمات کاندید تقسیم می‌شوند. به مثال زیر توجه کنید.

کاندید : مقالات مقالات مقالات مقالات مقالات مقالات مقالات مقالات مقالات مقالات.

رفرنس: مقالات مربوط به مزرعه در صفحه ای با عنوان مقالات مزرعه فهرست شده اند .

دقت unigram اصلاح شده در این مثال $\frac{۲}{۱۰}$ است در حالی که این دقت با استفاده از unigram $\frac{۱۰}{۱۰}$ است .

دقت n-gram اصلاح شده برای هر n قابل محاسبه است: تمامn-gramهای کاندید شمارش می‌شوند و رفرنس بیشینه مطابق با آن‌ها جمع می‌شود. شمارش کاندیداها بوسیله مقدار بیشینه مطابق آن‌ها جمع و بر تعداد کل n-gramهای کاندید تقسیم می‌شود. n-gramهای بیشینه درمثال ذکر شده برای دقت

bigram اصلاح شده برابر صفراست. این نوع از مقداردهی دقت n-gram اصلاح شده به دنبال جلب شایستگی وکفایت ترجمه است. تطبیق‌ها درn-gramهای بزرگتر برای حفظ روان ساختگی ترجمه به کار گرفته می‌شوند.

نتایج حاصل از ارزیابی با معیار BLEU در جدول ۷-۴ آمده است. همان‌طور که دیده می‌شود، هرچه پیکره آموزش رشد می‌کند، مقدار BLEU نیز افزایش می‌یابد. به دلیل آنکه فرکانس حاصل از رخداد کلمات چندبخشی مشابه، افزایش یافته است. البته با افزایش حجم داده‌های آموزش اگر افزایش تنوع کلمات چندبخشی را به‌جای افزایش فرکانس کلمات چندبخشی داشته باشیم، شناسایی کلمه به‌درستی صورت نمی‌گیرد و این معیار کمی افت می‌کند.

جدول ۷-۴ . نتایج حاصل از ارزیابی با استفاده از معیار BLEU

	پزشکی	علمی	ورزشی	فرهنگی و هنری
۱۰Kwords	۰/۸۳۱۳	۰/۸۹۶۱	۰/۹۰۲۷	۰/۹۱۵۵
۲۰Kwords	۰/۸۸۱۵	۰/۸۹۱۸	۰/۹۴۷۴	۰/۹۲۸۰
۳۰Kwords	۰/۸۸۲۷	۰/۸۹۳۲	۰/۹۵۰۴	۰/۹۲۸۲
۴۰Kwords	۰/۸۸۱۱	۰/۹۰۳۱	۰/۹۴۷۶	۰/۹۲۷۶
۵۰Kwords	۰/۸۸۰۹	۰/۸۹۱۶	۰/۹۴۸۴	۰/۹۲۵۳

۷.۳.۳ ارزیابی روش با استفاده از مثبت واقعی[1]، مثبت کاذب[2] و منفی کاذب[3]

با استفاده از مثبت واقعی می‌توان به بررسی کلمه‌های چندبخشی که انتظار می‌رود روش بتواند شناسایی و ویرایش کند، می‌پردازد. کلمه‌های چندبخشی که در متن تست با فاصله ازهم آمده‌اند و بعد از دیکد کردن آن با نیم‌فاصله جایگزین شده‌اند، محاسبه می‌شود. نتایج حاصل از محاسبه مثبت واقعی برروی چهار دسته داده تست و با پنج حجم متفاوت از داده‌های تست در جدول ۷-۵ آمده است.

بررسی خطای سیستم با استفاده از مثبت کاذب و منفی کاذب انجام می‌گیرد. در منفی کاذب کلمه‌ها چندبخشی که انتظارش را می‌رفته که شناسایی شوند، اما شناسایی نشده‌اند، محاسبه می‌شود. که این مقدار جهت نرمال سازی به مجموع کلمات محاسبه شده توسط مثبت واقعی ، منفی کاذب و مثبت کاذب تقسیم می‌شود. نتایج منفی کاذب در جدول۷-۶ آمده است.

