ارائه یک سامانهٔ ترجمهٔ ماشینی ترکیبی بر پایهٔ رمزگشای یک‌نوا

حسین خاتمی*، حکیمه فدائی و هشام فیلی

دانشکدهٔ مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی، دانشگاه تهران، تهران، ایران

چکیده

در این مقاله یک مترجم خودکار متون انگلیسی به فارسی با استفاده از معماری ترکیبی قاعده‌مند و آماری ارائه شده است. این معماری ترکیبی به‌منظور بهبود نتایج هر دو مترجم، خروجی مترجم ماشینی قاعده‌مند و آماری را ترکیب کرده و سعی می‌کند یک خروجی برتر از هر دو سامانه ایجاد کند. در این راستا از یک رمزگشای یک‌نوا با پیچیدگی زمانی چندجمله‌ای استفاده می‌شود. مترجم‌های ماشینی قاعده‌مند عمل ترجمه را بر اساس مجموعه‌ای از قواعد زبانی انجام می‌دهند. به‌طور معمول نتایج آنها از نظر ترتیب کلمات و ساختار نحوی، کیفیت بهتری نسبت به نتایج مترجم‌های آماری دارند؛ ولی عملکرد این مترجم‌ها در زمینه انتخاب لغات مناسب و روانی ترجمه، ضعیف‌تر از مترجم‌های ماشینی آماری است. ازاین‌رو در این معماری، ترجمه اولیه به‌وسیله مترجم ماشینی قاعده‌مند صورت می‌گیرد؛ سپس با استفاده از مترجم ماشینی آماری ترجمه آن بهبود داده می‌شود. به این منظور، ترتیب واژگان در ترجمه نهایی بر اساس ترجمه مترجم ماشینی قاعده مند صورت می‌گیرد؛ سپس عمل ترجمه و انتخاب لغات توسط رمزگشای یک‌نوا، با درنظرگرفتن ترجمه‌های نامزدهای ارائه‌شده توسط مترجم قاعده‌مند و آماری و همچنین با استفاده از مدل زبانی، انجام می‌شود. آزمایش‌های انجام‌شده نشان می‌دهند که کیفیت نتایج به‌دست‌آمده از معماری ترکیبی در معیار بلو، به‌طورتقریبی پنج واحد بهتر از نتایج مترجم ماشینی قاعده‌مند است. همچنین کیفیت این نتایج نسبت به نتایج مترجم ماشینی آماری در معیار بلو، یک واحد بهتر است.

واژگان کلیدی: مترجم ماشین، معماری ترکیبی، رمزگشای یکنوا، ترتیب کلمات ترجمه، انتخاب لغات

A Hybrid Machine Translation System Based on a Monotone Decoder

Hosein Khatami*, Hakimeh Fadaei & Heshaam Faili

School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

Abstract

In this paper, a hybrid Machine Translation (MT) system is proposed by combining the result of a rule-based machine translation (RBMT) system with a statistical approach. The RBMT uses a set of linguistic rules for translation, which leads to better translation results in terms of word ordering and syntactic structure. On the other hand, SMT works better in lexical choice. Therefore, in our system, an initial translation is generated using RBMT. Then the proper lexical for the resulted sentence is chosen by using a decoder algorithm which is inspired by SMT architecture.

In the pure SMT approach, decoder is responsible for selecting proper final lexical during the translation procedure. Normally this method deals with lexical choice as well as reordering and required exponential order in time complexity. By fixing the word order in the output, a polynomial version of this method, named monotone decoding, is used in this paper. Monotone decoder algorithm selects the best lexical from a candidate list by maximizing the language model of resulted sentence. The candidate list is gathered from the outputs of both pure RBMT and pure SMT systems.

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال: ۹۶/۰۷/۲۱ ● تاریخ پذیرش: ۹۸/۰۴/۱۹ ● نوع مطالعه: پژوهشی

The experiments of proposed hybrid method on English-Persian language pair show significant improvements over both RBMT and SMT results. The results show that the proposed hybrid method gains an improvement of almost +5 units over RBMT and about one unit over SMT in BLEU score.

Keywords: Machine translation, hybrid architecture, monotone decoder, translation reordering, lexical choice

۱- مقدمه

ترجمه ماشینی[1] یکی از مهم‌ترین و پرکاربردترین شاخه‌های پردازش متن است، در این زمینه پژوهش‌های زیادی در راستای بهبود عملکرد آنها صورت گرفته است. از پرکاربردترین و رایج‌ترین کارهای صورت‌گرفته در این حوزه، به روش‌های قاعده‌مند[2] و آماری[3] می‌توان اشاره کرد، که هر یک از آنها دارای نقاط قوت و ضعف هستند. یکی از نخستین رویکردها در حوزه مترجم‌های ماشینی، روش قاعده‌مند است. در این روش عمل ترجمه بر پایه مجموعه‌ای از قواعد زبان که می‌تواند توسط متخصصان زبان تولید شود، صورت می‌گیرد. این قواعد زبانی بر اساس زبان مبداء و مقصد، قواعد انتقال نحوی و لغوی از یک زبان به زبان دیگر را پیاده‌سازی می‌کنند. این نوع مترجم‌ها به‌دلیل استفاده از این قواعد و اطلاعات زبان‌شناسی، پرهزینه ولی از نظر تطابق فعل و فاعل، ترجمه مناسب شناسه فعل، خصوصیات ساخت‌واژی[4] و به‌طور کلی از نظر ساختار نحوی[5] عملکرد خوبی دارند. در مقابل، این نوع مترجم‌ها ترجمه روان و سلیسی ندارند و در انتخاب لغات[6] مناسب، یک‌پارچگی واژگان و نزدیکی آنها به لغاتی که انسان برای ترجمه استفاده می‌کند، ضعیف هستند [5].

یکی دیگر از متداول‌ترین روش‌ها در حوزه ترجمه ماشینی، روش‌های آماری مبتنی بر عبارت[7] است. این نوع مترجم‌ها بر اساس مجموعه‌ای از پیکره موازی[8] که می‌تواند توسط ترجمه انسانی تولید شده باشد، عمل ترجمه را انجام می‌دهند. مترجم‌های آماری به‌دلیل استفاده از واژگان و عبارات استخراج‌شده از این پیکره و همچنین استفاده از مدل زبانی[9]، نتایج بهتری نسبت به مترجم‌های قاعده‌مند در زمینه انتخاب لغات مناسب و نزدیکی به ترجمه انسان دارند. این مترجم‌ها از نظر جابه‌جایی‌های[10] نزدیک و ترجمه‌های در سطح عبارت، عملکرد خوبی دارند؛ ولی وابستگی‌های نحوی و ساخت‌واژی بین عبارات[11]، کمتر در ترجمهٔ آنها رعایت می‌شود [5]. علاوه‌بر این، مشکل دیگری که در مترجم‌های آماری وجود دارد، هزینه زیاد رمزگشای[12] آنها است. این نوع مترجم‌های ماشینی با وجود هرس بخشی از فضای جستجوی خود، در یک فضای به‌نسبه بزرگ، بخش زیادی از حالات ممکن برای ترجمه جمله مورد نظر را بررسی می‌کنند، که این عمل باعث افزایش هزینه ترجمه آنها می‌شود.

همان‌گونه که مشخص است، مترجم‌های ماشینی قاعده‌مند و آماری دارای نقاط قوت و ضعفی هستند که به‌نحوی تکمیل‌کننده یکدیگرند. ازاین‌رو در سال‌های اخیر توجه به معماری‌های ترکیبی[13] که با ارائه روش‌هایی، این دو نوع مترجم ماشینی را با یکدیگر ترکیب می‌کنند، بیشتر شده است. در این روش‌ها سعی می‌شود که با استفاده از نقاط قوت یک نوع مترجم، نقاط ضعف نوع دیگر پوشش داده شود تا معماری جدیدی با بهره‌مندی از نقاط قوت هر دو نوع مترجم ماشینی ایجاد شود. معماری‌های ترکیبی عمل ترجمه را به‌صورت موازی یا متوالی انجام می‌دهند. در روش‌های موازی، خروجی چندین نوع مترجم ماشینی با یکدیگر ترکیب می‌شوند تا ترجمه نهایی سامانه حاصل شود. در روش‌های متوالی نیز، عمل ترجمه بر پایه یک نوع مترجم ماشینی است؛ و نوع دیگر وظیفه تکمیل و بهبود ترجمه اولیه را دارد. در روش‌هایی که ترجمه اصلی بر پایه مترجم ماشینی قاعده‌مند است، به‌نحوی سعی می‌شود تا که با استفاده از مترجم آماری، نقطه ضعف انتخاب لغات مناسب و روانی ترجمه مترجم قاعده‌مند رفع شود. همچنین در روش‌هایی که ترجمه بر پایه مترجم آماری است، سعی می‌شود با استفاده از مترجم قاعده‌مند ترجمه‌ها از نظر نحوی و ترتیب واژگان بهبود داده شوند [13].

