Using Computational Methods for Persian Collocations Identification and Extraction

Zainabolhoda Heshmati*
PhD In Telecommunication Engineering; Assistant Professor; School of Intelligent Systems; University of Tehran; Tehran, Iran;
Email: zheshmati@ut.ac.ir

Mina Maleki Vika
Master of Computational Linguistics; University of Tehran; Tehran, Iran Email: mina.maleki.vika@ut.ac.ir

Mahmood Bijankhan
PhD in Linguistics; Professor; Faculty of Literature and Humanities; University of Tehran; Tehran, Iran;
Email: mbjkhan@ut.ac.ir

Hadi Veisi
PhD in Computer Engineering; Associate Professor; School of Intelligent Systems; University of Tehran; Tehran, Iran;
Email: h.veisi@ut.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 40 | No. 2 | pp. 577-604
Winter 2025
https://doi.org/10.22034/jipm.2024.2030932.1646

Received: 03, Jun. 2024 **Accepted: 02, Nov. 2024**

Abstract: This article explores the recognition of collocations in Persian language. Previous research in this field has primarily been statistical and comparative in nature. The objective of this study is to identify collocations using a corpus-based and computational approach. To this end, the Persian language database is utilized as the research corpus. Additionally, due to the absence of a comprehensive collocation dictionary for Persian, a dataset of collocations has been constructed based on the Advanced Learners' Persian Dictionary. Using FastText embedding vectors, a language model is trained with a Long Short-Term Memory (LSTM) network. Furthermore, by fine-tuning ParsBert, the performance of this language model is evaluated using lists of a thousand collocations and non-collocations. Finally, a comparative analysis of collocation recognition is conducted using Google Translate by translating a thousand Persian sentences into English, each containing at least one collocation. The results indicate that the ParsBert model achieves recall rates of 93.95% and 85.8% for collocation and non-collocation recognition, respectively. In contrast, the LSTM-based language model achieves recall rates of 6.6% and 0% for collocation and non-collocation recognition, respectively. The comparative analysis of Google Translate accuracy in translating

* Corresponding Author

collocations yielded three key findings: 1) The collocation was correctly recognized and translated; 2) The collocation was not correctly recognized, resulting in a literal, word-for-word translation; and 3) The collocation is not recognized, leading to an incorrect translation.

Keywords: Collocation, ParsBert, Long Short-Term Memory, Computational Linguistics, Persian Language

شناسایی و استخراج باهمایی‌های زبان فارسی با استفاده از روش‌های رایانشی

زینب‌الهدی حشمتی
دکتری مخابرات؛ استادیار؛ دانشکده سامانه‌های هوشمند؛ دانشگاه تهران؛ تهران، ایران؛
پدیدآور رابط zheshmati@ut.ac.ir

مینا ملکی ویکاء
دانش‌آموخته کارشناسی ارشد زبان‌شناسی رایانشی؛ دانشگاه تهران؛ تهران، ایران؛
mina.maleki.vika@ut.ac.ir

محمود بی‌جن‌خان
دکتری زبان‌شناسی؛ استاد؛ دانشکده ادبیات و علوم انسانی؛ دانشگاه تهران؛ تهران، ایران؛
mbjkhan@ut.ac.ir

هادی ویسی
دکتری مهندسی کامپیوتر؛ دانشیار؛ دانشکده سامانه‌های هوشمند؛ دانشگاه تهران؛ تهران، ایران h.veisi@ut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۳/۰۳/۱۴ | پذیرش: ۱۴۰۳/۰۸/۱۲ | مقاله برای اصلاح به مدت ۲۰ روز نزد پدیدآوران بوده است.

چکیده: در این مقاله به بازشناسی باهمایی‌ها در زبان فارسی پرداخته می‌شود. پژوهش‌های صورت‌گرفتهٔ زبان فارسی در این زمینه عمدتاً آماری و مقابله‌ای بوده است. هدف این پژوهش بازشناسی باهمایی‌ها به روش پیکره‌بنیاد و رایانشی است. برای این منظور از پایگاه داده زبان فارسی به‌عنوان پیکره پژوهش استفاده شده است. همچنین به‌علت نداشتن لغت‌نامهٔ باهمایی‌ها در زبان فارسی مجموعه داده‌ای از باهمایی بر اساس کتاب فرهنگ زبان‌آموز پیشرفته فارسی ساخته شده است. با استفاده از بردارهای تعبیهٔ fasttext مدل زبانی با شبکهٔ حافظه کوتاه‌مدت ماندگار آموزش داده می‌شود. همچنین با تنظیم دقیق «پارس‌برت» فراخوانی این مدل زبانی با استفاده از لیست‌های هزارتایی باهمایی‌ها و ناباهمایی‌ها محاسبه شد. در انتها، بررسی مقابله‌ای بازشناسی باهمایی در موتور ترجمه گوگل با استفاده از ترجمه هزار جمله فارسی به انگلیسی که هر یک از جملات دارای یک باهمایی است، انجام شد. نتایج نشان می‌دهد که مدل «پارس‌برت» با فراخوانی ۹۵/۹۳ درصد و ۸۵/۸ درصد به‌ترتیب، به بازشناسی باهمایی و ناباهمایی و مدل زبانی آموزش‌دیده با شبکهٔ حافظه کوتاه‌مدت ماندگار به‌ترتیب باهمایی و ناباهمایی را با فراخوانی ۶/۶ درصد و ۰ درصد بازشناسی کرد. همچنین بررسی مقابله‌ایِ دقت ترجمه موتور

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۴۰ | شمارهٔ ۲ | ص‌ص ۵۷۷-۶۰۴
زمستان ۱۴۰۳
https://doi.org/10.22034/jipm.2024.2030932.1646

گوگل در ترجمه باهمایی‌ها سه نتیجه را دربرداشت: ۱) باهمایی به‌درستی بازشناسی و ترجمه شد، ۲) باهمایی به‌درستی بازشناسی نشد و ترجمه به‌صورت تحت‌اللفظی و واژه‌به‌واژه است، و ۳) باهمایی بازشناسی نشد و ترجمهٔ غلطی صورت پذیرفته است.

کلیدواژه‌ها: باهمایی، پارس‌برت، حافظه کوتاه‌مدت ماندگار، زبان‌شناسی رایانشی، زبان فارسی

۱. مقدمه

یکی از بخش‌های مهم از درک زبان، دانستن مفاهیم و چگونگی استفاده از کلمات در مجاورت هم است. کلماتی را که در یک رابطه دستوری مشخصی باشند -مانند اسم و صفت یا مفعول و فعل- و وقوع آن‌ها در کنار یکدیگر متفاوت با سایر کلمات باشد، باهمایی[1] می‌نامند. برای مثال «آب سیاه»، به‌عنوان یک بیماری چشم و «چشم سفید» به معنای گستاخ و لجوج، در قالب اسم+صفت؛ و «آستین بالا زدن» و «کمر بستن» به معنای آماده شدن برای انجام کار، در قالب مفعول+فعل باهمایی هستند. باهمایی‌ها در بعضی از زبان‌ها در یک واژگان مستقل گردآوردی شده‌اند که از آن‌ها برای یادگیری زبان دوم و ترجمه استفاده می‌شود. اما امروزه، روش‌های یادگیری ماشین جایگزین روش‌های سنتی در پردازش زبان طبیعی شده و الگوریتم‌ها دانش زبانی انسان را از روی حجم انبوه داده‌های یک پیکره زبانی می‌آموزند و در بازشناسی الگوهای زبانی، مانند باهمایی‌ها، به‌کار می‌گیرند. حوزه‌هایی که بازشناسی و استخراج باهمایی‌ها برای آن‌ها ضرورت بنیادین دارد، عبارت‌اند از: یادگیری زبان مادری، یادگیری زبان دوم[2] (Ramos et al. 2010)، ترجمه ماشینی، رفع ابهام معنایی[3] (Maru et al. 2019)، تجزیه و ترجمه ماشینی[4] (Seretan 2013)، درک و تولید زبان طبیعی[5] (Smadja and McKeown 1991) و بازیابی اطلاعات[6].

باهمایی‌ها در این مقاله در دو بخش بررسی می‌شوند: بخش اول، به بازشناسی باهمایی‌ها در مدل‌های زبانی یادگیری عمیق می‌پردازد. حل مسائل زبانی در مباحث زبان‌شناسی رایانشی، در گذشته عمدتاً با روش‌های آماری انجام می‌شد، اما امروزه با پیشرفت سخت‌افزار و پردازنده‌های قوی و با استفاده از شبکه‌های عصبی عمیق مدل‌های زبانی قدرتمندی ایجاد شده است. ازآنجاکه در پیشینه پردازش رایانشی زبان فارسی،

1. collocation
2. second language learning
3. word sense disambiguation
4. parsing and machine translation
5. natural language generation
6. information retrieval

بررسی باهمایی‌ها به روش آماری و مقابله‌ای بوده است، برای بار نخست در زبان فارسی، بازشناسی باهمایی‌ها در مدل‌های زبانی ساخته‌شده با شبکه‌های عصبی مورد بررسی قرار می‌گیرد.

بخش دوم، به دقتِ موتورِ ترجمه گوگل در ترجمهٔ باهمایی‌ها در جملات فارسی به جملات انگلیسی می‌پردازد. ازآنجاکه موتور ترجمه گوگل یکی از ابزارهای پرکاربردِ دنیا در پردازش زبان طبیعی در بخش ترجمه است و در پشت صحنه آن از داده‌های زبانی و روش‌های گوناگون رایانشی استفاده می‌شود و ازآنجاکه یکی از معیارهای دقت یک موتور ترجمه، ترجمهٔ صحیح باهمایی‌ها و اصطلاحات زبانی است، به‌دست آوردن دقت ترجمه باهمایی‌های فارسی در ترجمه جملات به انگلیسی دارای اهمیت است.