اما در خروجی کلماتی وجود داشتند، که چندبخشی نبوده اما به اشتباه چندبخشی شناسایی‌شده و فاصله میان آن‌ها با نیم‌فاصله جایگزین شده‌اند. این‌گونه از خطاها با استفاده از مثبت کاذب محاسبه می‌شوند. در این محاسبه نیز پس از تعیین تعداد این رخدادها جهت نرمال‌سازی به مجموع مثبت واقعی ، منفی کاذب و مثبت کاذب تقسیم می‌شوند. نتایج حاصل از آن در جدول ۷-۷ آمده است.

[1] True Positive (TP)
[2] False Positive (FP)
[3] False Negetive (FN)

جدول۷-۵. ارزیابی مربوط به مثبت واقعی بر روی دسته داده‌های تست متفاوت با حجم داده‌های آموزش متفاوت

	پزشکی	علمی	فرهنگی و هنری	ورزشی
۱۰Kwords	۰/۶	۰/۷۵	۰/۵۸	۰/۴۲
۲۰Kwords	۰/۹۴	۰/۷۰	۰/۶۸	۰/۹۲
۳۰Kwords	۰/۹۶	۰/۷۷	۰/۷۳	۰/۹۴
۴۰Kwords	۰/۹۵	۰/۸۲	۰/۶۸	۰/۸۱
۵۰Kwords	۰/۹۰	۰/۷۲	۰/۶۹	۰/۹۵

جدول۷-۶. ارزیابی با استفاده از منفی کاذب برروی داده‌های تست با دسته‌های متفاوت و با حجم متن‌های آموزش متفاوت

	پزشکی	علمی	فرهنگی و هنری	ورزشی
۱۰Kwords	۰/۳۸	۰/۲۱	۰/۲۲	۰/۵۴
۲۰Kwords	۰/۰۵	۰/۲۵	۰/۱۸	۰/۰۲
۳۰Kwords	۰/۰۳	۰/۲۲	۰/۱۹	۰/۰۳
۴۰Kwords	۰/۰۱	۰/۱۲	۰/۱۴	۰/۱۴
۵۰Kwords	۰/۰۴	۰/۲۲	۰/۲۰	۰/۰۵

جدول ۷-۷. ارزیابی مثبت کاذب برروی داده‌های متفاوت با حجم داده‌های آموزش متفاوت

	پزشکی	علمی	فرهنگی و هنری	ورزشی
۱۰Kwords	۰/۱۶	۰/۰۴	۰/۱۹	۰/۰۳
۲۰Kwords	۰/۰۰۸	۰/۰۵	۰/۱۳	۰/۰۶
۳۰Kwords	۰/۰۰۷	۰/۰۴	۰/۰۸	۰/۰۳
۴۰Kwords	۰/۰۳	۰/۰۷	۰/۱۸	۰/۰۴۵
۵۰Kwords	۰/۰۵	۰/۰۵	۰/۱۰	۰/۰۰۷

همان‌طور که دیده می‌شود، نتایج بسیار چشمگیری با افزایش حجم پیکره آموزش حاصل شده است. به‌ویژه در داده‌های تست مربوط به ورزشی و پزشکی. به دلیل آنکه کلمه‌های چندبخشی که در این دو مجموعه مورد تست قرار گرفته‌اند، از فرکانس رخداد کافی در ساختارهای مشابه برخوردار و این رخدادها با افزایش حجم پیکره آموزش افزایش یافته‌اند.

در جاهایی که افزایش حجم داده آموزش منجر به کاهش مثبت واقعی شده، به دلیل افزایش تنوع کلمات چندبخشی به جای افزایش فرکانس این کلمات است.