در این مقاله، به ارائه یک معماری ترکیبی متوالی می‌پردازیم که به منظور بهره‌مندی از نقاط قوت مترجم قاعده‌مند، عمل ترجمه را بر پایه این نوع مترجم انجام می‌دهد و سپس با استفاده از یک رمزگشای یکنوا[14]، ترجمه قاعده‌مند را به‌وسیلهٔ مترجم آماری بهبود می‌بخشد. رمزگشای یک‌نوا،

[1] Machine translation
[2] Rule-based
[3] statistical
[4] morphology
[5] syntax
[6] Lexical choice
[7] Phrase base
[8] Parallel corpus
[9] Language model
[10] reordering
[11] phrase
[12] decoder
[13] Hybrid architect
[14] Monotone decoder

عمل جستجو در فضای حالات کل فرضیه‌های ترجمه ماشینی را که در آن امکان جابه‌جایی واژگان مبداء در نظر گرفته نمی‌شود، انجام می‌دهد. درواقع اگر فرض شود که ترتیب عبارات جمله مبداء با مقصد یکسان باشد، از این رمزگشا می‌توان بهره برد. این الگوریتم بر پایه برنامه‌نویسی پویا[1] تعریف می‌شود و دارای هزینه زمانی $O(n^3)$ است که n تعداد عبارات جمله مبداء است.

در این سامانه ابتدا جمله ورودی به‌وسیله یک مترجم قاعده‌مند ترجمه شده و سپس از ترتیب واژگان آن برای ترجمه نهایی استفاده می‌شود. علاوه‌براین، برای ترجمه عبارات جمله توسط رمزگشای یک‌نوا، از ترجمه‌های نامزد[2] مترجم قاعده‌مند و آماری، با به‌کارگیری مدل زبانی، استفاده می‌شود. درنتیجه ترجمه نهایی از نظر ساختار نحوی و روانی ترجمه، دارای کیفیت بهتری در معیارهای مربوطه است. همان‌طور که گفته شد در این معماری، برای عمل ترجمه از رمزگشای یک‌نوا با هزینه ترجمه چندجمله‌ای[3] استفاده شده است. همچنین به‌دلیل اینکه ابتدا جمله توسط یک مترجم قاعده‌مند با هزینه ترجمه چندجمله‌ای، ترجمه می‌شود، هزینه ترجمه نهایی معماری ترکیبی، چندجمله‌ای است که دارای هزینه کمتری نسبت به مترجم‌های آماری است.

در این مقاله آزمایش‌ها بر روی جفت زبان فارسی-انگلیسی انجام شده است که به‌ترتیب زبان‌های مبداء و مقصد هستند و نتایج نشان‌دهنده بهبود عملکرد نسبت به هر دو مترجم قاعده‌مند و آماری است. نتایج معماری ترکیبی در مقایسه با مترجم قاعده‌مند به‌طورتقریبی دارای افزایش پنج واحدی در معیار بلو[4] است که این بهبود در مقایسه با مترجم آماری دارای افزایش به‌طورتقریبی یک واحدی در این معیار است.

ادامه مقاله به این ترتیب ارائه شده است: بخش ۲ به مرور کارهای مرتبط پیشین پرداخته است. در بخش ۳ ابتدا به معرفی معماری ترکیبی می‌پردازیم؛ سپس به توضیح رمزگشای یک‌نوا پرداخته و در ادامه نحوه احتمال‌دهی به ترجمه‌های مترجم ماشینی قاعده‌مند خود و نحوه ترکیب و انتخاب بهترین ترجمه‌ها از میان ترجمه‌های نامزد مترجم قاعده‌مند و آماری را شرح می‌دهیم. در بخش ۴ به معرفی سامانه‌ها، آزمایش‌ها و نتایج آنها پرداخته و در بخش آخر نیز به نتیجه‌گیری و کارهایی که می‌توان در آینده انجام داد، اشاره‌ای می‌شود.

[1] Dynamic programming
[2] Candidate translation
[3] Polynomial complexity
[4] BLEU measure

۲- کارهای مرتبط

معماری‌های ترکیبی با دو روشِ متوالی و موازی، کار بهبود مترجم‌های ماشینی را انجام می‌دهند. هرچند هر دو روش در راستای بهبود عملکرد مترجم‌های ماشینی موفق بوده‌اند، ولی ایده اصلی معماری ترکیبی تا زمانی‌که تعامل و ارتباطی بین مترجم‌های ماشینی مختلف برقرار نباشد، محقق نمی‌شود [13]. روش‌های موازی، خروجی سامانه‌های مختلف را با یکدیگر ترکیب می‌کنند؛ به‌گونه‌ای که جمله ورودی به‌صورت موازی، توسط چندین مترجم ماشینی ترجمه شده و سپس بخش‌های مختلف خروجی آنها به‌عنوان ترجمه نهایی، در کنار یکدیگر قرار می‌گیرند؛ ازاین‌رو به این روش، ترکیب سامانه‌ها نیز گفته می‌شود. پارک و همکاران [16] با ارائه یک طبقه‌بند[5]، خروجی سه مترجم آماری و دو مترجم قاعده‌مند را با یکدیگر ترکیب کرده‌اند. مقاله یادشده با استفاده از مجموعه‌ای از ویژگی‌ها[6] که بر اساس اطلاعات زبانی و نحوی زبان‌های مبداء و مقصد هستند، یک طبقه‌بند ارائه داده است که توانایی انتخاب بهترین ترجمه برای یک عبارت را دارد. ترجمه‌های هر عبارت که توسط مترجم‌های ماشینی تولید شده‌اند، به‌عنوان ورودی به طبقه‌بند داده شده و خروجی آن، بهترین ترجمه برای آن عبارت است. با این ایده بخش‌های مختلف ترجمه‌های مترجم‌های ماشینی متفاوت، با یکدیگر ترکیب می‌شوند تا ترجمه نهایی سامانه ایجاد شود.

ایده ترکیب خروجی مترجم‌های ماشینی، به‌طورمعمول با ساخته‌شدن شبکه ترجمه[7] همراه است. در شبکه ترجمه هر گره به‌عنوان یک عبارت و هر یال به‌عنوان یک ترجمه نامزد با احتمال ترجمه‌اش است، هدف یافتن بهترین مسیر با بیشترین احتمال از گره نخست به آخر است. ماچری و همکارانش [14] در مقاله خود نیز با درنظرگرفتن خروجی یک مترجم ماشینی به‌عنوان ترجمه اصلی، با استفاده از هم‌ترازی[8] بین جمله مبداء و ترجمه هر مترجم، ترتیب واژگان خروجی مترجم‌های ماشینی را بر اساس ترتیب آنها در ترجمه اصلی تغییر داده و بار دیگر با استفاده از هم‌ترازی بین جمله مبداء و ترجمهٔ مترجم‌های ماشینی، شبکهٔ ترجمه را ایجاد می‌کند. در این شبکه، عبارات مختلف جمله مبداء به‌عنوان گره و ترجمه‌های نامزد آن که از مترجم‌های ماشینی مختلف تولید شده‌اند، به‌عنوان یال خروجی از آن گره شناخته می‌شوند. مقاله معرفی‌شده با توجه به احتمالی که مترجم‌های

[5] classifier
[6] features
[7] Translation network
[8] alignment

ماشینی مختلف به ترجمه خود داده‌اند، برای هر عبارت بهترین ترجمه را انتخاب می‌کند و درنهایت بهترین مسیر با بیشترین احتمال ترجمه در این شبکه یافت می‌شود.

تفاوت اصلی کارهای معرفی‌شده با مترجم ماشینی ارائه‌شده در این مقاله در روش کار آنها است. درواقع کارهای معرفی‌شده از روش موازی برای عمل ترجمه استفاده می‌کنند؛ درحالی‌که روش ارائه‌شده در این مقاله بر پایه روش‌های متوالی است. در روش متوالی از الگوریتمی برای انتخاب بهترین ترجمه از میان خروجی چندین مترجم ماشینی استفاده نمی‌شود، بلکه مرحله‌به‌مرحله ترجمه اولیه بهبود داده می‌شود. در ادامه به بررسی این روش و کارهای انجام‌شده در این زمینه می‌پردازیم.