به‌طور خلاصه، پرسش‌های پژوهش به‌شرح زیر است:

۱. با استفاده از مدل زبانی یادگیری عمیق، بازشناسی باهمایی‌ها با چه میزان دقت امکان‌پذیر است؟
۲. موتور ترجمه گوگل با چه دقتی باهمایی‌های زبان فارسی را در ترجمه جملات فارسی به انگلیسی بازشناسی کرده و آن‌ها را ترجمه می‌کند؟

ساختار مقاله به‌شرح زیر است: ابتدا در بخش ۲، پیشینه پژوهش شامل آرای زبان‌شناسان در باب مفهوم باهمایی‌ها و پژوهش‌های انجام‌شده در این زمینه مرور می‌شود. سپس در بخش ۳، ساختار پژوهش، توضیحات داده‌های مورد استفاده و روش‌های اجرایی شرح داده می‌شود. در بخش ۴، نتایج حاصل از روش‌ها آورده شده، و سرانجام، مقاله با بخش نتیجه‌گیری و پیشنهادات آتی به پایان می‌رسد.

۲. پیشینه پژوهش

«فرث» به‌عنوان اولین کسی که اصطلاح باهمایی را مطرح می‌کند، معتقد است که در باهمایی، معنای هر واژه به کمک معنای واژه‌هایی که به‌طور معمول، با آن همراه می‌شوند، مشخص می‌شود و این متمایز از باهمایی نحوی است. وی این پدیدهٔ زبانی را معنابنیاد فرض کرده است و نه دستوری. از نگاه «فرث»، معنای حاصل از باهمایی نحوی همان معنای دستوری است. برای مثال، در زبان انگلیسی، آمدنِ ing بعد از واژهٔ fancy (به‌عنوان یک فعل) نوعی باهمایی نحوی است (Firth 1957). «هلیدی و حسن» معتقدند

که تکرار[1] و باهمایی ابزار انسجام واژگانی هستند (Halliday and Hasan 1986). «کاوی» به بررسی باهمایی و اصطلاحات در فرهنگ اصطلاح‌های رایج انگلیسی آکسفورد[2] می‌پردازد. وی باهمایی را به دو گروه «باهمایی آزاد»[3] و «باهمایی محدود»[4] تقسیم می‌کند. در باهمایی آزاد، هر دو عنصر همنشین از معنای حقیقی برخوردارند، ولی در باهمایی محدود، یکی از عناصر از معنای حقیقی و دیگری از معنای مجازی برخوردار است (COWIE 1981). «نتینگز و دیکاریکو» باهمایی را به دو دسته «زایا» و «غیرزایا» تقسیم می‌کنند. باهمایی‌های زایا اصطلاح‌های واقعی هستند؛ چرا که معنای کلِ عبارت از معنای اجزای سازنده‌اش قابل تشخیص نیست؛ مانند by and large (به معنای کلاً). از سوی دیگر، باهمایی‌های غیرزایا از لحاظ معنایی کاملاً شفاف هستند؛ مانند see the river (به معنای رودخانه را دیدن). به گفتهٔ «وود»، باهمایی‌هایی که در آن‌ها به‌جای یک واژهٔ واژگانی خاص، یک مقولهٔ دستوری خاص داریم، «باهمایی نحوی» نام دارد. وی برای نمونه از عبارت off with his head (سرش را از تن جدا کنید) مثال می‌زند. وی این توالی واژگانی را به‌صورت پیوستاری در نظر می‌گیرد که یک طرف آن توالی‌های آزاد و طرف دیگر آن، اصطلاحات وجود دارد که ساختار آن در شکل ۱، مشاهده می‌شود (Nattinger and DeCarrico 1992 نقل از Wood).

توالی‌های آزاد – باهمایی‌های نحوی – باهمایی‌ها – اصطلاح‌ها

شکل ۱. پیوستار توالی‌های واژگانی (Nattinger and DeCarrico 1992)

«هاوسمن» باهمایی را *ترکیب دوتاییِ واژه‌ها*[5] می‌داند. این ترکیب از دو جزء «پایه»[6] و «باباهمایی»[7] تشکیل شده است. واژه‌ای که معنای مستقل دارد «پایه»، و واژهٔ دیگر «باباهمایی» نامیده می‌شود. باباهمایی معنای خود را از هم‌نشینی با پایه به‌دست می‌آورد. در عبارت pay attention، attention «پایه» است و pay که معنایش را از هم‌نشینی با attention به‌دست می‌آورد، «باباهمایی» است (Bischof and Klausurtagung 2004). «ملچوک» دیدگاه موجود نسبت به همایند را محدودتر کرد. وی همایندها را در چارچوب نقش‌های واژگانی[8] (در نظریهٔ معنا–متن[9]) جای داد (Seretan 2003). در نظریهٔ معنا–متن معنا، مفهوم و هرآنچه گوینده قصد

1. reiteration
2. Oxford Dictionary of Current Idiomatic English (ODCIE)
3. open collocation
4. restricted collocation
5. binary word combinations
6. base
7. collocate
8. lexical function (LF)
9. meaning-text theory (MTT)

انتقال آن را دارد، به متن مربوط می‌شود. هستهٔ اصلی نظریهٔ معنا–متن یعنی جایی که بخش اعظم داده‌های زبان در آن ذخیره می‌شود، *واژگان معنا–بنیاد صوری‌شده*[1] است که *فرهنگ ترکیبی تبیینی*[2] نامیده می‌شود. واحد مورد نظر در این فرهنگ، یک واژه یا مجموعه‌ای از عبارت‌هاست. کلیهٔ اشتقاق‌های معنایی[3] یک واژه یعنی روابط واژگانی جانشینی و هم‌نشینی آن به‌وسیلهٔ مجموعه‌ای از توابع واژگانی مشخص می‌شود. توابع واژگانی به‌صورت فرمول $f(x) = y$ نمایش داده می‌شود که در اینجا x واژهٔ اصلی[4] و y ارزش[5] آن محسوب می‌شود. خود نقش واژگانی، یک رابطه معنایی–نحوی است که کلمه یا عبارتی را به مجموعه‌ای از کلمات یا عبارات پیوند می‌دهد. نقش‌های واژگانی بیان‌کننده این حقیقت هستند که در زبان، کلمات یا عباراتی وجود دارد که کاربرد آن‌ها محدود به کلمه یا عبارت دیگر در زبان است. در حدود ۵۰ نقش واژگانی کوچک و متفاوت در مدل معنا–متن وجود دارد که بعضی از آن‌ها دارای روابط معنایی و بعضی دیگر دارای روابط نحوی و بعضی دیگر دارای هم‌وقوعی واژگانی محدود هستند. «کروز» همایند را توالی واحدهای واژگانی می‌داند که به‌طور معمول، با هم واقع می‌شوند و هر یک از سازه‌های واژگانی، یک سازه معنایی محسوب می‌شود. وی به‌عنوان مثال، عباراتی نظیر high winds و fine weather را نمونه‌هایی از همایند دانسته است (Cruse 1986). «چرچ» و همکاران به بررسی نقش آمار تحلیل‌های واژگانی پرداختند. از نظر آن‌ها باهمایی‌ها در رویکرد آماری، در یک پیکره زبانی توالی‌هایی تکراری هستند که احتمال وقوعشان بیش از تصادف صرف است (Church et al. 1991). «مک‌کوهن و رادوف» باهمایی‌ها را نقطه‌ای واقع در یک پیوستار واژگانی می‌دانند که در یک‌سو، اصطلاحات ثابت و در سوی دیگر آن، روابط آزاد کلمات وجود دارد (McKeown and Radev 2000). «سرتان» باهمایی را «ترکیب دل‌بخواهی و همیشگی کلمات» می‌داند. دل‌بخواهی بودن باهمایی‌ها ناشی از این حقیقت است که آن‌ها تحت هیچ‌گونه قواعد نحوی یا معنایی کلی درنمی‌آیند. افزون بر این، باهمایی‌ها حوزه وابسته، زبان وابسته، گویش وابسته، و حتی زمان وابسته‌اند. کاربرد درست باهمایی‌ها به طبیعی بودن متن کمک می‌کند (Seretan 2003 به نقل از بنسون). «عاصی» معیارهای بازشناسی ترکیبات باهمایی را در پنج گروهِ ساختاری، معنایی، واژگانی، کاربردی، و نحوی دسته‌بندی

1. formalized semantically-oriented lexicon
2. explanatory combinatorial dictionary (ECD)
3. semantic derivations
4. keyword
5. value

کرده است. وی دیگر ملاک‌های بازشناسی باهمایی‌ها را شمّ زبانی اهل زبان و معادل‌های ترجمه‌ای آن‌ها از زبان‌های دیگر می‌داند و معتقد است که معیارهای ذکرشده بر تمام ترکیبات منطبق نیستند (۱۳۷۱).

پی بردن به چگونگی باهمایی واژه‌ها و معانی حاصل از آن‌ها در زبان گفتار و نوشتار در کاربردهای پردازش زبان طبیعی بسیار دارای اهمیت است. در زبان فارسی پژوهش‌هایی که در زمینهٔ باهمایی‌ها و روش پیکره-بنیاد صورت گرفته، عبارت‌اند از: پیکره دو-زبانه در حوزهٔ آموزش زبان فارسی (Harati Mokhtari, Ghafar Samar & Kiany 2016)، بررسی مقابله‌ای باهمایی‌های فارسی و انگلیسی (جوادی ۱۳۸۵)، بررسی باهمایی واژگانی در زبان فارسی و فرانسوی (چوپان ۱۳۹۰)، رفع ابهام معنایی (Riahi and Sedghi 2016) استخراج آماری باهمایی‌ها از پیکره دو-زبانه (Atui, Faili & Atuie 2012)، و بررسی باهمایی در زبان فارسی به روش آماری (مدرس خیابانی ۱۳۸۶).

یکی از روش‌های رایج پیاده‌سازی پروژه‌های پردازش زبان طبیعی در سال‌های اخیر، روش یادگیری ماشین مانند شبکه‌های عصبی و یادگیری عمیق است. انواع بردار نمایش معنای کلمات داخل متن که جدیدترین و پرکاربردترین آن‌ها تعبیه کلمات است، به‌عنوان داده‌های ورودی این روش‌ها مورد استفاده قرار می‌گیرند. کارهایی که با استفاده از یادگیری ماشین در زبان‌های دیگر در زمینهٔ باهمایی‌ها صورت گرفته، عبارت‌اند از: ایجاد پایگاه داده‌های باهمایی‌ها (Khokhlova 2020)، استخراج خودکار باهمایی‌ها با روش یادگیری عمیق (Kordoni 2017)، ساختن پیکرهٔ موازی از باهمایی‌های چندزبانه (Fisas et al. 2020) با استفاده از ویژگی‌های نحوی مانند برچسب نحوی کلمات، نظریه‌های معنایی مانند نظریه معنا – متن، بازیابی و دسته‌بندی باهمایی‌های دارای برچسب معنایی (LF) از پیکره (Anke and Codina-Filbá 2021) که در این پژوهش مدل‌سازی زبانی با روش یادگیری ماشین بدون نظارت[1] و برای دسته‌بندی و استخراج باهمایی از روش بانظارت[2] استفاده می‌شود.