۷.۳.۴ ارزیابی روش با استفاده از Percision و Recall

روش ارائه‌شده با استفاده از Precision و Recall مورد ارزیابی قرار گرفته است. این دو معیار ارزیابی، به‌صورت گسترده‌ایی در پردازش زبان طبیعی مورد استفاده قرار گرفته است. این معیارها توسط [۸۵] برای ارزیابی ترجمه ماشینی به‌کارگرفته شد و به مقایسه میان ترجمه کاندید و ترجمه مرجع می‌پردازد.

معیار Precision با معادله ۱-۷ قابل محاسبه است و گویای این مطلب است که چقدر از کلمات چندبخشی که در ترجمه کاندید ویرایش شده‌اند، چندبخشی بوده‌اند. به عبارت دیگر، چقدر در ویرایش کلمات چندبخشی در مکان‌های صحیح انجام داده است. چون تعدادی از کلمه‌های ویرایش‌شده به اشتباه ویرایش شده‌اند و باتوجه به مطالب قبلی، این کلمات با مثبت کاذب محاسبه شده‌اند. بنابراین به کمک مثبت کاذب و مثبت واقعی می‌توان Precision را محاسبه کرد. نتایج حاصل از آن در جدول ۷-۸ آمده است.

$$Precision = \frac{TP}{TP+FP} \qquad (۷-۱)$$

جدول ۷-۸. ارزیابی Precision بر روی چهار دسته متن تست

	پزشکی	علمی	فرهنگی و هنری	ورزشی
۱۰Kwords	۰/۷۹	۰/۹۵	۰/۷۵	۰/۹۳
۲۰Kwords	۰/۹۹	۰/۹۳	۰/۸۴	۰/۹۴
۳۰Kwords	۰/۹۹	۰/۹۵	۰/۹۰	۰/۹۷
۴۰Kwords	۰/۹۷	۰/۹۲	۰/۷۹	۰/۹۵
۵۰Kwords	۰/۹۵	۰/۹۳	۰/۸۷	۰/۹۹

همان‌طور که دیده می‌شود، سیستم توانسته به صورت کارایی اغلب کلمات چندبخشی را شناسایی کند و با افزایش سایز پیکره آموزش بهبود یافته است.

معیار Recall نیز جهت ارزیابی اینکه، چقدر در شناسایی کلمات چندبخشی که در زمان تست دیده، موفق بوده است. این معیار توسط معادله ۲-۷ محاسبه می‌شود. نتایج حاصل از آن در جدول۷-۹ آمده است.

معادله ۲-۷ $$Recall = \frac{TP}{TP+FN}$$

جدول ۷-۹. ارزیابی Recall بر روی داده‌های متفاوت تست

	پزشکی	علمی	فرهنگی و هنری	ورزشی
۱۰Kwords	۰/۶۱	۰/۷۸	۰/۷۲	۰/۴۴
۲۰Kwords	۰/۹۴	۰/۷۴	۰/۷۹	۰/۹۸
۳۰Kwords	۰/۹۶	۰/۷۸	۰/۷۹	۰/۹۷
۴۰Kwords	۰/۹۹	۰/۸۷	۰/۸۳	۰/۸۵
۵۰Kwords	۰/۹۶	۰/۷۶	۰/۸۷	۰/۹۵

همان‌طور که از ارزیابی متریک Recall بر روی داده‌های تست مشاهده می‌شود، روش پیشنهادی توانسته در شناسایی و ویرایش کلمات چندبخشی به صورت کارایی عمل کند.

