نشریه پژوهش‌های زبان‌شناسی تطبیقی
سال چهارم- شماره ۸- پاییز و زمستان ۱۳۹۳

پیکرهٔ متنی تطبیقی فارسی-انگلیسی حوزهٔ تخصصی فاوا

شکوفه دشتبانی[۱]
محرم منصوری‌زاده[۲]
محمد نصیری[۳]

چکیده

در زبانشناسی، پیکره انباره‌ای از داده‌های متنی است. در این مقاله، تمرکز ما بر طراحی و ساخت خودکار پیکرهٔ دو زبانه فارسی-انگلیسی است. ما نرم افزاری برای ساخت پیکره طراحی کرده‌ایم که هزینه و زمان ساخت پیکره را کاهش می‌دهد؛ به‌علاوه نرم‌افزار ارائه شده قابلیت مدیریت پیکره را نیز برای کاربران فراهم می‌کند. در این مقاله، روشی برای ترازبندی جمله های پیکره فارسی تخصصی حوزهٔ فاوا و جملات انگلیسی پیکرهٔ تخصصی حوزهٔ فاوا ارائه شده است. هدف ما طراحی یک سیستم ترازبندی برای استخراج جمله‌های متناظر دو زبان است. در این روش، ما با استفاده از یک لغت‌نامهٔ دو زبانی که خود مؤلفان ایجاد کرده‌ایم و با استفاده از تکنیک پیشنهاد شده، امتیاز شباهت دو جمله را محاسبه می‌کنیم. آزمایشات نشان می‌دهد که این تکنیک علاوه بر اینکه از نظر دقت بسیار قوی است، تعداد جمله‌های کاندید را نیز کاهش می‌دهد.

کلید واژه‌ها: زبانشناسی رایانشی ، مدیریت پیکره ، ترازبندی جمله، پیکرهٔ تطبیقی، بلندترین زیر دنبالهٔ مشترک (LCS)

۱- دانشجوی کارشناسی ارشد دانشگاه بوعلی‌سینا ✉| Shokoofeh_dashtbani@yahoo.com
۲- استادیار دانشگاه بوعلی‌سینا (نویسنده مسؤول) ✉| mansoorm@basu.ac.ir
۳- استادیار دانشگاه بوعلی‌سینا ✉| m.nassiri@basu.ac.ir

تاریخ دریافت: ۹۳/۰۳/۰۵ تاریخ پذیرش: ۹۳/۰۷/۱۳

۱. مقدمه

در تحقیقات و مطالعات زبان‌شناسی که بر روی یک حوزهٔ خاص صورت می‌گیرد، لازم است داده‌های حوزه-ای که مورد مطالعه قرار گرفته است، از نمونه‌های طبیعی باشند. پس از گردآوری این مجموعه می‌توان آن را در حوزهٔ مورد مطالعه برای تحلیل و توصیف زبان استفاده کرد. به مجموعه‌ای از داده‌های متنی، پیکره(corpus) متنی می‌گویند. در علم زبان‌شناسی، شاخه‌ای به نام زبان‌شناسی پیکره‌ای (عاصی، ۱۳۹۱) وجود دارد که هدف آن مطالعات علمی بر روی زبان طبیعی است. پیکره‌ها سفارش‌های مختلفی را می توانند بپذیرند، بدین معنا که می‌توانند حاوی داده‌های گفتاری، مقولهٔ دستوری خاص مانند فعل، محاورات عامیانه و غیره باشند. در صورتی که مجموعه‌ای از داده‌های قابل اطمینان برای محققان فراهم باشد، نتایج دقیق‌تری حاصل خواهد شد. به‌عبارت دیگر، پیکره مجموعه‌ای از داده‌های متنی سازمان یافته است. پیکره ممکن است حاوی متن‌ها، نقل قول‌ها، فهرست‌ها و یا حتی لغات باشد. پیکره‌ها برای اهداف مختلفی ایجاد می‌شوند، تمامی پروژه‌های پردازش زبان‌های طبیعی، پروژه‌هایی که مربوط به حوزهٔ زبان‌شناسی هستند و غیره، از پیکره استفاده می‌کنند. علاوه بر این، برای ایجاد پایگاه داده‌های زبانی، از پیکره‌های آن زبان استفاده می-گردد. پیکره‌های متنی برحسب زبان اسناد تشکیل دهندهٔ پیکره به دو دسته پیکره‌های تک زبانی و پیکره‌های چند زبانی تقسیم می‌شوند. پیکرهٔ دو زبانی رایج‌ترین نوع پیکره است. در اکثر کشورها، به دلیل بین‌المللی بودن زبان انگلیسی، با استفاده از اسناد زبان رسمی آن کشور و اسناد زبان انگلیسی، پیکره‌های دو زبانی ایجاد می‌گردد. امروزه، پیکره‌های دو زبانی در زمینه‌هایی همچون ترجمهٔ ماشینی، توسعهٔ ابزارهای پردازش زبان‌های طبیعی، داده‌کاوی، وردنت، موتور جستجوی دو زبانه، توسعهٔ لغت‌نامه‌ها، مطالعات و تحقیقات میان زبانی، بازیابی و استخراج اطلاعات، تشخیص صوت و غیره استفاده می‌گردند. از این‌رو، ساخت پیکره‌های دو زبانی و ترازبندی اسناد بین دو زبان مورد توجه محققان قرار گرفته است. اگر پیکره دو زبانی یا چند زبانی باشد، به الگوریتم‌هایی برای ترازبندی اسناد نیاز است. منظور از ترازبندی در اینجا پیدا کردن داده-هایی است که موضوع یکسان ولی زبان‌های متفاوتی دارند. یک پیکره ترازبندی شده از این رو برای مترجم‌های انسانی مفید است که مترجم‌ها می‌توانند به ازای هر جملهٔ زبان منبع، تمامی جمله‌های معادل آن در زبان مقصد را در پیکره جستجو کنند. بدین ترتیب، مترجمان می‌توانند با مشاهدهٔ نمونه‌های زبان طبیعی، کلمه‌ها و عبارات را به درستی به زبان‌های دیگر ترجمه کنند. پیکره‌های چند زبانی از نظر شیوهٔ ترازبندی به دو دسته پیکره‌های موازی و پیکره‌های تطبیقی تقسیم می‌گردند. امروزه، با توجه به رشد تکنولوژی و افزایش پروژه‌های تخصصی حوزه‌ی فاوا (Information Communication Technology) در ایران، نیاز به پیکرهٔ متنی به زبان فارسی برای حوزهٔ فاوا بیش از پیش احساس می‌گردد. به همین دلیل، ما برآن شدیم تا پیکرهٔ دو زبانهٔ فارسی– انگلیسی را ایجاد نماییم. پیکرهٔ دو زبانه فارسی– انگلیسی از دو مجموعهٔ کاملاً مجزا به زبان فارسی و انگلیسی تشکیل شده است که این دو مجموعه با کمک یک فرهنگ لغت با یکدیگر ترازبندی می‌شوند. به‌دلیل غنای منابع موجود در زبان انگلیسی، استفاده از اصطلاحات تخصصی حوزهٔ فاوا به زبان انگلیسی در کشور و نبود معادل فارسی مناسب برای اکثر واژه‌های تخصصی، مجموعه اسناد فارسی

پیکره از روی مجموعه اسناد انگلیسی گسترش داده می‌شود. بنابراین، برای هر سند انگلیسی تخصصی حوزهٔ فاوا یک سند فارسی معادل به‌وجود می‌آید که می‌تواند نیاز محققان این حوزه را تأمین کند. در این مقاله ما به‌دنبال تکنیکی برای ترازبندی پیکرهٔ تطبیقی انگلیسی–فارسی فاوا هستیم. در بخش بعدی، علاوه بر مرور چند نمونه از پیکره‌های موجود، ترازبندی‌های انجام شده بر روی پیکره‌های فارسی–انگلیسی موجود را نیز به تفصیل بیان کرده‌ایم. در بخش سوم، به معرفی و نحوهٔ ساخت پیکرهٔ حوزهٔ فاوا، منابع ساخت پیکره، ابزارهای مورد استفاده و تکنیک ترازبندی پیکره انگلیسی–فارسی تخصصی فاوا پرداخته‌ایم. پس از بیان مشخصات پیکرهٔ حوزهٔ فاوا در بخش چهارم، نتایج تکنیک ارائه شده، مقایسه، تحلیل و ارزیابی می‌گردد.

مروری بر کارهای انجام شده

در این بخش، چند نمونه از پیکره‌های موجود به‌ویژه پیکره‌های دو زبانی انگلیسی– فارسی را بررسی کرده و سپس با توجه به تمرکز مقاله بر ترازبندی اسناد پیکره، ترازبندی پیکرهٔ تطبیقی فارسی–انگلیسی را با تفصیل بیشتری مطالعه می‌کنیم.