۳. روش پیشنهادی

برای این پژوهش از یکی از پیکره‌های موجود در زبان فارسی یعنی پایگاه دادگان زبان فارسی، بخش معاصر[3] آن (Assi 2019) استفاده شده است. کلمات، ابتدا با استفاده از بردار

1. unsupervised machine learning
2. supervised

۳. در ادامه، به‌طور اختصار از پیکره زبان فارسی معاصر استفاده می‌شود.

تعبیه کلماتِ از پیش آموزش‌دیدهٔ fasttext، به بردارهای عددی ۳۰۰ بعدی تبدیل شدند و سپس با استفاده از این بردارها و شبکه حافظه کوتاه‌مدت ماندگار[۲] (که در ادامهٔ متن در برخی موارد اختصار آن، LSTM به‌کار می‌رود) مدل‌سازی زبانی انجام شد. با استفاده از تنظیم دقیق[۳]، وجود مجموعه داده باهمایی‌ها در مدل زبانی «پارس‌برت»[۴] نیز مورد بررسی قرار گرفته است. در آخرین بخش نیز دقت موتور ترجمه گوگل[۵] برای باهمایی‌های زبان فارسی، به روش مقابله‌ای بررسی می‌شود.

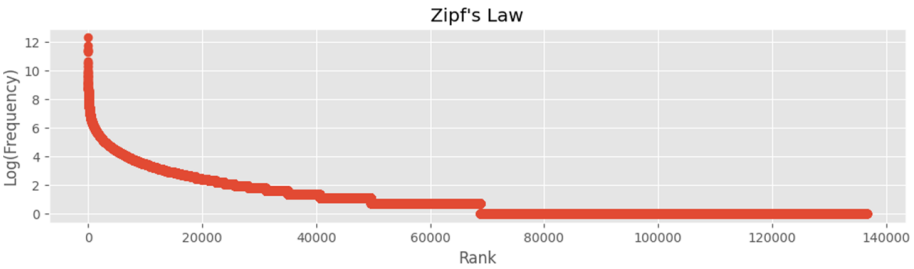
۳-۱. پیکره

برای مدل‌سازی زبانی با شبکه LSTM از پیکره زبان فارسی معاصر استفاده شده است. این پیکره شامل متون معاصر از آثار شعری، داستانی، غیرداستانی، نمایشنامه و فیلمنامه، ادبیات کودکان، نشریه‌های ادواری و مجلات علمی، تخصصی و ادبی می‌شود. حجم این پیکره ۵,۴۶۸,۹۳۲ کلمه است. برای اینکه از سلامت محتوای پیکره اطمینان حاصل شود، منحنی قانون «زیف»[۶] به‌دست آورده شد. بر اساس قانون «زیف»، فراوانی[۷] هر کلمه با رتبه[۸] آن رابطه معکوس دارد؛ یعنی کلمات در یک پیکره با فراوانی زیاد، رتبه یک و کلمات با فراوانی کم، رتبه بیشتر را خواهند داشت. بنابراین، اگر فراوانی را با f و رتبه را با r نشان دهیم، خواهیم داشت:

$$f \propto \frac{1}{r} \rightarrow f.r = k$$

فرمول (۱)

در این فرمول، K به‌طور تقریبی عددی ثابت خواهد بود. برای نشان دادن بهتر این توزیع از لگاریتم این رابطه استفاده می‌شود. شکل ۲، نمودار توزیع فراوانی کلمات[۹] بر اساس رتبه آن‌ها در قانون «زیف» را نشان می‌دهد.



شکل ۲. نمودار قانون «زیف» پیکره زبان فارسی

۳-۲. مجموعه داده‌ی باهمایی‌ها

برای انجام این پژوهش، به‌دلیل نبودِ فرهنگ واژگان باهمایی‌ها در زبان فارسی و نیاز به داشتن لیستی از برخی باهمایی‌ها، مجموعه داده‌ای از باهمایی‌ها به‌تعداد حدوداً ۷۳۰۰ باهمایی با استفاده از کتاب فرهنگ زبان‌آموز پیشرفته فارسی (عاصی ۱۴۰۱)، ایجاد شد تا بتوان از این مجموعه داده در بررسی باهمایی آن‌ها در مدل‌های زبانی جدید استفاده نمود. این تعداد باهمایی‌ها بر اساس تعاریف و ساختار باهمایی‌هایِ معنایی، به‌صورت دستی انتخاب شده است. بعضی از باهمایی‌های دادگان عبارت‌اند از: «پارتی کلفت، چراغ سبز، سازمان صدا و سیما، نازک نارنجی، بوق سگ، خرچنگ قورباغه، دست کج، آب مروارید».

۳-۳. تعبیه کلمات

برای استفاده از متن به‌عنوان ورودی شبکه عصبی در مدل‌سازی‌های زبانی باید کلمات متن به بردارهای عددی تبدیل شود. برای این‌کار با استفاده از روش‌های پیش‌آموزش‌دیده[1]، مجموعه‌های بزرگ کلمات تولید شده است (Vechtomova & Vineet 2017). به‌عنوان مثال، Word2vec بر روی دادگان Google News شامل ۱۰۰ میلیارد کلمه، GloVe بر روی دادگانی متشکل از ۶ میلیارد کلمه و fastText بر روی یک دادگان ۱۶ میلیارد کلمه‌ای آموزش داده شده‌اند. به این ترتیب، این مدل‌ها کلمات زیادی را پوشش داده و برای هر کلمه بردار مناسب ایجاد می‌کنند. بنابراین، word2vec دارای ۳ میلیون کلمه، GloVe دارای ۴۰۰ هزار کلمه و fastText دارای ۱ میلیون کلمه است (Dehghani 1399).

کتابخانهٔ fasttext را «بایونوفکسی»[2] در سال ۲۰۱۶ در آزمایشگاه تحقیقاتی هوش مصنوعی«فیسبوک»[3] توسعه داد. این کتابخانه روشی برای ایجاد بردارهای تعبیه مستقل از متن است که در این تعبیه برای هر کلمه، مستقل از متنی که در آن قرار دارد، بُرداری یکسان وجود دارد. دو مزیت مهم fasttext عبارت‌اند از: ۱) توجه به چندتایی‌های[4] نویسه‌ای[5] در یافتن بردار تعبیه که برای حل مشکل کلمات خارج از دایره لغات (مثلاً غلط‌های املایی) کمک می‌کند؛ و ۲) سرعت بسیار بالا در آموزش و به‌خصوص استنتاج نتایج (اخوان ۱۳۹۸).

1. pre-trained
2. Piotr Bojanowski
3. Facebook
4. N-gram
5. Character

۳-٤. شبکه عصبی حافظهٔ کوتاه‌مدت ماندگار

شبکه عصبی حافظه کوتاه‌مدت ماندگار نوع خاصی از شبکه عصبی بازگشتی است که مشکل حافظهٔ بلندمدت شبکهٔ عصبی بازگشتی[1] را حل می‌کند. شبکهٔ عصبی حافظه کوتاه‌مدت ماندگار از سازوکارهای داخلی به اسم دروازه[2] استفاده می‌کند. این دروازه‌ها جریان اطلاعات را کنترل می‌کنند؛ همین‌طور مشخص می‌کنند چه داده‌هایی در توالی مهم هستند و باید همچنان حفظ بشوند و چه داده‌هایی باید حذف شوند. به این شکل، شبکهٔ اطلاعات مهم را در طول زنجیرهٔ توالی عبور می‌دهد تا خروجی مد نظر را داشته باشد (Ma et al. 2019). شبکه LSTM چندین عملیات مختلف در ساختار داخلی خود دارد. مفهوم این شبکه همان حالت سلول[3] و دروازه‌های همراهش است. در واقع، حالت سلول مانند آزادراهی عمل می‌کند که اطلاعات را در طول زنجیرهٔ توالی جلو می‌برد. می‌توانیم به حالت سلول به‌عنوان حافظه شبکه نگاه کنیم. دروازه‌ها اطلاعات را در حالت سلول به‌روز نگه می‌دارند. این دروازه‌ها شبکه‌های عصبی مختلفی هستند که تصمیم می‌گیرند چه اطلاعاتی به حالت سلول وارد بشوند. این دروازه‌ها در طول آموزش شبکه یاد می‌گیرند چه اطلاعاتی باید حفظ یا فراموش شود (Oquab et al. 2014).

۳-۵. پارس‌برت

افزایش مدل‌های زبانیِ از پیش آموزش‌دیده با هدف ساخت مدل‌های غنی زبانی سبب شده است که فصل نوینی در زمینهٔ پردازش زبان طبیعی آغاز شود. در میان این مدل‌ها، مدل‌های مبتنی بر انتقال‌دهنده[4]، مانند «برت»[5] که توسط Devlin et al. (2019) معرفی شده، به‌دلیل عملکرد عالی محبوبیت بیشتری پیدا کرده است. با این حال، این مدل‌ها به‌طور معمول، بر روی زبان انگلیسی متمرکز بوده و زبان‌های دیگری که منابع محدودی دارند، از مدل‌های چندزبانه بهره می‌برند. «پارس‌برت» یک مدل «برت» تک‌زبانه برای زبان فارسی است که عملکرد بهتری در مقایسه با معماری‌ها و مدل‌های چندزبانه از خود نشان داده است. با وجود مجموعه داده‌های بسیار محدود در زبان فارسی در حوزهٔ پردازش زبان طبیعی، یک پیکرهٔ زبانی پرحجم با بیش از ۳ میلیارد واژه که از منابع مختلف جمع‌آوری‌شده، برای آموزش این مدل استفاده شده است (Farahani et al. 2020).

