Journal of Researches in Linguistics
10(2), 15-36
Received: 11.02.2018 Accepted: 23.06.2018

Extracting Parallel English/Persian Sentences from Comparable Corpora using Syntactic Information

Rezvan Motavalian
Assistant Professor , Department of Linguistics,University of Isfahan, Iran
Amirhasan Monajemi
Associate Professor, Department of Computer Engineering, University of Isfahan, Iran
Ibrahim Qudsollahi
M.A. Student, Department of Linguistics, University of Isfahan, Iran
**Corresponding author, e-mail: rezvan_motavalian@yahoo.com*

Abstract
Parallel corpora have always been considered among the richest resources in the field of natural language processing. These corpora include translated texts from two or more languages that are usually aligned at the different levels of word, clause, or sentence. Notwithstanding the many uses of these corpora in different studies such as linguistic researches, statistical machine translation, and cross language information retrieval; unfortunately parallel corpora have always been rare and limited in number and quality. Constructing such corpora using human power is a time consuming and expensive task. When huge comparable corpora are available, it is important to try to exploit these resources to extract parallel corpora from them. Accordingly, in this paper an automatic method for extracting parallel sentences from comparable resources is introduced which exploits syntactic information. In this method, by using syntactic information of the sentences, an alignment model is trained. The highest practical accuracy of the alignment model on the test set (208 pairs of sentences) was measured to be 77% and the highest precision on the training set (830 pairs of sentences) was 97.7%. Considering the tiny size of the golden corpora, n-fold cross validation technique was used in all training algorithms. To attain higher precision, a new similarity search algorithm was implemented which increased the practical accuracy on the test set from77% to 85.15%. The final outcome of this research was an alignment toolkit and framework which was named "Isfahan University Parallel Corpus Framework" or IPCF, which can be used by the researchers in the field of computational processing of Persian language to construct standard parallel corpora.
Keywords: Automated Extraction, Parallel Sentences, Corpus Linguistics, Comparable Corpus, Parallel Corpus, Classification.

Copyright©2019, University of Isfahan. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/BY-NC-ND/4.0), which permits others to download this work and share it with others as long as they credit it, but they can't change it in any way or use it commercially.

نشریه پژوهش‌های زبان‌شناسی
سال دهم، شماره دوم، شماره ترتیبی ۱۹، پاییز و زمستان ۱۳۹۷
تاریخ وصول:۱۳۹۶/۱۱/۲۲
تاریخ اصلاحات: ۱۳۹۷/۱/۱۰، ۱۳۹۷/۳/۱۷
تاریخ پذیرش: ۱۳۹۷/۴/۲
صص ۱۵ - ۳۶

استخراج خودکار جملات هم‌تراز انگلیسی-فارسی از متون مقایسه‌ای با بهره‌برداری از اطلاعات نحوی

رضوان متولیان *
امیرحسن منجمی **
ابراهیم قدس اللهی ***

چکیده

پیکره‌های موازی همواره از غنی‌ترین منابع در مباحث پردازش زبان طبیعی محسوب می‌شوند. این نوع پیکره‌ها شامل متون ترجمه‌شدهٔ دو یا چند زبان هستند که در سطوح مختلف کلمه، عبارت و یا جمله هم‌ترازشده‌اند. علیرغم کاربرد فراوان این نوع پیکره‌ها در مطالعات مختلف از جمله پژوهش‌های زبانی، ترجمهٔ ماشینی آماری و سامانه‌های خودکار بازیابی اطلاعات میان زبانی، متأسفانه همواره پژوهشگران با کمبود پیکره‌های موازی مواجه بوده‌اند. در این راستا، در پژوهش حاضر سعی شده است به‌منظور تولید پیکره موازی با بهره‌گیری از اطلاعات نحوی، روشی خودکار برای استخراج جملات هم‌ترازانگلیسی/فارسی از متون مقایسه‌ای ارائه شود. در این روش، با ساخت بردار ویژگی با بهره‌گیری از اطلاعات نحوی جملات، یک مدل هم‌ترازی آموزش داده می‌شود. دقت مدل هم‌ترازی،در بهترین حالت، به شکل عملیاتی روی داده‌های آزمون (۲۰۸ عدد جفت جمله) ۷۷٪ و روی داده‌های آموزشی (۸۳۰ عدد جفت جمله) ۹۷/۷ ٪ محاسبه شد. از آنجایی که حجم داده‌های طلایی بسیار کوچک بود روش n-fold cross validation در مورد تمام الگوریتم‌های آموزش مورد استفاده قرار گرفت. به‌منظور افزایش دقت، از یک الگوریتم

* استادیار گروه زبانشناسی دانشگاه اصفهان (نویسنده مسؤول) r.motavallian@fgn.ui.ac.i
**دانشیار گروه کامپیوتر دانشگاه اصفهان monadjemi@eng.ui.ac.ir
*** کارشناسی ارشد زبانشناسی رایانشی دانشگاه اصفهان abrahamqudsollahi@yahoo.com

Copyright©2019, University of Isfahan. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/BY-NC-ND/4.0), which permits others to download this work and share it with others as long as they credit it, but they can’t change it in any way or use it commercially.

جست‌وجوی شباهت لغوی جملات نیز استفاده شد که دقت را روی داده‌های آزمون از ۷۷٪ به ۸۵/۱۸٪ افزایش داد. پژوهش حاضر، با به‌کارگیری مدل هم‌ترازی به‌دست‌آمده، به تولید ابزار هم‌ترازی دانشگاه اصفهان منجر شد، که می‌تواند به‌منظور خودکفایی در تولید پیکره‌های موازی مورداستفاده محققین حوزه پردازش زبان فارسی قرار گیرد.

کلیدواژه‌ها

استخراج خودکار، جملات همتراز، زبانشناسی پیکره‌ای، پیکره مقایسه‌ای، پیکره موازی.

۱. مقدمه

پیکره‌های موازی[1] به‌صورت عام در مباحث زبانشناسی، و به‌صورت خاص در حوزه پردازش زبان طبیعی، به‌ویژه در حیطه ترجمه ماشینی، از غنی‌ترین و ارزشمندترین منابع محسوب می‌شوند. پیکره‌های موازی، به بیان مختصر، پیکره‌هایی هستند که در آن‌ها جملات هم‌ترجمه از دو یا چند زبان مختلف معادل یکدیگر قرارگرفته‌اند. این پیکره‌هامی‌توانند در سطوح کلمه، عبارت، جمله و یا پاراگراف هم‌تراز شده باشند (جورافسکی و مارتین[2]، ۲۰۰۹).

پیکره‌های موازی کاربردهای بسیاری در مطالعات زبانی و پژوهش‌های مربوط به پردازش زبان طبیعی دارند. یکی از مهم‌ترین کاربردهای پیکره‌های موازی دوزبانه در بحث ساخت ماشین ترجمه به روش آماری است (میتکو[3]، ۲۰۰۵). اما کاربرد پیکره‌های موازی دوزبانه تنها به ساخت ماشین ترجمه با روش‌های آماری محدود نمی‌شود و می‌توان از این پیکره‌ها در مطالعات زبانی و میان زبانی، تصحیح ترجمه انسانی، مطالعات و ارائه روش‌ها و ابزارهای بهبود ترجمه انسانی، سامانه‌های حافظه ترجمه[4]، ارزیابی کیفیت ترجمه انسانی، و سامانه‌های خودکار بازیابی اطلاعات میان زبانی (CLIR)[5] و غیره استفاده کرد (چه‌آن و یانگ جونگ[6]، ۲۰۱۷).

تولید پیکره‌های موازی دو یاچندزبانه از پرهزینه‌ترین امور تولید محتوا برای پردازش زبان طبیعی محسوب می‌شود (مک انری و زایو[7]، ۲۰۰۷؛ براون، ۲۰۰۵). ازآنجایی که برای تولید و گسترش پیکره‌های موازی مناسب، لازم است معیارهای زیادی مدنظر قرار گیرد، در عمل تعداد پیکره‌های موازی استاندارد به خصوص در جفت‌های زبانی که یک عضو آن‌ها زبان فارسی است (مانند جفت زبانی فارسی-انگلیسی)، بسیار پایین است.

. در تولید پیکره‌های موازی می‌بایست معیارهایی مانند به‌روز بودن داده‌ها، حجم مناسب (هان و سایرین، ۲۰۰۹)، دقت در هم‌ترازی (محمدی و قاسم آقایی، ۲۰۱۰)، کیفیت داده‌ها (زاری و صدرالدینی، ۱۳۹۲)، توجه به نوع ترجمه (مفهومی، تحت‌اللفظی، کلمه به کلمه، و غیره)، نویز[8] پایین (تقی پور و سایرین،۲۰۱۰؛ پیله‌ور و سایرین، ۲۰۱۱) را مدنظرقرارداد. مسئله نویز در پیکره موازی به‌خصوص هنگامی اهمیت بیشتری پیدا می‌کند، که هدف استفاده از پیکره موازی ایجادشده در تولید ماشین ترجمه باشد. هنگامی‌که یک مترجم انسانی متنی را ترجمه می‌کند، در مواردی ناچار

[1] Parallel corpus
[2] D. Jurafsky & J. H.Martin
[3] R. Mitkov
[4] translation memory systems
[5] Cross Language Information Retrieval
[6] J. Cheon & K. O. Youngjoong
[7] A. McEnery & R. Xiao
[8] noise

به حذف و اضافه کردن قسمت‌هایی از متن است، به این حذف و اضافه‌ها که می‌تواند برای یادگیری مدل ترجمه توسط ماشین ترجمه گیج‌کننده باشد، نویز گفته می‌شود (تقی پور و سایرین، ۲۰۱۰).

برای تولید پیکره‌های موازی می‌بایست جملات هم‌ترجمه از دو زبان، هم‌تراز شوند. به عمل جداسازی جملات موازی (هم ترجمه) از متون دوزبانه (ترجیحاً ترجمه‌شده) و قرار دادن آن‌ها در مقابل یکدیگر، هم‌ترازی جملات گفته می‌شود(جورافسکی و مارتین، ۲۰۰۹) و به این جملات که ترجمه یکدیگر محسوب می‌شوند و در برابر هم دیگر قرار دارند، جملات هم‌تراز می‌گویند (گیل و چرچ[1]، ۱۹۹۳). جملات هم‌تراز می‌توانند به انواع ترجمه‌ای ۰-۱ (جمله حذف‌شده)، ۱-۰ (جمله اضافه‌شده توسط مترجم)، ۱-۲ (ترجمه دو جمله به یک جمله)، ۲-۱ (ترجمه یک جمله به دو جمله)، ۱-۱ (ترجمه یک جمله به یک جمله) تقسیم‌بندی شوند.

