The 9th International Conference on Web Research
ICWR2023
May 3th-4th, 2023 ;Tehran, Iran

Solving Persian Crosswords with Natural Language Processing Techniques

[1]Mohammadreza Pakzadian*, [2]Mehrnoush Shamsfard

[1] Bachelor of Computer Engineering, Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
pakzadianmrp@gmail.com

[2] Associate Professor, Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
m-shams@sbu.ac.ir

Abstract

In this article, we introduce solutions for solving crossword puzzles by machine using natural language processing techniques. This task is divided into two subtasks of finding possible answers for each table description and then selecting the target word and placing it in the table. The first subtask, which is dedicated to finding the word from its description, has many other uses as in text generation and paraphrasing. For this purpose, we used a combination of different methods, including searching and finding semantic similarities on the data of previously solved tables, searching in dictionary and Wikipedia articles, using a masked language model, and finding related words in Farsnet and the Farsiyar tool. The results show that the combination of these methods has a better result (82% recall) compared to their individual implementation.

In the next subtask, we give the list of possible answers to a constraint-satisfaction search algorithm to choose the correct answer that can be placed in the table, taking into account the constraints of the table, and fill the empty cells in the best way and solve the crossword. The overall evaluation shows 80.22% precision and 68.86% recall in solving the crossword puzzle.

Keywords: crossword solving, constraint satisfaction problem, natural language processing, description understanding

The 9th International Conference on Web Research
ICWR2023
May 3th-4th, 2023, Tehran, Iran

حل جدول کلمات متقاطع فارسی با تکنیک‌های پردازش زبان طبیعی

محمدرضا پاکزادیان*[1]، مهرنوش شمس‌فرد[2]

[1] کارشناسی مهندسی کامپیوتر، دانشکده مهندسی وعلوم کامپیوتر، دانشگاه شهید بهشتی، تهران
pakzadianmrp@gmail.com

[2] دانشیار، گروه هوش مصنوعی، رباتیک و رایانش شناختی، دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران
m-shams@sbu.ac.ir

چکیده

در این مقاله، به معرفی راهکارهایی برای حل جدول کلمات متقاطع توسط ماشین با استفاده از تکنیک‌های پردازش زبان طبیعی می‌پردازیم. این کار به دو زیروظیفه یافتن کلمات محتمل برای هر شرح جدول و سپس انتخاب کلمه هدف و جایگذاری در جدول تقسیم می شود. زیروظیفه اول که به یافتن کلمه از روی توصیف آن اختصاص دارد دارای کاربردهای متعدد دیگر نیز هست. به این منظور از ترکیبی از روش‌های مختلف شامل جستجو و شباهت‌یابی معنایی روی دادگان جداول حل شده قبلی، جستجو در فرهنگ لغات و دانشنامه ویکی‌پدیا، استفاده از مدل زبانی ماسک شده و یافتن کلمات مرتبط در فارس‌نت و ابزار فارسی‌یار استفاده نمودیم. نتایج نشان می‌دهند ترکیب این روش‌ها نسبت به اجرای تک تک آن‌ها نتیجه بهتری (فراخوانی ۸۲٪) در برداشته است.

در زیروظیفه بعد لیست پاسخ‌های محتمل را به یک الگوریتم جستجو با ارضاء قیود می‌دهیم تا با درنظر گرفتن قیود جدول از میان پاسخ‌ها، پاسخ درستی که می‌تواند در جدول قرار گیرد را انتخاب کرده و خانه‌های خالی را به بهترین شکل پر کند و جدول را حل کند. ارزیابی کل نشانگر دقت ۸۰.۲۲٪ و فراخوانی ٦٨.٨٦٪ در حل جدول کلمات متقاطع است.

کلمات کلیدی

حل جدول کلمات متقاطع، جستجو با ارضاء قیود، پردازش زبان طبیعی، فهم تعریف .

۱- مقدمه

حل کردن جدول کلمات همواره به عنوان یک سرگرمی جذاب برای افراد شناخته می‌شود و بسیاری از افراد در اوقات فراغت خود به حل جدول می‌پردازند. همچنین پیشرفت‌های فناوری در زمینه هوش مصنوعی نشان داده است که کامپیوترها می‌توانند در بازی‌هایی مانند شطرنج، انسان‌ها را به چالش بکشند و حتی برنده شوند. از این رو حل کردن جدول کلمات متقاطع توسط کامپیوترها می‌تواند یک مسئله جذاب و چالش برانگیز برای افرادی که در حوزه علوم کامپیوتری کار می‌کنند، باشد.