1. recurrent neural network (RNN)
2. gate
3. cell state
4. transformer
5. Bert

مدل «برت» یک مدل زبانی یادگیری عمیق است. این مدل و مدل‌های مشابه آن، مانند «المو»[1] (Peters et al. 2018)، «جی‌پی‌تی»[2] (Radford and Narasimhan 2018)، مدل‌های تعبیه‌سازیِ مبتنی بر بافتِ از پیش آموزش‌دیده هستند. در این مدل‌ها، تعبیهٔ واژه‌ها با توجه به جمله‌ای که در آن قرار گرفته مشخص می‌شود. بنابراین، بازنمایی‌های متفاوتی از یک واژه با توجه به بافت جایگاهی واژه در جمله‌های متفاوت به‌دست خواهد آمد. ازاین‌رو، ویژگی ابهام واژگانی چالشی برای این‌گونه بردارسازی نخواهد بود؛ چرا که هر واژه با توجه به بافت شامل بردارهای مختلف است؛ درحالی‌که در شیوهٔ بردارسازی با word2vec هر واژه فقط یک بردار دارد و ابهام واژگانی در بردار واژه همچنان مستتر است. مدل‌های تعبیه‌ساز مبتنی بر بافت برای وظیفه‌هایی مانند مدل‌سازی زبان بر روی مجموعه داده‌های متنی کلان، آموزش داده می‌شوند. غنی بودن مدل «برت» به این است که از دوازده لایهٔ کدگذار انتقال‌دهنده ساخته شده است. این مدل با دو تابع مختلف آموزش می‌بیند که عبارت است از تهیهٔ مدل زبانی پوشیده و پیش‌بینی جملهٔ بعدی. مدل «برت» با استفاده از انتقال‌دهنده توانسته است محدودیت ترتیب پردازش دنباله‌های ورودی را از سر راه بردارد. همچنین، استفاده از انتقال‌دهنده سرعت آموزش مدل را نسبت به سایر مدل‌ها افزایش داده و امکان آموزش الگوریتم با پیکره‌های بزرگ‌تر را میسر ساخته است (Devlin et al. 2019).

در تابع مدل زبانی پوشیده تعدادی از واحدهای زبانی در جملهٔ ورودی با واحد [Mask] پوشانده شده و سپس توسط مدل حدس زده می‌شود. در تابع پیش‌بینی جملهٔ بعدی، دو جمله الحاق‌شده با واحد[SEP][3] که بیانگر مرزنمای جمله است، به‌عنوان ورودی به مدل داده می‌شود و سپس، این مسئله که کدام جمله می‌تواند ادامهٔ جمله اول باشد، توسط مدل مشخص می‌شود. کارکردهای مختلف این مدل در پردازش زبان طبیعی سبب می‌شود که مقدار دقیق وزن‌های مدل تنظیم شود.

٤. پیاده‌سازی و نتایج

این بخش شامل بازشناسی باهمایی‌ها، پیاده‌سازی روش‌های آن، و دقت آن‌هاست.

1. embeddings from language models (ELMO)
2. generative pre-trained transformer (GPT)
3. SEPerator (SEP)

٤-١. روش‌های بازشناسی باهمایی‌ها

یکی از روش‌های رایانشی بازشناسی باهمایی‌ها انواع روش‌های آماری است. در این پژوهش روش‌های آماری مانند اطلاعات متقابل نقطه‌به‌نقطه[1]، آزمون خی دو[2] و آزمون تی[3] برای استخراج باهمایی‌ها از پیکره به‌منظور ایجاد مجموعه دادگان آزمون[4] برای مدل زبانی استفاده شده است. روش دیگری که برای بازشناسی باهمایی‌ها وجود دارد، استفاده از آموزش مدل زبانی با استفاده از شبکه‌های عصبی عمیق است. برای این منظور از مجموع داده‌های باهمایی‌هایی که دارای برچسب معنایی بر اساس نظریه معناشناسی معنا-متن است، به‌عنوان ورودی مدل زبانی آموزش‌دیده استفاده می‌شود. مدل زبانی با روش بدون ناظر آموزش می‌بیند و با روش با ناظر به پیش‌بینی برچسب معنایی می‌پردازد. از آنجا که این نظریه معناشناسی در زبان فارسی پیاده‌سازی نشده، نمی‌توان به پیش‌بینی برچسب معنای باهمایی‌ها پرداخت. در این پژوهش بعد از آموزش مدل زبانی برای آزمودن بازشناسی باهمایی مدل زبانی، از دو گروه داده‌های باهمایی‌ها و ناباهمایی‌ها به‌عنوان داده‌های آزمون استفاده شده است و مقدار احتمال باهم‌آمدن باهمایی‌ها و ناباهمایی‌ها در این مدل محاسبه گردیده است. مدل‌های زبانی در آموزش خود، احتمال باهمایی کلمات را یاد می‌گیرند. این احتمالات بر اساس نوع بردارهای تعبیهٔ آموزشی، تعداد لایه‌های شبکه، میزان حجم پیکره‌ای که شبکه با آن آموزش می‌بیند، محاسبه می‌شود. در مدل‌هایی که دادهٔ آزمون برچسب‌دار وجود دارد، احتمال باهمایی داده‌های آزمون در مدل آموزش‌دیده محاسبه می‌شود و مقدار برچسب آن مورد بررسی قرار می‌گیرد و اگر مقدار احتمال آن‌ها بیشتر از صفر باشد، برچسب آن‌ها پیش‌بینی می‌شود. به‌دلیل اینکه برچسبی برای پیش‌بینی وجود ندارد، احتمال باهم‌آمدن آن‌ها از روش‌های زیر برای مجموعه داده‌های آزمون مورد استفاده واقع شد:

۱. در مرحله اول، یک مدل زبانی با شبکه LSTM بر روی پیکره مورد استفاده در پژوهش، آموزش داده شد و سپس پیش‌بینی باهمایی‌ها و محاسبه احتمال باهم‌آمدن باهمایی‌ها و ناباهمایی‌ها با روش‌های زیر انجام گردید:

الف) با استفاده از مجموعه دادهٔ باهمایی‌ای که از روش آماری استخراج شد، اولین کلمه از باهمایی‌های دو-کلمه‌ای به مدل زبانی ساخته‌شده داده شد تا دومین کلمه از آن را پیش‌بینی کند؛

1. pointwise mutual information (PMI)
2. Pearson's Chi-Square Test
3. T-Score
4. test data

ب) برای هر یک از جفت باهمایی‌های لیست قسمت (الف)، چند جفت ناباهمایی تهیه شد. برای تهیه ناباهمایی از کلمات موجود در پیکره زبانی، دو-تایی‌های پیکره و کلمات نزدیک به یکی از کلمات باهمایی در پایگاه داده زبان فارسی استفاده شده است. احتمال باهم‌آمدن باهمایی‌ها و ناباهمایی‌ها در این مدل محاسبه گردید. همه کلمات این لیست‌ها در پیکره زبانی مورد مطالعه وجود داشت؛

ج) ابتدا از دادگان باهمایی‌ها، هزار باهمایی به‌طور تصادفی انتخاب شد و سپس، لیستی از هزار ناباهمایی به‌عنوان دادگان آزمون نیز تهیه گردید. در مدل زبانی آموزش‌دیده با شبکه LSTM مثبت صحیح[1] هزار باهمایی و هزار ناباهمایی مورد بررسی قرار گرفت؛ یعنی احتمال باهمایی این جفت کلمات در مدل زبانی آموزش‌دیده بررسی شد. در شبکه‌های عصبی عمیق، مثبت صحیح حالتی است که در آن یک مدل به‌درستی یک نمونه مثبت را شناسایی می‌کند. مثبت صحیح نشان‌دهندهٔ تعداد مواردی است که مدل به‌درستی پیش‌بینی کرده است.

۲. در روش دوم، با تنظیم دقیق «پارس‌برت» و با استفاده از دادگان باهمایی و ناباهمایی قسمت «۱/۳» یعنی دو دادهٔ هزارتایی از باهمایی و ناباهمایی به‌عنوان داده آزمون استفاده شد و احتمال باهم‌آمدن این دو گروه داده برای آزمودن این مدل زبانی استفاده شد.

در انتها بازشناسی باهمایی‌ها در ترجمهٔ برخط مورد پژوهش قرار گرفت و دقت مترجم گوگل که سامانه ترجمهٔ برخطی است که در دنیا پرکاربرد است و زبان‌های زیادی را پوشش می‌دهد و دارای داده‌های بسیار است، انتخاب شد. برای این کار از هزار باهمایی‌ای که در قسمت «۱/۳» به‌صورت تصادفی از مجموعه دادهٔ باهمایی ایجاد شده[2] بود، مورد استفاده قرار گرفت.

۴-۲. پیاده‌سازی روش‌های بازشناسی باهمایی‌ها

برای ساخت داده‌های آزمون با استفاده از روش‌های آماری نام‌برده‌شده در بخش ۴-۱ باهمایی‌های دو-تایی از پیکره مورد استفاده پژوهش استخراج شدند و از بین برگزیدگان آن‌ها، ۱۳۵ باهمایی برای ورودی این مدل آموزش‌دیده انتخاب گردید. همچنین برای ساخت مدل زبانی با شبکه LSTM، از دادگان زبان فارسی بخش ۳ که پیکره مورد استفاده

1. TP - True Positive

۲. از آن‌ها برای دقت مدل‌های زبانی استفاده شده است.

در این پژوهش است، پس از پیش‌پردازش آن، با استفاده از بردارهای تعبیه fasttext، متن پیکره را به بردارهای مربوط تبدیل کرده و این بردارهای ۳۰۰ بُعدی ورودی شبکهٔ عصبی عمیق هستند. در ادامه، شبکه‌ای با دو لایه پنهان[1]، با روش گره[2] ساخته شد. تعریف اشتراک‌گذاری وزن[3] با گره زدن (به‌اشتراک‌گذاری) وزن لایه‌های تعبیه‌شده و softmax عملکرد مدل‌های زبان را بهبود می‌بخشد. این روش همچنین تعداد کل پارامترها را در مدل‌های زبانی که روی آن اعمال می‌شود، کاهش می‌دهد (Press & Wolf 2016). همان‌طور که در بخش ۳-۵ توضیح داده شده است، مدل زبانی «پارس‌برت» با استفاده از بیش از ۳ میلیارد واژه با استفاده از شبکه عصبی «برت» آموزش دیده است. در این پژوهش با استفاده از تنظیم دقیق این مدل آموزش‌دیده برای بازشناسی باهمایی‌ها استفاده شده است. برای دقت ترجمه موتور ترجمه گوگل طبق توضیحات داده‌شده در بخش ۴-۱، هر یک از هزار باهمایی در یک جمله فارسی مورد استفاده قرار گرفته است. هر جمله با موتور ترجمه گوگل به انگلیسی برگردانده شده و سپس ترجمه باهمایی‌ها در جملات انگلیسی توسط افراد خبره بررسی گردید.