در روش‌های متوالی ترجمه نهایی بر اساس ترجمه یک نوع مترجم ماشینی ایجاد می‌شود و نوع دیگر وظیفه کامل‌کردن این ترجمه و بهبود آن را دارد. در این روش، ترجمه نهایی می‌تواند بر پایه مترجم آماری و یا قاعده‌مند باشد. مقاله [7] با تکیه بر مترجم آماری، سعی کرده است که با ترجمه پیکره خود با استفاده از چندین مترجم قاعده‌مند، جدول عبارات[1] مترجم آماری را کامل‌تر کند. هدف از این کار نیز قرارگرفتن ترجمه‌های قاعده‌مند در کنار ترجمه‌های مترجم آماری است که باعث کامل‌ترشدن جدول عبارات و در نهایت بهبود عملکرد مترجم آماری می‌شود. مقاله [4] نیز با تکیه بر مترجم آماری این کار را انجام می‌دهد. در این مقاله واژگان و عبارات جمله مبداء بر اساس زبان مقصد جابه‌جا می‌شوند، سپس با استفاده از جدول عباراتی که این جابه‌جایی در آن اعمال شده است، عمل ترجمه با استفاده از مترجم آماری یک‌نوا انجام می‌شود که باعث بهبود ترجمه نهایی از لحاظ ترتیب واژگان می‌شود. سانچز و همکاران [17] نیز جدول عبارات مترجم آماری را با استفاده از یک انتقال سطح پایین قاعده‌مند[2] و فرهنگ لغت[3] بهبود بخشیده‌اند. در این فرآیند هر عبارت در جدول عبارات با به‌کارگیری مجموعه‌ای از قوانین زبانی و نحوی تغییراتی در آن اعمال و سپس بررسی می‌شود که آیا عبارت جدید ترجمه‌ای در فرهنگ لغت مترجم قاعده‌مند دارد یا خیر، و در صورت وجود، ترجمه فرهنگ لغت را به جدول عبارات مترجم آماری اضافه می‌کند. تفاوت کارهای انجام‌شده در این زمینه با روش بیان‌شده در این مقاله، در انتخاب مترجم ماشینی پایه است. در کارهای معرفی‌شده از مترجم آماری به‌عنوان مترجم پایه استفاده و با استفاده از اطلاعات مترجم قاعده‌مند ترجمه بهبود داده شده است؛ درحالی‌که در روش ارائه‌شده در این مقاله از مترجم قاعده‌مند به‌عنوان مترجم پایه استفاده شده است و با استفاده از اطلاعات مترجم آماری و رمزگشای یک‌نوا، ترجمه اولیه بهبود داده می‌شود.

مقالات [1] و [10] در کارهای خود فرهنگ لغت مترجم قاعده‌مند را با استفاده از جدول عبارات بهبود بخشیده‌اند. در مقالات نامبرده چون ترجمه‌های عبارات فرهنگ لغت مترجم قاعده‌مند به‌طورمعمول ترجمه‌های روانی نیستند، سعی شده است با استفاده از روش‌هایی، معادل آن عبارات و یا خود آنها را در جدول عبارات پیدا کرده و ترجمه روان آن را به فرهنگ لغت اضافه کنند. تفاوت عملکرد این روش با روش ارائه‌شده در این مقاله در معماری و نحوه عملکرد رمزگشا است. در روش معرفی‌شده درواقع از مترجم قاعده‌مند استفاده شده که فرهنگ لغت آن بهبود یافته است. درحالی‌که در مقاله ارائه‌شده معماری و نحوه عملکرد رمزگشا به‌طورکامل متفاوت است که برتری آن نسبت به مقالات مطرح‌شده است، زیرا از یک رمزگشای یک‌نوا استفاده شده است که باعث روان‌ترشدن ترجمه نسبت به ترجمه‌های قاعده‌مند می‌شود. مقاله [2] نیز در کار خود به پس‌ویرایش[4] خروجی مترجم قاعده‌مند پرداخته است. در این ایده یک پیکره موازی ارائه شده که یک سمت آن خروجی مترجم قاعده‌مند و سمت دیگر آن ترجمه بهبودیافته آن است. با استفاده از این پیکره یک جدول عبارات به‌وجود می‌آید و ترجمه خروجی مترجم قاعده‌مند به‌وسیله آن پس‌ویرایش و تغییراتی در آن ایجاد می‌شود که این کار باعث بهبود نتایج می‌شود. تفاوت این روش با روش ارائه‌شده در این مقاله نیز در معماری و نحوهٔ عملکرد رمزگشا است. در این روش درواقع از مترجم ماشینی آماری با بهره‌مندی از یک جدول عبارات خاص و در روش ارائه‌شده در این مقاله از یک رمزگشای یک‌نوا با معماری متفاوت استفاده شده است.

معماری ترکیبی ارائه‌شده توسط اسپانا بونت و همکارانش در مقاله [8] به‌عنوان نزدیک‌ترین کار به فعالیت ما و شبیه به معماری ترکیبی ارائه‌شده در این مقاله است. مقاله یادشده ابتدا عمل ترجمه را با استفاده از مترجم قاعده‌مند انجام می‌دهد؛ سپس بر اساس درخت جمله ترجمه‌شده که از خروجی مترجم قاعده‌مند به‌دست آمده است، ترجمه عبارات مختلف جمله را تقویت می‌کنند. در مقاله معرفی‌شده ابتدا مرزهای عبارات مختلف را بر اساس خروجی مترجم قاعده‌مند به‌دست می‌آورند؛ سپس برای هر عبارت علاوه‌بر ترجمه

[1] Phrase table
[2] Shallow transfer rule
[3] Dictionary
[4] Post editing

مترجم قاعده‌مند، ترجمه‌های نامزدی از جدول عبارات انتخاب می‌شود؛ درنهایت از بین کلیه ترجمه‌های نامزد بر اساس مجموعه‌ای از ویژگی‌ها مانند مدل زبانی و امتیاز هر ترجمه، بهترین ترجمه انتخاب می‌شود. این فرآیند توسط رمزگشای مترجم ماشینی موزز[1] [12] انجام می‌شود و مجموعه‌ای از ویژگی‌ها مانند احتمال ترجمه استخراج‌شده از فرهنگ لغت، از مترجم قاعده‌مند به آن اضافه شده است. این فرآیند بدون درنظرگرفتن مدل جابه‌جایی و به‌صورت یک‌نوا انجام می‌شود؛ ولی جابه‌جایی در ترتیب واژگان در جمله ورودی صورت نمی‌گیرد. همان‌طور که گفته شد، روش معرفی‌شده نزدیک‌ترین روش به روش ارائه‌شده در این مقاله است. تفاوت اصلی در جابه‌جایی عبارات جمله مبداء بر اساس ترجمه قاعده‌مند است. در مقاله [8] تغییری در ترتیب عبارات در جمله ورودی وجود ندارد و فقط از رمزگشای موزز با حذف مدل جابه‌جایی استفاده می‌شود؛ درحالی‌که در روش ارائه‌شده در این مقاله، ترتیب عبارات جمله ورودی بر اساس ترجمه قاعده‌مند تغییر می‌کند و سپس با استفاده از رمزگشای یک‌نوا عمل ترجمه انجام می‌شود که باعث می‌شود ترجمه نهایی از نظر ترتیب واژگان و عبارات در معیارهای مربوطه دارای عملکرد بهتری باشد.

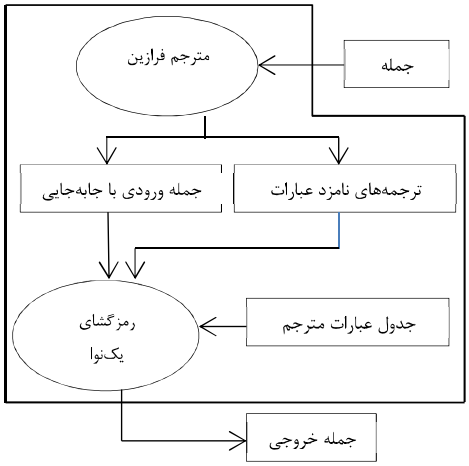
کاری که ما در این مقاله انجام داده‌ایم، استفاده از یک رمزگشاهای یک‌نواخت برای عمل ترجمه و بهبود ترجمه مترجم قاعده‌مند با استفاده از جدول عبارات است. این رمزگشا در مقایسه با رمزگشای مترجم ماشینی موزز پیچیدگی زمانی کمتری دارد؛ همچنین برای احتمال‌دهی به ترجمه‌های مترجم قاعده‌مند خود که دارای احتمال ترجمه نیستند، روش‌هایی ارائه شده است که در بخش‌های بعد به‌تفصیل به آن می‌پردازیم.

۳- معماری سامانه

معماری ترکیبی ارائه‌شده بر اساس یک مترجم قاعده‌مند و یک مترجم آماری است. مترجم قاعده‌مند استفاده‌شده در این معماری، مترجم انگلیسی به فارسی فرازین[2] است که توسط آزمایشگاه پردازش متن دانشگاه تهران تهیه و توسعه داده شده است. این مترجم ماشینی بر اساس قواعد نحوی و ساخت‌واژی، جمله ورودی را ترجمه می‌کند و خروجی آن علاوه‌بر ترجمه نهایی، حاوی اطلاعاتی شامل بخش‌بندی‌ها و ترتیب عبارات و هم‌ترازی بین جمله مبداء و مقصد در سطح عبارت است. علاوه بر این موارد، برای هر جمله مجموعه‌ای از اطلاعات دیگر مانند برچسب‌های اجزای کلام[3]، عبارات باهم‌آیی و پرکاربرد[4] که نتیجه پردازش این مترجم است، نیز وجود دارد. مترجم آماری مبتنی بر عبارت استفاده‌شده در این معماری نیز، مترجم ماشینی موزز است که با استفاده از پیکره افک[5] [9] آموزش داده شده و جدول عبارات آن استخراج شده است.