پیکره‌های زبان انگلیسی و فارسی

تاکنون پیکره‌های زیادی به زبان انگلیسی ایجاد شده‌اند. از جمله پیکره‌های انگلیسی معروف می‌توان پیکرهٔ Penn (M.A.Marcinkiewicz,B.Santorini, M.P.Marcus, ,1993,313-330) را نام برد. این پیکره یک پیکرهٔ حاشیه‌نویسی (Annotation) شده مشهور است و هنوز هم دارای خطاهای نشانه‌گذاری است که رفع نشده است. این پیکرهٔ حاوی ۴.۵ میلیون کلمه به زبان انگیلیسی آمریکایی است. پیکرهٔ آکسفورد (OEC) یک پیکرهٔ بسیار حجیم به زبان انگلیسی است (Bijankhan corpus,2011). این پیکره برای ساخت فرهنگ لغت آکسفورد استفاده شده است. پیکرهٔ Brown (H. K. W. N. Francis, 2011) نیز به زبان انگلیسی است که در دانشگاه Brown تهیه شده است و حاوی یک میلیون لغت است. از جمله پیکره‌های معروف دیگر در زبان انگلیسی می‌توان پیکرهٔ BNC (British National corpus) را نام برد که به زبان انگلیسی بریتانیایی است و حاوی ۴۰۰۰ دادهٔ متنی و دادهٔ صوتی است. لغات پیکره بیشتر از ۱۰۰ میلیون کلمه است. داده‌های آن مربوط به قرن ۲۰ به بعد است. آخرین نسخه این پیکره در سال ۲۰۰۷ منتشر شده است. ۹۰ درصد از داده‌های این پیکره، داده‌های نوشتاری و ۱۰ درصد آن هم داده‌های صوتی هستند. BNC یکی از پیکره‌های مهم زبان انگلیسی است (G.Leech, R.Garside, M.Bryant,1994, 622-628) پیکره‌های دیگری از قبیلCOCA (M.Davies,2009,159-190), ANC (N.Idel, C.Macload, 2001,831-836) و غیره نیز ایجاد شده‌اند.

برای زبان فارسی چند پیکرهٔ مشهور تک زبانی ارائه شده است. پیکره دکتر محمود بی جن خان (Bijankhan corpus,2011) یک پیکرهٔ دستوری است که برچسب‌گذاری هم شده است. این پیکره بر اساس استاندارد ایگلز (بی‌جن‌خان، ۱۳۸۶) برچسب گذاری شده است. این پیکره شامل ۲۵۹۸۲۱۵ واژه و ۵۵۰ برچسب می‌باشد که به طور دستی برچسب زده شده است. هرچند که پیکرهٔ دکتر بی‌جن‌خان از نظر حجم و

گستردگی بسیار غنی است، اما به دلیل استاندارد نبودن متن اسناد، در بسیاری از موارد در محیط دیجیتال برای محققان قابل استفاده نیست. پیکرهٔ همشهری(Hamshahri Collection,2011) یکی از معتبرترین منابع در زبان فارسی است. این پیکره تک زبانی است و به صورت نیمه خودکار ساخته شده است. منبع ساخت آن اخبار خبرگزاری همشهری بوده است. این پیکره حاوی حدود سیصد هزار سند است. ایراد وارد به این پیکره نیز همانند پیکرهٔ بی‌جن‌خان، استاندارد نبودن متن اسناد پیکره است. دادگان زبان فارسی دکتر مصطفی عاصی (عاصی،۱۳۸۸)، (پایگاه اطلاع رسانی حوزه، ۲۰۱۱) یک پیکرهٔ محاوره‌ای است. آثار گردآوری شده در این فهرست ۱۵۰۰ مورد بوده است و روش ساخت آن به صورت دستی انجام شده است. پیکره‌های تک زبانی دیگری هم در زبان فارسی ایجاد شده است و ما در اینجا فقط چند نمونه از مشهورترین آنها را معرفی نمودیم.

پیکره‌های دو زبانی فارسی–انگلیسی که تاکنون ایجاد گردیده است، غالباً پیکره‌های موازی هستند و از روش‌های ترازبندی جمله که غالبا مبتنی بر طول جمله‌ها است، استفاده کرده‌اند. به عنوان مثال، پیکره تطبیقی فارسی–انگلیسی UTPECC (A.Shakery, H.Faili, H.Baradarn, 2010,29-37) از دو معیار تاریخ انتشار مقاله و نمرهٔ شباهت برای ترازبندی اسناد استفاده کرده است. ترازبندی این پیکره در سطح سند بوده است و این پیکره در سطح جمله و کلمه ترازبندی نشده است. بنابراین، تاریخ انتشار می‌تواند معیار مناسبی برای ترازبندی در سطح سند باشد. اما این معیار برای ترازبندی در سطح جمله و کلمه معیار مناسبی نیست. نمرهٔ شباهت نیز بر اساس کلمات کلیدی سند محاسبه می‌گردد که تمام جمله‌های یک متن ممکن است کلمات کلیدی مشابهی داشته باشند. از این‌رو، کلمات کلیدی یک سند نیز به تنهایی معیار مناسبی برای ترازبندی در سطح جمله نیست.

پیکره‌هایی همچون پیکرهٔ ویکی‌پدیا (N.QasemAghaee, M.Mohammadi, 2010,19-21)، پیکره TEP (H. Faili, A. H.Pilevar,M.T.Pilevar, 2011,68-75)، پیکره اطلاعات پزشکی کالیفرنیای جنوبی(W.May, S.Narayanan, P.Georgiou, SH.Ganjavi, R.S.Belvin,2004)، پیکره شیراز (H.MansouriRad, K.Megerdoomian, R.ZajacJan, W. Amtrup, 2000)، پیکره PEN ((PEN: Parallel English Persian News Corpus,2011)) و پیکرهٔ ELRA/ELDA (European Language Resource Association)(ELRA,2012) پیکره‌های موازی فارسی–انگلیسی هستند. در پیکره‌های موازی، از آنجا که به ازای هر سند فارسی دقیقاً یک سند انگلیسی معادل آن همزمان به پیکره افزوده شده است، پیدا کردن جمله‌های هم‌ارز از جفت اسناد فارسی–انگلیسی ساده‌تر و دارای دقت بالاتری است. علاوه بر این، حجم این پیکره‌ها به دلیل کم بودن منابع زبان فارسی چندان قابل توجه نیست.

پیکرهٔ تطبیقی فارسی–انگلیسی دیگری که در ایران ساخته شده است، پیکرهٔ تطبیقی فارسی–انگلیسی کریمی(s. Karimi,2008) نام دارد که این پیکره به صورت دستی و از ۱۱۰۰ سند خبری BBC ایجاد شده است. اسناد انگلیسی این پیکره به فارسی ترجمه شده‌اند و در پیکره قرار گرفته‌اند.

تاکنون، چندین الگوریتم برای ترازبندی اسناد پیکره ارائه شده است. به عنوان مثال، الگوریتم ترازبندی به روش Brown(H. Kucera W. N. Francis, 1979) وابسته به تعداد واژگان هر سند است. ایدهٔ Brown این بود که دو جمله که دربارهٔ یک موضوع یکسان هستند، اما زبانشان با هم متفاوت است، به احتمال زیاد دارای طول یکسان هم هستند. این روش برای پیکره‌های بزرگ چندان مناسب نیست. زیرا مرتبهٔ زمانی اجرای این الگوریتم توان دوم طول جملات است که برای پیکره‌های بزرگ زمان زیادی است. الگوریتم دیگری که برای ترازبندی پیکره‌ها بسیار رایج است، الگوریتم Gale-church (Kenneth W. Church ,William A. Gale, 1993) نام دارد. در این روش، به جای شمردن تعداد کلمه‌های یک جمله، از شمردن تعداد کاراکترها استفاده می‌شود. دو جمله به زبان‌های مختلف که در مورد یک موضوع یکسان هستند، به احتمال زیاد دارای طول کاراکتری یکسان هم خواهند بود. این روش و روش قبلی هر دو به محتوای کلمه در یک جمله توجهی ندارند. از این‌رو، این دو الگوریتم چندان نتایج دقیقی ندارند. ترازبندی با استفاده از شکل دستوری یک کلمه(Stanley F. Chen, 1993) نیز یکی دیگر از تکنیک‌های ترازبندی پیکره است. در این روش، از ترجمهٔ کلمه به کلمه استفاده می‌شود. برای اینکه معادل یک جمله را در سند زبان دیگر پیدا کنیم، جمله، کلمه به کلمه ترجمه می‌شود. دقت این روش بسیار وابسته به دقت ترجمه است. هر اندازه که ترجمهٔ جمله دقیق‌تر باشد، نتیجهٔ ترازبندی نیز دقیق‌تر خواهد بود. از این‌رو، استفاده از یک ابزار دقیق مشخص کنندهٔ نقش دستوری یا استفاده از یک سیستم ترجمهٔ ماشینی دقیق ضروری است. روش‌های دیگری هم برای ترازبندی متون پیکرهٔ دو زبانه وجود دارد که الگوریتم‌های ذکر شده، چند نمونه رایج از آنها بودند.

ترازبندی پیکرهٔ تطبیقی فارسی-انگلیسی

ترازبندی یک گام ضروری در ساخت پیکره است. به‌طور کلی، سه روش برای ترازبندی اسناد وجود دارد، ترازبندی در سطح واژگان یا عبارت، ترازبندی در سطح جمله و ترازبندی در سطح سند. ترازبندی می‌تواند در سطح پاراگراف نیز انجام شود ولی این روش چندان مرسوم نیست.

ترازبندی جمله عبارت است از ارزیابی جمله یا جمله‌های زبان منبع با جمله یا جمله‌های زبان مقصد. هدف یافتن شباهت بین جمله‌های زبان منبع و جمله‌های زبان مقصد است. همان‌طور که در شکل (۱) مشاهده می‌شود در پیکرهٔ تطبیقی به ازای هر جمله در یک سند، ممکن است چندین جمله معادل در چندین سند وجود داشته باشد.