٤-۳. شاخص ارزیابی

یکی از شاخص‌های ارزیابی مدل‌ها، فراخوانی[4] است که نشان‌دهندهٔ توانایی یک مدل در شناسایی نمونه‌های مثبت واقعی است. این شاخص به‌صورت نسبت تعداد مثبت‌های صحیح به مجموع تعداد مثبت‌های واقعی محاسبه می‌شود. فرمول ۲، محاسبهٔ معیار فراخوانی را نشان می‌دهد. در روش‌های آماری نیز این شاخص برای محاسبه تعداد نمونه‌های مثبت شناسایی‌شده کاربرد دارد.

$$Recall = \frac{TP}{TP+FN}$$ **فرمول (۲)**

در این پژوهش، طبق آنچه در بخش ۴-۱ آمده است، شاخص ارزیابی بازشناسی باهمایی‌ها در مدل‌های زبانی آموزش‌دیده، به‌شرح زیر است:

۱. فراخوانیِ پیش‌بینی ۱۳۵ باهمایی و ناباهمایی‌ای که با روش آماری از پیکره مد نظر استخراج شده در مدل زبانی آموزش‌دیده با شبکه LSTM؛
۲. محاسبه احتمال باهم‌آمدن تعدادی باهمایی و ناباهمایی در مدل زبانی آموزش‌دیده با شبکه عصبی LSTM؛

1. hidden layers 2. tie 3. weight tying 4. recall

۳. فراخوانیِ پیش‌بینی هزار باهمایی و ناباهمایی‌ای که از دادگان باهمایی انتخاب شده در دو مدل زبانی آموزش‌دیده با شبکه عصبی LSTM و «پارس‌برت».

در بخش آخر، دقت بازشناسی باهمایی‌ها در مترجم گوگل نیز با فرمول فراخوانی محاسبه شده است.

٤-٤. دقت روش‌های بازشناسی باهمایی‌ها

در بخش ۴-۱ روش بازشناسی باهمایی، در بخش ۴-۲ چگونگی پیاده‌سازی این روش‌ها و بخش ۴-۳ شاخص ارزیابی بیان شد. در این بخش دقت روش‌های پیاده‌سازی‌شده بیان می‌شود و به‌تفکیک به‌شرح زیر است:

۱. روش مدل زبانی آموزش‌دیده با شبکه LSTM: همان‌طور که در بخش ۴-۱ بیان شده، با سه روش مختلف به بازشناسی باهمایی پرداخته شده است.

۱-۱. فراخوانیِ پیش‌بینی ۱۳۵ باهمایی و ناباهمایی‌ای که اولین کلمه از باهمایی‌های دو-کلمه‌ای به مدل زبانی آموزش‌دیده با شبکه LSTM داده می‌شود که دومین کلمه از آن را پیش‌بینی کند. در این روش از ۱۳۵ باهمایی‌ای، ۹ باهمایی پیش‌بینی شد و مقدار فراخوانی آن ۶/۶ درصد است؛

۱-۲. پس‌محاسبه احتمال باهم‌آمدن، تعدادی باهمایی و ناباهمایی در مدل زبانی آموزش‌دیده با LSTM مقادیر عددی احتمال به‌دست‌آمده را می‌توان به چهار دسته تقسیم کرد که در جدول ۱، مثال‌هایی از خروجی احتمال باهم‌آمدن‌ها مشاهده می‌شود که در این جدول باهمایی‌ها و ناباهمایی‌ها مشخص شده است و ردیف‌های همرنگ شامل یک باهمایی و چند ناباهمایی است. در ادامه، چهار دسته مورد نظر مشاهده می‌شود.

۱-۲-۱. بخش اول

جدول ۱. احتمالات باهم‌آمدن

نوع باهم‌آمدن	باهمایی	احتمال
باهمایی	خدای نخواسته	3.2877e-08
ناباهمایی	خدای مهربان	7.6783e-09
ناباهمایی	خدای عادل	5.1998e-09
ناباهمایی	خدای بصیر	5.1143e-09
ناباهمایی	خدای رحمان	5.1138e-09
باهمایی	سازمان ملل	0.0003
ناباهمایی	سازمان داخلی	2.2966e-05
ناباهمایی	سازمان خارجی	2.5925e-05
ناباهمایی	سازمان بشر	2.1630e-05
ناباهمایی	سازمان انقلابی	2.1296e-05

۱-۲-۱. بخش اول- احتمال باهمایی کلمات باهمایی با اختلاف بسیاری از ناباهمایی‌ها قرار داشت. برای مثال، در جدول ۲، «بارک‌الله» یک باهمایی در زبان فارسی است که در معنای تشویق کردن است، «الله»، «تعالی»، «دیده» اسم و «مبارک» اسم پسر هستند. همان‌طور که مشاهده می‌شود، احتمال باهمایی ترکیب «بارک» با این کلمات با وجود اینکه باهمایی نحوی هستند و در جدول ۲، ردیف دوم به بعد آمده است، صفر و احتمال باهمایی کلمه «بارک» و «الله» که باهمایی هستند در این مدل زبانی آموزش‌دیده ۱۰۰ درصد شده است.

جدول ۲. احتمال باهم‌آمدن باهمایی و ناباهمایی، بخش اول

نوع باهم‌آمدن	باهمایی	احتمال
باهمایی	بارک الله	1.
ناباهمایی	بارک تعالی	0.
ناباهمایی	بارک مبارک	0.
ناباهمایی	بارک خط	0.
ناباهمایی	بارک دیده	0.

در جدول ۳، ردیف اول «قتل عام»، ترکیب اضافی است که باهمایی است. احتمال باهمایی ترکیبات اضافی دیگری مانند «قتل کودکان»، «قتل مردان»، «قتل زنان»، «قتل درختان»، «قتل خون» بررسی شد که در این مدل آموزش‌دیده احتمال باهمایی آن‌ها صفر به‌دست آمد، ولی احتمال باهمایی واژه «قتل» و واژه «عام» مقدار $۱/۴۰۱۳\times۱۰^{-۴۵}$ محاسبه شد.

جدول ۳. احتمال باهم‌آمدن باهمایی و ناباهمایی، بخش اول

نوع باهم‌آمدن	باهمایی	احتمال
باهمایی	قتل عام	1.4013e-45
ناباهمایی	قتل کودکان	0.
ناباهمایی	قتل مردان	0.
ناباهمایی	قتل زنان	0.
ناباهمایی	قتل درختان	0.

۱-۲-۲. بخش دوم- احتمال باهمایی کلمات باهمایی با اختلاف کمی از ناباهمایی‌ها قرار داشت. برای مثال، «مواد مخدر» که یک باهمایی است که از ترکیب اسم و صفت درست شده است، ولی معنای حاصل از آن از ترکیب معنای کلمه اول و دوم حاصل نمی‌شود. معنای واژه مخدر در فارسی «سست‌کننده» است، ولی مواد مخدر در فارسی به معنای «مواد روانگردانی که اعتیادآور است» استفاده می‌شود. ترکیبات ردیف دوم به بعد «مواد مقوی»، «مواد مختلف»، «مواد تازه»، «مواد کهنه» و «مواد سوختی» ترکیبات وصفی است که معنای ترکیب از روی معنای ترکیب واژگان ساخته می‌شود. احتمال باهمایی دو واژه «مواد» و «مخدر» بر اساس جدول ۴، مقدار $۶/۵۳۶۸\times۱۰^{-۵}$ و احتمال باهمایی ترکیبات دیگر ردیف‌های بعدی کمتر از ترکیب باهمایی بود و اختلاف کمی با آن داشتند.

جدول ٤. احتمال باهم‌آمدن باهمایی و ناباهمایی، بخش دوم

نوع باهم‌آمدن	باهمایی	احتمال
باهمایی	مواد مخدر	6.5368e-05
ناباهمایی	مواد مقوی	2.0098e-05
ناباهمایی	مواد مختلف	3.2326e-05
ناباهمایی	مواد تازه	4.5463e-05
ناباهمایی	مواد کهنه	2.6657e-05

۱-۲-۳. بخش سوم- احتمال باهمایی‌ها و ناباهمایی‌ها صفر شده است (که احتمالاً یا بردار تعبیه آن در مدل fasttext وجود نداشت، یا مدل زبانی نیازمندآموزش در حجم گسترده‌تری بود). برای مثال، «مرغ ماهیخوار» یک ترکیب وصفی و باهمایی است، و ترکیبات وصفی دیگر «مرغ چرب»، «مرغ نامرغوب»، «مرغ پرکنده» و «مرغ پخته» ناباهمایی است. زیرا ترکیبات وصفی را می‌توان گسترش داد و یا ترکیبات وصفی دیگری با صفت آن ساخت، ولی مرغ ماهیخوار این‌گونه نیست. برای مثال، به‌جای «مرغ چرب» می‌توان «ماهی چرب» ساخت یا «برنج‌های نامرغوب» یا «کودک فربه»؛ اما نمی‌توان ترکیبی شبیه «اسب ماهیخوار» یا «کودک ماهیخوار» ساخت.

جدول ٥. احتمال باهم‌آمدن باهمایی و ناباهمایی، بخش سوم

نوع باهم‌آمدن	باهمایی	احتمال
باهمایی	مرغ ماهیخوار	0.
ناباهمایی	مرغ چرب	0.
ناباهمایی	مرغ نامرغوب	0.
ناباهمایی	مرغ پرکنده	0.
ناباهمایی	مرغ پخته	0.