فصل هشتم: جمع‌بندی

استفاده از سیتم ترجمه ماشینی آماری، روش جدیدی است که در این پایان‌نامه برای ویرایش متن فارسی ارائه‌شده است. هدف آن استخراج ساختار زبان وکلمه‌های چندبخشی در پیکره موازی آموزش و بکارگیری آن در زمان تست است. برای آنکه سیستم بتواند گرامرهای مربوط به فاصله و نیم‌فاصله را از پیکره موازی ویرایش‌نشده-ویرایش‌شده آموزش استخراج کند، از نشانه "&" برای مشخص‌کردن نیم‌فاصله و نشانه "*" برای فاصله استفاده‌شده است. برهمین اساس پنج پیکره موازی با تعداد کلمات ۱۰۰۰۰، ۲۰۰۰۰ ،۳۰۰۰۰،۴۰۰۰۰ و ۵۰۰۰۰ کلمه‌ای جهت آموزش، ۱۲۰۴۴ برای Tune و نیز از برای مدل زبانی پیکره ۱۱۷۲۷ کلمه ایجاد شده است.

اطلاعات زبان‌شناختی کلمه‌های چندبخشی را با استفاده از استخراج گرامرهای مستقل از متن همگام و استخراج ویژگی‌های گرامرها براساس روش Log-linear و سپس تنظیم وزن‌های هر ویژگی توسط الگوریتم MERT صورت گرفته است. روش ارائه شده به‌گونه‌ای است که حتی کلمه‌های چندبخشی را که در پیکره آموزش ندیده‌اند، اما ساختار مشابه آن را به تعداد رخداد کافی دیده باشد، می‌تواند به صورت کارایی ویرایش کند. به دلیل آنکه اطلاعات زبان‌شناختی مربوط به فاصله و نیم‌فاصله را استخراج کرده است. از طرفی این روش می‌تواند کلمات چندبخشی که از اجزاء معنی‌دار و کلمات مستقل تشکیل شده‌اند و با جایگزینی فاصله به جای نیم فاصله و برعکس معنای آن تغییر می‌کند، ویرایش کند.

تست برروی پنج دسته مختلف از متن‌ها انجام گرفته است. نتایج آن با معیارهای BLEU ، مثبت واقعی، مثبت کاذب، منفی کاذب و نیز معیارهای Precision و Recall ارزیابی شده است. نتایج حاصل نشان‌دهنده کارایی سیستم و توانایی آن برای شناسایی بیشتر کلمه‌های چندبخشی در متن و ویرایش صحیح آن است.

نتایج حاصل نشان دادند که هرچه تعداد تکرارهای کلمه‌های چندبخشی با ساختار مشابه در پیکره موازی آموزش بیشتر باشد، بسیار برروی نتایج حاصل تاثیرگذار است.

مراجع

1. Sag A.I., Baldwin T., Bond F., Copestake A., Flickinger D., (2002). "Multiword Expressions: A Pain in the Neck for NLP", Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, P1-15, London, UK

2. Shamsfard M., Jafari H.S., Ilbeygi M., (2010) "Step-1: A set of fundamental tools for Persian text processing", Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta

3. Rasooli M.S., Kholy A.E. and Habash N., (2013) "Orthographic and Morphological Processing for Persian-to-English Statistical Machine Translation", Proceedings of the Sixth International Joint Conference on Natural Language Processing, P1047, Nagoya, Japan

4. Jelinek F., Lafferty J.D. and Mercer R.L. (1992) "Basic methods of probabilistic context free grammars" **Speech Recognition and Understanding: Recent Advances, Trends, and Applications**, 75, pp 345

5. Azmia A. M. and Alsaiarib A. (2014) "A calligraphic based scheme to justify Arabic text improving readability and comprehension" **Computers in Human Behavior**, 35, pp 177

6. Amtrup J.W., Mansouri Rad H., Megerdoomian K. and Zajac R. (2000) "Persian-English machine translation: An overview of the Shiraz project", Computing Research Laboratory, New Mexico State University

7. Brunning J., (August 2010), PhD. thesis, "Alignment Models and Algorithms for Statistical Machine Translation", Cambridge University Engineering Department, Cambridge University

8. Schwartza L. and Callison-Burchb Ch., (2010) "Hierarchical Phrase-Based Grammar Extraction in Joshua Suffix Arrays and Prefix Trees", the Prague Bulletin of Mathematical Linguistics, P157, Prague