لازمه حل جدول کلمات متقاطع، درک معنای شرح‌های داده شده و یافتن کلمه ای با محدودیت‌های مشخص طول و حروف معلوم شده، متناظر با شرح است.

از جمله نرم افزارهای مشابهی که در زمینه حل جدول کلمات وجود دارند می‌توان به WebCrow[3]، Dr.fill[4]، Proverb[5] و Berkeley Crossword Solver[6] اشاره کرد که همگی برای حل جدول به زبان انگلیسی طراحی شده‌اند. Dr.fill از یک روش شبیه به TFIDF برای یافتن پاسخ شرح‌ها استفاده می‌کند اما Proverb و WebCrow از تعدادی پیمانه[1] سفارشی برای این منظور استفاده می‌کنند. WebCrowهمچنین از یک موتور جستجوی وب نیز برای این کار استفاده می‌کند. همچنین Berkely از یک مدل پرسش و پاسخ[2] برای یافتن پاسخ‌ها استفاده می‌کند. برای بخش حل جدول، Proverb و WebCrow از الگوریتم انتشار باورهای حلقه‌ای[3] در ترکیب با الگوریتم جستجوی A* استفاده می‌کنند در حالی که Dr.fill از یک الگوریتم جستجوی عمق اول[4] اصلاح شده استفاده می‌کند و پس از آن با استفاده از یک الگوریتم جستجوی محلی پاسخ‌های قرار گرفته در جدول را اصلاح می‌کند و با استفاده از یک تابع مکاشفه[5] به این جدول‌های اصلاح شده یک امتیاز نسبت می‌دهد. همچنین Berkely با استفاده از الگوریتم انتشار باورهای حلقه‌ای به همراه جستجوی محلی به پر کردن جدول می‌پردازد. در سال ۲۰۲۱ برای اولین بار در مسابقات جدول کلمات متقاطع آمریکایی، ترکیبی از سیستم‌های Berkely و Dr.fill توانست از تمام رقبای انسانی خود بهتر عمل کند.

۲- طرح مسأله

مسألهای که ما به دنبال حل آن هستیم این است که راهکارهایی برای حل جدول کلمات متقاطع فارسی توسط ماشین ارائه کنیم. برای این منظور ابتدا باید ساختاری برای تعریف جدول مشخص کنیم و سپس با به کارگیری تکنیکها و روشهای هوش مصنوعی و پردازش زبانهای طبیعی پاسخ مورد نظر برای هریک از شرحهای جدول را پیدا کرده و با انتخاب کلمه هدف و جایگذاری آن، جدول را کامل کنیم. نمونهای از جدول کلمات متقاطع مورد نظر در شکل (۱) قابل مشاهده است.

۳- راه حل پیشنهادی

حل جدول شامل دو مرحله اصلی است. (۱) یافتن کلمات کاندیدا به عنوان پاسخ محتمل با توجه به هر شرح (۲) انتخاب کلمه صحیح از میان کاندیداها با توجه به جدول و کلمات دیگر موجود درآن. شمای کلی مراحل مورد استفاده در حل جدول در شکل (۲) قابل مشاهده است.

۳-۱- تهیه دادگان مورد نیاز

دادگان مورد نیاز را به دو دسته تقسیم می کنیم. دسته اول دادگان آموزش و آزمون برای جدول کلمات متقاطع است. به این منظور، با نوشتن یک برنامه خزشگر[۶] برروی سایت جدول‌یاب[۱]، تعداد ۱۰۸۸ عدد جدول از این سایت دریافت و ذخیره شدند. تهیه این مجموعه داده از دو جهت حائز اهمیت است. اول اینکه از بخشی از این دادهها میتوان به عنوان مجموعه دادگان آزمون برای بررسی عملکرد نرم‌افزار استفاده کرد و دوم اینکه در بسیاری از جدول‌ها سوالات تکراری وجود دارند و از این مجموعه داده می‌توان به عنوان دادههای آموزش برای یافتن پاسخ اینگونه سوالات استفاده کرد. جداول استخراج شده به صورت تصادفی به دو مجموعه ۱۰۰۰ تایی برای آموزش و ۸۸ تایی برای آزمون تقسیم شدند. نمودارهای فراوانی شرحها برحسب تعداد کلمات موجود در آن و فراوانی پاسخها برحسب طول کلمه در این مجموعه دادگان در شکل‌های (۳) و (۴) آمده است. در این مجموعه بلندترین شرح حاوی ۳۲ کلمه و بلندترین کلمه پاسخ به طول ۱۶ است.