شکل ۱: ترازبندی اسناد پیکره تطبیقی

وجود ویژگی‌های بی‌قاعده در متون فارسی، ترازبندی اسناد زبان فارسی را دشوار می‌کند. بنابراین، ما به دنبال روشی برای ترازبندی بودیم تا در حد امکان این ویژگی‌ها را در حین ترازبندی در نظر نگیرد. یکی از ویژگی‌های زبان فارسی که فرایند ساخت پیکره را سخت‌تر می‌کند، این است که فارسی یک زبان با قابلیت صرف بالاست و برخی از کلمه‌ها چند شکل نوشتاری متفاوت دارند. نوشتن حروف زبان فارسی برخلاف زبان انگلیسی از راست به چپ است که بعضی از حروف همانند زبان عربی به هم می‌چسبند. با اتصال پسوندها و پیشوندها به کلمه‌ها، شکل‌های متفاوتی از واژگان ایجاد می‌گردد. علاوه‌بر این، زبان فارسی دارای منابع غنی برای ساخت پیکره نیست. پیکره‌های فارسی موجود یا به صورت دستی و یا به صورت نیمه خودکار گردآوری و پردازش شده‌اند. پیکره‌هایی که به صورت نیمه خودکار ایجاد شده‌اند نیز از نظر اندازه کوچک هستند. در پیکره‌های تطبیقی، به ازای هر سند ممکن است که هیچ سندی در زبان دوم وجود نداشته باشد و یا اینکه ممکن است اسناد زیادی که از نظر محتوا شبیه به هم هستند، نیز وجود داشته باشند. به عبارت دیگر، در پیکره‌های تطبیقی چند زبانی اسناد نامتقارن زیادی وجود دارد؛ بدین معنا که بخش اعظم داده‌ها دارای روابط ۱به ۰ ، ۰ به ۱، ۱ به چند و چند به ۱ است. به عنوان مثال برای ترازبندی در سطح جمله، اگر یک جمله از زبان منبع هیچ معادلی در زبان مقصد نداشته باشد، یک رابطه ۱ به ۰ بین اسناد دو زبان برای این جمله برقرار است. همین امر سبب شده که ترازبندی پیکره‌های تطبیقی دشوارتر از پیکره‌های موازی باشد. هدف ما یافتن تکنیکی برای ترازبندی اسناد نامتقارن و متقارن زبان انگلیسی و فارسی است.

۲. ساخت پیکره

هر چند جمع‌آوری داده‌ها و ساخت پیکره‌های بزرگ فرایند ساده‌ای نیست اما باید دقت شود تا پیکره‌ای که ایجاد می‌شود، کاملاً سازگار باشد. بدین معنا که اسناد موجود حاوی اطلاعات متناقض با یکدیگر نباشند، داده‌های موجود در پیکره باید دقیق، غیر تکراری و حاوی اطلاعات مفید باشند. اگر اضافه کردن یک سند به پیکره، دانش جدیدی را به مجموعه اضافه نمی‌کند و یا حاوی اطلاعاتی است که قبلاً در پیکره موجود بوده است، این سند نباید به پیکره اضافه گردد. پیکرهٔ فاوا حاوی اسناد حوزه‌ی فناوری ارتباطات و اطلاعات است. این پیکره را برای ساخت نرم‌افزار وردنت فارسی مربوط به حوزهٔ فاوا استفاده می‌کنیم. در این پیکره، امکاناتی از قبیل سطوح مختلفی از تفسیر متن مثل برچسب‌گذاری، ریخت شناختی، تجزیه (parse) کردن

وابستگی‌های نحوی یا گرامری را انجام داده‌ایم. پیکرۀ فاوا را می‌توان به‌عنوان منبعی برای استخراج خودکار روابط معنایی، ترجمۀ ماشینی، نگاشت خودکار وردنت تخصصی حوزۀ فاوا و سایر پروژه‌های این حوزه استفاده نمود.

منابع ساخت پیکره

اولین مرحله ساخت پیکره جمع‌آوری داده است. برای فراهم کردن متون پیکره در حوزۀ فاوا از منابع مختلفی استفاده کرده‌ایم که این منابع عبارتند از:

- کتاب های تخصصی و نیمه‌تخصصی حوزه فاوا ، برق، کامپیوتر و یا فناوری اطلاعات به زبان فارسی و انگلیسی.
- وب سایت ها مثل سایت انجمن کامپیوتر ایران، انجمن رمز ایران، انجمن انفورماتیک ایران، انجمن مهندسی برق و الکترونیک، باشگاه مهندسان ایران
- مجموعه واژگان مصوب فرهنگستان ادب و زبان فارسی در حوزۀ فاوا (فهرست واژه‌ها بر اساس حوزه، ۱۳۹۱)
- مقالات مربوط به حوزۀ فاوا مثل کنفرانس «فناوری اطلاعات و ارتباطات،دستاوردها و پیشرفت‌ها»، بانک دانش که بانک مقالات علمی و کنفرانس‌های کشور است.
- کنفرانس‌ها و همایش‌های بین المللی حوزۀ فاوا، برق، کامپیوتر و یا فناوری اطلاعات در ایران و در سایر کشورها.
- مستندات پروژه‌های انجام شدۀ فاوا، برق، کامپیوتر و یا فناوری اطلاعات که در دسترس هستند.

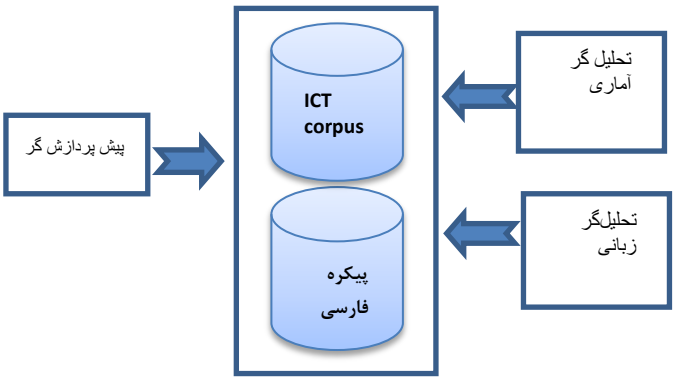
این منابع با استفاده از نرم‌افزار ساخت و مدیریت پیکره (management Corpus) پردازش می‌گردند. از این‌رو، ساخت پیکره خودکار است. تعداد اسناد جمع‌آوری شده در حدود ۶۰۰۰ سند است که همۀ این اسناد مختص حوزۀ فاوا هستند. هرچند با افزایش وب‌سایت‌های حوزۀ فاوا، برگزاری کنفرانس‌ها و چاپ مقالات مربوط به حوزۀ فاوا و غیره، این منابع در حال افزایش هستند. بنابراین، بروزرسانی پیکره حوزۀ فاوا همچنان ادامه دارد.

سیستم مدیریت پیکره

با افزایش نیاز به سیستمی برای ذخیره سازی متون، سامان‌دهی، شاخص‌گذاری به منظور دستیابی سریع تر به داده‌ها، شاخه‌ای در زبان‌شناسی به نام زبان‌شناسی رایانشی (linguistics Computational) ایجاد شده است. به عنوان مثال، شما ممکن است که به متن‌های مربوط به یک موضوع خاص از پیکره نیاز داشته باشید، خواندن کلیه متون موجود در پیکره توسط نیروی انسانی هم یک کار زمان بر و پرهزینه است. در این صورت، شما به یک سیستم کامپیوتری نیاز خواهید داشت که کلیه متون را جستجو کند و همچنین ارتباط بین متن‌ها را با استفاده از تکنیک‌های هوش مصنوعی استخراج کند. بنابراین به ابزاری نیاز داریم تا جستجو در پیکره، اضافه کردن داده‌های جدید، فهرست گیری، اصلاح خطاها و غیره را آسان کند.

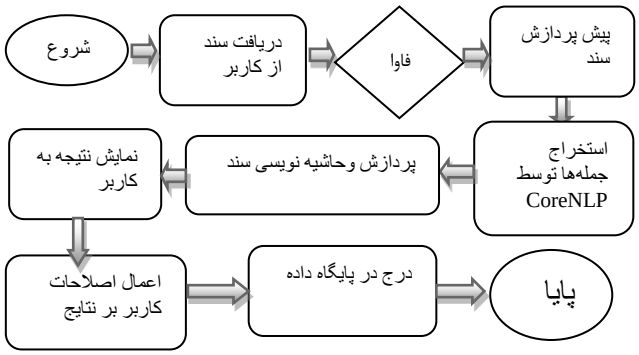
سیستم مدیریت پیکره‌ای که ما توسعه داده‌ایم، دارای دو بخش اصلی است که بخش اول مربوط به ساخت پیکره است و بخش دوم مربوط به استخراج اطلاعات از پیکره می‌باشد. نرم‌افزاری که برای مدیریت پیکره متون ایجاد کرده‌ایم، حاشیه‌نویسی اسناد، جستجو در پیکره و تصحیح خطا را آسان می‌کند. با ایجاد نرم‌افزار ساخت و مدیریت پیکره، فرایند ساخت پیکره آسان شده است. همانطور که در شکل (۲) مشاهده می‌کنید، هر سند قبل از شروع فرایند پردازش اصلی توسط نرم‌افزار پیش پردازش می‌شود. پیش پردازش سند به منظور انتخاب جمله‌های درست و بامفهوم برای پردازش اصلی است. علاوه‌بر آن، در صورتی که در سند کاراکترهای بی‌مفهوم وجود داشته باشند، در فرایند پیش پردازش از سند حذف می‌گردند. ویرایش پیکره، اضافه کردن متن‌های جدید به پیکره، اندیس گذاری، حاشیه نویسی پیکره و غیره از جمله کارهایی است که این سیستم انجام می دهد. بخش دوم در واقع یک موتور جستجوی پیکره است که هدف از طراحی آن مدیریت کردن مجموعه‌های عظیمی از متون است.