پیکره‌های موازی را می‌توان به سه روش دستی، نیمه‌خودکار و خودکار ایجاد نمود. در روش دستی با استفاده از نیروی انسانی و متخصص جملات هم‌ترجمه ایجاد یا انتخاب می‌شوند و در پیکره نهایی قرار می‌گیرند؛ پیکره‌هایی که به این شکل ایجاد می‌شوند پیکره‌های موازی طلایی محسوب می‌شوند (جورجیو و همکاران[2]، ۲۰۰۶). در روش نیمه‌خودکار با به‌کارگیری روش‌های آموزش ماشینی و مدل‌سازی، جملات هم‌ترجمه اولیه توسط ماشین انتخاب می‌شوند. سپس این جملات توسط یک تیم کوچک انسانی مورد بازبینی قرارگرفته و جملات نامطلوب حذف می‌شوند. این عمل منجر به تولید پیکره‌های موازی نقره‌ای می‌شود. در روش‌های خودکار با استفاده از روش‌های رایانشی مانند یادگیری ماشینی، سنجش میزان شباهت، روش‌های احتمالاتی و دیگر روش‌ها به کشف و استخراج جملات موازی از منابع زبانی پرداخته می‌شود. امروزه این روش‌ها به دلیل سرعت قابل‌توجه خود در پردازش حجم انبوهی از منابع متنی، موردتوجه هستند. یکی از منابعی که در روش‌های خودکار موردتوجه است، استفاده از منابع پیکره‌های مقایسه‌ای به‌منظور استخراج جملات هم‌تراز است. پیکره‌های مقایسه‌ای شامل متونی از دو زبان جداگانه اما در یک حوزه خاص هستند. جملات موجود در پیکره‌های مقایسه‌ای الزاماً ترجمه یکدیگر نیستند، اما می‌توان انتظار داشت که در این نوع از پیکره‌ها جملات هم‌ترجمه کشف شوند (راف و شونک[3]، ۲۰۱۱). همان‌طور که مشخص است، به دلیل سادگی تولید این پیکره‌ها، و حجم بالای آنها (بارزیلی و الحداد[4]، ۲۰۰۳)، پیکره‌های مقایسه‌ای منابع بسیار خوبی برای استخراج جملات هم‌تراز محسوب می‌شوند.

تفاوت پیکره‌های موازی و پیکره‌ها یا منابع مقایسه‌ای در این است که در مورد پیکره‌های موازی این اطمینان وجود دارد که منابع در سطوح مورد نظر مانند کلمه، عبارت، جمله، بند، یا مقاله، بدون تردید ترجمه یکدیگر هستند و از معیارهای مورد نظر پیروی می‌کنند، اما در مورد منابع یا پیکره‌های مقایسه‌ای این اطمینان وجود ندارد. هرچند ممکن است بتوان در منابع مقایسه‌ای مقالاتی پیدا کرد که در سطح جمله هم‌ترجمه باشند، اما این الزام وجود ندارد که این ترجمه‌ها دقیق باشند، یا از استانداردهای مورد نظر برخوردار باشند. به بیان دقیق‌تر، از طریق همترازسازی منابع مقایسه‌ای

[1] W.A. Gale & K.W. Church
[2] P.Georgiou et al.
[3] S.A. Rauf & H. Schwenk
[4] R. Barzilay & N. Elhadad

این اطمینان حاصل می‌شود که جملات دقیقاً ترجمه یکدیگر هستند و از معیارها و استانداردهایی که مورد نظر هستند پیروی می‌کنند.

تاکنون پژوهش‌های بسیاری به‌منظورارائه روش‌های استخراج جملات هم‌تراز انجام‌گرفته است (محمدی و قاسم آقایی، ۲۰۱۰؛ تقی پور و سایرین؛ پیله‌ور و سایرین؛ جکیان طوسی، ۱۳۹۱؛ رحیمی و همکاران[1]، ۲۰۱۲) اما بسیاری از این پژوهش‌ها به ارائه یک ابزار کاربردی متناسب با ویژگی‌های زبان فارسی منجر نشده‌اند. همچنین پیکره‌های حاصل از بسیاری از این پژوهش‌ها یا به دلایل مختلف مانند عدم انتشار در دسترس عموم پژوهشگران قرار ندارند (مانند پیکره امیرکبیر (جباری و ضیاع‌بری[2]، ۲۰۱۲)) یا از استانداردهای لازم به‌منظور کاربرد در پردازش زبان فارسی برخوردار نیستند (مانند نویز بالا در پیکره میزان). مقاله حاضر به معرفی روشی برای استخراج جملات هم‌تراز از منابع مقایسه‌ای می‌پردازد. در این راستا، مقالات ترجمه‌شده دانشنامه آزاد ویکی‌پدیا[3] به علت در دسترس بودن و عدم نیاز به کسب مجوزهای خاص به‌عنوان پیکره مقایسه‌ای مورداستفاده قرار گرفته است. وجود جملات ترجمه‌شده با نرخ رخداد بالا در این منبع و همچنین توجه نگارندگان ویکی‌پدیا به رعایت مصوبات نگارشی فرهنگستان در مقالات این دانشنامه، احتمال کشف جملات هم‌تراز با کیفیت مطلوب در این منبع را افزایش می‌دهد. لازم به ذکر است که در طراحی الگوریتم نهایی دقت لازم به‌عمل‌آمده است که جملاتی که نویز پایینی دارند از منابع استخراج شوند؛ بنابراین جملاتی که، به‌عنوان‌مثال، به شکل آزاد در دانشنامه ویکی‌پدیاترجمه‌شده‌اند، توسط ابزار و مدل هم‌ترازی ارائه‌شده در این روش در خروجی دیده نخواهند شد. روش حاضر با بهره‌گیری از اطلاعات نحوی و زبانی جملات، یک مدل هم‌ترازی برای جفت‌های زبانی انگلیسی و فارسی ایجاد می‌کند. این مدل به شکل کاربردی وروی داده‌های آزمون جملات هم‌تراز را با دقت ۷۷٪ استخراج می‌کند. گفتنی است براساس یافته‌های پژوهش حاضر با استفاده از الگوریتم‌های جست‌وجوی شباهت بر اساس داده‌های یک فرهنگ لغت، دقت مدل را می‌توان تا بیش از ۸۵٪ افزایش داد. علاوه بر این، در پژوهش حاضر، در نهایت با استفاده از مدل هم‌ترازی به‌دست‌آمده، چارچوب و ابزاری برای استخراج پیکره موازی از متون مقایسه‌ای ایجاد شد که می‌تواند نیاز پژوهشگران حوزه پردازش زبان فارسی درزمینهٔ تولید پیکره‌های موازی را تا حد قابل قبولی مرتفع نماید.

۲. سوال تحقیق و روش‌کار

با در نظر داشتن مشکلات پیش روی بحث پردازش زبان فارسی، و با توجه به پیشرفت علم پردازش زبان طبیعی در مورد سایر زبان‌ها و همچنین با عنایت به اولویت‌بندی نیازهای امور پردازش زبان فارسی با توجه به مسیر آینده، ساخت ابزارهایی که بتوانند به شکل خودکار و در حجم و کیفیت قابل قبول، منابع زبانی را برای زبان فارسی فراهم کنند یک نیاز بنیادین محسوب می‌شود. یکی از این منابع پیکره‌های موازی هستند که کاربردهای پردازشی بی‌شماری دارند. در این راستا سوال تحقیق مطرح در این مقاله این است که، از چه طریق می‌توان به شکل خودکار به ایجاد پیکره‌های موازی از پیکره‌های مقایسه‌ای اقدام نمود و در عین حال از سرعت روش ارائه شده و رعایت معیارهای لازم در این روش اطمینان حاصل نمود؟

[1] Z. Rahimi et al.
[2] M. FattanehJabbari & S. M. Ziabary
[3] Wikipedia

بعد از بررسی پیشینه پژوهشی، و در نظر داشتن میزان موفقیت روش‌های مختلف، هزینه (مالی و پردازشی) این روش‌ها، پیچیدگی این راهکارها و قابلیت پیاده‌سازی و در صورت نیاز گسترش سریع و آسان آن‌ها، این فرضیه مطرح شد که می‌توان با بهره‌برداری از اطلاعات نحوی جملات در دو زبان، به شمارش برچسب‌های دستوری جملات اقدام نموده و با استفاده از روش‌های یادگیری ماشینی یک مدل همترازی ایجاد نمود و از این مدل برای استخراج جملات همتراز از منابع مقایسه‌ای استفاده کرد.

استفاده از روش ارائه شده در این پژوهش، میزان موفقیت استخراج جملات همتراز از دانشنامه آزاد ویکی‌پدیا در پژوهش‌های پیشین را از رقم ۲۱٪ (محمدی و قاسم‌آقایی، ۲۰۱۰) و ۴۹/۱۲ ٪ (انصاری و سایرین، ۲۰۱۴) (مربوط به آستانه ۰/۱ که آزادی بیشتری به الگوریتم می‌دهد) به بیش از ۷۷ ٪ افزایش داده است.

۳. پیشینه پژوهش

اکثر روش‌های خودکار تشخیص جملات هم‌تراز، با بهره‌گیری از سنجش میزان شباهت جملات در دو زبان مبدأ و مقصد کار می‌کنند. درعین‌حال، در یک دسته‌بندی کلی می‌توان روش‌های تشخیص جملات هم‌تراز را به دودسته «مبتنی بر طول» و «مبتنی بر اطلاعات زبانی» تقسیم کرد.

از اولین روش‌های مبتنی بر اطلاعات زبانی روش چن (۱۹۹۳) است. کار چن ساخت یک مدل ترجمه ساده کلمه به کلمه در حین جست‌وجو برای جملات موازی است. ادعای چن این است که روش او، برخلاف سایر روش‌های کشف جملات موازی، به اطلاعات زبانی توجه کافی دارد و درعین‌حال از سرعت قابل قبولی برخوردار است. در روش چن آن دسته از جملات هم‌ترازی در پیکره نهایی قرار می‌گیرند که احتمال ایجاد پیکره موازی را بیشینه نمایند. مدل چن با استفاده از ۱۰۰ جمله که به شکل دستی مرتب‌شده‌اند خود راه‌اندازی[1] می‌شود. بعدازاینکه پارامترهای مدل او روی ۲۰۰۰۰ جمله موازی دیگر و با استفاده از الگوریتم EM[2] باز تخمین زده شد، در یک مرحله، جملات موازی استخراج می‌شوند. خود چن دقت روش خود را ۹۹/۶ ٪ گزارش می‌کند، اما بررسی‌های آینده نشان داد که دقت روش او پایین‌تر از این رقم است (منینگ و شوتز[3]، ۱۹۹۹).

روش گیل و چرچ (۱۹۹۳) از اولین روش‌های استخراج جملات هم‌تراز مبتنی بر طول است. اساس کار روش گیل و چرچ سنجش میزان شباهت جملات بر اساس طول آن‌هابرحسب تعداد نویسه‌های[4] به‌کاررفته در هر جمله می‌باشد. به عبارتی، در نظر گیل و چرچ جملاتی که از یک زبان به زبان دیگر ترجمه می‌شود به احتمال زیاد دارای طول یکسانی هستند. روش گیل و چرچ دقت ۹۶ ٪ را گزارش می‌کند.