۱-۲-۴. بخش چهارم- تنها دو مورد یکی از ناباهمایی‌ها از خود باهمایی دارای احتمال بیشتری بود. در مثال زیر، «سجده شکر» یک سجده مستحب است و اسم یکی از اعمال دینی اسلام است و ترکیب دیگری مانند «قنوت

شکر» یا «رکوع شکر» نمی‌توان ساخت. با این حال «سجده بردی» که فعل «بردی» در ادبیات قدیم به معنای انجام دادن کاری است و معنای این ترکیب به معنای انجام دادن فعل سجده است. احتمال باهمایی این ترکیب در مدل زبانی آموزش‌دیده در جدول ۶ با مقدار $۱/۰۶۹۱\times۱۰^{-۹}$ از احتمال باهم‌آمدن باهمایی «سجده شکر» با مقدار $۹/۶۸۶۷\times۱۰^{-۱۰}$ بیشتر بود، اما باقی ترکیبات ناباهمایی، احتمال باهمایی کمتری از باهمایی داشتند.

جدول ٦. احتمال باهم‌آمدن باهمایی و ناباهمایی، بخش چهارم

نوع باهم‌آمدن	باهمایی	احتمال
باهمایی	سجده شکر	9.6867e-10
ناباهمایی	سجده بردی	1.0691e-09
ناباهمایی	سجده ابلیس	9.5388e-10
ناباهمایی	سجده نماز	9.5594e-10
ناباهمایی	سجده نیایش	9.5378e-10
ناباهمایی	سجده ترس	1.0150e-09

طبق آنچه بیان شد و محاسبات انجام‌شده، مثال‌هایی از چهار دسته بالا در پیوست ۱، آمده است.

۱-۳. فراخوانیِ پیش‌بینی هزار باهمایی و ناباهمایی در مدل زبانی آموزش‌دیده با شبکه LSTM که از هزار باهمایی، ۲ باهمایی پیش‌بینی شده است و فراخوانی آن ۰/۰۲ درصد است. از هزار ناباهمایی، هیچ‌یک پیش‌بینی نشده است و فراخوانی پیش‌بینی ناباهمایی‌ها در مدل زبانی آموزش‌دیده ۰ درصد است.

۲. روش تنظیم دقیق پارس‌برت»: در بخش ۴-۱ توضیحات کامل بیان شده است. فراخوانی باهمایی‌هایی که در این مدل مورد آزمایش قرار گرفتند، ۹۵/۹۳ درصد بود؛ یعنی ۹۵/۹۳ درصد از هزار باهمایی در مدل تنظیم دقیق‌شده با بردارهای «پارس‌برت» پیش‌بینی شد. فراخوانی هزار باهمایی که باهمایی نیستند، ۸۵/۸ درصد بود. بنابراین، مدل زبانی تنظیم دقیق‌شده با بردارهای «پارس‌برت»، ۸۵/۸ درصد از دادگان آزمون هزار ناباهمایی را پیش‌بینی کرد.

۳. نحوه بازشناسی باهمایی‌ها در ترجمه برخط مترجم گوگل در بخش‌های ۴-۱ و ۴-۲ بیان شده است. از ترجمه هزار باهمایی در هزار جمله فارسی به انگلیسی در مترجم گوگل، در

۵۱۰ جمله باهمایی به‌درستی بازشناسی شد و ترجمهٔ درستی از آن صورت پذیرفت. بنابراین، فراخوانی بازشناسی درست باهمایی و ترجمه آن در هزار جمله فارسی به انگلیسی در مترجم گوگل ۵۱ درصد است. نتایج این بررسی‌ها به‌صورت زیر است:

۳-۱. باهمایی به‌درستی بازشناسی شد. ترجمه صحیحی از آن در جمله انگلیسی صورت گرفته که در جدول ۷، مثال‌هایی از ترجمه درست باهمایی بیان شده است. در این جدول باهمایی، ترجمه جمله فارسی دارای باهمایی و ترجمه انگلیسی که توسط مترجم گوگل صورت گرفته، مشاهده می‌شود؛

۳-۲. باهمایی به‌درستی بازشناسی نشد. ترجمه‌ای تحت‌الفظی و کلمه‌به‌کلمه از آن به جمله انگلیسی صورت پذیرفت. در جدول ۸، مثال‌هایی از عدم بازشناسی باهمایی دیده می‌شود. این جدول شامل باهمایی، ترجمه آن، جمله فارسی دارای باهمایی و ترجمه آن به انگلیسی با مترجم گوگل است؛

۳-۳. باهمایی به‌درستی بازشناسی نشد. از ترجمه همه یا بخشی از عبارت در جملات انگلیسی چشم‌پوشی می‌شود. در جدول ۹، مثال‌هایی برای عدم بازشناسی باهمایی وجود دارد. در این جدول مانند جدول‌های ۷ و ۸، باهمایی، ترجمه آن، جمله فارسی دارای باهمایی و ترجمه آن با مترجم گوگل به انگلیسی وجود دارد.

جدول ۷. ترجمه درست باهمایی‌ها از فارسی به انگلیسی در مترجم گوگل

باهمایی	ترجمه باهمایی	جمله فارسی	ترجمه جمله به انگلیسی
سرفه سیاه	Whooping cough	طبق گزارش‌های مرکز کنترل و پیشگیری از بیماری‌ها (CDC) هر ساله ۱۰ تا ۴۰ هزار مورد سرفه سیاه گزارش می‌شود.	According to the Center for Disease Control and Prevention (CDC), 10,000 to 40,000 cases of whooping cough are reported every year.
دل گرم	Hopeful/ cheer up	هیچی مثل تغییر دادن یه نفر دیگه منو دل گرم نمیکنه.	Nothing cheers me up like seeing someone else change.
فلج اطفال	Polio	اولین واکسن فلج اطفال از نوع ویروس غیرفعال بود.	The first polio vaccine was an inactivated type of virus.
پوست کنده	Frankly	اگر مزاحم باشید رک و پوست کنده بهتان می‌گویم.	If you are a nuisance, I will tell you frankly.
آفتاب خورده	Sunburned	صورتش آفتاب‌خورده و تیره شد.	His face was sunburned and dark.

جدول ۸. ترجمه نادرست باهمایی‌ها از فارسی به انگلیسی در مترجم گوگل

باهمایی	ترجمه باهمایی	جمله فارسی	ترجمه جمله به انگلیسی
ترگل ورگل	/ smart/ groomed	سه دوشیزه خانم ترگل و ورگل را نشان داد.	He showed the three young ladies, Tergel and Vergel.
بازار گرمی	Sales talk	و بی‌آنکه احتیاجی به بازار گرمی باشد مردم آن را می‌قاپند، ملاحظه بفرمایید.	And without the need for a hot market, people grab it, please note.
بالای چشم	with pleasure	به به، استاد، خوش آمدید، قدمتان بالای چشم!	Bye bye, teacher, welcome, your step above the eye!
آهنگ ملایم	Mild music/ relaxing music	یک آهنگ ملایم گذاشتم و چراغ خواب قرمزم را روشن کردم	I put on a soft song and turned on my red night light
ته تغاری	The last child	این فسقلی دختر ته‌تغاری و آخرین بچه من هستند.	This Fasqali is Te Taghari's daughter and my last child.

جدول ۹. ترجمه نادرست و نادیده‌گرفتن باهمایی‌ها از فارسی به انگلیسی در مترجم گوگل

باهمایی	ترجمه باهمایی	جمله فارسی	ترجمه جمله به انگلیسی
خلق تنگ	Bad-tempered	هر روز صبح با خلق تنگ متوجه میشه که شب پیش چه حماقت‌هایی از او سر زده است.	Every morning, he realizes what stupid things he has done the night before.
نخود آش	inquisitive/ nosy	وقتی جورج نخود آش بشه یعنی عقلش را از دست داده.	When George becomes a pea, it means he has lost his mind.
قند حبه‌ای	Sugar cube	من به تجارت علاقه دارم؛ و به همین دلیل قراردادهایی با کشورهای همسایه امضا کرده‌ام که به ما قند حبه‌ای بفروشند.	I am interested in business; And for this reason, I have signed contracts with neighboring countries to sell us granulated sugar.
غیبت کبری	Long absence	جدایی این دو مکتب به دوران پس از غیبت کبری بازمی‌گردد.	The separation of these two schools will also return after the occultation.
زانوی غم	Sorrow	چرا زانوی غم بغل کردی؟	Why did you hug the knees?

۵. نتیجه‌گیری

پژوهش‌های صورت‌گرفته در زبان فارسی در مبحث باهمایی‌ها عمدتاً مقابله‌ای، نظری، و کاربردهای خاص آن به‌صورت آماری بوده است که در بخش ۲، توضیحات آن بیان شد. مهم‌ترین و جامع‌ترین پژوهشی که در زمینهٔ باهمایی‌ها صورت گرفته بود، رسالهٔ دکتری

«مدرس خیابانی» است که در آن به بررسی باهمایی و تعاریف مختلف آن از زبان‌شناسان ایرانی و غیرایرانی پرداخته می‌شود و در انتها با چند روش آماری به استخراج باهمایی‌ها در یکی از پیکره‌های زبان فارسی می‌پردازد.

روش‌هایی که در بررسی بازشناسی باهمایی در این مقاله انجام شد، با پژوهش‌های پیشین زبان فارسی و غیرفارسی متفاوت است. در این مقاله، دادگانی از باهمایی‌ها تهیه و ایجاد می‌شود که در زبان فارسی وجود نداشته است. همچنین در پژوهش‌های صورت‌گرفته در زبان‌های دیگر برای بازشناسی باهمایی‌ها، به آموزش مدل زبانی با استفاده از شبکه‌های عصبی عمیق بدون نظارت و سپس با استفاده از دادگان باهمایی‌هایی که دارای برچسب معنایی هستند، به پیش‌بینی برچسب باهمایی‌ها به‌روش باناظر پرداخته می‌شود.