حاشیه‌ها، اطلاعات اضافی هستند که برای گویاتر شدن پیکره به آن اضافه می گردد. به این اطلاعات اضافی که به پیکره افزوده می‌شود، ابرداده (Meta data) می‌گویند. جستجو کردن هرآنچه که نیاز کاربر باشد، از طریق ابرداده‌ها صورت می‌گیرد. حاشیه‌نویسی پیکرهٔ فاوا با استفاده از نرم‌افزار ساخت و مدیریت پیکره به صورت خودکار انجام می‌شود. هر سند پس از فرایند حاشیه‌نویسی توسط نرم‌افزار، قبل از اینکه به پیکره اضافه گردد، به متخصص انسانی نمایش داده می‌شود تا در صورتی که خطایی در برچسب‌گذاری، تجزیه کردن و سایر مراحل حاشیه نویسی وجود داشته است، ویرایش شود. بنابراین با استفاده از نرم‌افزار ارائه شده، ساخت پیکره به صورت خودکار و تحت نظارت متخصصین انجام می‌شود. خروجی نرم‌افزار، اسناد حاشیه‌نویسی شده با فرمت XML است که در پیکره قرار می‌گیرند. بنابراین، پیکره حاوی متن‌های حاشیه-نویسی شده‌ای است که در زبان‌شناسی رایانشی به آنها پیکرهٔ حاشیه‌نویسی‌شده (Corpora annotated) می‌گویند.



شکل ۲: بلوک دیاگرام پیکره

پس از آنکه پیکره ساخته شد، برای مدیریت آن به ابزارهایی نیاز داریم. یکی از این ابزار، موتور جستجو در پیکره است. کار این موتور جستجو این است که کلید جستجو را از کاربر بگیرد و بر اساس این کلید پیکره را جستجو کند و بر اساس آن اطلاعات مورد نیاز کاربر را از پیکره استخراج نماید. کاربر می‌تواند بر اساس جمله، تک‌واژه، تک‌واژه همراه با نقش دستوری (Part Of Speech) کلمه همراه با ریشه‌یابی، کلمه به همراه ریشه‌یابی و هم به همراه نقش آن کلمه ، جستجو را انجام دهد. پس از استخراج نتایج از پیکره، نرم-افزار برای هر جملهٔ استخراج شده درخت تجزیه (Parse tree) را رسم می‌کند. از این طریق تحلیل نتایج و مشاهده‌ی ارتباط میان واژگان یک جمله برای متخصصین در پروژه‌های زبان‌شناسی تسهیل می‌گردد. همان‌طور که در شکل (۳) نشان داده شده است، پیکره قابلیت پذیرفتن داده‌های جدید را برای کاربران فراهم می‌کند. بنابراین، یکی از وظایف ابزار مدیریت پیکره این است که قبل از اضافه کردن سند به پیکره از غیرتکراری بودن سند اطمینان حاصل کند. اگر سندی که کاربر قصد افزودن آن به پیکره را دارد، در پیکره موجود باشد، نرم‌افزار پس از تائید کاربر سند جدید را جایگزین سند قبلی می‌کند.



شکل ۳: فلوچارت اجرای افزودن سند به پیکره

ترازبندی جمله‌ها در پیکرهٔ تطبیقی انگلیسی-فارسی فاوا

همانطور که در بخش دوم اشاره شد، تعداد کمی پیکره‌های تطبیقی فارسی-انگلیسی ایجاد شده است. در بین پیکره‌های موجود تاکنون هیچکدام به صورت تمام خودکار ایجاد نشده‌اند و هیچ یک در سطح جمله و کلمه ترازبندی نشده‌اند. ما در این مقاله روشی برای ترازبندی جمله‌ها در یک پیکره تطبیقی با استفاده از یک سیستم بازیابی اطلاعات دو زبانی ارائه داده‌ایم. روند کار بدین صورت است که در ابتدا ما یک سیستم مدیریت پیکرهٔ فاوا را طراحی نمودیم. هدف از طراحی این نرم‌افزار، ساخت پیکرهٔ دو زبانی تخصصی فاوا بوده است. برای ساخت پیکرهٔ تطبیقی ما دو مجموعه مستقل فارسی و انگلیسی از اسناد تخصصی فاوا را جمع-آوری کردیم. سپس با استفاده از نرم‌افزار مدیریت پیکره‌ای که طراحی نموده‌ایم، این اسناد را پردازش کردیم. این سیستم حاوی یک طبقه‌بند (Classifier) برای پذیرش اسناد حوزهٔ فاوا، برچسب‌گذار نقش دستوری واژگان (tager Part Of Speech) (محمدحسین الهی منش،۱۳۹۱) و یک تجزیه‌کننده (Parser) برای

اسناد زبان فارسی و همچنین یک تجزیه‌کننده، برچسب‌گذار نقش دستوری، ریشه‌یاب (Stemmer) برای اسناد زبان انگلیسی است. اسناد تخصصی فاوا با کمک این سیستم حاشیه‌نویسی می‌شوند و علاوه بر آن اطلاعات پردازش شدهٔ اسناد در پایگاه داده پیکره ذخیره می‌شوند. الگوریتم ما از یک مدل ترجمه‌ای کلمه به کلمه و تکنیک بزرگ‌ترین زیر دنباله مشترک برای ترازبندی استفاده می‌کند. در ادامه، گام‌های ترازبندی پیکرهٔ فارسی– انگلیسی فاوا شرح داده شده است.

پردازش سند فارسی و استخراج جمله‌ها و واژگان سند

پس از آنکه اسناد تخصصی انگلیسی حوزه فاوا جمع‌آوری شد و توسط سیستم مدیریت پیکرهٔ فاوا حاشیه نویسی گردید، این نرم‌افزار از اطلاعات اسناد انگلیسی، یک پایگاه داده به نام پایگاه داده انگلیسی فاوا ایجاد می‌کند. پس از اتمام ساخت پیکرهٔ انگلیسی، ما تصمیم گرفتیم تا ترازبندی متون فارسی و انگلیسی را در حین ساخت پیکرهٔ فارسی انجام دهیم. در پیکره‌های تطبیقی دیگر مثل پیکره UTPECC و پیکرهٔ کریمی ابتدا دو مجموعه انگلیسی و فارسی به طور مستقل ایجاد شده‌اند و در نهایت یک فرایند ترازبندی بین دو مجموعه صورت گرفته است اما از آنجا که هدف ما ساخت پیکره‌ای بود که همواره قابلیت افزودن اسناد جدید و بروزرسانی اطلاعاتش وجود داشته باشد، فرایند ترازبندی اسناد انگلیسی و فارسی را به امکانات سیستم مدیریت پیکره اضافه نمودیم.

بنابراین، اولین گام برای ترازبندی، پردازش سند فارسی و استخراج اطلاعات سند است. پس از آنکه مرز جمله‌ها و واژگان متن مشخص گردید، برچسب‌گذار، برچسب دستوری هر کلمه را تعیین می‌کند. پس از پایان فرایند پردازش سند فارسی، تمامی جمله‌های سند برای انجام گام بعدی ترازبندی حاضر هستند.

ترجمهٔ کلمات کلیدی جملهٔ فارسی به انگلیسی

پس از استخراج جمله‌ها و کلمه‌های زبان فارسی نوبت به ترجمه‌ی این جمله‌ها و واژگان می‌رسد. ما برای ترجمه به یک لغت‌نامه انگلیسی–فارسی تخصصی فاوا نیاز داشتیم. اکثر واژگان تخصصی فاوا در زبان فارسی به همان شکل انگلیسی رایج هستند. به عنوان مثال، وب‌سایت، وبلاگ، ویندوز و غیره کلمه‌هایی هستند که در زبان فارسی بیشتر به همان صورت انگلیسی رایج شده‌اند.

ما در ابتدا یک مجموعه از ۳۳۰۰۰ کلمه‌ی انگلیسی جمع‌آوری کردیم که برخی از این واژگان در مجموعه واژگان فرهنگستان ادب و زبان فارسی ترجمه شده‌اند. بخش اعظم آن را نیز از اسناد انگلیسی تخصصی فاوا استخراج کردیم و به فارسی ترجمه نمودیم. همان‌طور که در شکل (۲) مشاهده می‌شود، فرایند ترجمهٔ کلمه-ها و ترازبندی در سطح واژگان با استفاده از این لغت‌نامه صورت می‌گیرد.