۳-۱- معماری ترکیبی

معماری ترکیبی ارائه‌شده بر پایه مترجم ماشینی فرازین است. شکل (۱) شمای[6] کلی معماری را نشان می‌دهد. بر اساس این شکل، ابتدا جمله ورودی توسط مترجم فرازین ترجمه می‌شود و سپس خروجی آن مورد پردازش قرار می‌گیرد. در این پردازش ابتدا از بخش‌بندی عبارات ترجمه فرازین استفاده و مرز عبارات برای ترجمه مشخص می‌شود؛ سپس بر اساس ترتیب ترجمه عبارات، در ترجمه فرازین، این عبارات در کنار یکدیگر قرار می‌گیرند و ترتیب آنها با جمله مقصد یکسان می‌شود؛ درنتیجه می‌توان عمل ترجمه را با استفاده از یک رمزگشای یک‌نوا که در آن جابه‌جایی واژگان در نظر گرفته نمی‌شود، با هزینه ترجمه چندجمله‌ای انجام داد.



(شکل-۱): شمای کلی معماری ترکیبی
(Figure-1): hybrid architecture's schema

شکل (۲) یک نمونه از این تغییر ترتیب واژگان را نشان می‌دهد. هر یک از عبارات جمله مبداء بر اساس موقعیت قرارگیری ترجمه‌شان در ترجمه فرازین، در کنار یکدیگر قرار می‌گیرند.

[1] moses
[2] http://www.faraazin.ir/
[3] Port of speech tag
[4] Simple phrase
[5] AFEC
[6] schema

(شکل-۲): تغییر ترتیب کلمات جمله ورودی
(Figure -2): reordering of the input sentence

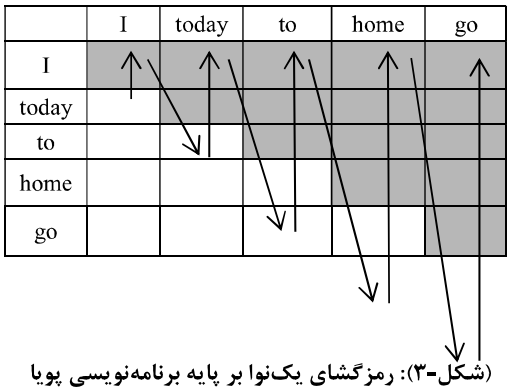
جمله مبدأ	I go to home today .
ترجمه مقصد فرازین	من امروز به خانه می‌روم.
جمله مبدأ با جابه‌جایی بر اساس ترجمه مقصد	I today to home go .

به‌عنوان مثال، چون ترجمه عبارت "go" در انتهای ترجمه آمده است، پس در جابه‌جایی جملهٔ مبدأ نیز، این عبارت به آخر جمله منتقل می‌شود. با ترجمه جمله جدید، درواقع ترتیب واژگان ترجمه نهایی بر اساس ترجمه فرازین صورت می‌گیرد. پس از جابه‌جایی جمله ورودی، فعالیت رمزگشای یک‌نوا آغاز می‌شود. در این مرحله، برای هر عبارت تشخیص‌داده‌شده توسط فرازین، از میان ترجمه‌های نامزد مترجم قاعده‌مند و آماری، بر اساس احتمال ترجمه و مدل زبانی بهترین ترجمه‌ها انتخاب می‌شود.

درواقع سامانه با استفاده از یک رمزگشای یک‌نوا که در بخش بعدی بیشتر آن را توضیح می‌دهیم، در هر مرحله، ترجمه‌های فرازین را با ترجمه‌های جدول عبارات ترکیب کرده و از بین ترجمه‌های نامزد، بهترین‌ها را بر اساس امتیاز هر ترجمه که در ادامه توضیح داده می‌شود، انتخاب می‌کند؛ و این کار را تا زمانی ادامه می‌دهد که در مرحله آخر بهترین ترجمه برای جمله مورد نظر بر اساس معیارهای معرفی‌شده در بخش ۲-۴ به‌دست آید.

۲-۳- رمزگشای یک‌نوا

ایده اصلی رمزگشای یک‌نوا برای نخستین‌بار در [18]ارائه شد. در این ایده برای جفت‌زبان‌هایی که از نظر ترتیب واژگان به هم نزدیک هستند، می‌توان با استفاده از برنامه‌نویسی پویا رمزگشایی طراحی کرد که عمل ترجمه را با هزینه چندجمله‌ای انجام دهد. معماری ترکیبی ما نیز بعد از ترجمه جمله توسط مترجم فرازین و در نخستین مرحله پردازش خود، جابه‌جایی جمله انگلیسی را با توجه به ترتیب واژگان ترجمه فرازین انجام می‌دهد. در نتیجه هم‌ترازی عبارات جمله مبداء و مقصد یک‌نواخت می‌شود و می‌توان از رمزگشای یک‌نوا برای عمل ترجمه استفاده کرد. همان‌گونه که در شکل (۳) نیز قابل مشاهده است، برای عمل ترجمه از یک ماتریس دوبعدی با اندازهٔ n*n استفاده شده است که هر بعد آن بیان‌گر تعداد عبارات جمله مبداء تشخیص داده‌شده توسط فرازین است. این ماتریس درواقع یک ماتریس بالامثلثی است که خانه‌های پایین قطر اصلی در فرآیند ترجمه، ترجمه نمی‌شوند و نحوهٔ کامل‌کردن آن به‌صورت ستونی، از پایین به بالا است.



(شکل-۳): رمزگشای یک‌نوا بر پایه برنامه‌نویسی پویا
(Figure-3) : monotone decoder based on dynamic programming

هر خانه [i,j] بیان‌گر ترجمه عبارت i ام تا j ام تشخیص داده‌شده توسط فرازین است که احتمال آن بر اساس رابطه (۱) از ضرب وزن‌دار احتمال ترجمه و مدل زبانی محاسبه می‌شود. در این رابطه α و β فاکتورهای وزنی پارامترهای رابطه هستند که در بخش ۴ در رابطه با مقداردهی آنها صحبت می‌شود. در انتها نیز بهترین ده ترجمه به‌عنوان ترجمه عبارت مورد نظر در خانه [i,j] ذخیره می‌شوند.

انتخاب بهترین ده ترجمه برای هر خانه بر اساس روش‌های مکاشفه‌ای[1] انجام شده است. دلیل انجام این کار این است که ترکیب ترجمه‌های نامزد غیر از بهترین ترجمه در هر خانه نیز بررسی شود تا بهترین ترکیب ترجمه‌ها به‌دست آید. نتایج آزمایش‌ها نشان می‌دهد که انتخاب بیش از ده ترجمه نامزد برای هر خانه تأثیری در نتایج ندارد.

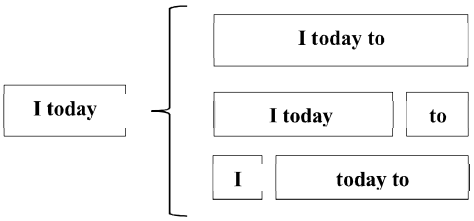
$$p[i,j] = \max_{i\le k<j} [\ (\text{احتمال ترجمه})[i,k] * (\text{احتمال ترجمه})[k+1, j] * (\text{احتمال مدل زبانی})(\text{ترجمه}[i,k] + \text{ترجمه}[k+1,j])^{\beta}\ ] \quad (۱)$$

برای ترجمه هر خانه با اندیس[2] [i,j]، عبارت مورد نظر به دو زیرعبارت[3] در محل k شکسته می‌شود و ترجمه‌های خانه‌های [i,k] و [k+1,j] با هم ترکیب می‌شوند. مقدار k نیز برابر $i \le k < j$ است. درواقع ترکیب‌های مختلف زیرعبارات، عبارت مورد نظر بررسی می‌شود تا ترکیب با بیشترین احتمال یافت شود. احتمال ترجمه این ترکیب نیز برابر با حاصل‌ضرب احتمالات ترجمه دو خانه ترکیب شده است و احتمال مدل زبانی نیز جداگانه محاسبه می‌شود. خانه‌های سطر نخست بیان‌گر ترجمه کل جمله از عبارت نخست تا عبارت مشخص‌شده هستند، و خانه [0,n] حاوی ترجمه نهایی کل

[1] heuristic
[2] index
[3] Sub-phrase

جمله است. درواقع در یک ستون زمانی که از قطر اصلی به سمت بالا پیش می‌رویم، به‌صورت عقب‌گرد از عبارت پایانی به سمت نخست جمله حرکت می‌کنیم.