آنچه در این مقاله باعث شد روش اجرایی آن همانند پژوهش‌های مشابه که در دنیا انجام می‌شود صورت نگیرد، عدم پیاده‌سازی نظریه معنا-متن و فقدان برچسب‌های معنایی برای باهمایی‌های زبان فارسی است. در این مقاله برای بررسی بازشناسی باهمایی، مدل زبانی با شبکه LSTM آموزش، و دیگری با مدل زبانی «پارس‌برت» تنظیم دقیق انجام شد. در شبکه LSTM چون حجم پیکره آموزشی و بردارهای تعبیه کمتر از شبکه «پارس‌برت» است و همچنین ساختارهای تعبیه در شبکه «برت» به‌گونه‌ای است که چالش ابهام واژگانی در آن وجود ندارد. پس، مدل زبانی قوی‌تری است و پیش‌بینی و فراخوانی باهمایی‌ها و ناباهمایی‌ها در آن بسیار بالاتر از شبکه آموزش‌دیده با LSTM است. در LSTM ۶/۶ درصد باهمایی‌ها و ۰ درصد از ناباهمایی‌ها پیش‌بینی شد. به دلایل بیان‌شده پیش‌بینی باهمایی در «پارس‌برت» ۹۵/۹۳ درصد است که از پیش‌بینی ناباهمایی‌ها به‌مقدار ۸۵/۸ درصد بیشتر است.

آموزش یک مدل زبانی یعنی احتمال وقوع کلمات یک پیکره به‌درستی به شبکه یاد داده شود. از این رو، اگر مدل زبانی درست آموزش ببیند باید احتمال باهمایی کلمات در آن از احتمال باهمایی صرف بالاتر باشد. در بخش بررسی احتمالات در مدل زبانی آموزش‌دیده LSTM، به بررسی احتمال باهم‌آمدن باهمایی‌ها و ناباهمایی‌ها نیز پرداخته شد که در خروجی احتمالات باهمایی‌ها و ناباهمایی‌ها احتمال باهم‌آمدن باهمایی‌ها بالاتر از کلماتی است که به‌صورت ناباهمایی کنار یکدیگر قرار می‌گیرند. البته همان‌طور که مشاهده شد، در موارد کمی نیز هیچ احتمالی برای باهمایی‌ها وجود نداشت که علت آن

کم بودن حجم داده‌های زبانی برای آموزش مدل زبانی در شبکه LSTM یا وجود نداشتن بردار تعبیه آن کلمه است.

از آنچه انجام شد، می‌توان نتیجه گرفت که مدل‌های زبانی مانند «پارس‌برت» بسیار قدرتمند است و در پردازش زبان و چگونگی قرارگیری کلمات کنار یکدیگر تبحر دارد. در ادامه، نتایج به‌صورت مقابله‌ای به‌دقت مترجم گوگل در بازشناسی و ترجمه باهمایی زبان فارسی به انگلیسی پرداخته شد. تاکنون در زبان فارسی برای بازشناسی باهمایی در مترجم‌های برخط پژوهشی صورت نگرفته است. نتایج این بررسی شامل بازشناسی درست باهمایی و ترجمه درست آن، عدم بازشناسی باهمایی و ترجمه کلمه‌به‌کلمه از آن و بازشناسی نادرست باهمایی و ترجمه نادرست و نادیده‌گرفتن باهمایی است. میزان دقت ۵۱ درصد، مقدار قابل قبولی برای مترجم برخط گوگل نیست.

۶. کارهای آتی

به‌دلیل اینکه باهمایی‌ها از نظر معنایی بسیار به یکدیگر وابستگی دارند و گروهی بزرگ هستند که در عین حال، تعریف منسجمی برای آن وجود ندارد و محدودهٔ آن پیوستاری است و همچنین در زبان‌های دیگر در سال‌های اخیر در زمینه استخراج باهمایی‌ها از داده‌های برچسب معنایی‌دار LF به‌عنوان راهنمای بازشناسی نوع باهمایی و استخراج آن‌ها استفاده می‌شود؛ و اینکه بتوان از روش‌های مرسوم سال‌های اخیر در زمینه بازشناسی باهمایی‌ها با استفاده از روش‌های یادگیری عمیق درکاربردهای مختلف پردازش زبان طبیعی استفاده کرد.

پیشنهاد می‌شود نظریه معناشناسی معنا-متن برای زبان فارسی پیاده‌سازی شود تا بتوان داده‌هایی با برچسب معنایی برای باهمایی‌های زبان فارسی تهیه نمود. از این داده‌ها برای بازشناسی باهمایی، پیکره‌های موازی و رفع ابهام معنایی در شبکه معنا نیز می‌توان استفاده کرد. همچنین در زبان فارسی، موتورهای ترجمه برخط وجود دارد. پیشنهاد می‌شود، دقت این مترجم‌های برخط در بازشناسی باهمایی‌ها از فارسی به انگلیسی مورد سنجش قرار گیرد تا بتوان مقایسه‌ای برای بالا بردن کیفیت ترجمه آن‌ها صورت گیرد.

فهرست منابع

اخـوان مهـدوی، رسـول. ۱۳۹۸. ارزیابـی روش‌هـای پـردازش متـن بـر روی داده‌هـای آگهـی دیـوار. بازیابـی از:
https://virgool.io/@rasoulam/%D8%A7%D8%B1%D8%B2%DB%8C%D8%A7%D8%A8%DB%8C-%D8%B1%D9%88%D8%B4%D9%87%D8%A7%DB%8C-%D9%BE%D8%B1%D8%AF%D8%A7%D8%B2%D8%B4-%D9%85%D8%AA%D9%86-%D8%A8%D8%B1-%D8%B1%D9%88%DB%8C-%D8%AF%D8%A7%D8%AF%D9%87%D9%87%D8%A7%DB%8C-%D8%A2%DA%AF%D9%87%DB%8C-%D8%AF%DB%8C%D9%88%D8%A7%D8%B1-clianu3d719w. (دسترسـی در ۴/۱۹/ ۱۴۰۲)

افراشی، آزیتا. ۱۳۷۸. نگاهی به مسئله باهم‌آیی واژگان. *متن‌پژوهی ادبی* ۷-۸ (۲): ۷۳-۸۲.

باطنی، محمدرضا. ۱۳۴۸. *توصیف ساختمان دستوری زبان فارسی*. انتشارات امیرکبیر.

جـوادی، فـروزان. ۱۳۸۵. بررسـی مقابلـه‌ای باهمایی‌هـای زبـان فارسـی و انگلیسـی و معادل‌هـای ترجمـهٔ آن. پایان‌نامـه کارشناسـی ارشـد. دانشـکده ادبیـات و علـوم انسـانی. دانشـگاه تهـران.

چوپـان، سـیما. ۱۳۹۰. بررسـی باهمایـی واژگانـی در زبان‌هـای فارسـی و فرانسـه. پایان‌نامـه کارشناسـی ارشـد. دانشـکده زبان‌هـای خارجـی. دانشـگاه آزاد اسـلامی واحـد تهـران مرکـزی.

عاصـی، مصطفـی. ۱۳۷۱. نقـش ترکیـب در گسـترش واژگان زبـان فارسـی بـا نگرشـی بـر آثـار نظامـی گنجـوی. *فرهنـگ* (۱۰). تهـران: پژوهشـگاه علـوم انسـانی و مطالعـات فرهنگـی.

_____. ۱۴۰۱. *فرهنگ زبان‌آموز پیشرفته فارسی*. تهران: انتشارات سمت.

مـدرس خیابانـی، شـهرام. ۱۳۸۶. بررسـی باهمایـی واژگانـی در زبـان فارسـی. پایان‌نامـه دکتـری. دانشـکده زبان‌شناسـی. دانشـگاه علامـه طباطبایـی.

مصطفـوی، مهدیـه. ۱۳۹۵. باهمایـی دسـتوری در زبـان فارسـی. پایان‌نامـه کارشناسـی ارشـد. دانشـکده ادبیـات و علـوم انسـانی. دانشـگاه الزهرا (س).

References

Anke, Luis Espinosa, Codina-Filbá, & Leo Joan Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Kyiv, Ukraine: Association for Computational Linguistics.

Assi, M. 2019. Persian linguistic database (PLDB). In: Tehran: Institute for Humanities and Cultural Studies. Retrieved from < https://pldb.ihcs.ac.ir/>. (accessed December 17, 2022)

Atui, Kavosh Asadi, Heshaam Faili, & Kaveh Assadi Atuie. 2012. *Collocation extraction using parallel corpus*. Proceedings of COLING 2012: Posters.

Bischof, Beatrice, Klausurtagung Kleinwalsertal. 2004. *The collocation in French*. Retrieved from <http://www.ilg.uni-stuttgart.de/gk/aktivitaeten/dokumente/2004/bischof.pdf> (accessed October 1, 2022)

Church, K. W., W. A. Gale, P. Hanks, & D. Hindle. 1991. *Using Statistics in Lexical Analysis*, in *Lexical Acquisition Exploiting On-line Resources To Build A Lexicon*, U. Zernik, Editor. 1991: Englewood Cliff: ?.

Cowie, A. P. 1981. The Treatment of Collocations and Idioms in Learners' Dictionaries. *Applied Linguistics* II (3). https://doi.org/10.1093/applin/II.3.223

Cruse, D. Alan. 1986. Lexical semantics. Cambridge: Cambridge University Press.

Dehghani, M. 2020. *Embedding*. Retrieved from https://data-hub.ir/word-embedding-%DA%86%DB%8C%D8%B3%D8%AA. (accessed July 08, 2023)

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.18653/v1/n19-1423.

Farahani, Mehrdad, Mohammad Gharachorloo, Marzieh Farahani, & Mohammad Manthouri. 2021. *ParsBERT: Transformer-based Model for Persian Language Understanding***.** *Neural Processing Letters*, *53* (6), 3831–3847. https://doi.org/10.1007/s11063-021-10528-4

Firth, John R. 1957. *Modes of meaning, papers in linguistics*. Oxford: Oxford University Press.

Fisas, Beatríz, Joan Codina-Filbà, Anke Luis Espinosa, & L. Wanner. 2020. *CollFrEn: Rich Bilingual English–French Collocation Resource*. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova, & A. Savary (Eds.), Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (pp. 1–12). Association for Computational Linguistics. [Held online]. https://aclanthology.org/2020.mwe-1.1/

Halliday, Michael Alexander Kirkwood, & Ruqaiya Hasan. 1986. *Cohesion in english*. London: Longman.