سایر روش‌های استخراج جملات هم‌تراز به‌نوعی از روش‌های اشاره‌شده در بالا الگوبرداری می‌کنند. روش سیمارد و سایرین[5] (۱۹۹۳) بر این ایده استوار است که جملات هم ترجمه به شکل معمول بیشترین میزان اشتراک کلمات هم‌ریشه را دارند. در نظر آن‌ها کلماتی هم‌ریشه هستند که در جمله مبدأ و جمله مقصد قطعاً وجود خواهند داشت و به

[1] bootstrapping
[2] expectation maximization
[3] C. D. Manning & H. Schütze
[4] characters
[5] M. Simard et al.

یک بیان ترجمه می‌شوند. به‌عنوان‌مثال کلمه «تهران» همواره به کلمه «Tehran» ترجمه می‌شود و برعکس. یا یک جمله سؤالی همواره به یک جمله سؤالی ترجمه خواهد شد. دقت الگوریتم آن‌ها ۹۰/۴٪ گزارش‌شده است. در پژوهشی دیگر پاپیریدیس و همکاران[1] (۲۰۰۰) مهم‌ترین عامل در شناسایی جملات هم‌تراز را کلماتی می‌دانند که بار معنایی جملات را منتقل می‌کنند. در تعریف آن‌ها این کلمات شامل گروه‌های دستوری اسم، فعل، قید و صفت می‌شود و به‌احتمال بسیار زیاد تعداد این کلمات در یک جمله از زبان مبدأ و در ترجمه همین جمله در زبان مقصد دارای رابطه نزدیکی هستند. آن‌ها با استفاده از رگرسیون خطی چند متغیره (MLR) و با حدود ۳۰۰۰ جمله برای جفت زبانی انگلیسی-یونانی به دقت ۹۹٪ دست یافتند. روش آن‌ها در سایر تحقیقات مورد ارزیابی و بازسازی علمی قرار گرفت و در مورد جفت زبانی پرتغالی-انگلیسی دقت ۹۷/۸۴ ٪ گزارش شد (کاسلی و نانس[2]، ۲۰۰۳).

مونتانا و مارکو[3] (۲۰۰۵) با به‌کارگیری یک طبقه‌بندی آنتروپی بیشینه[4] روشی را برای شناسایی جملات هم‌ترازارائه می‌کنند. در روش آن‌ها در مقالات ورودی، با استفاده از مدل‌های زبانی و با استفاده از یک فرهنگ لغت جملات نماینده انتخاب می‌شوند. سپس با به‌کارگیری یک طبقه بند ME جملات نهایی استخراج می‌شوند. یکی از معیارهای آن‌ها استفاده از میزان شباهت کلمات بر اساس فرهنگ لغت است. روش آن‌ها برای جفت زبانی عربی-انگلیسی دقت ۸۶٪ - ۹۴٪ را گزارش می‌کند. فتاح و همکاران (۲۰۰۶) با ترکیب داده‌های زبانی مختلف و استفاده از یک طبقه‌بندی جدید جملات هم‌تراز نوع ۱-۱ را با دقت ۹۸/۱٪ در پیکره نهایی خود قراردادند. استفانسکو و همکاران[5] (۲۰۱۲) روشی متفاوت برای استخراج جملات هم‌تراز از پیکره‌های مقایسه‌ای ارائه می‌کنند. آن‌ها با استفاده از روشی مبتنی بر CLIR به جست‌وجوی جملات موازی اقدام می‌کنند. آن‌ها در طی چند مرحله و با روش‌های جست‌وجوی مختلف، فضای جست‌وجوی خود برای کشف جملات موازی را کوچک‌تر می‌کنند. سپس در فضای جست‌وجو با اندازه مناسب با استفاده از معیار «شباهت ترجمه» جملات موازی را انتخاب می‌کنند. معیار شباهت ترجمه، درواقع ترکیبی از تعداد کلمات مشترک، طول جملات، و داده‌های فرهنگ لغت است. میزان دقت در روش آن‌ها ۸۰ ٪ گزارش شده است.

محمدی و قاسم آقایی (۲۰۱۰) الگوریتمی را برای استخراج جملات هم‌تراز از دانشنامه آزاد ویکی‌پدیا ارائه می‌کنند. روش آن‌ها دقت ۲۱٪ را گزارش می‌کند که آن‌ها دلیل پایین بودن دقت را املاهای متفاوت زبان فارسی مانند نویسه‌های عربی «ک» و «ی» عنوان می‌کنند. روش محمدی و قاسم آقایی مشابه روش گیل و چرچ (۱۹۹۳) است. جکیان طوسی (۱۳۹۱) با استفاده از بردار ویژگی که شامل برچسب‌های دستوری و طول جملات و داده‌های فرهنگ لغت می‌شود به ارزیابی الگوریتم‌های طبقه‌بندی مختلف می‌پردازد. شاخص F روش جکیان طوسی (۱۳۹۱) به شکل میانگین ۹۰/۳۵٪ گزارش‌شده است.

انصاری و همکاران (۲۰۱۴) با ارائه یک سیستم جدید بازیابی اطلاعات مبتنی بر ترجمه ماشینی، به استخراج جملات موازی از ویکی‌پدیا پرداخته‌اند. در روش پیشنهادی آن‌ها ابتدا با استفاده از ابزار دیکودر موزز[6] جملات زبان

[1] S. Piperidis et al.
[2] H. M. Caseli and M. G.V. Nunes
[3] D. S. Munteanu & Marcu
[4] maximum entropy classifier
[5] D. Ștefănescu et al.
[6] Moses Decoder

مبدأ را به جملات زبان مقصد ترجمه می‌کنند، و سپس با سنجش میزان شباهت جملات ترجمه‌شده با جملات موجود در زبان مقصد و تعیین یک عنصر آستانه (به‌عنوان‌مثال درصورتی‌که طول جمله بیشتر از حد مورد نظر باشد) جملات موازی را استخراج می‌کنند. دقت سیستم آن‌ها در بهترین حالات ۹۴٪ (آستانه حساسیت ۰/۸ که تعداد زیادی از جملات هم‌ترجمه را نیز حذف می‌کند و تنها ۷۳ جمله در خروجی ارائه میکند) و در پایین‌ترین حالت ۴۹/۱۲ ٪ (آستانه حساسیت ۰/۱ با خروجی ۲۸۹ جمله) گزارش شده است. یکی از مشکلات روش ارئه شده توسط انصاری و سایرین (۲۰۱۴) پیچیدگی پیاده‌سازی و راحتی کاربرد است. زاری و صدرالدینی (۱۳۹۲) در کنار استفاده از روش طبقه‌بندی آنتروپی بیشینه به معرفی چهار دسته از ویژگی‌ها برای جملات هم‌تراز پرداخته و در هر مرتبه الگوریتم هم‌ترازی را با استفاده از یکی از این ویژگی‌ها پیاده‌سازی می‌کنند. در بهترین حالت دقت الگوریتم آن‌ها ۸۹/۲۲ ٪گزارش شده است.

همان‌طور که مشاهده می‌شود، برخی از روش‌های استخراج خودکار جملات هم‌تراز یا در مورد زبان فارسی پیاده‌سازی نشده‌اند، یا از دقت بالایی برای زبان فارسی برخوردار نیستند؛ مانند روش محمدی و قاسم آقایی (۲۰۱۰) برای استخراج جملات هم‌تراز از دانشنامه ویکی‌پدیا که دقت ۲۱٪ را گزارش می‌کند. در صورت گزارش دقت‌های بالا نیز این موارد صرفاً در محیط آزمایشگاهی پیاده‌سازی و ارزیابی‌شده‌اند، و تنها در شرایط خاص می‌توان به این دقت‌ها دست‌یافت و این گزارش‌ها مورد بازسازی علمی-پژوهشی و ارزیابی در شرایط جدید قرار نگرفته‌اند. برخی روش‌ها نیز دارای پیچیدگی‌های پردازشی بالا و مراحل رایانشی متعدد هستند. همچنین مواردی که از دقت قابل‌قبول برخوردار هستند و قابلیت پیاده‌سازی راحتی دارند به ارائه ابزار نهایی منجر نشده‌اند. در این مقاله به معرفی روشی با پیچیدگی زمانی خطی[1] و دقت قابل‌قبول می‌پردازیم که می‌توان از آن به‌راحتی در تولید ابزار کاربردی تولید پیکره‌های موازی استفاده نمود. همچنین با توجه به کارایی‌ها و قابلیت‌های ابزارها و روندهای پردازشی مورد استفاده در تولید ابزار نهایی، در طراحی ابزار ارائه‌شده در پژوهش حاضر، توجه کافی به ویژگی‌های خاص زبان فارسی مانند شناسایی نسبی با هم‌آیی‌ها، به کار بردن نویسه‌های عربی به‌جای نویسه‌های فارسی و سایر موارد شده است.

٤. مبانی نظری

در ابتدا به ارائه تعریف صوری عمل هم‌ترازی جملات پرداخته می‌شود. اگر جملات زبان مبدأ $f_1, \dots, f_{n_f}$ فرض شوند و ترجمه این جملات در زبان مقصد، $e_1, \dots, e_{n_e}$ فرض شود، لیست S مجموعه جملات موازی می‌باشد به‌صورت: $S_1, \dots, S_n$. هر جفت جمله S_i را می‌توان به‌صورت زیر تعریف کرد (کوئن[2]، ۲۰۰۹) :

$$s_i = (\{f_{start-f(i)}, \dots, f_{end-f(i)}\}, \{e_{start-e(i)}, \dots, e_{end-e(i)}\}) \quad (۱)$$

فرض بر این است که جملات به ترتیب ترجمه یکدیگر هستند. یا به بیان ریاضی:

$$Start - f(i) = end - f(i-1) + 1 \quad (۲)$$

[1] linear time complexity
[2] P.Koehn

$$Start - e(i) = end - e(i-1) + 1 \quad (۳)$$

همچنین داریم:

$$\begin{aligned} &start - f(1) = 1 \\ &start - e(1) = 1 \\ &end - f(n) = n_f \\ &end - e(n) = n_e \\ &start - f(i) \leq end - f(i) \\ &start - e(i) \leq end - e(i) \end{aligned} \quad (۴)$$

نوع هم‌ترازی، با توجه به اینکه چه تعداد جمله در هم‌ترازی وجود دارند و چند جمله از زبان مبدأ به چند جمله در زبان مقصدترجمه شده‌اند، می‌تواند به شکل زیر تعریف شود:

$$type = \left|\{f_{start-f(i)}, \dots, f_{end-f(i)}\}\right| - \left|\{e_{start-e(i)}, \dots, e_{end-f(i)}\}\right| = end - f(i) - start - f(i) + 1 - end - e(i) - start - e(i) + 1 \quad (۵)$$

طبق تعریف در رابطه ۵، نوع ترجمه ۱-۱ ترجمه‌ای است که در آن یک جمله از زبان مبدأ به یک جمله از زبان مقصد ترجمه می‌شود. در یک ترجمه ۱-۲ دو جمله از زبان مبدأ به یک جمله از زبان مقصد ترجمه می‌شود و الی‌آخر. در الگوریتم‌های هم‌ترازی، باید تمامی جملات پردازش شوند و هر جمله باید تنها یک‌بار در پیکره نهایی ذکر شود. به بیان کلی مجموعه جملات موازی $S = \{S_1 , \dots , S_n\}$ مورد جست‌وجو است که رابطه ۶ را ارضاء کند:

$$score(S) = \prod_{i}^{n} match(s_i) \quad (۶)$$

این تابع را می‌توان با به‌کارگیری روش‌های مختلفی پیاده‌سازی نمود. در مقاله حاضر تابع match بر اساس معیار شباهت جملات بر اساس اطلاعات نحوی آن‌ها شامل داده‌های تعداد اسامی، افعال، قیود و صفت‌ها در دو جمله انگلیسی و فارسی تعریف شده است که خروجی مدل هم‌ترازی می‌باشد.