خانه‌های غیر از قطر اصلی، نشان‌گر ترجمه ترکیب چندین عبارت هستند که برای ترجمه آنها از ترکیب ترجمه‌های زیرعبارات آن عبارت که درقبل محاسبه شده است، و جدول عبارات استفاده می‌شود. این فرآیند در شکل (۴) قابل مشاهده است. به‌عنوان مثال، برای ترجمه عبارت "i today to" که از سه زیرعبارت "i"، "today" و "to" تشکیل شده است، می‌توان کل عبارت را توسط جدول عبارات ترجمه کرد و یا می‌توان از ترکیب زیرعبارات "i" و "today to" با هم و یا ترکیب "i today" و "to" استفاده کرد که هر کدام از این خانه‌ها در مراحل قبل توسط بهترین ترجمه‌های آن عبارت تکمیل شده‌اند. می‌توان این فرآیند را به‌عنوان پیداکردن بهترین مسیر از یک رأس به رأس دیگر در شبکه ترجمه در نظر گرفت. در بخش ۲ به شبکه ترجمه اشاره شد؛ فرآیند پیداکردن بهترین مسیر در شبکه ترجمه به‌گونه‌ای است که اگر بین دو رأس یال مستقیمی وجود داشته باشد، برای پیمودن مسیر بین این رئوس، می‌توان از آن یال استفاده و همچنین می‌توان از رأس‌های بین این دو رأس برای پیمودن این مسیر نیز استفاده کرد.



(شکل-۴): حالات مختلف ترکیب زیرعبارات
(Figure –4): different way of combine sub-phrases

خانه‌های قطر اصلی این ماتریس نیز نشان‌گر ترجمه یک عبارات به‌تنهایی هستند؛ درنتیجه نمی‌توان برای ترجمه آنها از زیرعبارات استفاده کرد و فقط از ترجمه‌های فرازین و جدول عبارات استفاده می‌شود. احتمال خانه‌های قطر اصلی نیز بر اساس رابطه (۲) از ضرب وزن‌دار احتمال ترجمه و احتمال مدل زبانی به شرط دو کلمه ماقبل محاسبه می‌شود.

$$p[\mathrm{i},\mathrm{i}] = \max_{z \in \text{کلمات مرزی ستون قبل}} [\ (\text{احتمال ترجمه } i)^{\alpha} * (\text{احتمال مدل زبانی } [i \text{ ترجمه عبارت } | z])^{\beta}\ ] \qquad (۲)$$

برای محاسبه احتمال مدل زبانی نیز، به ازای هر ستون، در تمامی خانه‌های آن، دو کلمه آخر بهترین ده ترجمه به‌عنوان واژگان مرزی ترجمه تا آن عبارت، ذخیره می‌شوند. درواقع می‌توان ماتریس ترجمه را یک ماتریس سه‌بعدی فرض کرد که بُعد سوم آن مربوط به مدل زبانی است و اندازهٔ آن برای هر ستون برابر تعداد واژگان مرزی ستون قبل در ماتریس ترجمه است. برای ترجمه هر خانه نیز، ترجمه را نسبت به تمامی واژگان مرزی ستون قبل بررسی می‌کنیم تا بهترین ترکیب با بیشترین احتمال یافت شود.

با توجه به شکل (۳)، به‌عنوان مثال برای شروع ابتدا خانه [0,0] را ترجمه می‌کنیم. این خانه بیان‌گر ترجمه نخستین عبارت تشخیص‌داده‌شده توسط فرازین یعنی عبارت "i" به شرط نخست جمله است که با فرضیات بخش قبل، ترجمه می‌شود و بهترین ده ترجمه آن و دو کلمه آخر آنها به‌عنوان واژگان مرزی ستون نخست ذخیره می‌شوند؛ سپس خانه [1,1] ترجمه می‌شود. در این خانه به ترجمه دومین عبارت تشخیص داده شده توسط فرازین، عبارت "today" به‌شرط دو کلمه ماقبل یعنی دو کلمه آخر ترجمه‌های نامزد عبارت "i" می‌پردازیم. در خانه [0,1] نیز به ترکیب ترجمه عبارات نخست و دوم تشخیص‌داده‌شده توسط فرازین یعنی ترکیب ترجمه‌های خانه‌های [0,0] و [1,1] یعنی عبارات "i" و "today" پرداخته می‌شود و علاوه بر این، اگر کل عبارت خانه مورد نظر یعنی عبارت "i today" نیز در جدول عبارات وجود داشته باشد، از ترجمه آن استفاده می‌شود و این کار تا محاسبه بهترین ترجمه خانه [0,n] ادامه پیدا می‌کند.

۳-۳- احتمال‌دهی و انتخاب ترجمه‌ها

در فرآیند ترجمه هر عبارت، ابتدا در صورت وجود، ده ترجمه نخستع آن عبارت از جدول عبارات استخراج می‌شود. بعد از آن نیز، در صورت وجود، ده ترجمه نخست فرازین، پس از محاسبه احتمال ترجمه برای آنها به ترجمه‌های استخراج‌شده از جدول عبارات اضافه می‌شوند. درنهایت از میان این ترجمه‌ها بر اساس احتمال ترجمه و مدل زبانی، بهترین ده ترجمه به‌عنوان ترجمه‌های آن عبارت انتخاب می‌شوند. ترجمه‌های فرازین دارای احتمال ترجمه نیستند و برای احتمال‌دهی به آنها از روش‌های مکاشفه‌ای زیر استفاده می‌کنیم :

- نخستین روشی که برای احتمال‌دهی به ترجمه‌های فرازین استفاده می‌شود، روش تطابق بین ترجمه فرازین و ترجمه‌های عبارت مورد نظر در جدول عبارات است. درصورتی‌که ترجمه فرازین در ترجمه‌های عبارت مورد نظر در جدول عبارات وجود داشته باشد از تطابق دقیق[1] برای

[1] Exactly matching

احتمال‌دهی به ترجمه فرازین استفاده می‌شود و احتمال آن به ترجمه فرازین اختصاص می‌یابد. درصورت عدم تطابق دقیق از روش کم‌ترین فاصله ویرایش[1] برای تطابق جزئی[2] استفاده می‌شود. در این روش نیز ترجمه‌ای که دارای کمترین فاصله ویرایش است، احتمالش به ترجمه فرازین اختصاص داده می‌شود.

- درصورتی‌که ترجمه‌ای از فرازین مطابقتی در جدول عبارات پیدا نکند از تعدادی فرضیات مکاشفه‌ای برای احتمال‌دهی به ترجمه‌های فرازین استفاده می‌شود. به‌عنوان مثال برای احتمال‌دادن به ترجمه دوم فرازین از احتمال ترجمه دوم، ترجمه‌های آماری استخراج‌شده از جدول عبارات برای عبارت مورد نظر استفاده می‌شود و یا مجموع احتمالات ترجمه‌های آماری استخراج‌شده برای عبارت مورد نظر به‌صورت یکسان بین ترجمه‌های فرازین تقسیم می‌شود.

همان‌طور که درقبل اشاره شد، برای هر عبارت، بهترین ده ترجمه از بین ترجمه‌های نامزد آن عبارت انتخاب و برای این منظور، از روش رأی‌گیری[3] استفاده می‌شود. در این روش که به‌طورمعمول در مترجم‌های ماشینی ترکیبی با معماری موازی استفاده می‌شود، ترجمه‌هایی که در چندین مترجم ماشینی، به‌صورت یکسان وجود دارند، یک ترجمه درنظر گرفته می‌شوند و احتمال ترجمه همگی آنها با هم جمع می‌شود. به‌عنوان مثال در شبکه ترجمه که پیش‌تر به آن اشاره کردیم، یال‌هایی که دارای ترجمه یکسان هستند، به یک یال با احتمال ترجمه‌ای برابر با جمع احتمالات ترجمه آنها تبدیل می‌شوند. هدف از این کار نیز این است که اگر یک ترجمه، توسط چندین مترجم ماشینی تولید شده باشد، به‌احتمال ترجمه مناسبی است و باید احتمال بالاتری نسبت به بقیه ترجمه‌های نامزد به آن اختصاص داده شود؛ ازاین‌رو احتمالات ترجمه‌ای که مترجم‌های ماشینی مختلف به آن ترجمه اختصاص داده‌اند، با هم جمع می‌شوند. گفتنی است در این حالت امکان دارد احتمال ترجمه به امتیاز ترجمه تبدیل شود.

به این منظور ما نیز زمانی که ترجمه‌های جدول عبارات استخراج شدند و ترجمه‌های فرازین نیز احتمال ترجمه‌شان محاسبه شد، این عمل را انجام داده و سپس بر اساس احتمال ترجمه‌های موجود و احتمال مدل زبانی آنها نسبت به دو کلمه قبل از خود با استفاده از یک مدل زبانی ۳-گرام[4]، بهترین ده ترجمه را به‌عنوان ترجمه آن عبارت در نظر می‌گیریم و ذخیره می‌کنیم.

[1] Minimum edit distance
[2] Partial matching
[3] voting
[4] trigram

۴- آزمایش‌ها

به‌منظور استفاده از جدول عبارات مترجم آماری، نیاز به یک پیکره موازی برای آموزش مترجم موزز است. پیکره استفاده‌شده برای این آموزش، پیکره دو زبانه افک است که اطلاعات این پیکره در جدول (۱) ارائه شده است.

از طرفی مجموع دادگان آزمون[5] و توسعه[6] ما نیز، یک مجموعه دادگان ۸۱۸ جمله‌ای با دامنه خبر است. به‌طورتقریبی نیمی از این مجموعه به‌منظور تنظیم پارامترهای[7] مربوط به رابطه (۱ و ۲)، یعنی وزن‌های احتمال ترجمه و احتمال مدل زبانی، به‌عنوان دادگان توسعه و نیم دیگر آن به‌عنوان دادگان آزمون استفاده شده است. اطلاعات این مجموعه دادگان نیز در جدول (۱) قابل مشاهده است. این مجموعه دارای ۴ ترجمه مرجع[8] است که برای محاسبه معیارهای ارزیابی از آنها استفاده شده است.