در ادامه ابتدا به دادگان مورد استفاده پرداخته و سپس روش به کار گرفته شده برای هریک از دو مرحله فوق را شرح می دهیم.

دسته دوم دادگان پایگاه‌های داده و دانش در مورد کلمات، توصیف آن‌ها و مثال‌هایی از کاربردشان است. جمع آوری اطلاعات از لغت‌نامه دهخدا[۲]، ویکی‌پدیای فارسی، فارس‌نت[7] و فارسی‌یار[8] در این دسته قرار دارند.

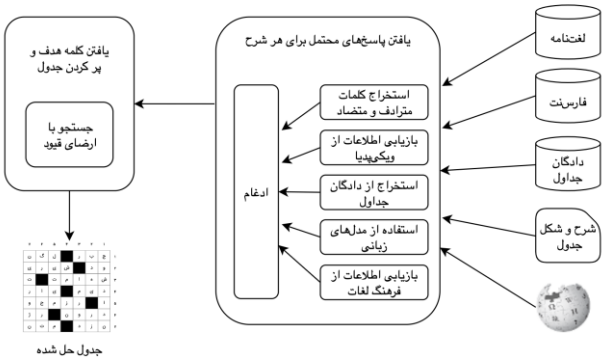
در این راستا با خزش بر روی سایت لغت‌نامه دهخدا تعداد ۳۴۳۳۱۸ کلمه و معنی متناظر با آن جمع آوری شد. از فارس‌نت (وردنت[9] فارسی) نیز ۱۰۰ هزار مدخل واژگانی با توضیح، مثال و روابط (ترادف، تضاد، شمول و ...) آن‌ها با کلمات دیگر استخراج شد. همچنین ازمجموعه دادگان ویکی‌پدیا فارسی که حاوی بیش از یک میلیون مقاله فارسی است نیز استفاده شد. علاوه بر این دادگان، خواهیم دید که از ابزار استخراج رابطه معنایی کلمات در جعبه ابزار فارسی‌یار نیز بهره گرفته شده است.

۷	۶	۵	۴	۳	۲	۱	
ن	گ	ل	■	ر	ب	ج	۱
ی	ر	ی	ش	■	د	و	۲
ت	■	ت	م	ا	ه	ش	۳
ر	ا	ی	■	م	ی	د	۴
و	ج	م	ز	ر	■	ا	۵
ژ	ر	■	ن	و	ر	د	۶
ن	ت	م	■	د	ز	ن	۷

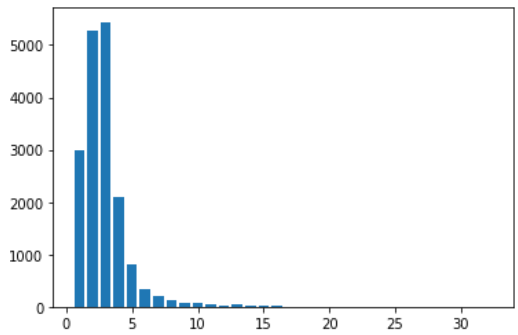
افقی:
۱- زور - یار آفتابه
۲- دوستی - رنگ سفید مات
۳- شجاعت و دلیری
۴- کشت بارانی - دوست
۵- جنگاور
۶- داخل - از لوازم آرایش
۷- پیش و جلو - درون حاشیه

عمودی:
۱- پیوند دادن
۲- قرض و دین - برنج
۳- گلابی
۴- حس بویایی - یار مرد
۵- عنصر شیمیایی فلزی
۶- مخفف اگر - مزد
۷- ازت

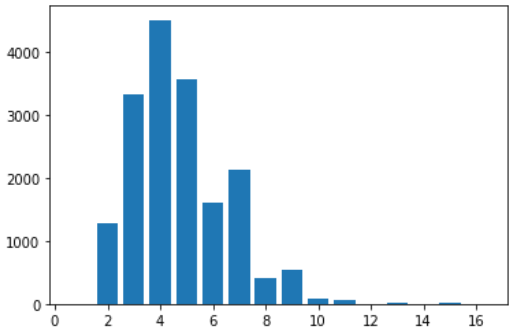
شکل (۱): نمونه‌ای از یک جدول کلمات متقاطع فارسی