اساس نظری روش ارائه‌شده توسط پاپیریدیس و همکاران (۲۰۰۰) برای استخراج جملات هم‌تراز بر این ایده نظری استوار است که هدف اصلی مترجم در هنگام ترجمه هر جمله‌ای حفظ و انتقال پیام است. بر همین اساس در نظر پاپیریدیس و همکاران (۲۰۰۰)، طبقه کلمات باز که با برچسب‌های دستوری فعل، اسم، صفت و قید مشخص می‌شوند، بیشترین وزن و اهمیت را در انتقال مفهوم دارند. در همین راستا، بار معنایی[1] یک جمله را می‌توان مجموع تعداد تمامی کلماتی در آن جمله دانست که برچسب‌های دستوری اسم، فعل، صفت و قید به آن‌ها داده می‌شود. بنابراین می‌توان این

[1] semantic load

فرض را مطرح کرد که رابطه «هم‌ترازی» بین دو جمله صحیح منطقی است (به این معنی که رابطه هم‌ترازی بین دو جمله وجود دارد یا یک جمله هم ترجمه جمله دیگر است)، اگر و فقط اگر بار معنایی جمله در زبان مقصد به شکل معنی‌داری نزدیک بار معنایی جمله‌ای از زبان مبدأ باشد. یا به بیانی دیگر، اگر و فقط اگر مجموع تعداد کلمات خانواده‌های اسم، فعل، قید و صفت در دو جمله نزدیک هم باشند (پاپیریدیس و همکاران، ۲۰۰۰).

یکی از روش‌های پیاده‌سازی این دیدگاه استفاده از رگرسیون خطی چند متغیره[1]به‌منظور ساخت مدل هم‌ترازی است. این مدل را می‌توان با استفاده از داده‌های یک پیکره موازی که از قبل موجود است به دست آورد. برای پیاده‌سازی این روش، Y مجموع برچسب‌های اسم، فعل، قید و صفت، در جمله زبان مقصد فرض می‌شود و X_i برابر تعداد برچسب‌های یک گروه دستوری خاص (به‌عنوان مثال X_1 برابر تعداد کلمات دارای برچسب دستوری فعل در جمله زبان مبدأ) در جمله زبان مبدأ فرض می‌شود. ارتباط خطی Y و X_i را می‌توان با استفاده از رابطه ۷ بیان نمود. که در این رابطه X_1 مجموع تعداد تمامی کلماتی است که در جمله زبان مبدأ برچسب فعل دریافت کرده‌اند، X_2 تعداد کلمات با برچسب اسم، X_3 برچسب صفت و X_4 تعداد کلماتی است که برچسب قید دریافت کرده‌اند. تخمین میزان وزن‌ها یا همان b_i ها و همچنین میزان خطا در هنگام آموزش مدل رخ می‌دهد (پاپیریدیس و همکاران، ۲۰۰۰).

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + \varepsilon \qquad (۷)$$

باید توجه داشت که هرچند روش پاپیریدیس بسیار کارآمد است، اما در این پیاده‌سازی عملاً داده‌های مربوط به تعداد برچسب‌های مجزای اقسام دستوری در جمله زبان مقصد باهم جمع ریاضی شده و در Y ذخیره شده است و بردار ویژگی کوچک‌تر شده است و عملاً این داده‌های ارزشمند آموزشی حذف می‌شوند. بنابراین، می‌بایست از سایر روش‌های آموزش ماشینی مانند ماشین بردار پشتیبان[2] استفاده کرد که قابلیت استفاده از این داده‌ها را فراهم می‌کنند، که در پژوهش حاضر نیز به همین شکل عمل شده است.

آنچه باید در مورد الگوریتم‌های هم‌ترازی در نظر داشت این است که الگوریتم نهایی باید:

- رامپذیر[3] باشد. راه‌حل ارائه‌شده برای کشف جملات موازی باید دارای پیچیدگی زمانی خطی باشد و پیچیدگی فضایی آن مدیریت پذیر باشد.
- برای داده‌های فرامتنی مانند جدول‌ها، شکل‌ها، زیرنویس‌ها، و سایر مواردی که در متن ترجمه نمی‌شوند تمهیدات پردازشی لازم انجام شده باشد.
- قابلیت گسترش، تغییر و اضافه کردن مدل‌ها و زبان‌های جدید را داشته باشد.
- نیاز به ورودی‌های با پیش‌پردازش‌های فراوان نداشته باشد و از برون‌داد متن شلوغ، دارای نویسه‌های غیرمتعارف، نامنظم، غیرقابل‌خواندن توسط انسان و نیازمند به پس پردازش‌های فراوان خودداری کند (پاپیریدیس و همکاران، ۲۰۰۰).

[1] multiple linear regression
[2] support vector machine
[3] tractable

روش پیاده‌سازی شده در این مقاله دارای پیچیدگی خطی می‌باشد، ابزار نهایی تا حد امکان موارد فرامتنی را شناسایی و حذف می‌کند، به‌راحتی می‌توان مدل را برای سایر جفت‌های زبانی آموزش داد و ایجاد نمود، همچنین خروجی ابزار ایجادشده صرفاً جملات به‌صورت متن هستند و متن خروجی به‌راحتی توسط انسان و ماشین قابل‌خواندن است، و نیاز به پردازش‌های مجزا ندارد.

۵. روش پیشنهادی

با توجه به آنچه گفته شد، روش ارائه‌شده در این مقاله به این صورت است که در ابتدا به‌منظور آموزش الگوریتم‌های طبقه‌بندی مانند ماشین بردار پشتیبان، یک پیکره موازی آموزشی که دارای استانداردهای مورد نظر است تهیه می‌شود. به بیان دیگر، ورودی اصلی الگوریتم آموزش، یک پیکره موازی طلایی در حوزه علوم انسانی دارای ۱۰۳۸ جفت جمله است، که در سطح جمله هم‌تراز شده‌اند، و تنها جملاتی که به صورت تحت اللفظی ترجمه شده‌اند در آن قرار گرفته‌اند. اطلاعات این اسناد و این پیکره آموزشی در جدول (۲) ارائه‌شده است. سپس بردارهای ویژگی که شامل تعداد برچسب‌های دستوری گروه‌های نحوی اسم، فعل، قید، و صفت برای هر جمله فارسی و انگلیسی و همچنین طول جملات هر دو جمله در دو زبان می‌باشد، برای هر جفت جمله این پیکره آموزشی ایجاد می‌شوند (مانند داده‌های شکل (۱)). برچسب‌های دستوری به شکل کاملاً خودکار و با استفاده از ابزارهای برچسب‌گذاری دانشگاه فردوسی مشهد (FEP)، ابزار برچسب گذاری دانشگاه استنفورد (SPT)، و ابزار HunPos به اجزای سخن تخصیص داده می‌شوند. در مرحله بعد، با استفاده از این بردارهای ویژگی ایجادشده به آموزش و ارزیابی میزان موفقیت الگوریتم‌های طبقه‌بندی و یادگیری ماشینی مختلف پرداخته می‌شود. درنهایت با مقایسه میزان موفقیت الگوریتم‌های طبقه‌بندی مختلف در امر شناسایی جملات هم‌تراز، بهترین گزینه برای ایجاد مدل هم‌ترازی نهایی انتخاب می‌شود. این مدل در ابزار استخراج جملات هم‌ترازقرارگرفته و با استفاده از یک الگوریتم جست‌وجوی میزان شباهت جملات بر اساس داده‌های یک فرهنگ لغت دو زبانه فارسی-انگلیسی، دقت این مدل افزایش پیدا می‌کند.

دلیل عدم گنجاندن داده‌های فرهنگ لغت در مدل هم‌ترازی و بردار ویژگی‌هادر مرحله آموزش (به جای استفاده در گام فیلترینگ نهایی) این بوده است که درواقع هدف بکارگیری الگوریتم جست‌وجوی شباهت بر اساس داده‌های فرهنگ لغت، بیشتر استفاده از آن به‌عنوان یک پالایه در خروجی ابزار بوده است تا بتواند خروجی‌هایی که به دلیل ریاضیاتی بودن مدل، هم‌تراز در نظر گرفته می‌شوند را فیلتر کند. به‌عنوان‌مثال، اگر تنها میزان شباهت تعداد برچسب‌های دستوری (یک عنصر ریاضی و غیرزبانی) دو جمله «This is a pencil» و «این یک کتاب است» را در نظر بگیریم، مدل آماری، این دو جمله را به‌اشتباه هم‌تراز درنظر گرفته و در خروجی ارائه می‌کند. لذا با استفاده از سنجش میزان شباهت لغوی، این دو جمله در خروجی حذف شده و به‌عنوان جملات هم‌ترازارائه نمی‌شوند. استفاده از فرهنگ لغت به‌عنوان یک پالایه، به صورتی که در این پژوهش پیاده‌سازی و استفاده شده است، توجه به اطلاعات زبانی را بیشتر می‌کند و ابزار نهایی تنها به داده‌های آماری متکی نخواهد بود. هرچند استفاده از فرهنگ لغت به این شکل مشکلاتی مانند کلمات خارج از دامنه فرهنگ لغت[1] (کلماتی که در فرهنگ لغت وجود ندارند) یا عدم توانایی فرهنگ لغت در

[1] out of vocabulary

ترجمه عبارات یا ضرب‌المثل‌ها را به همراه دارد، اما به هر صورت همان‌طور که اشاره خواهد شد، استفاده از فرهنگ لغت دقت برنامه را افزایش می‌دهد.

در مرحله نخست، یک پیکره طلایی به شکل دستی و با استفاده از جملات مقالات ترجمه‌شده در زمینه‌های مختلف مانند زبانشناسی، ادبیات، حقوق، اقتصاد، روانشناسی، علوم اجتماعی و غیره تشکیل شد. در مرحله تولید پیکره، از انتخاب جملاتی که بیشتر به‌صورت مفهومی ترجمه‌شده بودند، خودداری شد. علت حذف جملات با ترجمه مفهومی به این دلیل است که جملاتی که به این شکل ترجمه می‌شوند، به دلیل حذف و اضافه‌های بسیار، تنها موجب ایجاد نویز در پیکره نهایی می‌شوند. جدول ۱ نمونه‌ای از جملات این پیکره طلایی را نشان می‌دهد. در جدول ۲ می‌توان مشخصات کلی این پیکره مانند تعداد جملات و کلمات موجود در این پیکره و توزیع برچسب‌های دستوری مختلف در این پیکره را مشاهده نمود.