(جدول-۱): اطلاعات پیکره
(Table-1): corpus information

	تعداد واژگان	تعداد واژگان یکتا	تعداد جملات	متوسط طول جملات (کلمه)
دادگان آموزش انگلیسی	۵/۱۴ میلیون	۱۸۵ هزار	۶۸۳ هزار	۲۱
دادگان آموزش فارسی	۴/۱۵ میلیون	۲۰۲ هزار	۶۸۳ هزار	۲۲
دادگان توسعه	۱۱/۶۲۶	۳/۰۵۱	۴۱۸	۲۸
دادگان آزمون	۱۰/۶۵۸	۲/۷۶۴	۴۰۰	۲۷

۴-۱- معرفی سامانه‌های ترکیبی

دو پارامتر تعیین‌کننده در کیفیت خروجی یک مترجم ماشینی، ارائه ترتیب مناسب برای واژگان و انتخاب معادل‌های مناسب برای واژگان و عبارات است. به‌منظور مقایسه عملکرد سامانه‌های مترجم ماشینی مختلف، سامانه‌هایی با ترکیب‌های مختلف این دو پارامتر پیاده‌سازی شده‌اند که اطلاعات آنها در جدول (۲) قابل مشاهده است. به‌منظور مقایسه عملکرد سامانه ترکیبی نسبت به دو سامانه قاعده‌مند و آماری، سامانه یک

[5] Test data
[6] development data
[7] Parameter tuning
[8] Gold data

به‌عنوان سامانه قاعده‌مند و سامانه دو به‌عنوان سامانه آماری پیاده‌سازی شده است.

(جدول –۲): اطلاعات سامانه‌ها
(Table–2): systems information

نام سامانه	ترتیب کلمات	انتخاب لغات
۱	مترجم قاعده‌مند	مترجم قاعده‌مند
۲	مترجم آماری	مترجم آماری
۳	مترجم قاعده‌مند	مترجم آماری
۴	مترجم قاعده‌مند	مترجم قاعده‌مند و آماری
۵	مترجم آماری	مترجم قاعده‌مند و آماری

بر اساس این جدول، سامانه یک، یک سامانه قاعده‌مند بوده که درواقع مترجم فرازین است. مترجم فرازین به‌منظور انتخاب لغات از واژگان داخلی خود و از روشی مبتنی بر قواعد نحوی برای ترتیب واژگان ترجمه استفاده می‌کند. در پیاده‌سازی سامانهٔ دو که یک سامانه آماری است، فقط از اطلاعات مترجم آماری مبتنی بر عبارت موزز استفاده شده است. این مترجم برای انتخاب لغات از جدول عبارتی که به‌صورت آماری از پیکره موازی استخراج کرده است و برای ارائه ترتیب ؤ ترجمه از مدل جابه‌جایی لغوی و مبتنی بر فاصله استفاده می‌کند. در این پیاده‌سازی ابتدا دادگان توسط مترجم موزز ترجمه شده است و سپس از هم‌ترازی در سطح عبارت ترجمه آن استفاده شده تا ترتیب واژگان ترجمه و مرز عبارات تشخیص داده شود. همچنین برای مرحله انتخاب لغات فقط از جدول عبارات این مترجم استفاده شده است.

به‌منظور مقایسه ترتیب واژگان در ترجمه‌های مترجم قاعده‌مند و آماری، سامانه ترکیبی، با بهره‌مندی از ترتیب واژگان مترجم قاعده‌مند در قالب سامانه چهار و با بهره‌مندی از ترتیب واژگان مترجم آماری در قالب سامانه پنج پیاده‌سازی شده، که در هر پیاده‌سازی از هم‌ترازی در سطح عبارت خروجی مترجم مربوطه، یعنی مترجم فرازین و موزز استفاده شده است.

برای مقایسه اینکه برای انتخاب لغات بهتر است از یک منبع استفاده شود یا هر دو، سامانه با ترتیب واژگان مترجم قاعده‌مند یک‌بار در حالتی پیاده‌سازی شده که برای انتخاب لغات فقط از اطلاعات مترجم آماری یعنی جدول عبارات مترجم موزز استفاده شده است که به‌عنوان سامانه سه معرفی شده، و بار دیگر در حالت ترکیبی، که همان سامانه چهار است، و از هر دو منبع مترجم قاعده‌مند و آماری، یعنی جدول عبارات مترجم موزز و ترجمه‌های فرازین استفاده می‌کند. با پیاده‌سازی این سامانه‌ها می‌توان سامانه قاعده‌مند را که فقط از ترجمه‌های مترجم قاعده‌مند برای انتخاب لغات استفاده می‌کند، با سامانه‌ای که برای این عمل فقط از ترجمه‌های مترجم آماری استفاده می‌کند و همچنین سامانه ترکیبی مقایسه کرد.

۲-۴- ارزیابی، نتایج و تحلیل

برای ارزیابی نتایج سامانه‌ها از سه معیار BLEU [15]، NIST [6] و LR-Score [3]، استفاده شده است. معیار بلو به‌عنوان رایج‌ترین معیار ارزیابی نتایج مترجم‌های ماشینی است که بر پایه تطابق چند-گرام‌های[1] متفاوت در ترجمه سامانه و ترجمه مرجع است. همچنین معیار NIST نیز بر پایه معیار بلو است، البته با تغییراتی در وزن‌دهی به چندگرام‌های متفاوت برای امتیازدهی نهایی. معیار LR-Score نیز با استفاده از هم‌ترازی ترجمه سامانه و ترجمه مرجع، موقعیت ترجمه هر عبارت در ترجمه سامانه را نسبت به ترجمه مرجع بررسی کرده و درنهایت با مقایسه موقعیت کلیه عبارات و اختلاف بین این موقعیت‌ها بین ترجمه سامانه و ترجمه مرجع، ترتیب واژگان ترجمه سامانه نسبت به ترتیب واژگان ترجمه مرجع را می‌سنجد.

به‌منظور احتمال‌دهی به ترجمه‌های فرازین در روش تطابق جزئی در بخش ۳-۳، برای ترجمه‌هایی که طول آنها کمتر از ۴ است، اختلاف یک و برای طول بیشتر از چهار، اختلاف ۴۰٪ طول ترجمه به‌عنوان فاصله ویرایش در روش کم‌ترین فاصله ویرایش، درنظر گرفته می‌شود. البته بررسی نتایج آزمایش‌ها نشان می‌دهد که تغییر میزان تطابق، تأثیر زیادی در نتایج سامانه ندارد؛ همچنین برای احتمال‌دهی به ترجمه نخست فرازین، اگر این ترجمه برابر با خود عبارت باشد، یعنی فرازین نتوانسته باشد عبارت مورد نظر را ترجمه کند، کمترین احتمال ترجمه‌های آماری و در غیر این‌صورت بیش‌ترین احتمال ترجمه‌های آماری به آن اختصاص داده می‌شود. اگر عبارتی در جدول عبارات وجود نداشته باشد نیز مجموع احتمالات را برابر ۰/۵، بیشترین احتمال را برابر ۰/۳ و کمترین احتمال را ۰/۲ در نظر می‌گیریم.

به‌منظور تنظیم‌کردن پارامترهای رابطه (۱)، سامانه‌های موجود را با مقادیر مختلف پارامترهای α و β بر

[1] N-gram

روی دادگان توسعه اجرا می‌کنیم. پارامترهای سامانه با بیشترین مقدار معیار بلو به‌عنوان پارامترهای اصلی هر سامانه انتخاب می‌شوند. وزن پارامترهای هر سامانه در جدول (۳) قابل مشاهده است.

(جدول-۳): وزن پارامترهای سامانه

(Table-3): weight of system parameters

نام سامانه	۲	۳	۴	۵
وزن احتمال ترجمه α	0.62	0.48	0.59	0.58
وزن احتمال مدل زبانی β	0.38	0.52	0.41	0.42

پس از تنظیم پارامترهای هر سامانه، آنها را بر روی دادگان آزمون اجرا کردیم. نتایج این آزمایش‌ها در جدول (۴) قابل مشاهده است. همان‌گونه که در این جدول مشخص است، در معیارهایی مانند بلو، که توجه بیشتر بر روی تطابق چندگرام‌های مختلف است، مشاهده می‌شود که سامانه آماری عملکرد بهتری نسبت به سامانه قاعده‌مند دارد و ترجمه آن به ترجمه انسانی نزدیک‌تر است. همچنین سامانه‌های ترکیبی که از هر دو منبع برای انتخاب لغات استفاده می‌کنند، عملکرد بهتری نسبت به سامانه‌هایی دارند که برای این کار فقط از یک منبع ترجمه مترجم قاعده‌مند و یا مترجم آماری استفاده می‌کنند. این برتری نسبت به سامانه آماری صادق نیست، ولی با این وجود سامانه ترکیبی با ترتیب واژگان مترجم قاعده‌مند در معیار بلو مقداری نزدیک به سامانه آماری دارد و در معیار NIST نیز عملکرد بهتری نسبت به سامانه آماری دارد و به‌عنوان بهترین سامانه شناخته می‌شود.