Harati Mokhtari, Parastoo, Reza Ghafar Samar, & Gholam Reza Kiany. 2016. Collocational Processing in Two Languages: A Psycholinguistic Comparison of Monolinguals and Bilinguals. *Iranian Journal of English for Academic Purposes* 5 (1): 69-52.

Khokhlova, Maria. 2020. *Quantitative Properties of Russian Adjective-Noun Collocations across Dictionaries and Corpora*. In A. M. Elizarov & N. V. Loukachevitch (Eds.), Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) (Vol. 2780, pp. 202–211). CEUR-WS.org. https://ceur-ws.org/Vol-2780/paper18.pdf

Kordoni, Valia. 2017. *Beyond Words: Deep Learning for Multiword Expressions and Collocations*. In M. Popović & J. Boyd-Graber (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts (pp. 15–16). Vancouver, Canada: Association for Computational Linguistics. https://aclanthology.org/P17-5005/

Ma, Xiaolei, Jiyu Zhang, Bowen Du, Chuan Ding, & Leilei Sun. 2019. Parallel Architecture of Convolutional Bi-Directional LSTM Neural Networks for Network-Wide Metro Ridership Prediction. *IEEE Transactions on Intelligent Transportation Systems* 20 (6): 2278–2288. https://doi.org/10.1109/TITS.2018.2867042

Maru, Marco, Federico Scozzafava, Federico Martelli, & Roberto Navigli. 2019. *SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). (pp. 3534–3540). Hong Kong, China: Association for Computational Linguistics. https://aclanthology.org/D19-1359/, https://doi.org/10.18653/v1/D19-1359

Mckeown, Kathleen R, & Dragomir R. Radev. 2000. *Collocations. Handbook of Natural Language Processing*. New York: Marcel Dekker, 23-1.

Nattinger, James R., & Jeanette S. Decarrico. 1992. *Lexical phrases and language teaching*. Oxford University Press.

Olah, Christopher. 2015. *Understanding LSTM Networks -- colah's blog.* http://colah.github.io/posts/2015-08-Understanding-LSTMs. (accessed Octobor 10, 2022)

Oquab, Maxime, Leon Bottou, Ivan Laptev, & Josef Sivic. 2014. *Learning and transferring mid-level image representations using convolutional neural networks*. Proceedings of the IEEE conference on computer vision and pattern recognition. (CVPR) (pp. 1717–1724). Columbus, OH, USA: IEEE. https://doi.org/10.1109/CVPR.2014.222Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, & Luke Zettlemoyer. 2018. Deep Contextualized Word Representations https://doi.org/10.18653/v1/n18-1202, https://aclanthology.org/N18-1202.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, & Luke Zettlemoyer. 2018. Deep Contextualized Word Representations https://doi.org/10.18653/v1/n18-1202, https://aclanthology.org/N18-1202.

Press, Ofir & Lior Wolf. 2016. *Using the output embedding to improve language models*. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Volume 2, Short Papers (pp. 157–163). Valencia, Spain: Association for Computational Linguistics. https://aclanthology.org/E17-2025/

Radford, Alec, & Karthik Narasimhan. 2018. *Improving Language Understanding by Generative Pre-Training*. OpenAI. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed ?)

Ramos, Margarita Alonso, Leo Wanner, Orsolya Vincze, Gerark Del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez, & Sabela Prieto González. 2010 May. *Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora*. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) Valletta, Malta.

Riahi, Noushin & Fatemeh Sedghi. 2016. Improving the Collocation Extraction Method Using an Untagged Corpus for Persian Word Sense Disambiguation. *Journal of Computational Chemistry*. *4*, 109–124. https://api.semanticscholar.org/CorpusID:8172246.

Seretan, Maria-Violeta. 2003. *Syntactic and Semantic Oriented Corpus Investigation for Collocation Extraction, Translation and Generation.* Ph. D. thesis, Language Technology Laboratory, Department of Linguistics, University of Geneva, Geneva, Switzerland.

Seretan, Violeta. 2013. On collocations and their interaction with parsing and translation. *Informatics,* 31–11 ,(1)1. https://doi.org/10.3390/informatics1010011

Smadja, Frank, & Kathleen Mckeown. 1991. *Using collocations for language generation 1*. *Computational Intelligence* 7 (4): 146-147, 229-239.

Vechtomova, Olga & John Vineet. 2017. UWat-Emote at EmoInt-2017: Emotion Intensity Detection using Affect Clues, Sentiment Polarity and Word Embeddings. In A. Balahur, S. M. Mohammad, & E. van der Goot (Eds.), Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA@EMNLP) (pp. 249–254). Copenhagen, Denmark: Association for Computational Linguistics. https://aclanthology.org/W17-5235/, https://doi.org/10.18653/v1/W17-5235

پیوست ۱: مثال‌هایی از احتمال باهم آمدن باهمایی و ناباهمایی

دوتایی‌های باهمایی و ناباهمایی	احتمال باهمایی	دوتایی‌های باهمایی و ناباهمایی	احتمال باهمایی	دوتایی‌های باهمایی و ناباهمایی	احتمال باهمایی
عربستان سعودی	0.	خلیج فارس	0.0245	تاج شاهی	0.
عربستان خشک	0.	خلیج سرد	5.0011e-08	تاج ایتالیایی	0.
عربستان فئودالی	0.	خلیج متصرف	4.9609e-08	داش غلام	0.8585
بفهمی نفهمی	1.2356e-10	روغن کرچک	5.6052e-45	داش عزیز	3.9559e-07
بفهمی مطلب	3.1005e-17	روغن سرد	1.4013e-45	داش پیر	3.9803e-07
بفهمی خورده	3.2522e-17	روغن داغ	2.8026e-45	استاد هرمز	0.0006
مدینه فاضله	9.3774e-11	تنبیه بدنی	4.2039e-45	استاد معروف	1.7001e-13
مدینه جهودان	3.6111e-16	تنبیه بچه	0.	استاد دانشگاه	1.2819e-13
مدینه وقت	3.6918e-16	تنبیه توبیخ	0.	صمیم قلب	1.3846e-12
طول عمر	1.7466e-06	دانشکده پزشکی	2.4123e-12	صمیم دست	5.3538e-13
طول خیابان	1.6439e-06	دانشکده بزرگ	1.7028e-14	صمیم دوستی	3.2095e-13
طول اتاق	1.5897e-06	دانشکده قدیمی	1.1125e-14	مواد مخدر	6.5368e-05
بریتانیای کبیر	0.3035	بارک‌الله	1.	مواد مقوی	2.0098e-05
بریتانیای نفتی	1.6619e-05	بارک تعالی	0.	مواد کهنه	2.6657e-05
بریتانیای مصرف‌کننده	1.6585e-05	بارک نظر	0.	صدر اعظم	0.0485
ممالک محروسه	0.0014	بلوک زهرا	6.5861e-44	صدر نشین	3.3470e-06
ممالک ایران	0.0001	بلوک آتلانتیک	5.4651e-44	صدر بازار	3.6908e-06
ممالک خارجه	2.5323e-05	بلوک فریده	5.4651e-44	سازمان ملل	0.0003
ممالک اطراف	2.5289e-05	سجده شکر	9.6867e-10	سازمان بشر	2.1630e-05
توران السلطنه	3.0419e-09	سجده ابلیس	9.5388e-10	سازمان انقلابی	2.1296e-05
توران شاه	2.2000e-09	سجده نماز	9.5594e-10	سازمان رقیب	2.0976e-05
توران شجاع	1.1626e-09	افسر نگهبان	4.5441e-08	قتل عام	1.4013e-45
شاطر حبیب	1.1464e-25	افسر جوان	3.9127e-08	قتل کودکان	0.
شاطر نانوائی	2.8026e-45	افسر شجاع	3.3098e-08	قتل مردان	0.
شاطر جوان	2.8026e-45	بسم‌الله	0.0001	قتل خون	0.
خلع سلاح	2.3507e-06	بسم محمد	2.3806e-05	شورای امنیت	5.2818e-05

دوتایی‌های باهمایی و ناباهمایی	احتمال باهمایی	دوتایی‌های باهمایی و ناباهمایی	احتمال باهمایی	دوتایی‌های باهمایی و ناباهمایی	احتمال باهمایی
خلع سپاه	1.6885e-06	بسم الرحمن	2.3806e-05	شورای کشور	2.4765e-05
خلع لباس	1.6864e-06	حوا سلطان	4.9611e-05	شورای زنانه	2.3930e-05
خلیج مکزیک	5.8496e-08	حوا جان	2.3932e-05	امام جمعه	1.6584e-11
خلیج ساحلی	4.8769e-08	حوا مریم	2.3811e-05	امام شنبه	6.2557e-15
خلیج گرم	5.2122e-08	تاج الملوک	0.	امام نماز	6.5627e-15

زینب‌الهدی حشمتی

دارای مدرک تحصیلی دکتری در رشته مخابرات از دانشگاه لیدز انگلستان است. ایشان هم‌اکنون استادیار گروه علوم و فناوری داده دانشگاه تهران است.
تحلیل شبکه‌های پیچیده از جمله تحلیل شبکه‌های اجتماعی و هوش مصنوعی از جمله علایق پژوهشی وی است.

مینا ملکی ویکا

متولد سال ۱۳۶۸، دارای مدرک کارشناسی ارشد زبان‌شناسی رایانشی از دانشگاه تهران است. ایشان هم‌اکنون دانش‌آموخته رشته زبان‌شناسی دانشگاه تهران است.
پردازش زبان طبیعی و زبان‌شناسی از جمله علایق پژوهشی وی است.

محمود بی‌جن‌خان

متولد سال ۱۳۳۷، دارای مدرک تحصیلی دکتری در رشته زبان‌شناسی از دانشگاه تهران است. ایشان هم‌اکنون استاد تمام گروه زبان‌شناسی دانشکده ادبیات و علوم انسانی دانشگاه تهران است.
آواشناسی، واج‌شناسی، زبان‌شناسی پیکره‌ای و رایانشی از جمله علایق پژوهشی وی است.

هادی ویسی

دانش‌آموخته مهندسی کامپیوتر از دانشگاه صنعتی شریف در سال ۱۳۹۰، و هم‌اکنون دانشیار دانشگاه تهران است.

یادگیری عمیق (هوش مصنوعی) و پردازش زبان طبیعی از جمله علایق پژوهشی وی است.