9. Germann U., Jahr M., Knight K., Marcu D. and Yamada K., (2001) "Fast Decoding and optimal decoding for machine translation", Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, P228, Stroudsburg, PA, USA

10. Fournier B., (2008), MSc. thesis, "Preprocessing on bilingual data for Statistical Machine Translation", Computer Science, University of Twente

11. Koehn P. (2010), “Statistical Machine Translation”, Cambridge University Press

12. Brown P.F., Della Pietra S.A., Della Pietra V.J. and Mercer R.L. (1993) “The mathematics of statistical machine translation: Parameter estimation” **Computational Linguistics**, 19, 2, pp 263

13. Knight K. (1999) “Decoding complexity in wordreplacement translation models” **Computational Linguistics**, 25, 4, pp 607

14. Brown P.F., Cocke J., della P.S.A., della P.V.J., Jelinek F., Lai J.C. and Mercer R.L (1996) “Method and system for natural language translation” **Expert Systems with Applications**, 11, 2, pp iv

15. Wang Ye-Yi and Waibel A. (1997) “Decoding algorithm in statistical machine translation”, 8th Conference of the European Chapter of the Association for Computational Linguistics, P366, Madrid, Spain

16. Bijankhan M., Sheykhzadegan J., Bahrani M. and Ghayoom M. (2011) “Lessons from building a Persian written corpus: Peykare” **Language Resources and Evaluation**, 45, 2, pp 143

17. Rasooli M.S., Kouhestani M. and Moloodi A., (2013) “Development of a Persian syntactic dependency treebank”, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, P306, Atlanta, Georgia

18. Chiang D. (2007) “Hierarchical phrase-based translation” **Computational Linguistics**, 33, 2, pp 201

19. Vogel S., Ney H. and Tillmann Ch., (1996) “HMM-based word alignment in statistical translation”, Proceedings of the 16th conference on Computational linguistics, P 836, USA

20. Och F.J and Ney H., (2000) “A comparison of alignment models for statistical machine translation”, Proceedings of the 18th conference on Computational linguistics, P1086, Saarbrücken, Germany

21. Och F.J and Ney H., (2000) “improved statistical alignment models”, Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, P 440, USA

22. Kneser R., Ney H., (1993) “Forming Word Classes by Statistical Clustering for Statistical Language Modelling”, the First International Conference on Quantitative Linguistics, P221, Trier, Germany

23. Moore R.C., (2004) "Improving IBM word-alignment model 1", Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, P519, Barccelona, Spain

24. Dyer Ch., Chahuneau V. and Smith N.A., (2013) "A Simple, Fast, and Effective Reparameterization of IBM Model 2", P644, USA

25. Schoenemann Th., (2010) "Computing Optimal Alignments for the IBM-3 Translation Model", 14th Conference on Computational Natural Language Learning, P98, Uppsala, Sweden

26. Weese J., Ganitkevitch J., Callison-Burch1 Ch. , Post M. and Lopez A., (2011) "Joshua 3.0: Syntax-based Machine Translation with the Thrax Grammar Extractor", Proceedings of the Sixth Workshop on Statistical Machine Translation, P478, Edinburgh, Scotland, UK

27. Vulić I., (March, 2010), MSc. thesis, "Term Alignment State of the Art Overview", Katholieke Universiteit Leuven

28. Koehn Ph., Och F.J and Marcu D., (2003) "Statistical phrase-based translation", Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics, P48, Edmonton, Alberta,Canada

29. Riedel S. and Clarke J., (2009) "Revisiting Optimal Decoding for Machine Translation IBM Model 4", P5, Boulder, Colorado

30. Specia L., (2010) "Fundamental and New Approaches to Statistical Machine Translation", International Conference on Computational Processing of the Portuguese Language, Porto Alegre, Brazil