شکل ٤: نگاشت واژگان دو مجموعه با استفاده از لغت‌نامه

بخشی از داده‌های هر سند ممکن است حاوی بار معنایی نباشند، مثلاً اطلاعات مربوط به انتشارات و یا نویسندهٔ سند باشند و یا ممکن است که داده‌های زائد (Noise) باشند. وجود این‌گونه از داده‌ها در سند موجب کاهش دقت ترازبندی می‌گردد. برای رفع این مشکل، ما فقط جمله‌هایی که ICT هستند را انتخاب نمودیم. برای این منظور ما از هر جمله کلمات فاوای آن را به عنوان کلمه کلیدی استخراج نمودیم. بدین ترتیب، فقط جمله‌هایی به گام بعدی ترازبندی راه می‌یابند که حاوی واژگان فاوا باشند. پس از استخراج این کلمات، آنها را با استفاده از لغت‌نامه‌ای که خودمان ایجاد کرده‌ایم، به زبان انگلیسی ترجمه می‌کنیم.

اجرای query بر روی پیکره‌ی انگلیسی

در این روش، ما پس از پردازش سند فارسی و استخراج جمله‌ها، به ازای هر یک جملهٔ سند فارسی یک پرس‌وجو (Query) به پایگاه داده ایجاد کردیم. پس از ترجمهٔ کلمه‌های کلیدی فاوا نوبت به ساخت پرس-وجو می‌رسد. حروف اضافه، کلمه‌های غیر مفهومی(رایج) (Stop words)، ارقام و غیره در ساخت پرس‌وجو لحاظ نمی‌شوند. به عبارت دیگر، پرس‌وجو را فقط بر اساس واژگان فاوا می‌سازیم.

همان‌طور که در بخش قبل بیان شد، غالب کلمات حوزه فاوا در زبان فارسی به شکل انگلیسی رایج هستند. بنابراین، ممکن است که یک جمله فارسی حاوی کلمه‌های انگلیسی فاوا باشد. این نوع کلمه‌ها دیگر نیازی به ترجمه ندارند و به همان شکل در پرس‌وجو قرار می‌گیرند. بعد از ایجاد پرس‌وجو به زبان انگلیسی، این پرس‌وجو بر روی تمام جمله‌های اسناد پیکره انگلیسی اعمال می‌گردد و تمام جمله‌های انگلیسی که حداقل یکی از کلمات کلیدی را داشته باشند، از پایگاه داده استخراج می‌گردند.

محاسبهٔ شباهت با تکنیک بلندترین زیر دنباله مشترک

تکنیکی که ما برای یافتن شباهت بین جمله‌های زبان فارسی و جمله‌های زبان انگلیسی استفاده می‌کنیم، تکنیک بلندترین زیر دنباله مشترک (Longest common subsequence (LCS)) نام دارد. LCS روشی است که برای پیدا کردن بلندترین زیردنباله مشترک بین دو دنباله یا دو رشته به کار می‌رود. هدف این روش مقایسه دو رشته و پیدا کردن شباهت بین آنها است. به عنوان نمونه، دو رشتهٔ زیر را در نظر بگیرید:

S1=(7,3,2,6,9)
S2=(2,3,6,7,9)

بلندترین زیردنباله مشترک S1 وS2 برابر است با

S3=(3,6,9)

بلندترین زیردنبالهٔ مشترک این طور تعریف می‌شود که دنباله‌ای مانند S3 است به طوری که حروف موجود در S3 با حفظ ترتیب درS1 وS2 موجود باشند. اما مطلقاً لزومی ندارد که متوالی باشند. همین خواص LCS سبب شد تا ما از این تکنیک برای پیدا کردن میزان شباهت دو جمله استفاده کنیم. برای این منظور روش LCS، دو جمله را به عنوان ورودی دریافت می‌کند و بلندترین زیردنباله از کلمه‌ها که اجزای جمله هستند با حفظ ترتیب استخراج می‌گردند.

در این گام، ما به ازای هر جملهٔ فارسی، مجموعه‌ای از جمله‌های انگلیسی کاندید را که از گام قبلی به-دست آمده‌اند را داریم. سپس با استفاده از تکنیک LCS میزان شباهت بین جملهٔ فارسی و مجموعه جملات

کاندید انگلیسی را به‌دست می‌آوریم. برای محاسبه شباهت از فرمول زیر استفاده نمودیم:

$$r = \frac{l_{s3} \times l_{s3}}{l_{s1} \times l_{s2}} \quad (۱)$$

در معادله (۱)، مقدار متغیر r میزان شباهت جمله‌ی فارسی و انگلیسی را نشان می‌دهد. متغییر l_{S_3} طول بزرگ‌ترین زیر دنباله مشترک دو جمله ، l_{S_1}طول جملهٔ اول و l_{S_2}طول جملهٔ دوم است.

از آنجا که ممکن است به ازای هر یک جمله فارسی، بیشتر از صد جمله انگلیسی به عنوان کاندید از گام قبلی استخراج شده باشد، لازم است تا ما با تعیین میزان آستانه شبیه‌ترین جمله‌ها را انتخاب کنیم. برای این منظور، جمله‌های انگلیسی را بر حسب عدد شباهت به صورت صعودی مرتب می‌کنیم و جمله‌هایی را که مقدار شباهت آنها از آستانه کم‌تر باشد حذف می‌کنیم. در نهایت، اطلاعات مربوط به نگاشت جملهٔ فارسی به جمله یا جمله‌های انگلیسی را در پایگاه داده ذخیره می‌کنیم.

ارزیابی پیکرهٔ تطبیقی فارسی-انگلیسی فاوا

ما برای ارزیابی کیفیت پیکره از ۵ معیار (M. Braschler, P. Scäuble, 1998,183-197) استفاده نمودیم. این معیارها عبارتند از:

۱. رخداد یکسان: هر دو داده در مورد یک رخداد باشند.
۲. رخداد مرتبط: دو داده دقیقا درباره یک رخداد نیستند اما به هم مرتبط هستند.
۳. جنبه مشترک: دو داده موضوع یکسان یا مرتبط ندارند اما در برخی جهات مشترک هستند مثل مکان رخداد یکسان یا افراد یکسان.
۴. اصطلاحات علمی مشترک: شباهت دو داده بسیار کم است اما اصطلاحات مشترکی بین دو حوزه وجود دارد.
۵. نامرتبط. هیچ ارتباطی بین آنها وجود ندارد.

ما برای ارزیابی کیفیت پیکره از این معیارها استفاده کردیم. در صورتی که ترازبندی پیکره معیار ۱ و ۲ را داشته باشد، ترازبندی عالی و در صورتی که معیار ۳و۴ را داشته باشد، ترازبندی خوب است.

ارزیابی و نتایج آزمایشات

در این بخش، ابتدا آماری از وضعیت کنونی پیکره فاوا ارائه می‌دهیم، پس از آن به ارزیابی کیفیت ترازبندی اسناد زبان فارسی و انگلیسی در پیکره فاوا می پردازیم.

ارزیابی پیکرهٔ فاوا و مقایسه با پیکره‌های موجود

پیکرهٔ حوزهٔ فاوا مجموعه‌ای از داده‌های مطلوب، مناسب و دور از نارسایی است که کاربران بر پایه نیاز و هدف پژوهشی خود می‌توانند اسناد خاص خود را با استفاده از نرم‌افزار ساخت و مدیریت پیکره، به آن اضافه نمایند. حتی پژوهندگان می‌توانند اسناد اختصاصی سایر حوزه‌ها را به پیکره اضافه نمایند و تنوع اسناد موجود در پیکره را افزایش دهند. پژوهندگان می‌توانند با استفاده از نرم‌افزار ایجاد شده، تحلیل‌ها و فهرست‌گیری-

های موردنظر خود را انجام دهند. در این نرم‌افزار، مدت زمان پردازش سندی که حاوی ۱۰ هزار جمله است، در حدود ۱۶ ساعت طول می‌کشد. در فرایند پردازش، محتویات این سند تجزیه می‌شود و برچسب‌های دستوری تمام کلمات به همراه رابطه‌ای که این کلمات با یکدیگر دارند، مکان کلمه در جمله، مکان جمله در سند، ریشهٔ تمامی کلمات موجود در سند استخراج می‌گردد. تعداد کلمه‌هایی که در جدول ۱ گزارش شده است، تعداد کلمه‌های منحصر به فرد در پیکره است و هیچ سند تکراری در پیکره وجود ندارد. تمامی این کلمه‌ها، تخصصی حوزهٔ فاوا نیستند و مجموعهٔ کلمه‌های پیکرهٔ حاوی حروف اضافه، اسم‌های سره و سایر کلمه‌هایی است که در اسناد حوزهٔ فاوا وجود دارند و نقش و مفهوم خاصی در جمله ندارند.

جدول ۱: مشخصات پیکره انگلیسی حوزه فاوا

نام پیکره	پیکره انگلیسی فاوا
نحوه ساخت پیکره	خودکار
تعداد اسناد جمع‌آوری شده	۶۰۰۰
تعداد اسناد پردازش شده	۱۶۶
تعداد کل جمله‌های پردازش شده	۶۷۳۰۲۱
تعداد جمله‌های منحصر به فرد پردازش شده	۶۰۴۹۲۴
تعداد کل کلمه‌های پردازش شده	۱۳۶۹۱۶۰۹
تعداد کلمه‌های منحصر به فرد پردازش شده	۱۴۳۵۷۸
مدت زمان ساخت پیکره	حدود۴۴ روز

لازم به ذکر است که پردازش اسناد جمع‌آوری شده هنوز ادامه دارد و آمار ارائه شده مربوط به پردازش ۱۶۶ سند از مجموعهٔ جمع‌آوری شده است. با توجه به تعداد جمله‌ها و کلمه‌هایی که تاکنون به پیکره اضافه شده است، پس از پایان اجرای برنامه پیکره‌ای با حجم بالایی از داده‌های فاوا خواهیم داشت.