شکل (۲): مراحل مورد استفاده در حل جدول



شکل (۳): نمودار فراوانی شرح‌ها برحسب تعداد کلمات



شکل (٤): نمودار فراوانی پاسخ‌ها برحسب طول کلمه

۳-۲- یافتن پاسخ‌های محتمل برای هر شرح

برای یافتن مجموعه پاسخ‌های محتمل برای هر شرح جدول از ترکیبی از چند روش استفاده کردیم. هر کدام از این روش‌ها علاوه بر پیدا کردن مجموعه پاسخ‌ها، به هر عضو این مجموعه یک عدد بین صفر و یک نسبت می‌دهند که نشان دهنده این است که این پاسخ با چه احتمالی می‌تواند پاسخ مورد نظر ما باشد. در ادامه به توضیح هرکدام از این روش‌ها که در قالب یک پیمانه در سیستم پیاده سازی شده اند، می‌پردازیم.

۳-۲-۱- استخراج کلمات مترادف و متضاد

این روش که مناسب یافتن پاسخ برای شرح‌هایی با یک کلمه یا یک عبارت ساده است، با استفاده از فارس‌نت و فارسی‌یار کلمات مترادف یا متضاد با کلمه یا عبارت شرح را پیدا می‌کند و پس از اعمال قید طول کلمه، مجموعه‌ای از پاسخ‌های محتمل را برمی‌گرداند. الگوی شرح های مورد پوشش به شکل زیر است:

- **<کلمه>** مثال: دوستی (ود)، مزد (اجرت)
- **<کلمه> و <کلمه>** مثال: شجاعت و دلیری (شهامت)
- **مخالف <کلمه>** مثال: مخالف کوچک (بزرگ)
- **متضاد <کلمه>** مثال: متضاد سریع (کند)

احتمال هرکدام از این پاسخ‌ها از طریق فرمول (۱) به دست می‌آید که در آن L برابر با تعداد کلمات موجود در مجموعه پاسخ‌های محتمل است.

$$p = \frac{1}{L} \qquad (۱)$$

۳-۲-۲- بازیابی اطلاعات از ویکی‌پدیا

بسیاری از پرسش‌های جدول به گونه‌ای هستند که پاسخدهی به آنها نیاز به دانش در حوزه خاصی مانند علوم، جغرافیا، تاریخ و... دارد. پاسخ اینگونه پرسش‌ها معمولا با جستجو در دانشنامه هایی مانند ویکی‌پدیا قابل استخراج است. در این بخش، از جستجو روی مجموعه دادگان ویکی‌پدیای فارسی با استفاده از ابزار الستیک سرچ[۷] که بر پایه موتور جستجوی لوسین [10] توسعه یافته است، برای یافتن پاسخ اینگونه پرسش‌ها استفاده نمودیم.

در این جستجو پرس و جوی[۸] مورد استفاده همان شرح جدول است. جستجو بر روی کل اسناد انجام میگردد و ۵- بهترین نتیجه جستجو بعنوان اسناد کاندیدا بازگردانده می شود. سپس از میان این اسناد در ویژگی‌های title ، RedirectList و links به دنبال کلماتی با طول مورد نظر می‌گردیم و تمامی این کلمات را در مجموعه پاسخ‌های احتمالی قرار می‌دهیم. برای محاسبه احتمال درستی هرپاسخ، از فرمول (۲) استفاده می‌کنیم.

$$p = \frac{score}{100} * c \qquad (۲)$$

که در آن score امتیازی است که ابزار الستیک سرچ به هر کدام از این اسناد نسبت می‌دهد و c ضریب اهمیت ویژگی مورد نظر در آن سند است و مقدار آن برای ویژگی title برابر ۰.۹۹ و برای سایر ویژگی‌ها برابر ۰.۹ است.