31. Rosenfeld R. (2000) "Two Decads of Statistical Language Modeling: Where do we go from here?" **Proceedings of IEEE**, 88, 8, pp 1270

32. Chen S.F. and Goodman J., (1998) "An Empirical Study of Smoothing Techniques for Language Modeling", Proceedings of the 34th annual meeting on Association for Computational Linguistics, P310, USA

33. Och F.J., (2003) "Minimum error rate training for statistical machine translation", Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, P160, Stroudsburg, PA, USA

34. Och F. J., Tillmann Ch. and Ney H., (1999) "Improved alignment models for statistical machine translation", In Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora, P440, Stroudsburg, PA, USA

35. Marcu D., Wong W., (2002) "A phrase-based, joint probability model for statistical machine translation", Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, P133, Philadelphia, Pennsylvania

36. Och F.J. and Ney H. (2004) "The alignment template approach to statistical machine translation" **Computational Linguistics**, 30, 4, pp 417

37. Koehn, Ph., Hieu Hoang, Birch A., Callison-Burch Ch., Federico M., Bertoldi N., Cowan B., Shen W., Moran Ch., Zens R., Dyer Ch., Bojar O., Constantin A. and Herbst E., (2007) "Moses: Open source toolkit for statistical machine translation", Proc. ACL-2007, Prague, Czech Republic

38. Zhifei Li, Callison-Burch Ch., Dyer Ch., Ganitkevitch J., Khudanpur S., Schwartz L., Thornton W., Weese J. and Zaidan O., (2009) "Joshua: an open source toolkit for parsing-based machine translation", Proceedings of the Fourth Workshop on Statistical Machine Translation, P135, Stroudsburg, PA, USA

39. Lopez A., (2007) "Hierarchical phrase-based translation with suffix arrays", Empirical Methods in Natural Language Processing and Computational Natural Language Learning, P976, Czech Republic

40. Chiang D., (2005) "A hierarchical phrase-based model for statistical machine translation", Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, P263, Ann Arbor, Michigan, USA

41. Callison-Burch Ch., Bannard C. and Schroeder J., (2005) "Scaling phrase-based statistical machine translation to larger corpora and longer phrases", Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, P255, Ann Arbor, Michigan, USA

42. Zhang Y. and Vogel S., (2005) "An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora", Proceedings of the 10th Annual Conference of the European Association for Machine Translation, P294, Budapest, Hungary

43. Manber U. and Myers G., (1990) "Suffix arrays: A new method for on-line string searches", The First Annual ACM-SIAM Symposium on Dicrete Algorithms, P319, San Francisco, CA, USA

44. Lopez A., (2008), PhD. thesis, "Machine Translation by Pattern Matching", University of Maryland

45. Fredkin E. (1960) "Trie memory" **Communications of the ACM**, 3, pp 490–499

46. Zhifei Li, Callison-Burch Ch., Dyer Ch., Ganitkevitch J., Khudanpur S., Schwartz L., Thornton W., Weese J. and Zaidan O., (2009) "Joshua: An open source toolkit for parsing-

based machine translation", Proceedings of the Fourth Workshop on Statistical Machine Translation, P135, Athens, Greece

47. Baker K., Bethard S., Bloodgood M., Brown R., Callison-Burch Ch., Coppersmith G., Dorr B., Filardo W., Giles K., Irvine A., Kayser M., Levin L., Martineau J., Mayfield J., Miller S., Phillips A., Philpot A., Piatko Ch., Schwartz L., Zajic D., (2009) "Semantically Informed Machine Translation (SIMT)", Summer Camp for Applied Language Exploration, Baltimore, Maryland

48. Zhifei Li and Khudanpur S., (2008) "Large-scale discriminative n-gram language models for statistical machine translation", 30th Annual Meeting & Symposium, Boston, Massachusetts, USA

49. Baker J.K., (1979) "Trainable grammars for speech recognition", Speech Communication Papers for the 97th Meeting of the Acoustical Society of America, P547, New York, USA