پس از بیان مشخصات پیکرهٔ ارائه شده، به ارزیابی پیکره در مقایسه با چند پیکرهٔ معروف که در ابتدای این مقاله معرفی شدند، می‌پردازیم.

همان طور که در جدول (۱) مشاهده می‌شود، روش ساخت پیکرهٔ Penn بدین صورت بوده است که در فاز اول این پروژه که ۳سال هم به طول انجامیده است، متون برچسب گذاری شده‌اند. روش ساخت پیکرهٔ فاوا با استفاده از نرم‌افزاری که ایجاد شده‌است، در مقایسه با پیکره‌ی Penn، از نظر زمانی بسیار به صرفه‌تر می‌باشد. پیکرهٔ آکسفورد یک پیکره بسیار جامع است و منابع آن همانند پیکرهٔ فاوا بسیار گسترده است. یک ایراد مهم پیکرهٔ آکسفورد این است که نه تنها متن‌های ویرایش شده و استاندارد از لحاظ املایی و گرامری در این پیکره قرار گرفته‌اند، بلکه پیکره حاوی اسنادی است که ممکن است از لحاظ املایی و گرامری دارای خطا باشند، مثل متن صفحه‌های وب و یا ایمیل‌ها. این اسناد بدون هیچ پیش‌پردازشی در پیکره قرار گرفته‌اند. برخلاف پیکرهٔ آکسفورد، پیکرهٔ فاوا حاوی اسناد استاندارد است؛ زیرا قبل از اضافه شدن به پیکره توسط

نرم‌افزاری که توسعه داده‌ایم، پیش‌پردازش می‌گردند و جملات بی‌مفهوم از اسناد حذف می‌گردند. علاوه بر این تعداد ۲ بیلیون کلمه‌ای که برای پیکرهٔ آکسفورد گزارش شده است، بدون احتساب کلمات تکراری است. زیرا پیکره، مجموعه‌ای از اسناد است که در هر سند احتمال وجود واژگان تکراری مثل کلمه‌ی "the" بسیار زیاد است. به عنوان مثال، کلمهٔ "the" در این پیکره تقریباً ۱۰۰ میلیون بار تکرار شده‌است. اما ما در پایگاه داده هر واژه را تنها یک بار ذخیره می‌کنیم. آماری که در جدول ۲ ارائه داده‌ایم، تعداد کلمه‌ها و جمله‌های غیرتکراری اسناد حوزهٔ فاوا است. مدت زمان ساخت پیکرهٔ آکسفورد را نیز به‌طور دقیق نمی‌توان تعیین کرد، زیرا این پیکره به تدریج و پس از چند بار ویرایش ایجاد شده‌است. همان طور که مشاهده می‌کنید پیکرهٔ فاوا هرچند که فقط حاوی اسناد حوزه‌ی فاوا است و مثل پیکرهٔ آکسفورد حاوی اسناد متنوع سایر حوزه‌ها نیست، اما از برخی جهات که پیش‌تر بیان گردید، بهینه‌تر از پیکرهٔ آکسفورد است.

پیکرهٔ حاشیه‌نویسی شده‌ای که ما در مقیاس بزرگ ایجاد کرده‌ایم، قابلیت تعمیم برای سایر زبان‌ها را نیز دارد، زیرا اسناد حوزهٔ فاوا حاوی متون علمی این حوزه است و عاری از گفتارهای عامیانه و محاوره‌ای است.

جدول ۲: مقایسهٔ چند پیکرهٔ موجود با پیکرهٔ انگلیسی حوزه‌ی فاوا

نام پیکره	نحوه‌ی ساخت	تعداد واژگان	مدت ساخت پیکره
پیکره Penn	نیمه خودکار	۴.۵ میلیون کلمه	۳ سال
پیکره Brown	دستی	۱میلیون کلمه	۱ سال
پیکره BNC	نیمه خودکار	۱۰۰ میلیون کلمه	۳ سال
پیکره oxford	نیمه خودکار	۲ بیلیون کلمه	–
پیکره حوزه‌ی فاوا	نیمه خودکار	۱۴۳۵۷۸ کلمه	۴۴ روز

پیکرهٔ Brown از آن دسته از پیکره‌هایی است که به‌صورت دستی ساخته شده است. نسخهٔ اولیهٔ این پیکره در سال ۱۹۶۴ از اسناد زبان انگلیسی آمریکایی ساخته شد. این پیکره اولین پیکره‌ای بود که برای زبان انگلیسی ساخته شد. ویراست‌های بعدی این پیکره، تگ‌گذاری شده‌اند، اما به دلیل وجود خطاهای انسانی، این پیکره چندان دقیق نیست. مزیت پیکره‌هایی که به‌صورت خودکار ساخته می‌شوند، این است که نه تنها که در هزینه‌های مربوط به زمان و نیروی کار انسانی صرفه‌جویی می‌کند بلکه فاقد از خطاهای انسانی است. پیکره‌ای که ما با استفاده از نرم‌افزارمان ایجاد کرده‌ایم، در مقایسه با پیکره‌ی Brown هم از نظر حجم و صحت اطلاعات و هم از نظر زمان ایجاد پیکره بهینه‌تر است.

روش ساخت پیکرهٔ BNC نیمه‌خودکار است و در حدود ۴۰۰۰ سند در زمینه‌های متنوع در آن وجود دارد اما همان طور که پیش‌تر گفته شد در پیکرهٔ فاوا ۶۰۰۰ سند مختص حوزهٔ فاوا جمع آوری شده است که از این بابت برای پروژه‌هایی که در این حوزه هستند، یک منبع غنی محسوب می‌گردد. علاوه بر این، مدت زمانی که برای جمع‌آوری و حاشیه‌نویسی ۶۰۰۰ سند صرف شده است، نسبت به مدت زمانی که برای جمع-

آوری و حاشیه‌نویسی ۴۰۰۰ سند برای پیکره BNC صرف شده است ، بهینه‌تر می‌باشد. علاوه بر این، تعداد واژگانی که برای پیکرۀ BNC گزارش شده‌است بدون در نظر گرفتن تکراری بودن واژگان است.

پیکرۀ حاشیه‌نویسی شده‌ای که ما در مقیاس بزرگ ایجاد کرده‌ایم، قابلیت تعمیم برای سایر زبان‌ها را نیز دارد، زیرا اسناد حوزۀ فاوا حاوی متون علمی این حوزه است و عاری از گفتارهای عامیانه و محاوره‌ای است.

مشکلی که ما در جمع‌آوری اسناد فارسی فاوا داشتیم، کمبود منابع فارسی حوزۀ فاوا بود. اکثر اسناد این حوزه به زبان انگلیسی هستند و یا اینکه به صورت کتاب‌های دیجیتالی موجود نیستند. فرایند جمع‌آوری اسناد فارسی همچنان ادامه دارد و آمار ارائه شده در جدول (۲) اسناد جمع‌آوری شده و پردازش شده تا زمان ارائۀ این مقاله است.

جدول ۳: مشخصات پیکرۀ فارسی حوزۀ فاوا

نام پیکره	پیکرۀ فارسی فاوا
نحوۀ ساخت پیکره	خودکار
تعداد اسناد جمع‌آوری شده	۱۰۰
تعداد اسناد پردازش شده	۱۰
تعداد کل جمله‌های پردازش شده	۳۰۲۴
تعداد جمله‌های منحصر به فرد پردازش شده	۲۸۴۳
تعداد کل کلمه‌های پردازش شده	۷۵۲۶۵
تعداد کلمه‌های منحصر به فرد پردازش شده	۷۹۶۵
مدت زمان ساخت پیکره	۴روز

ترازبندی اسناد فارسی و انگلیسی نیز همزمان با اضافه شدن اسناد فارسی انجام می‌گردد. همان‌طور که پیش‌تر نیز گفته شد، تاکنون هیچ پیکرۀ تطبیقی دو زبانی در ایران به‌صورت خودکار ساخته نشده است. بنابراین، پیکره‌ای که قابل مقایسه با پیکرۀ فاوا باشد، وجود ندارد.

پیکره‌های تک زبانی فارسی همانند پیکرۀ دکتر عاصی و پیکره دکتر بی‌جن‌خان به صورت دستی ساخته شده‌اند. پیکره‌هایی همچون پیکرۀ همشهری، محک و پیکره ویکی‌پدیا هم که به‌صورت نیمه‌خودکار ساخته شده است، چون هیچ پیش پردازشی بر روی اسناد انجام نشده، حاوی اسناد غیر استاندارد هستند. پیکرۀ TEP که یک پیکرۀ دو زبانی موازی محاوره‌ای است و مهم‌ترین ضعف آن وجود ابهاماتی است که به دلیل تفاوت زیاد میان واژگان موجود در زبان محاوره‌ای و زبان نوشتاری است. با توجه به ایرادات وارد به پیکره‌های موجود، پیکرۀ فاوا به دلیل پیش پردازش اسناد فارسی و خودکار بودن فرایند ساخت پیکره قابل استفاده در بسیاری از کاربردهای پردازش زبان فارسی است. سایر پیکره‌های موجود نیز یا در دسترس قرار نگرفته‌اند و یا بسیار کوچک هستند، همانند پیکرۀ شیراز، کالیفرنیای جنوبی، میانگاه و پیکرۀ PEN.