جدول ۱- جملات نمونه پیکره طلایی کوچک ایجادشده برای مدل‌سازی

جملات زبان انگلیسی	جملات زبان فارسی
A commonly cited management axiom is	یک از اصول متعارف مدیریت که به آن بسیار اشاره می‌شود این است.
What you measure is what you get	آنچه در پایان تحویل می‌گیرید همان چیزی است که اندازه‌گیری می‌کنید.
However, there is much evidence of their over-representation among disadvantaged groups of all kinds	اما، شواهد بسیاری گواه بر حضور بیشتر این افراد درگروهای آسیب‌پذیر جامعه است.
There have been few studies of children in care who are educationally successful	تحقیقات اندکی به بررسی کودکان تحت مراقبت که به موفقیت تحصیلی رسیده‌اند نیز وجود دارد.
These killings are often conducted in ways meant to ensure the secrecy of the killers' identities	این قتل‌ها اغلب طوری انجام می‌شوند که هویت قاتلان آن محرمانه بماند.
Obviously, the household cannot spend more than the extra dollar (without borrowing)	بدیهی است که آن خانوار نمی‌تواند بیشتر از یک دلار اضافی خرج کند (بدون استقراض).

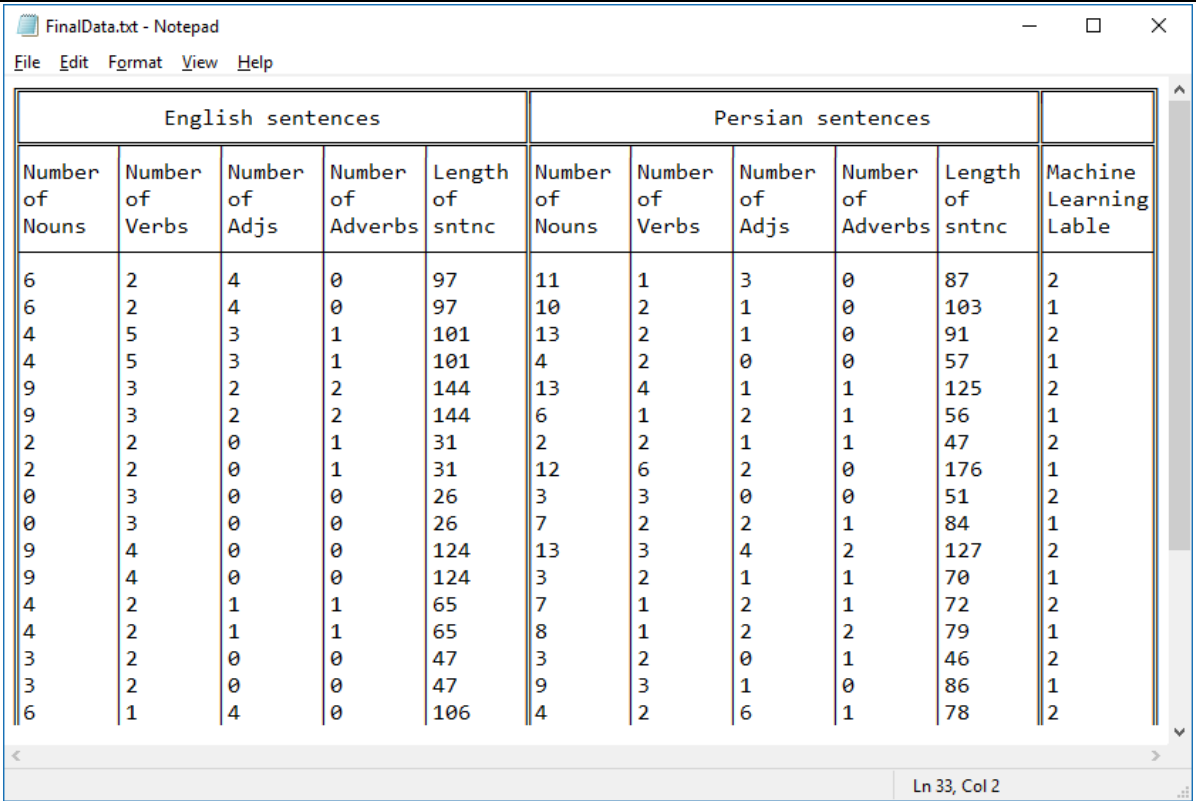
جدول ۲- آماره‌های پیکره طلایی ایجادشده به‌منظور آموزش الگوریتم‌های طبقه‌بندی

مشخصه	پیکره فارسی	پیکره انگلیسی
تعداد جملات	۱۰۳۸	۱۰۳۸
تعداد کلمات	۲۲۳۹۲	۱۹۲۰۷
میانگین طول جملات (به حرف)	۷۹/۸۶	۷۲/۹۴
تعداد برچسب‌های اسم (میانگین در جمله)	۹۰۷۴ (۷۴/۸)	۵۹۱۷ (۷/۵)
تعداد برچسب‌های فعل (میانگین در جمله)	۲۳۱۵ (۲۳/۲)	۳۰۴۶ (۹۳/۲)
تعداد برچسب‌های صفت (میانگین در جمله)	۲۱۲۱ (۰۴/۲)	۱۹۵۹ (۸۸/۱)
تعداد برچسب‌های قید (میانگین در جمله)	۴۵۸ (۴۴/۰)	۹۱۰ (۸۷/۰)

ازآنجایی‌که هدف این پژوهش ساخت بردار ویژگی بر اساس تعداد برچسب‌های دستوری جملات بود ابتدا باید از بین ابزارهای برچسب‌زنی موجود برای زبان‌های فارسی و انگلیسی، آنهایی که از دقت کافی برخوردار بودند انتخاب می‌شد که پس از بررسی‌های صورت گرفته ابزارهای HunPos با دقت ۹۶/۵۸ ٪ برای زبان انگلیسی (هالاکسی[1]، ۲۰۰۷) و با دقت ۹۶/۹ ٪ برای زبان فارسی (سراجی، ۲۰۱۱) و ابزار برچسب‌گذاری دستوری دانشگاه فردوسی با دقت ۹۷٪ برای زبان فارسی و ابزار دانشگاه استنفوردبا دقت ۹۷/۲۸ ٪ برای زبان انگلیسی (منینگ، ۲۰۱۴) مدنظر قرار گرفتند؛ البته در مورد این ابزارها نیز می‌بایست در صورت نیاز متن پیکره برای آنهاپ ش‌پردازش و آماده‌سازی می‌شد. ابزارهای دانشگاه فردوسی و استنفورد نیاز به پردازش‌های خاصی نداشتند، اما ابزار HunPos نیازمند ورودی‌های خاص بود. به‌عنوان‌مثال جملات باید با یک خط فاصله از هم جدا می‌شدند و تمامی کلمات هر جمله به‌صورت «هر خط یک کلمه» تبدیل می‌شدند، یا هر کلمه از هر جمله می‌بایست در خط جداگانه‌ای قرار می‌گرفت.

بعد از آماده‌سازی متن پیکره برای ابزارهای برچسب‌گذاری، می‌بایست داده‌های پیکره به این ابزارها داده می‌شد و اطلاعات آن‌هابه‌منظور کاربرد در یادگیری ماشینی استخراج می‌شدند. شکل (۱) نمونه‌ای از این داده‌ها را نشان می‌دهد که به‌صورت فایل متنی ذخیره شده است. در این شکل ۵ ستون اول مربوط به داده‌های جملات زبان انگلیسی و داده‌های ۵ ستون دوم مربوط به داده‌های به‌دست‌آمده از جملات زبان فارسی می‌باشد. این ستون‌ها به ترتیب برای هر دو زبان عبارت‌اند از تعداد برچسب‌های اسم، تعداد برچسب‌های فعل، تعداد برچسب‌های صفت، و تعداد برچسب‌های قید برای دو جمله. ستون پنجم در هر دو زبان طول جملات برحسب تعداد نویسه‌های جملات می‌باشد. ستون یازدهم ستون هدف است که در آن ۲ به معنی هم‌ترازی و ۱ به معنی عدم وجود هم‌ترازی است. تعداد داده‌های آموزش برای تمام الگوریتم‌ها ۸۰ ٪کل داده‌ها و ۱۶۶۰ عدد است. تعداد داده‌های آزمون ۴۱۶ عدد است.

[1] P. Halácsy



FinalData.txt - Notepad

File Edit Format View Help

English sentences					Persian sentences					
Number of Nouns	Number of Verbs	Number of Adjs	Number of Adverbs	Length of sntnc	Number of Nouns	Number of Verbs	Number of Adjs	Number of Adverbs	Length of sntnc	Machine Learning Lable
6	2	4	0	97	11	1	3	0	87	2
6	2	4	0	97	10	2	1	0	103	1
4	5	3	1	101	13	2	1	0	91	2
4	5	3	1	101	4	2	0	0	57	1
9	3	2	2	144	13	4	1	1	125	2
9	3	2	2	144	6	1	2	1	56	1
2	2	0	1	31	2	2	1	1	47	2
2	2	0	1	31	12	6	2	0	176	1
0	3	0	0	26	3	3	0	0	51	2
0	3	0	0	26	7	2	2	1	84	1
9	4	0	0	124	13	3	4	2	127	2
9	4	0	0	124	3	2	1	1	70	1
4	2	1	1	65	7	1	2	1	72	2
4	2	1	1	65	8	1	2	2	79	1
3	2	0	0	47	3	2	0	1	46	2
3	2	0	0	47	9	3	1	0	86	1
6	1	4	0	106	4	2	6	1	78	2

Ln 33, Col 2

شکل ۱- داده‌های نهایی به‌دست‌آمده از پردازش جملات پیکره

بعد از ایجاد داده‌های آموزش، باید به ارزیابی الگوریتم‌های یادگیری و طبقه‌بندی مختلف روی داده‌ها پرداخت. در ابتدا یک ارزیابی کلی از الگوریتم‌های طبقه‌بندی روی داده‌های به‌دست‌آمده از ابزارهای دانشگاه استنفورد و فردوسی انجام شد که نتایج آن در جدول (۳) ارائه‌شده است و همان‌طور که مشاهده می‌شود، الگوریتم ماشین بردار پشتیبان بالاترین میزان دقت را گزارش می‌کند. بعد از این ارزیابی اولیه، داده‌ها در ۲ مرحله کلی مورد ارزیابی قرار گرفتند. در فاز اول الگوریتم‌های طبقه‌بندی ماشین بردار پشتیبان و جنگل تصادفی[1] روی‌داده‌های به‌دست‌آمده از ابزار HunPos و همچنین ابزارهای دانشگاه فردوسی و استنفورد بدون در نظر داشتن معیار طول مورد ارزیابی قرار گرفتند. سپس داده‌های مربوط به معیار طول نیز به بردارهای ویژگی اضافه شدند و الگوریتم‌ها دوباره مورد ارزیابی قرار گرفتند. که نتایج این بررسی ها در شکل‌های (۲) و (۳) به شکل جزئی‌تر ارائه شده است. جدول (۳) میزان موفقیت الگوریتم‌های طبقه‌بندی مختلف روی داده‌هایی که با استفاده از ابزارهای دانشگاه فردوسی و استنفورد از روی پیکره به‌دست‌آمده است را نشان می‌دهد. اعداد این جدول با استفاده از الگوریتم‌های طبقه‌بندی که به شکل پیش‌فرض در بسته‌های نرم‌افزاری متلب[2] موجود هستند به‌دست‌آمده‌اند. هدف، انجام یک ارزیابی اولیه بوده است.