(جدول-۴): نتایج آزمایش‌ها

(Table-4): Result of the experiments

نام سامانه	۱	۲	۳	۴	۵
توضیحات سامانه	سامانه قاعده‌مند	سامانه آماری	سامانه قاعده‌مند با انتخاب لغات مترجم آماری	سامانه ترکیبی با ترتیب واژگان مترجم قاعده‌مند	سامانه ترکیبی با ترتیب واژگان مترجم آماری
BLEU	18.66	**22.50**	12.39	22.36	22.14
NIST	7.44	7.79	6.32	**8.13**	7.77
LR-Score	**0.52**	0.46	0.52	0.50	0.46

برای مقایسه ترتیب واژگان مترجم قاعده‌مند و آماری نیز، همان‌گونه که مشخص است در تمامی معیارها، به‌خصوص LR-Score نتایج ترتیب واژگان سامانه آماری ضعیف‌تر از قاعده‌مند است. همچنین در این معیار نتایج سامانه قاعده‌مند بیشترین مقدار را دارد و سامانه آماری دارای کمترین مقدار است. این موضوع نشان‌دهنده آن است که اگر برای انتخاب پارامتر ترتیب واژگان از مترجم قاعده‌مند استفاه شود، نتایج، بهتر از حالتی است که از مترجم آماری استفاده شود. انتخاب مترجم قاعده‌مند به‌عنوان سامانه پایه در معماری ترکیبی ارائه‌شده، انتخاب صحیحی است. همچنین سامانه ترکیبی که بر پایه ترتیب واژگان مترجم قاعده‌مند است، در این معیار دارای عملکرد بهتری نسبت به سامانه ترکیبی است که بر پایه ترتیب کلمات مترجم آماری است.

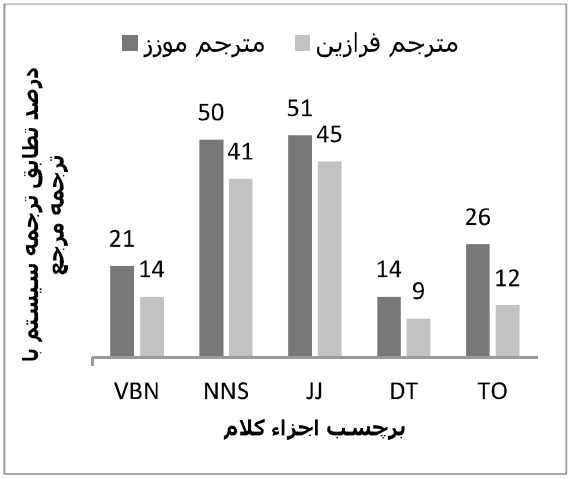
نتایج آزمایش‌ها نشان می‌دهد که ترتیب واژگان مترجم قاعده‌مند بهتر از مترجم آماری است و اگر برای انتخاب لغات از هر دو منبع استفاده شود، بهترین سامانه ترکیبی به‌دست می‌آید. به‌منظور ارتقای سامانه و دست‌یابی به سامانه‌ای بهتر از سامانه آماری، فرضیاتی را به سامانه ترکیبی اضافه می‌کنیم.

در این فرضیات با توجه به نقاط قوت عملکرد مترجم فرازین و موزز، بعضی عبارات فقط توسط یک مترجم، ترجمه می‌شود. از جمله عباراتی که فقط توسط مترجم فرازین ترجمه می‌شوند، عبارات باهم‌آیی و پرکاربرد و عبارات خاص هستند که با برچسب اجزای کلام اسم خاص[1] شناخته می‌شوند و عبارات باهم‌آیی و پرکاربرد نیز به‌طورمعمول از ترکیب چندین کلمه به‌وجود می‌آیند و به‌عباراتی اطلاق می‌شوند که فرازین برای آنها ترجمه مشخص و یک‌پارچه‌ای دارد، مانند واژگان خاص چندبخشی. در آزمایش‌هایی که انجام گرفت، مشخص شد که اگر برای ترجمه این عبارات فقط از ترجمه‌های مترجم فرازین استفاده شود، نتایج سامانه ترکیبی بهبود می‌یابد؛ همچنین برای بررسی نقاط قوت مترجم موزز، آزمایش‌هایی در زمینه برچسب‌های اجزای کلام انجام شد. در ابتدا خروجی مترجم فرازین، موزز و سامانه ترکیبی با استفاده از یک برچسب‌گذار اجزای کلام[2] فارسی، برچسب زده شدند. در

[1] Proper noun
[2] POS tagger

مرحله بعد برای هر برچسب[1]، بررسی شد که چند درصد از ترجمه‌های آن برچسب در هر سامانه، تطابق بیشتری با ترجمه‌های مرجع دارد. بررسی آزمایش‌ها نشان داد که مترجم موزز در بیش‌تر برچسب‌ها تطابق بیشتری با ترجمه‌های مرجع آن برچسب دارد. نتایج این آزمایش‌ها نشان‌دهندهٔ این است که عملکرد مترجم موزز در ترجمه این دسته از برچسب‌های اجزای کلام، بهتر از عملکرد مترجم فرازین است و اگر برای ترجمه آنها فقط از ترجمه‌های مترجم موزز استفاده شود، می‌توان عملکرد سامانه ترکیبی را بهبود داد. از میان این دسته برچسب‌های اجزای کلام، برچسب‌هایی که دارای فراوانی کمی بودند، به‌عنوان مثال تعداد آن برچسب در کل پیکره برابر ده عدد بود، حذف شده و برچسب‌هایی که دارای فراوانی بالایی بودند و علاوه‌براین عملکرد سامانه ترکیبی نیز در ترجمه آنها و تطابق با ترجمه مرجع پایین بود، انتخاب شدند. برای هر برچسب در این دسته درحالتی‌که برای ترجمه آنها فقط از ترجمه‌های مترجم موزز استفاده شود و حالت معمولی که همان حالت ترکیبی است، عملکرد سامانه ترکیبی بررسی شد. برچسب‌هایی که باعث بهبود عملکرد سامانه ترکیبی شدند به‌همراه میزان تطابق ترجمه آنها در هر سامانه با ترجمه مرجع در شکل (۵) قابل مشاهده است.

در نتیجه برای ترجمه این دسته از برچسب‌های اجزای کلام، فقط از ترجمه‌های جدول عبارات مترجم موزز استفاده شد. این برچسب‌ها شامل VBN برچسب مربوط به افعال در زمان گذشته، NNS برچسب اسامی جمع، JJ برچسب مربوط به صفات، DT برچسب مربوط به حروف اشاره و TO برچسب مربوط به حرف اضافه to است.



(شکل-۵): نتایج آزمایش‌های برچسب اجزاء کلام
(Figure-5): results of POS tag experiment

[1] tag

با اعمال این فرضیات بر روی سامانه ترکیبی، نتایج بدست آمده و مقایسه آن با رقیب اصلی یعنی سامانه آماری در جدول (۵) مشاهده می‌شود. همان‌گونه که مشخص است در تمامی معیارها سامانه ترکیبی بهتر از سامانه آماری و در بیش‌تر آنها از سامانه قاعده‌مند نیز بهتر است. همچنین آزمون معناداری[2] [11] بر روی نتایج این دو سامانه انجام شد و نتایج نشان داد که تفاوت این دو سامانه معنادار بوده است و وابسته به پیکره و مجموعه دادگان نیست.

(جدول-۵): نتایج نهایی و آزمایش معناداری
(Table-5): final results and significance test

سامانه	معیار BLEU	مقدار p-value آزمایش معناداری
سامانه ترکیبی با ترتیب کلمات مترجم قاعده‌مند	23.57	0.05
سامانه آماری	22.50	

(جدول-۶): نمونه‌ای از نتایج
(Table-6): result sample

سامانه‌ها	ترجمه	معیار بلو برای یک جمله
جمله انگلیسی	four afghans, including two students, were also killed, said hashmat stanikzai, spokesman for kabul's police chief.	
ترجمه مرجع	چهار افغان از جمله دو دانش‌اموز نیز طبق گفته‌های حشمت استانیکزای، سخنگوی رییس پلیس کابل کشته شده‌اند.	
ترجمه سامانه **قاعده‌مند**	چهار افغان‌ها از جمله دو تن از دانشجویانی، همچنین کشته شدند. ، حشمت stanikzai، سخنگو برای رییس پلیس کابل گفتند	0.11
ترجمه سامانه آماری	چهار افغان‌ها از جمله دو دانشجو کشته نیز گفت حشمت استانیکزی سخنگوی رییس پلیس کابل.	0.21
ترجمه سامانه ترکیبی	**گفت**: چهار افغان از جمله دو دانشجو، همچنین **کشته شدند**. ، حشمت استانیکزی، سخنگوی رییس پلیس کابل	0.36

در جدول (۶) نیز یک نمونه از ترجمه بهبودیافته توسط سامانه ترکیبی در مقایسه با سامانه آماری و قاعده‌مند با

[2] Significance test

معیاری مشابه معیار بلو برای یک جمله[1] را مشاهده می‌کنیم. همان‌گونه که مشخص است، سامانه آماری به‌دلیل اینکه تطابق بیشتری با ترجمه مرجع در چند-گرام‌های متفاوت دارد، عملکرد بهتری در معیار مربوطه نیز دارد، ولی همان‌گونه که مشخص است، عملکرد سامانه قاعده‌مند در زمینه ترتیب واژگان و ترجمه مناسب افعال، بهتر از عملکرد سامانه آماری است و سامانه ترکیبی که از نقاط قوت هر دو سامانه بهره برده است، دارای بهترین ترجمه است. در این ترجمه واژگان پررنگ از سامانه قاعده‌مند و بقیه واژگان و عبارات از سامانه آماری انتخاب شده است. همچنین ترتیب واژگان و عبارات نیز برگرفته از سامانه قاعده‌مند است.