ارزیابی ترازبندی پیکره تطبیقی فاوا

در این بخش ما نتایج ارزیابی کیفیت پیکره را ارائه داده‌ایم. ما با کمک نرم‌افزار مدیریت پیکره، اسناد تخصصی حوزه فاوا را پردازش نمودیم و اطلاعات اسناد را در پایگاه داده اسناد انگلیسی ذخیره کردیم. با استفاده از الگوریتم ترازبندی که در بخش قبل ارائه شد، جمله‌های فارسی و انگلیسی را به یکدیگر نگاشت دادیم. نگاشت بین واژگان نیز از طریق لغت‌نامه‌ای که ایجاد نمودیم، صورت گرفته است. مهم‌ترین اصل در ارزیابی یک پیکره تطبیقی کیفیت ترازبندی است.

ما برای ارزیابی سیستم ترازبندی که طراحی نموده‌ایم، از ۶۰۴ هزار جمله غیرتکراری انگلیسی تخصصی فاوا که در پیکره ذخیره نموده‌ایم، استفاده کردیم. سپس، برای آزمایش ، یک مجموعه سند تخصصی هوش-مصنوعی از سایت ویکی‌پدیا استخراج نمودیم. این مجموعه اسناد حاوی ۱۰۰۰ جملهٔ تخصصی فاوا است.

جدول٤: تعداد ترازبندی‌های تعلق گرفته به هر کلاس

	تعداد ترازبندی‌ها	درصد
کلاس ۱	۷۸۰۰	٪۲۶
کلاس۲	۷۲۰۰	٪۲۴
کلاس۳	۱۰۲۰۰	٪۳۴
کلاس٤	۴۳۰۰	٪۱۵
کلاس٥	۴۰۰	٪۱
مجموع	۲۹۹۰۰	٪۱۰۰

پس از چند بار آزمایش سیستم ترازبندی و ارزیابی عدد شباهت به‌دست آمده برای جفت جمله‌های فارسی و انگلیسی، مشاهده گردید که میانگین عدد شباهت برای جمله‌هایی که درست ترازبندی شده‌اند، حداقل حدود ۰.۰۱ است. بنابراین ما میزان آستانه را ۰.۰۱ در نظر گرفتیم. در صورتی که مقدار شباهت دو جمله کمتر از میزان آستانه باشد، این جمله‌ها به یکدیگر نگاشت داده نمی‌شوند. ما ۵ معیار ارزیابی را که در بخش قبل معرفی کردیم، تحت عنوان ۵ کلاس در نظر گرفتیم. در جدول (۴) نتایج ارزیابی ترازبندی سند آزمایشی که مربوط به حوزهٔ هوش مصنوعی است، نشان داده شده است. همان‌طور که در جدول (۴) مشاهده می‌کنید، ۱درصد از جمله‌های فارسی به جمله‌های نامرتبط ترازبندی شده‌اند. هرچند که میزان خطای ترازبندی ناچیز است اما ما با بررسی جمله‌های نامرتبط به این نکته پی بردیم که تعلق برخی از واژگان به چندین حوزه سبب به وجود آمدن این خطا در ترازبندی شده است. به عنوان مثال، کلمهٔ "شبکه" ممکن است هم در جمله‌های مربوط به حوزه شبکه‌های کامپیوتری و هم در حوزه شبکه‌های عصبی وجود داشته باشد. زمانی که تعدد واژگانی که به چندین حوزه تعلق دارند، در یک جمله زیاد باشد، احتمال خطا افزایش می‌یابد. به طور کلی، ۱۰۰ جمله از سند آزمایشی ما نگاشت داده نشده است.

نکتهٔ دیگری که ما آن را مزیت سیستم ترازبندی ارائه شده می‌دانیم، حذف جمله‌های زائد و بی‌مفهوم از فرایند ترازبندی است. نتایج ما نشان داد که نگاشت جمله‌هایی که حاوی کلمه‌های فاوا هستند، نه تنها که باعث افزایش دقت ترازبندی می‌گردد، بلکه به دلیل حذف جمله‌های زائد و یا نامرتبط به حوزه فاوا سرعت ترازبندی افزایش می‌یابد. بررسی نتایج به‌صورت دستی نشان داد که جمله‌های فارسی که به هیچ جملهٔ انگلیسی نگاشت داده نشده‌اند، یا دارای کلمه‌های فاوا نبودند و یا غالباً از کلمه‌های غیر رایج در حوزه فاوا استفاده کرده بودند که این واژگان در لغت‌نامهٔ ما موجود نبوده است. در جدول (۵) میانگینی از نتایج به‌دست آمده برای سند آزمایشی را محاسبه و ارائه نموده‌ایم.

جدول ۵: میانگین نمره شباهت حاصل از ترازبندی سند آزمایشی

۰.۰۷	**میانگین نمره شباهت کلاس ۱**
۰.۰۵	**میانگین نمره شباهت کلاس ۲**
۰.۰۴	**میانگین نمره شباهت کلاس ۳**
۰.۰۳	**میانگین نمره شباهت کلاس ٤**
۰.۰۰۲	**میانگین نمره شباهت کلاس ٥**

میانگین نمره شباهتی که به ازای جمله‌های ترازبندی شده در هر کلاس محاسبه کردیم، برای کلاس ۱ که بهترین حالت ترازبندی است، نسبت به بقیهٔ کلاس‌ها بیشتر است. مقدار این نمرهٔ شباهت برای سایر کلاس‌ها همان‌طور که در جدول (۵) مشاهده می‌کنید، به ترتیب کاهش یافته است. زیاد بودن نمره شباهت برای کلاس ۱ و کم بودن آن برای کلاس ۵، نشان دهنده صحت این معیار در فرایند ترازبندی است. در جدول (۶) چند نمونه از جمله‌هایی را که با هم ترازبندی شده‌اند، نمایش داده‌ایم.

جدول ٦: نمونه‌ای از ترازبندی صورت گرفته با سیستم ترازبندی فاوا

نوع کلاس	جملهٔ انگلیسی	جملهٔ فارسی
کلاس ۱	Natural language processing (NLP) is a field of computer science artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages .	پردازش زبان‌های طبیعی پردازش زبان‌های طبیعی یکی از زیرشاخه‌های بااهمیت در حوزه گسترده هوش مصنوعی و نیز در دانش زبان شناسی است.
کلاس ۱	They simply apply statistical methods to the words surrounding the ambiguous word.	آنها به سادگی روش‌های آماری را برای کلمات اطراف کلمه مبهم، اعمال می‌کنند.
کلاس۲	The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses.	

در ردیف اول جدول، دو جمله نگاشت شده به هم با داشتن ۳ کلمه فاوای مشترک کاملاً مشابه هم هستند، اما در ردیف دوم جمله فارسی با یکی از دو جمله شابهت زیادی دارد و با جملهٔ دیگر، هرچند که هر دو مربوط به یک حوزه هستند، کاملاً مشابه نیستند. با بررسی نمونه‌های دیگر به این نکته پی بردیم که هر چقدر واژگان فاوا تخصصی‌تر باشند، احتمال پیدا کردن جملهٔ مشابه بیشتر خواهد شد. به عنوان مثال، وجود کلماتی از قبیل "هوش مصنوعی"، "زبان شناسی" و"پردازش زبان های طبیعی" باعث استخراج جملهٔ انگلیسی صحیح از پیکره شده است.

جدول (۷) نشان می‌دهد که سیستم ترازبندی فاوا هر جمله فارسی را به‌طور میانگین به ۳۰ جمله انگلیسی نگاشت می‌دهد که از این تعداد به طور متوسط ۹ جمله به کلاس ۱، ۸ جمله به کلاس ۲ و ۱۱ جمله به کلاس ۳ تعلق دارد. اگر کلاس ۱،۲ ، ۳و۴ را به‌عنوان ترازبندی‌های خوب در نظر بگیریم. در این صورت، می‌توان ادعا کرد که سیستم ترازبندی ما دارای ۹۳٪ دقت در نگاشت جمله‌های فارسی و انگلیسی است.

جدول ۷: نتایج حاصل از ترازبندی از طریق سیستم ترازبندی فاوا

۳۰	میانگین تعداد جمله‌های انگلیسی ترازبندی شده به ازای هر جملهٔ فارسی
۸.۶	میانگین تعداد جملات کلاس ۱ به ازای هر جملهٔ فارسی
۸	میانگین تعداد جملات کلاس ۲ به ازای هر جملهٔ فارسی
۱۱.۳	میانگین تعداد جملات کلاس ۳ به ازای هر جملهٔ فارسی
۴.۷	میانگین تعداد جملات کلاس ۴ به ازای هر جملهٔ فارسی
۰.۵	میانگین تعداد جملات کلاس ۵ به ازای هر جملهٔ فارسی

از آنجا که در ترازبندی پیکرهٔ تطبیقی ممکن است حتی دو جمله دو حوزهٔ مختلف را پوشش دهند، نمی‌توان نتایج کلاس ۴ را جزء نتایج بد ارزیابی کرد. بنابراین ما نگاشت این جمله‌ها به یکدیگر را نیز در پایگاه داده ذخیره کردیم. منتها ما به منظور نشان دادن کم بودن شباهت این‌گونه جمله‌ها و زیاد بودن شباهت جمله‌های کلاس‌های قبل، نمره شباهت جمله‌ها را نیز ذخیره کردیم.