[1] Random Forest
[2] Matlab

جدول ۳- میزان دقت الگوریتم‌های یادگیری طبقه‌بندی روی داده‌های ابزارهای دانشگاه فردوسی و استنفورد

نام الگوریتم	دقت
Tree/Complex Tree	٪۹/۶۴
Tree/Medium Tree	٪۹/۶۰
Tree/Simple Tree	٪۱/۵۲
Linear Discriminant	٪۳/۴۷
Quadratic Discriminant	٪۹/۷۳
Logistic Regression	٪۲/۴۷
Linear SVM	٪۳/۴۸
Quadratic SVM	٪۷/۷۴
Cubic SVM	٪۸/۷۱
Fine Gaussian SVM	٪۷/۶۰
Medium Gaussian SVM	٪۴/۷۰
Coarse Gaussian SVM	٪۶/۵۷
Fine KNN	٪۰/۶۱
Medium KNN	٪۴/۶۷
Coarse KNN	٪۶/۶۶
Cosine KNN	٪۴/۶۹
Cubic KNN	٪۵/۶۶
Weighted KNN	٪۰/۶۶
Ensemble/Boosted Trees	٪۶/۶۶
Ensemble/Bagged Trees	٪۳/۶۶
Ensemble/Subspace Discriminant	٪۲/۴۷
Ensemble/Subspace KNN	٪۳/۵۴
Ensemble/RUSBoosted Trees	٪۸/۶۳

شکل (۲) آماره‌های آموزش الگوریتم‌های طبقه‌بندی روی داده‌های آموزش بدون معیار طول را گزارش می‌کند. در این شکل مشخصه‌های Precision, Accuracy, Sensitivity, Specificity, NECM, ECM, MR, Consistency, F-measure, Recall, و Practical Accuracy به ترتیب عبارت‌اند از میزان طبقه‌بندی اشتباه، میزان هزینه طبقه‌بندی اشتباه، نرخ نرمال شده هزینه طبقه‌بندی اشتباه، نرخ تمایز، میزان حساسیت، صحت، دقت، فراخوانی مجدد، شاخص F ، معیار توازن، و دقت عملیاتی. در این شکل عبارت «FPT+SPT» منظور داده‌های بدست آمده از ابزارهای برچسب‌گذاری دستوری دانشگاه فردوسی برای زبان فارسی و ابزار برچسب‌گذاری دستوری دانشگاه استنفورد برای زبان انگلیسی است. همچنین عبارت‌های «RF» و «SVM» به ترتیب به معنی «الگوریتم جنگل تصادفی» و «ماشین بردار پشتیبان» هستند. شکل (۳) آماره‌های آموزش الگوریتم‌های طبقه‌بندی روی داده‌های پیکره آموزشی با در نظر داشتن معیار طول است. توضیحات این شکل مانند توضیحات شکل (۲) است.

	HunPos		FPT+SPT	
	RF	SVM	RF	SVM
MR	0.314	0.126	0.326	0.259
ECM	0.47	0.14	0.485	0.388
NECM	0.235	0.07	0.242	0.194
Specificity	0.686	0.944	0.682	0.742
Sensitivity	0.686	0.831	0.667	0.741
Accuracy	0.686	0.874	0.674	0.741
Precision	0.686	0.961	0.669	0.743
Recall	0.686	0.831	0.667	0.741
F-measure	0.686	0.891	0.668	0.742
Consistency	0.373	0.552	0.345	0.48
Practical Accuracy	0.627	0.534	0.735	0.71

شکل ۲- مقایسه عملکرد الگوریتم‌های طبقه‌بندی روی داده‌های پیکره بدون در نظر داشتن معیار طول جملات

	HunPos		FPT+SPT	
	RF	SVM	RF	SVM
MR	0.243	0.119	0.31	0.23
ECM	0.372	0.119	0.467	0.352
NECM	0.186	0.059	0.234	0.176
Specificity	0.74	0.966	0.686	0.712
Sensitivity	0.776	0.831	0.695	0.873
Accuracy	0.757	0.881	0.69	0.77
Precision	0.723	0.977	0.654	0.633
Recall	0.776	0.831	0.695	0.873
F-measure	0.748	0.898	0.673	0.734
Consistency	0.58	0.54	0.434	0.8
Practical Accuracy	0.661	0.52	0.77	0.77

شکل ۳- مقایسه عملکرد الگوریتم‌های طبقه‌بندی روی داده‌های پیکره با در نظر داشتن معیار طول جملات

در جدول فوق مخفف MR (Misclassification Rate) عبارت است از نرخ کلاس‌بندی اشتباه، ECM (Expect Cost of Misclassification) یا هزینه کلاس‌بندی اشتباه، و NECM (Normalized Expect Cost of Misclassification) عبارت است از نرمال‌شده هزینه کلاس‌بندی اشتباه. این آماره‌ها توسط ماژول‌های ارزیابی دقت مدل در نرم افزار متلب ارزیابی و ارایه می‌شوند. این آماره‌ها در ارزیابی کیفیت مدل و کارایی آن از اهمیت بالایی برخوردار هستند.

به لحاظ پژوهشی می‌بایست نتایج روش جکیان طوسی (۱۳۹۱) با نتایج ارایه شده در شکل‌های (۲) و (۳) در مورد پژوهش حاضر نیز مقایسه شوند اما از این جهت که روش حل مسئله جکیان طوسی (۱۳۹۱) و همچنین مدل سازی جکیان طوسی (۱۳۹۱) با روش ارایه شده در این پژوهش تفاوت‌های پایه‌ای و اساسی دارد از ارایه آماره‌های روش جکیان طوسی (۱۳۹۱) در این شکل ها خودداری می‌شود. چرا که متاسفانه اساساً در هیچ یک از مستندات ارایه شده توسط جکیان طوسی (۱۳۹۱) این آماره‌ها ارایه نمی‌شوند و از آنجایی که دستیابی به داده‌های جکیان طوسی (۱۳۹۱) فراهم نبود امکان بازسازی و گزارش این آماره‌ها فراهم نشد. اما می‌توان با مقایسه دقت و شاخص F این دو پژوهش و همچنین توجه به مدل سازی و روش برخورد با مسئله، تفاوت میان این روش ها را مشاهده نمود. همچنین پژوهش حاضر بیشتر به جنبه کاربردی بودن و ارایه محصول کاربردی برای جامعه زبانشناسی رایانشی زبان فارسی توجه داشته است در حالی که روش جکیان طوسی (۱۳۹۱) بیشتر جنبه آزمایشگاهی و آزمون ایده بوده است. جدول (۶) می‌تواند این تفاوت ها را آشکار کند.

همان‌طور که نتایج نشان می‌دهد بهترین گزینه برای ایجاد مدل هم‌ترازی، الگوریتم ماشین بردار پشتیبان و با استفاده از داده‌های ابزارهای برچسب‌گذاری دانشگاه فردوسی و دانشگاه استنفورد و با در نظر داشتن معیار طول است. هرچند شاید داده‌های ابزار HunPos در مرحله آموزش میزان موفقیت ۹۷٪ را نشان بدهد اما همان‌طور که مشاهده می‌شود

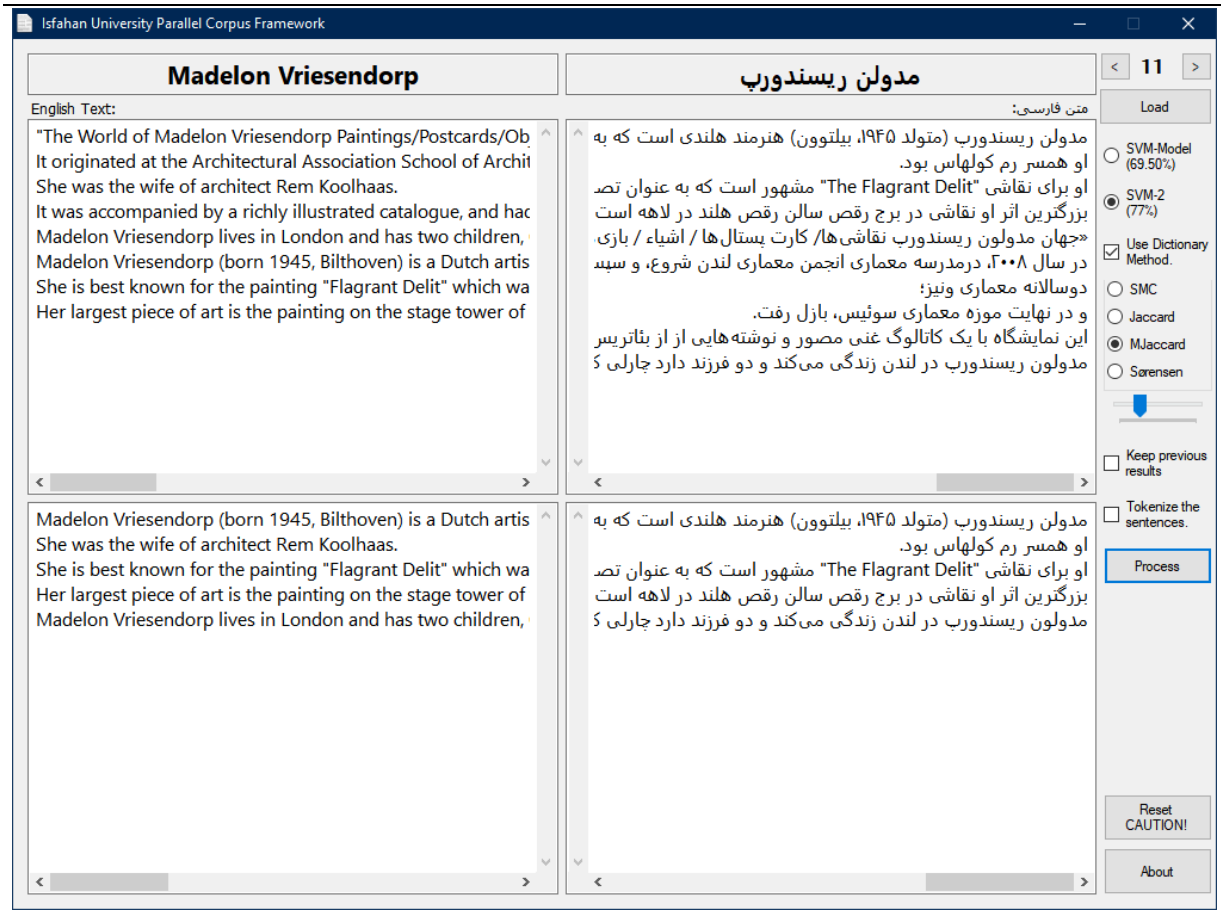
موفقیت عملیاتی مدل‌های هم‌ترازی ایجادشده با استفاده از این داده‌ها بسیار پایین است و این مدل‌ها قدرت تمیز بالایی روی داده‌های آزمون ندارند و نمی‌توان از آن‌ها به شکل کاربردی استفاده کرد.