۵- نتیجه‌گیری

در این مقاله با استفاده از یک رمزگشای یکنوا با هزینه چند جمله‌ای، یک مترجم ماشینی ترکیبی طراحی کردیم. این مترجم بر پایه یک مترجم قاعده‌مند است که ترتیب واژگان در ترجمه آن بر اساس مترجم قاعده‌مند و انتخاب لغات آن به‌صورت ترکیبی از مترجم‌های قاعده‌مند و آماری است. آنچه که از نتایج آزمایش‌ها مشخص است، این است که مواردی در ترجمه مانند ترتیب واژگان که مربوط به قواعد نحوی زبان است، بهتر است، توسط مترجم‌های ماشینی قاعده‌مند پوشش داده شوند. در معیارهایی که موارد نحوی مانند ترتیب واژگان را بررسی می‌کنند، این مترجم‌ها عملکرد بهتری نسبت به مترجم‌های آماری دارند؛ ولی در مقابل این مترجم‌ها در به‌کارگیری لغات مناسب و نزدیک به ترجمه انسانی ضعیف هستند و در معیارهایی مانند بلو عملکرد بدتری نسبت به مترجم‌های آماری دارند؛ درنتیجه برای رسیدن به سامانه بهینه باید برای انتخاب لغات، از ترجمه‌های نامزد مترجم قاعده‌مند و آماری با هم استفاده کرد؛ پس می‌توان با ترکیب این دو نوع معماری به مترجم ماشینی رسید که در همه معیارها بتواند فاصله اختلاف این دو نوع مترجم ماشینی را کاهش دهد و از نقاط قوت هر دو نوع بهره ببرد.

معماری ترکیبی را می‌توان با استفاده از رمزگشاهای دیگری مانند رمزگشای مترجم ماشینی موزز با ویژگی‌های بیشتر و کامل‌تر، پیاده‌سازی کرد. همچنین می‌توان با افزایش تعداد مترجم‌های ماشینی و انتخاب توابع هوشمندتر برای احتمال‌دهی به ترجمه‌های مترجم‌های قاعده‌مند و انتخاب ترجمه مناسب از بین ترجمه‌های نامزد موجود عملکرد معماری‌های ترکیبی را افزایش داد. برای انتخاب ترجمه مناسب نیز می‌توان علاوه‌بر احتمال ترجمه و مدل زبانی، ویژگی‌های دیگری مانند فرضیات اضافه شده به سامانه ترکیبی، برای رمزگشا تعریف کرد تا مرحله انتخاب ترجمه مناسب نیز هوشمندانه‌تر شود.

[1] Sentence level BLEU

۶- مراجع 6- References

[1]A. Antonova and A. Misyurev, "Improving the precision of automatically constructed human-oriented translation dictionaries", in Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL, 2014, pp. 58–66.

[2]H. Béchara, R. Rubino, Y. He, Y. Ma, and J. van Genabith, "An Evaluation of Statistical Post-Editing Systems Applied to RBMT and SMT Systems.," in COLING, 2012, vol. 21, pp. 5–230.

[3]A. Birch and M. Osborne, "LRscore for evaluating lexical and reordering quality in MT," *in Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, 2010, pp. 327–332.

[4]Y. Chen and A. Eisele, "Hierarchical hybrid translation between english and german," *in Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, 2010, pp. 90–97.

[5]M. R. Costa-Jussa, M. Farrús, J. B. Marino, and J. A. R. Fonollosa, "Study and Comparison of Rule-Based and Statistical Catalan-Spanish Machine Translation Systems.," Comput. informatics, vol. 31, no. 2, pp. 245–270, 2012.

[6]G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," *in Proceedings of the second international conference on Human Language Technology Research*, 2002, pp. 138–145.

[7]A. Eisele, C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann, and Y. Chen, "Hybrid Architectures for Multi-Engine Machine Translation," Proc. Transl. Comput. 30. ASLIB Transl. Comput. Conf. (TATC), 30th, Novemb. 27-28, London, United Kingdom, 2008.

[8]C. España Bonet, L. Màrquez Villodre, G. Labaka, A. de Ilarraza Sánchez, and K. Sarasola Gabiola, "Hybrid machine translation guided by a rule-based system," *in Machine translation summit XIII: proceedings of the 13th machine translation summit*, September 19-23, 2011, Xiamen, China, 2011, pp. 554–561.

[9]S. FattanehJabbari and S. M. M. Ziabary, "Developing an open-domain English-Farsi translation system using AFEC: Amirkabir bilingual Farsi-English corpus," *in The Fourth*

Workshop on Computational Approaches to Arabic Script-based Languages, 2012, pp. 17.

[10] N. Habash, B. Dorr, and C. Monz, "Symbolic-to-statistical hybridization: extending generation-heavy machine translation," Mach. Transl., vol. 23, no. 1, pp. 23–63, 2009.

[11] P. Koehn, "Statistical Significance Tests for Machine Translation Evaluation," in EMNLP, 2004, pp. 388–395.

[12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, and others, "Moses: Open source toolkit for statistical machine translation," *in Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 2007, pp. 177–180.

[13] G. Labaka, C. España-Bonet, L. Màrquez, and K. Sarasola, "A hybrid machine translation architecture guided by syntax," Mach. Transl., vol. 28, no. 2, pp. 91–125, 2014.

[14] W. Macherey and F. J. Och, "Consensus translations from multiple machine translation systems." Google Patents, 2012.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," *in Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 311–318.

[16] E. Park, O. Kwon, K. Kim, and Y. Kim, "A Classification-based Approach for Hybridizing Statistical Machine Translation and Rule-based Machine Translation," pp. 1–9, 2015.

[17] V. M. Sánchez-Cartagena, F. Sánchez-Mart'\inez, J. A. Pérez-Ortiz, and others, "Integrating shallow-transfer rules into phrase-based statistical machine translation," 2011.

[18] C. Tillmann, S. Vogel, H. Ney, and A. Zubiaga, "A DP based search using monotone alignments in statistical translation," *in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997, pp. 289–296.

حسین خاتمی دانشجوی کارشناسی ارشد رشته مهندسی نرم‌افزار در دانشگاه تهران از سال ۱۳۹۳ است. همچنین مقطع کارشناسی را از دانشگاه صنعتی سجاد مشهد در سال ۱۳۹۲ در رشته مهندسی نرم‌افزار اخذ کرده است. زمینهٔ پژوهشی ایشان پردازش زبان طبیعی و به‌طور خاص مترجم ماشینی است.

نشانی رایانامهٔ ایشان عبارت است از:

h.khatami@ut.ac.ir

حکیمه فدایی دانشجوی مقطع دکترای رشته مهندسی نرم‌افزار در دانشگاه تهران است. همچنین ایشان مدرک کارشناسی و کارشناسی ارشد خود را در رشته مهندسی نرم‌افزار در دانشگاه شهید بهشتی کسب کرده‌اند و زمینه پژوهشی ایشان پردازش زبان طبیعی و به‌طور خاص مترجم ماشینی است.

نشانی رایانامه ایشان عبارت است از:

h.fadaei@ut.ac.ir

هشام فیلی تحصیلات خود را در مقطع کارشناسی مهندسی نرم‌افزار در دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شریف با رتبهٔ یک در سال ۱۳۷۶ به پایان رساند؛ سپس مقاطع کارشناسی ارشد نرم‌افزار و دکترای هوش مصنوعی را به‌ترتیب در سال‌های ۱۳۷۸ و ۱۳۸۵ در همان دانشکده تکمیل کرد. از سال ۱۳۸۷ تاکنون عضو هیأت علمی دانشکدهٔ مهندسی برق و کامپیوتر دانشکدهٔ فنی دانشگاه تهران است. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش هوشمند متن و زبان طبیعی، ترجمهٔ ماشینی، داده‌کاوی، بازیابی اطلاعات و شبکه‌های اجتماعی هستند.

نشانی رایانامهٔ ایشان عبارت است از:

hfaili@ut.ac.ir