تاکنون پیکرهٔ فارسی–انگلیسی که هم تطبیقی باشد و هم در سطح جمله و کلمه ترازبندی شده باشد، ساخته نشده است. پیکره‌های تطبیقی موجود که در بخش دوم معرفی شدند نیز یا در سطح سند ترازبندی شده‌اند و یا دقتی برای فرایند ترازبندی‌شان گزارش نکرده‌اند. از آنجا که در پیکره‌های موازی اسناد دو زبان ترجمه یکدیگر هستند، ارزیابی انجام شده در جدول (۴)، (۵) و (۷) با نتایج ارزیابی پیکره‌های موازی قابل مقایسه نیست؛ زیرا در پیکرهٔ تطبیقی اسناد ترجمه همدیگر نیستند و حتی ممکن است که به ازای یک جمله هیچ جمله‌ی معادل وجود نداشته باشد و یا چندین جملهٔ معادل به ازای یک جمله وجود داشته باشد. از طرفی، در ترازبندی پیکرهٔ تطبیقی الزامی برای اینکه دو جمله دقیقاً مثل هم باشند، وجود ندارد. بلکه یکسان بودن مفهوم دو جمله کفایت می‌کند. به عنوان مثال دقت ترازبندی پیکرهٔ موازی فارسی–انگلیسی ویکی‌پدیا ۴۳٪ گزارش شده است. در پیکرهٔ ویکی‌پدیا از آنجا که برخی از صفحات ویکی‌پدیا درباره اشخاص هستند و

این اسناد در بدنهٔ پیکره قرار گرفته‌اند، نتیجه حاصل از ترازبندی از طریق ترجمهٔ جمله‌ها، به نسبت خوب بوده است. اما در سیستم ترازبندی که ما طراحی کردیم از آنجا که محور ترازبندی فقط جمله‌های فاوا هستند، بسیاری از این جمله‌ها در گام سوم حذف می‌گردند. حذف جمله‌های زائد و غیرمرتبط به حوزهٔ فاوا یکی دیگر از علل افزایش دقت سیستم ترازبندی فاوا است. ترازبندی‌های کلاس ۳،۲ و ۱ را تعداد ترابندهای درست فرض کردیم، هرچند که ترازبندهای کلاس ۴ هم مناسب هستند و ما آنها را با درجهٔ شباهت‌شان در پایگاه داده ذخیره نمودیم اما در محاسبهٔ دقت ترازبندی آنها جز ترازبندی‌های نادرست در نظر گرفتیم. ترازبندی با استفاده از محاسبهٔ شباهت از طریق LCS و امتیاز کلمات ۸۳٪ به‌دست آمد.

نتیجه‌گیری

پیکره‌های حاشیه‌نویسی شده که مقیاس بزرگی دارند، نه تنها که در تحقیقات حوزهٔ زبان‌شناسی مؤثر هستند، بلکه در توسعهٔ سیستم‌های پردازش زبان طبیعی نیز مؤثر هستند. تمام مؤسسات، شرکت‌ها و دانشگاه‌هایی که کار ترجمه انجام می‌دهند و یا پروژه‌هایی که نیاز به پیکره دارند، می‌توانند ار نتیجه این کار استفاده کنند. هدف ما ایجاد پیکره دو زبانهٔ فارسی–انگلیسی حوزهٔ فاوا است که تاکنون ساخت پیکرهٔ انگلیسی آن میسر شده است و ساخت مجموعهٔ فارسی این پیکرهٔ دو زبانه همچنان ادامه دارد. آزمایشات ما نشان داد که این تکنیک برای اسناد گوناگون و حتی نویزی نیز دقت قابل قبولی داشته است. ما پی بردیم که تکنیک ما یک روش مستقل از زبان است و برای ترازبندی اسناد سایر زبان‌ها نیز قابل استفاده است. همچنین، این تکنیک را می‌توان برای ساخت پیکره‌های چند زبانی نیز استفاده کرد. در این تکنیک، فقط ترتیب و مکان کلمات کلیدی دو جمله برای ترازبندی در نظر گرفته می‌شود و از این طریق ما تأثیر کلمه‌های بی‌مفهوم و زائد را کاهش داده‌ایم. به طور کلی، روش ترازبندی ما دارای دقت قابل قبولی بوده است.

سپاسگزاری

طراحی و ایجاد پیکرهٔ متنی دوزبانه برای حوزهٔ تخصصی فاوا از مجموعه زیرفعالیت‌های پروژهٔ توسعهٔ وردنت فارسی فاوا به مجری‌گری دانشگاه بوعلی سینا و با حمایت مرکز تحقیقات ارتباطات و فناوری اطلاعات (Research Institute for ICT (ITRC)) می‌باشد. همچنین، از جناب آقای محمدحسین الهی منش و پژوهشگاه نور به دلیل در اختیار گذاشتن برچسب گذار متون فارسی تشکر می‌نماییم.

منابع

– عاصی ,د .م,. (۱۳۹۱)،" انجمن زبان فارسی"
– Available: http://persianlanguage.ir/interviews/392.
– Marcus, Marcinkiewicz, Santorini, June 1993, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics - Special issue on using large corpora*, vol. 19, no. 2, 313-330.
– 2011, "Oxford Dictionary", Available: www.oxforddictionaries.com.
– H. K. W. N. Francis, 2011, Available: http://dilbilim.info/yukseklisans/Corpus%20Based/Brown_Corpus_Manual.pdf.
– Leech G, Graside R, Brayant M, 1994, "CLAWS4: the tagging of the British National Corpus" in *15th conference on Computational linguistics*, 622-628.
– Davies M, 2009, "The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights", *International Journal of Corpus Linguistics*, vol. 19,159-190 .
– Idle N, Macload C, 2001, "The American National Corpus: A standardized resource for American English", *Proceedings of Corpus Linguistics 2001*, 831-836.
– Bijankhan, 12 2011, "Biankhan corpus", Available: http://ece.ut.ac.ir/dbrg/bijankhan/.
– بی‌جن‌خان، ر. ر. م (۱۳۸۶)، "به کار گیری یک نظام برچسب‌دهی برای تعبیر و تفسیر یک پیکره متنی زبان فارسی" در *مجموعه مقالات هفتمین همایش زبان‌شناسی ایران*.
– "Hamshahri Collection," 12 2011. Available: http://ece.ut.ac.ir/dbrg/hamshahri/.
– عاصی ,د. م. ۱۳۸۸, Available: http://linguist87.blogfa.com/post-515.aspx.
– ۲۰۱۱ ۱۲, Available: http://www.hawzah.net.
– Shakery Azadeh, Faili Heshaam, Baradaran Hashemi Homa, 2010, "Creating a Persian-English Comparable Corpus," *Springer,Computer Science*, vol. 6360/2010, 27-39.
– QasemAghaee Naser, Mohammadi Mehdi, March 2010, "Building bilingual parallel corpora based on wikipedia," Computer Engineering and Applications (ICCEA), *Second International Conference on*, no. 19-21.
– Pilevar M.T, Pilevar A.H, Faili H, 2011, "TEP: Tehran English-Persian Parallel Corpus", *12th International Conference, CICLing*, pp 68-79.
– May Win, Narayanan Shrikanth, Georgiou Panayiotis, Ganjavi Shadi, S. Belvin Robert, 2004, "Creation of a Doctor-Patient Dialogue Corpus Using Standardized Patients", *LREC Conference*.
– Mansouri Rad Hamid, Megerdoomian Karine, Rémi Zajac , Jan W. Amtrup, 2000, "Persian-English Machine Translation: An Overview of the Shiraz Project", *Memoranda in Computer and Cognitive Science*.
– Farajian Mohammad amin, 2011, "PEN: Parallel English Persian News Corpus".
– "ELRA" ,2011, Dec , Available:http://catalog.elra.info/index.php.
– Karimi Sarvnaz, 2008, "Machine Transliteration of Proper Names between English and Persian," Melbourne, Victoria, Australia, *Ph.D. thesis*.
– Kucera H, Francis W. N, July 1979, "A Standard Corpus of Present-Day Edited American English, for use with Digital Computers," *Brown University*.
– Kenneth W. Church William A. Gale, March 1993, "A Program for Aligning Sentences in Bilingual Corpora," *Computational Linguistics*, MIT Press Cambridge, MA, USA , vol. 19, no. 1.

– Stanley F. Chen, 1993, "Aligning sentences in bilingual corpora using lexical information," in *ACL '93 Proceedings of the 31st annual meeting on Association for Computational Linguistics.*

– ۱۳۹۱, "فهرست واژه ها براساس حوزه," Available: http://www.persianacademy.ir.

– الهی منش محمدحسین ، دکتر مینایی بهروز (۱۳۹۱)،"برچسب گذار ادات سخن متون فارسی با استفاده از مدل مخفی مارکوف "*فناوریهای پردازش هوشمند متون اسلامی.*

– Braschler M, Scäuble peter, 1998, " Multilingual information retrieval based on document alignment techniques", In: *ECDL*. pp. 183-197.