این روش برای یافتن پاسخ پرسش‌هایی مشابه پرسش‌های زیر عملکرد خوبی از خود نشان می‌دهد:

- تنظیم کننده قند خون (انسولین)
- از شهرهای استان فارس (استهبان)
- عنصر شیمیایی فلزی (لیتیم)

۳-۲-۳- استخراج از دادگان جداول

در بسیاری از مواقع سوال‌های جدول تکراری هستند و افرادی که جدول‌های زیادی حل می‌کنند با دانستن این موضوع می‌توانند عملکرد بهتری در حل جدول داشته باشند. وظیفه این پیمانه این است که با استفاده از مجموعه داده‌ای که از جدول‌های سایت جدول‌یاب تهیه شد پاسخ اینگونه سوال‌های تکراری را پیدا کند و به حل سریع‌تر جدول کمک کند. این مجموعه داده حاوی ۱۷۶۶۶ ردیف داده متمایز که از ۱۰۰۰ عدد جدول استخراج شده‌اند می‌باشد. این روش با جستجو برروی این مجموعه داده، شرح‌های مشابه با شرح مورد نظر را پیدا می‌کند و پاسخ آن‌ها را به عنوان پاسخ احتمالی برای این شرح در نظر می‌گیرد. پیاده سازی این پیمانه به دو روش انجام شد که در ادامه به توضیح هرکدام از روش‌ها می‌پردازیم.

یکی از روش‌هایی که می‌توان برای پیاده سازی این پیمانه استفاده کرد، استفاده از ابزارهای جستجوی متنی مانند الستیک سرچ برای یافتن شرح‌های مشابه است. برای این کار ابتدا ویژگی‌های سند مربوطه را در این ابزار تعریف کردیم و سپس داده‌های مورد نظر را وارد این ابزار کردیم.

سپس یک پرس و جو بر روی این اسناد انجام می‌دهیم و از میان اسنادی که در پاسخ این پرس و جو داده می‌شوند ۱۰- بهترین نتیجه جستجو را انتخاب می‌کنیم و پاسخ آن‌ها را به عنوان لیست پاسخ‌های احتمالی در نظر می‌گیریم. سپس این پاسخ‌ها را برحسب طول کلمه‌ای که به دنبال آن هستیم، محدود می‌کنیم و برای محاسبه اینکه هرکدام از این پاسخ‌ها با چه احتمالی می‌توانند پاسخ مورد نظر ما باشند، از فرمول (۳) استفاده می‌کنیم.

$$p_i = \frac{score_i}{\sum score} \qquad (۳)$$

که در آن score امتیازی است که ابزار الستیک سرچ به هر کدام از این اسناد نسبت می‌دهد.

روش دیگری که با استفاده از آن می‌توان این جستجو را انجام داد، استفاده از مدل‌های زبانی[9] برای پیدا کردن جملات مشابه با استفاده از جستجوی معنایی[10] است. برای این منظور می‌توان از مدل SBERT[11] که بر پایه مدل زبانی BERT[12] و برای یافتن شباهت جملات ساخته شده است، استفاده کرد. مزیت این روش نسبت به روش قبل این است که در روش قبل معنای کلمات در جستجو نقشی ندارند و فقط شکل نوشتاری آن‌ها در نظر گرفته می‌شود اما در مدل‌های زبانی علاوه بر شکل نوشتاری، معنای کلمات نیز در جستجو نقش دارند.

با توجه به اینکه ساخت و آموزش یک مدل زبانی نیاز به مقدار زیادی داده و صرف منابع و زمان زیادی دارد، برای این کار از مدل‌های زبانی از پیش آموزش داده شده استفاده کردیم. به همین منظور مدل ParsBERT[13] را روی مجموعه دادگان جداول ریزتنظیم[11] کردیم. سپس با استفاده از مشابهت یابی معنایی، شرح‌های مشابه با شرح مورد نظر را پیدا می‌کنیم. از میان تمام شرح‌های مشابه، ۱۰ عدد از بهترین شرح‌هایی که شباهت بالای ۷۵ درصد با شرح مورد نظر ما دارند و همچنین قید طول پاسخ را ارضا می‌کنند را انتخاب می‌کنیم تا پاسخ آن‌ها به عنوان پاسخ محتمل در نظر گرفته شود. برای نسبت دادن احتمال به هرکدام از این پاسخ‌ها، از همان فرمول ۳ استفاده می‌کنیم و مقدار score در این فرمول برابر میزان شباهت در نظر گرفته می‌شود.