50. Booth T.L. and Thompson R.A. (1973) "applying probability measures to abstract languages" **IEEE Transactions on Computers**, 22, 5, pp 442

51. Charniak E. (1994), "statistical language learning", MIT Press, MA, USA

52. Baker J.K., (1979) "Trainable grammars for speech recognition", In Proceedings of the Spring Conference of the Acoustical Society of America, P547, Boston, MA

53. Jelinek F., Lafferty J.D. and Mercer R.L. (1992) "Basic Methods of Probabilistic Context Free Grammars" **Computer and system science**, 75, pp 345

54. Kupiec J., (1991) "A Trellis-based algorithm for estimating the parameters of a hidden stochastic context-free grammar", Proceedings of the workshop on Speech and Natural Language, P241, New York, USA

55. Lari K. and Young S.J. (1990) "the estimation of stochastic context-free grammars using the inside-outside algorithm" **Computer Speech & Language**, 4, 1, pp 35

56. Suppe P. (1970) "probabilistic grammars for natural language" **Synthese**, 22, pp 59

57. Zollmann A., Venugopal A., (2006) "Syntax Augmented Machine Translation via Chart Parsing", Workshop on statistical machine translation, P138, New York, USA

58. Yamada K. and Knight K., (2001) "A syntax-based statistical translation model", In ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, P523, USA

59. Yamada K. and Knight, K., (2002) “A decoder for syntax-based statistical MT”, In ACL ’02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, P303, NJ, USA

60. Koehn Ph., (2004) “Pharaoh: A beam search decoder for phrase-based statistical machine translation models”, Proceedings of the 6th Biennial Conference of the Association for Machine Translation in the Americas, P115, Washington DC, USA

61. Tillmann Ch., (2003) “A projection extension algorithm for statistical machine translation”, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, P1, Sapporo, Japan

62. Venugopal A., Vogel S. and Waibel A., (2003) “Effective phrase translation extraction from alignment models”, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, P319, Sapporo, Japan

63. Zollmann A., Venugopal A., (2006) “Syntax augmented machine translation via chart parsing”, In Proceeding StatMT '06 Proceedings of the Workshop on Statistical Machine Translation, P138, Stroudsburg, PA, USA

64. Baker K., Bloodgood M., Brown R., Callison-Burch Ch., Coppersmith G., Dorr B., Filardo W., Giles K., Irvine A., Kayser M., Levin L., Martineau J., Mayfield J., Miller S., Phillips A., Philpot A., Piatko Ch., Schwartz L., Zajic D., (2009) “Semantically Informed Machine Translation (SIMT)”, Summer Camp for Applied Language Exploration, Valletta, Malta

65. Baum L.E., (1972) “An inequality and associated maximization technique in statistical estimation for probabilistic function of Markov process”, In equalities, 3, P1

66. Onaizan A., Curine J., Jahr M., Knight K., lafferty J., Melamed I. D., Och F. J., Pardy D., Smith N. A. and Yarowsky D., (1999) “Statistical Machine Translation”, JHU Summer Workshop, Department of Computational Linguistics, University of Heidelberg

67. Zaidan O., (2009) “Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems”, The Prague Bulletin of Mathematical Linguistics, P79, Czech Republic

68. Och F.J and Ney H., (2002) “Discriminative training and maximum entropy models for statistical machine translation”, In Proceedings of the 40th Annual Meeting of the ACL, P295, Philadelphia, PA

69. Eggers sh., (2003), “Johns Hopkins 2003 Summer Workshop on Syntax and Statistical Machine Translation”, USA

70. Chung T. and Galley M., (2012) “Direct Error Rate Minimization for Statistical Machine Translation”, Proceedings of the Seventh Workshop on Statistical Machine Translation, P468, Montrèal, Canada

71. Koehn Ph., Axelrod A., Birch A., Callison-Burch Ch., Osborne M., Talbot D., (2005) “Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation”, International Workshop on Spoken Language Translation Pittsburgh, PA, USA