٦. ساخت ابزار نهایی

در گام نهایی این پژوهش سعی شد تا با استفاده از مدل هم‌ترازی به‌دست‌آمده، یک ابزار کاربردی با در نظر داشتن معیارهای ایجاد پیکره موازی استاندارد و توجه به ویژگی‌های زبان فارسی به‌منظور استخراج جملات هم‌تراز و پیکره‌های موازی ایجاد شود. در این مرحله بعد از ساخت مدل‌های مختلف هم‌ترازی، و آزمودن آن‌ها به لحاظ عملیاتی بودن، بهترین مدل برای کاربرد در ابزار نهایی انتخاب شد. این ابزار تحت عنوان «چارچوب ایجاد پیکره موازی دانشگاه اصفهان» یا به‌اختصارIPCF[1] ارائه‌شده است. در ساخت این ابزار از زبان‌های برنامه‌نویسی متلب، VB.Net، و C# استفاده شده است. در حال حاضر مدل هم‌ترازی با دقت ۷۷٪ در این ابزار قرار دارد. ذکر این نکته ضروری است که این دقت، دقت عملیاتی مدل هم‌ترازی است. به عبارتی، همان‌طور که نتایج آزمون‌ها نیز نشان می‌دهد، دقت مدل هم‌ترازی در شرایط آزمایشگاهی و در مرحله آموزش به ۹۷٪ نیز می‌تواند برسد، اما ازآنجایی‌که هدف ایجاد مدلی با کاربرد عملیاتی بوده است، ارزیابی مدل روی داده‌های آزمون مختلف نیز انجام شده و دقت ٪۷۷ حاصل شده است. به‌بیان‌دیگر، منظور از دقت عملیاتی این است که می‌توان از این مدل هم‌ترازی انتظار داشت که بتواند با دقت ٪۷۷ در شرایط غیر آزمایشگاهی به استخراج جملات موازی از متون مقایسه‌ای بپردازد. همچنین امکان فیلتر کردن نتایج نهایی ابزار با استفاده از الگوریتم جست‌وجوی شباهت با استفاده از داده‌های فرهنگ لغت نیز در این برنامه وجود دارد که دقت ابزار و مدل هم‌ترازی را از ٪۷۷ به بیش از ٪۸۵ افزایش می‌دهد. شکل (۴) نمای کلی ابزار ایجادشده را نشان می‌دهد. همانطور که در شکل دیده می‌شود، ابزار توانسته جملات هم‌تراز را از یک متن مقایسه‌ای استخراج نماید.

مقالات مورد استفاده که در این ابزار و در شکل (۴) مشاهده میشوند. مقالات ترجمه شده دانشنامه ویکی‌پدیا هستند که در ابزار قرار گرفته اند و تنها جنبه مثال دارند. به این معنی که کاربر میتواند مقالات و جملات خود را به برنامه اضافه نماید. این مقالات به شکل کاملا تصادفی از دانشنامه ویکی‌پدیا انتخاب شده‌اند و تنها وجه مشترک بین مقالات این است که مقالات مربوط به حوزه علوم انسانی هستند. طول این مقالات برحسب تعداد جمله بین ۱ تا ۶۰ جمله متغیر است.

[1] Isfahan University Parallel Corpus Framework



شکل ٤- نمای کلی ابزار استخراج جملات موازی

۷. الگوریتم فرهنگ لغت

همان‌طور که قبلاً اشاره شد هدف استفاده از فرهنگ لغت، حذف موارد ناخواسته‌ای است که به دلیل ماهیت آماری مدل هم‌ترازی به شکل ناخواسته هم‌تراز تشخیص داده می‌شوند. نحوه کار الگوریتم جست‌وجوی شباهت بر این اساس است که ابتدا ۱۰ جمله از متن ورودی در هر دو زبان خوانده می‌شود (اگر جملات بیش از ۱۰ مورد باشد در پنجره‌های بعدی پردازش می‌شوند). علت انتخاب پنجره‌های با اندازه ۱۰ جمله این مسئله است که انتخاب پنجره‌های کوچک‌تر علاوه بر افزایش بار پردازشی، درصورتی‌که فاصله جملات هم ترجمه زیاد باشد، دقت ابزار را کاهش می‌دهد. با انتخاب پنجره‌های بزرگ‌تر از ۱۰ جمله دقت کار ابزار به دلیل احتمال انتخاب نامزدهای هم‌ترازی بیشتر کاهش پیدا می‌کند. بعد از خواندن ۱۰ جمله از ورودی، این جملات برچسب دهی دستوری می‌شوند، و پردازش‌های لازم روی آن‌ها صورت می‌گیرد. سپس به ازای هر جمله فارسی، تمامی ۱۰ جمله انگلیسی به مدل هم‌ترازی داده می‌شوند و تمامی جملاتی که توسط مدل هم‌تراز جمله فارسی تشخیص داده می‌شوند استخراج می‌شوند. بعدازاین مرحله، تمامی این جملات منتخب توسط داده‌های فرهنگ لغت مورد ارزیابی قرار می‌گیرند. جمله‌ای که بیشترین میزان شباهت به لحاظ کلمات مشترک در فرهنگ لغت را داشته باشد به‌عنوان جمله هم ترجمه و هم‌تراز انتخاب می‌شود. این میزان بیشینه باید از حساسیت تعین شده توسط کاربر بیشتر باشد. این الگوریتم استاندارد «تنها یک جمله» را مدنظر قرار می‌دهد. به‌بیان‌دیگر این الگوریتم به این مسئله توجه دارد که یک جمله تنها می‌تواند به یک جمله ترجمه شود و نمی‌تواند چندین ترجمه داشته باشد.

الگوریتم‌های سنجش میزان شباهت لغوی جملات که در این ابزار پیاده‌سازی شده‌اندالگوریتم‌های Simple Matching Algorithm, Jaccard, Modified Jaccard و sorensen هستند. الگوریتم Modified Jaccard از رابطه ۸ استفاده می‌کند. سایر الگوریتم‌ها از رابطه‌های مرسوم و معیار استفاده می‌کنند.

$$\text{شباهت} = \frac{\text{تعداد لغات مشترک}^2}{\text{تعداد لغات انگلیسی} \times \text{تعداد لغات فارسی}} \qquad (۸)$$

۸. ارزیابی تأثیر استفاده از فرهنگ لغت

به جهت سنجش میزان تأثیر استفاده از الگوریتم جست‌وجوی میزان شباهت لغوی جملات، می‌توان با آزمودن برنامه روی داده‌های پیکره‌های مطمئن و طلایی از صحت کار برنامه مطلع شد. به همین دلیل تعداد ۱۰۰ جفت جمله که از هم‌تراز بودن آن‌ها به شکل انسانی و دستی اطمینان حاصل شد از پیکره میزان استخراج شدند. سپس تعداد ۱۰۰ جمله که نمی‌توانستند ترجمه یکدیگر باشند اما در پیکره میزان وجود داشتند نیز از پیکره میزان استخراج شدند.این جملات هم‌تراز و غیر هم‌تراز به شکل تصادفی باهم ترکیب شدند تا یک پیکره آزمون نهایی حاصل شود. بعد از اجرای برنامه روی داده‌های این پیکره آزمون نتایج جدول ۴ حاصل شد که نشان‌دهنده میزان تأثیر استفاده از الگوریتم جست‌وجوی میزان شباهت لغوی جملات در بهبود دقت مدل هم‌ترازی است.

جدول ٤- میزان عملکرد مدل هم‌ترازی و تأثیر استفاده از فرهنگ لغت در بهبود دقت نهایی ابزار

	SVM2	SVM2+Dictionary
Precision	52.38%	85.18%
Recall	77%	69%
F-Measure	62.34%	76.24%

۹. ارزیابی و مقایسه

هرچند مقایسه روش‌های مختلف همترازی، از آنجایی که از نظر اهداف و روش‌های پیاده سازی با یکدیگر تفاوت های بنیادین دارند به شکل کامل امکان پذیر نباشد اما در اینجا به منظور مقایسه، به بررسی یکی از کارهایی که از جهت روش با مقاله حاضر قرابت بسیاری دارد پرداخته می‌شود.

روش جکیان طوسی (۱۳۹۱)، از آنجایی که تنها روشی است که از برچسب‌های دستوری در مورد جفت زبانی فارسی-انگلیسی به منظور شناسایی جملات همتراز استفاده کرده است، مطرح می‌شود. جکیان طوسی (۱۳۹۱) از الگوریتم‌های آموزش ماشینی بسیاری برای ساخت مدل استفاده می‌کند. به عنوان مثال جدول (۵) نشان دهند میزان موفقیت الگوریتم سیستم MDT در مورد پیکره FEP6000 است که نسخه ویرایش شده پیکره TEP می‌باشد.

جدول ۵- موفقیت یکی از پیاده‌سازی‌های جکیان طوسی (۱۳۹۱)

انواع بردار ویژگی	FEP6000				TEP10000			
	W-AVG	2-1	1-2	1-1	W-AVG	2-1	1-2	1-1
LPT	0.9035	0.70	0.65	0.96	0.7702	0.63	0.40	0.89
L	0.8999	0.64	0.66	0.96	0.7709	0.63	0.42	0.89
LP	0.8834	0.67	0.64	0.94	0.7686	0.63	0.41	0.89
CLP	0.9008	0.71	0.55	0.96	0.7622	0.60	0.40	0.88
CLPT	0.8874	0.59	0.54	0.95	0.7403	0.52	0.33	0.87
T	0.8538	0.34	0.24	0.94	0.7090	0.39	0.23	0.86

در جدول فوق روش ساخت مدل‌های آموزش ماشینی بر اساس بردارهای ویژگی را نشان می‌دهد. حرف L نماینده استفاده از معیار شباهت طول، حرف P نشانه مدل مبتنی بر معیار شباهت برچسب های دستوری و حرف T نشانه مدل مبتنی بر میزان شباهت دو جمله از لحاظ ترجمه کلمه به کلمه میباشد. حال مدل‌های ترکیبی LPT به عنوان نمونه مدلی است که از هرسه معیار L و P و T استفاده کرده است. حرف C نشانه ایجاد مدل‌های ترکیبی وزنی معیارهای اشاره شده است. از آنجایی که مدل مورد استفاده در روش جکیان طوسی (۱۳۹۱) با مدل ایجاد شده در این پژوهش تفاوت‌های بسیار اساسی و مهم دارد مقایسه و ارایه آماره‌های مربوط به نتایج به دست آمده در این پژوهش در این جدول معنی‌دار نخواهد بود چرا که روش‌های پیاده‌سازی و مدل‌سازی تفاوت جدی دارند.

هرچند شاید در نگاه اول روش جکیان طوسی (۱۳۹۱) روش موفق‌تری نسبت به روش پژوهش حاضر به نظر برسد (شاخص F در بهترین حالت در مقاله حاضر ۸۹/۸٪ بوده است)، اما باید در نظر داشت که در روش این مقاله،تمام سعی بر این بوده که از ابزار های کاملاً خودکار در تمامی مراحل استفاده شود. به بیان دیگر جکیان طوسی (۱۳۹۱) برای ایجاد داده‌های آموزش، برچسب گذاری دستوری و اصلاحات نگارشی از روش‌های دستی استفاده کرده است، اما در پژوهش حاضر این مراحل به شکل خودکار انجام می‌شوند که منجر به تولید ابزاری خودکار و یکپارچه شده است که نیازی به پیش پردازش و ویرایش‌های خاص ندارد. جدول (۶) به شکل کلی به مقایسه این روشها می‌پردازد.

جدول ٦- مقایسه روش جکیان طوسی (۱۳۹۱) با روش ارائه شده در پژوهش حاضر

ویژگی	روش طوسی (۱۳۹۱)	روش پژوهش حاضر
خودکار بودن	نیمه‌خودکار	خودکار
ارائه ابزار کاربردی	ندارد	دارد
راحتی کار با ابزار	-	دارد
شاخص Fگزارش‌شده	٪۹۰/۳۵	٪۷۶/۲۴
مشخصه دقت	آزمایشگاهی	عملیاتی-کاربردی
پیکره خروجی ابزار	ندارد	دارد
ویژگی ورودی ابزار	متن باید در سطح پاراگراف هم‌تراز شود	نیازی به ویرایش خاص ندارد. فیلترهای در نظر گرفته شده به شکل خودکار متن را ویرایش می‌کنند.
حوزه خاص	ندارد	دارد (در ویرایش فعلی ابزار، تنها علوم انسانی)
قابلیت تشخیص جملات ترجمه‌شده به شکل محاوره‌ای	ندارد	ندارد
استفاده از دیکشنری	الگوریتم Sorensen	Sorensen Modified Jaccard Jaccard SMP
قابلیت بسط برای سایر زبان‌ها	دارد	دارد
تشخیص انواع هم‌ترازی	۱-۱ ۲-۱ ۱-۲ ۲-۲	۱-۱ ۱-۲ ۲-۱ ۲-۲

۱۰. خلاصه و نتیجه‌گیری

در این مقاله به پیاده‌سازی و بررسی یک روش کاربردی به‌منظور استخراج پیکره‌های موازی از متون ترجمه‌شده و مقایسه‌ای پرداخته شد. اساس کار مدل هم‌ترازی در این روش شباهت دستوری و نحوی جملات در دو زبان است و درصورتی‌که بتوان ابزارهای برچسب‌گذاری برای زبان‌های دیگر ایجاد نمود می‌توان این ابزار و روش را به سایر جفت‌های زبانی به‌غیراز فارسی و انگلیسی نیز گسترش داد. برای پیاده‌سازی و ارزیابی این روش ابتدا یک پیکره طلایی کوچک حاوی ۱۰۳۸ جفت جمله هم‌تراز ایجاد شد. سپس با برچسب‌گذاری دستوری جملات این پیکره با استفاده از ابزارهای HunPos و ابزارهای برچسب‌گذاری دانشگاه فردوسی و دانشگاه استنفورد داده‌های آموزش ایجاد شد. که ۱۶۶۰ عدد از این داده‌هابه‌منظور آموزش الگوریتم‌های طبقه‌بندی و ۴۱۶ عدد به‌منظور آزمون موفقیت این الگوریتم‌ها اختصاص داده شدند. بعد از ارزیابی‌های مختلف بهترین مدل هم‌ترازی مدل ماشین بردار پشتیبان روی داده‌های ابزارهای برچسب‌گذاری دانشگاه فردوسی و دانشگاه استنفورد با موفقیت ۷۷٪ تشخیص داده شد. البته سایر مدل‌ها میزان موفقیت ۹۷٪ در حین آموزش را نیز گزارش می‌کردند که البته دقت عملیاتی این مدل‌ها پایین تشخیص داده شد. در پایان با استفاده از بهترین مدل و همچنین با طراحی یک الگوریتم سنجش میزان شباهت لغوی جملات، دقت ابزار هم‌ترازی به ۸۵٪ افزایش یافت.

یکی از چالش‌های پیش روی پژوهش حاضر وجود کلمات خارج از دامنه فرهنگ لغت است. به این معنی که کلماتی که در فرهنگ لغت وجود نداشته باشند اما در جملات ورودی کاربر موجود باشند میزان شباهت لغوی جملات را پایین آورده و در کار الگوریتم سنجش شباهت لغوی ایجاد اشکال می‌کنند. همچنین به شکل کلی برچسب‌گذاری دستوری باهم‌آیی‌ها و عبارت یکی از چالش‌های ابزارهای برچسب‌گذاری محسوب می‌شود. به‌عنوان مثال هیچ‌وقت نمی‌توان با ترجمه کلمه به کلمه عبارت انگلیسی «kick the bucket» به عبارتی دست‌یافت که بیشترین میزان اشتراک برچسب‌های دستوری با عبارت اصلی را داشته باشد. به همین جهت ایجاد ابزارهای برچسب‌گذاری که بتوانند باهم‌آیی‌ها را تشخیص دهند و همچنین ارائه الگوریتم‌هایی برای مرتفع کردن مشکلات کلمات خارج از دامنه فرهنگ لغت از کارهایی است که برای آینده پژوهشی پیشنهاد می‌شود.

کتابنامه

زاری، علیمه؛ صدرالدینی، محمد (۱۳۹۲). شناسایی جملات هم ترجمه با استفاده از طبقه بند آنتروپی بیشینه، دوازدهمین کنفرانس ملی سیستم‌های هوشمند، انجمن سیستم‌های هوشمند ایران. (صص ۷۱۶-۷۲۱)

جکیان طوسی، سید احمد (۱۳۹۱). ارائه رهیافتی جدید برای تولید پیکره موازی انگلیسی-فارسی، پایان‌نامه دوره کارشناسی ارشد، دانشکده مهندسی دانشگاه فردوسی

Ansari, E., Sadreddini, M. H., Tabebordbar, A., & Wallace, R. (2014). Extracting Persian-English parallel sentences from document level aligned comparable corpus using bi-directional translation. *Advances in Computer Science: an International Journal, 3*, 59-65.

Barzilay, R., & Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 25-32. Stroudsburg: Association for Computational Linguistics.

Brown, K. (2005). *Encyclopedia of Language and Linguistics, 14-Volume Set.* Elsevier Science.

Caseli, H. M. and Nunes, M. G. V. (2003). Evaluation of sentence alignment methods on portuguese-english parallel texts. *Scientia, 14*, 1-14.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 9-16. Stroudsburg: Association for Computational Linguistics.

Cheon, J., & Youngjoong, K. O. (2017). Automatically Extracting Parallel Sentences from Wikipedia Using Sequential Matching of Language Resources. *IEICE Transactions on Information and Systems* .100 (2),405-408.

Fattah M.A., Ren F., Kuroiwa S. (2006) Text-Based English-Arabic Sentence Alignment. In DS. Huang, K. Li & G.W. Irwin (eds.) *Computational Intelligence. ICIC 2006. Lecture Notes in Computer Science*, vol 4114. 748-753. Berlin, Heidelberg: Springer.

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

Georgiou, P., Sethy, P., Shin, J., & Narayanan, S. (2006). An English-Persian Automatic Speech Translator: Recent Developments in Domain Portability and User Modeling. *Proceedings of the International Conference on Intelligent Systems and Computing*, Cyprus: ISYC.

Halácsy, P., Kornai, A., & Oravecz, C. (2007). HunPos: an open source trigram tagger. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 209-212. Prague, Czech Republic: Association for Computational Linguistics (ACL).

Han, X., Li, H., & Zhao, T. (2009). Train the machine with what it can learn: corpus selection for SMT. In P. Fung, P. Zweigenbaum & R. Rapp (eds.) *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, 27-33. Singapore: Association for Computational Linguistics (ACL)

Jabbari, F. & Ziabary, M. (2012). Developing an open-domain English-Farsi translation system using AFEC: Amirkabir bilingual Farsi-English corpus. *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages proceedings*, 17-24. San Diego: AMTA

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition).* Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Koehn, P. (2009). *Statistical Machine Translation.* Cambridge University Press.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge MA: The MIT Press.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In K. Bontcheva Z. Jingbo (eds.) *Proceedings of 52nd Annual Meeting of the Association for computational linguistics: system demonstrations*, 55-60. Baltimore, Maryland: Association for Computational Linguistics (ACL).

McEnery, A., & Xiao, R. (2007). Parallel and comparable corpora: What are they up to? In G. James and G. Anderman (eds.) *Incorporating Corpora: The Linguist and the Translator*, 18-31. Clevedon, UK: Multilingual Matters.

Mitkov, R. (ed.) (2005). *The Oxford handbook of computational linguistics.* Oxford University Press: New York.

Mohammadi, M., & GhasemAghaee, N. (2010). Building bilingual parallel corpora based on wikipedia. *Second International Conference on Computer Engineering and Applications*

(ICCEA), 264-268. Bali Island, Indonesia: Institute of Electrical and Electronics Engineers (IEEE).

Munteanu, D. S., & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics, 31*, 477-504.

Pilevar M.T., Faili H., Pilevar A.H. (2011) TEP: Tehran English-Persian Parallel Corpus. In A. Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science*, vol 6609. pp. 68-79. Berlin, Heidelberg: Springer.

Piperidis, S., Papageorgiou, H., & Boutsis, S. (2000). From sentences to words and clauses. In J. Véronis (ed.) *Parallel text processing: Alignment and Use of Translation Corpora*. 117-138. Dordrecht: Springer.

Rahimi, Z., Taghipour, K., Khadivi, S., & Afhami, N. (2012). Document and sentence alignment in comparable corpora using bipartite graph matching. *2012 Sixth International Symposium on Telecommunications (IST),* 817-821. Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE).

Rauf, S. A., & Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine translation. 25*, 341-375.

Seraji, M. (2011). A statistical part-of-speech tagger for Persian. In B. S. Pedersen, G. Nešpore and I. Skadiņa (eds.). *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA,* 340-343. *Riga, Latvia*: Northern European Association for Language Technology (NEALT).

Simard, M., Foster, G. F., & Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, 1071-1082. Toronto: IBM Press

Ştefănescu, D., Ion, R., & Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. *Proceedings of the 16th Conference of the European Association for Machine Translation*, 137-144.Trento: Fondazione Bruno Kessler (FBK).

Taghipour, K., Afhami, N., Khadivi, S., & Shiry, S. (2010). A discriminative approach to filter out noisy sentence pairs from bilingual corpora. *2010 5th International Symposium on Telecommunications(IST),* 537-541. Institute of Electrical and Electronics Engineers (IEEE): Curran Associates, Inc.