72. Hayashi K., Tsukada H., Sudoh K., Duh K. and Yamamoto S., (2010) “Hierarchical Phrase-based Machine Translation with Word-based Reordering Model”, Proceedings of the 23rd International Conference on Computational Linguistics, P439, Beijing

73. Koehn Ph., Och F.J. and Marcu D., (2003) “Statistical phrase-based translation”, Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics, P48, Canada

74. Chappelier J. and Rajman M., (1998) “A Generalized CYK Algorithm for Parsing Stochastic CFG”, Tabulation in Parsing and Deduction, P133, France

75. Dekai Wu, (1995) “Stochastic inversion transduction grammars with application to segmentation, bracketing, and alignment of parallel corpora”, 14th international joint conference on Artificial intelligence, P1328, USA

76. Graham S.L., Harrison M.A. and Ruzzo W.L. (1980) “An Improved Context-Free Recognizer” **ACM Transactions on Programming Languages and Systems**, 2, 3, pp 415

77. Sikkel K. and Nijholt A. (1997), “parsing context free grammar”, Springer, USA

78. Stolcke A. (1995) “An efficient probabilistic context-free parsing algorithm that computes prefix probabilities” **Computational Linguistics**, 21, 2, pp 165

79. Post M., Ganitkevitch J., Orland L., Weese J., Cao Y. and Callison-Burch Ch., (2013) “Joshua 5.0: Sparser, better, faster, server”, Workshop on Statistical Machine Translation, P206, Sofia, Bulgaria

80. Zajac R., Helmreich S., Megerdoomian K., (2000) “Black-Box/Glass-Box Evaluation in Shiraz”, Proceedings of the Workshop on Machine Translation Evaluation at LREC, Greece

81. Bijankhan M. (2004) “The role of the corpus in writing a grammar: An introduction to a software” **Iranian Journal of linguistics**, 19, 2

82. Zhifei L., Callison-Burch Ch., Khudanpur S. and Thornton W., (2009) “Decoding in Joshua Open Source, Parsing-Based Machine Translation”, the Prague Bulletin of Mathematical Linguistics, P47, Czech Republic

83. Stolcke A., (2002) “SRILM - an extensible language modeling toolkit”, Proceedings of the International Conference on Spoken Language Processing, Colorado

84. Och F.J, Ney H. (2003) “A systematic comparison of various statistical alignment” **Computational Linguistics**, 29, 1, pp 19

85. Melamed I.D., Green R., Turian J.P., (2003) “Precision and recall of machine translation”, Proceedings of the 2003 Conference of the North American, P 61, USA

Abstract

Persian language consists of words that are made up of multi-parts and these multi-part words have standard script, it means that it employs half-space between the parts of multi-part words to keep integrity of the parts of the multi-part words. According to this standard Persian script, half-space has an important role in readability of the text and the reader can understand the meaning of the text. Moreover, in natural language processing, including machine translation, words boundary detection has a considerable impact on system performance.

In this thesis, a new statistical method according to statistical machine translation is provided for Persian text editing. In this method spaces between the parts of multi-part words are replaced with the half-space with the aid of statistical machine translation. Linguistic information extracted from the parallel corpus and then this information is used to identify and edit multipart words. In this method a parallel corpora is needed to train in which an unedited corpora is on one side and the edited one is another side that is created in this thesis.

The results show the efficiency of the method in more accurate detecting and editing space between the parts of the multi-part words with half-space.

Keywords: *Persian Multi-Part Words, Spacing Rules, Statistical Machine Translation, Persian Parallel Corpora, Combining Persian Multi-part Words*

University of Shahrood

Faculty of Computer engineering and Information Technology

A statistical approach to combining multi-part words

Arezoo Arjomandzadeh

Supervisor:

Dr. Morteza Zahedi

Adviser:

Marzieh Rahimi

February 2015