۳-۲-٤- استفاده از مدل‌های زبانی

یکی دیگر از روش‌هایی که با استفاده از آن می‌توان پاسخ یک شرح را به دست آورد استفاده از تکنیک‌های پرامپتینگ و مدل‌های زبانی بافتاری است. در این روش با استفاده از ۱۰ الگوی از پیش تعریف شده، از روی شرح جدول یک جمله می‌سازیم و پاسخ مورد نظر را به شکل جای خالی در شرح قرار می‌دهیم و از این مدل زبانی می‌خواهیم تا با توجه به معنای جمله جای خالی را با کلمه مناسب پر کند. چند نمونه از الگوها و جملات ساخته شده با استفاده از آن‌ها به شکل زیر است:

- <پاسخ> <شرح> است.

مثال: [MASK] سازمان آموزشی، علمی و فرهنگی سازمان ملل متحد است. (یونسکو)

- <شرح> <پاسخ> است.

مثال: آنچه مورد لزوم و احتیاج باشد، [MASK] است. (ضروری)

- <پاسخ> معادل <شرح> است.

مثال: مروت معادل [MASK] است. (جوانمردی)

با توجه به اینکه هزینه آموزش یک مدل زبانی از پایه بسیار زیاد است، در این بخش ما از مدل زبانی ParsBERT که از قبل برروی حجم زیادی از متون فارسی آموزش داده شده است، استفاده کردیم و این مدل را بر اساس مجموعه داده‌ای که داشتیم ریزتنظیم کردیم. سپس ۲۰ نماینده برتر برای پر کردن جای خالی را که قید طول قید طول کلمه مورد نظر را ارضا می‌کنند، به عنوان مجموعه پاسخ‌های محتمل انتخاب می‌کنیم و با استفاده از فرمول ۳ و امتیازی که مدل زبانی به هر کدام از نماینده‌ها نسبت می‌دهد، برای هر کدام از پاسخ‌ها احتمال را محاسبه می‌کنیم.

۳-۲-۵- بازیابی اطلاعات از فرهنگ لغات

وظیفه این بخش، بازیابی اطلاعات از فرهنگ لغات است. در این پیمانه اطلاعات لغت‌نامه دهخدا در ابزار الستیک سرچ وارد شد تا با استفاده از این ابزار بتوانیم برروی معنی کلمات در لغت‌نامه دهخدا جستجوی متنی انجام دهیم و پاسخ مورد انتظار را پیدا کنیم.

در این جستجو پرس و جوی مورد استفاده همان شرح جدول است. جستجو بر روی کل اسناد انجام میگردد و ۱۰- بهترین نتیجه جستجو بعنوان اسناد کاندیدا بازگردادنده می شود و کلمه‌ای که سند در مورد آن است به شرطی که قید طول کلمه را ارضا کند، به عنوان پاسخ احتمالی در نظر گرفته می‌شود. همچنین از فرمول ۳ برای محاسبه احتمال هر کدام از پاسخ‌ها استفاده می‌کنیم.

۳-۲-٦- ادغام پاسخ‌های پیشنهادی

پس از آنکه هر کدام از پیمانه‌ها لیست پاسخ‌های پیشنهادی خود برای هرکدام از شرح‌های جدول را ارائه کردند، این لیست‌ها با یکدیگر ادغام می‌شوند و مقدار احتمالی که برای هر پاسخ ارائه شده بود نیز برابر با بیشینه مقدار احتمال ارائه شده برای این پاسخ در پیمانه‌ها می‌شود.

۳-۳- پر کردن پاسخ‌ها در جدول

در این مرحله برای پرکردن پاسخ‌های احتمالی در جدول و پیدا کردن بهترین حالت پر کردن جدول، از الگوریتم جستجو با ارضای قیود[12] استفاده کردیم. در این الگوریتم از یک تابع ارزیابی برای مقایسه و ارزیابی حالت‌های مختلف جدول‌های پر شده و یک تابع مکاشفه‌ای برای پیش بینی امتیاز حالت نهایی جدول، برای جدول‌هایی که در حین جستجو به دست می‌آیند استفاده کردیم. فرمول (۴) تابع ارزیابی و فرمول (۵) تابع مکاشفه‌ای را نشان می‌دهد.

$$f(x) = a * \frac{N_{fq}}{N_q} + b * \frac{N_{cb}}{N_b} + c * \frac{\sum p_{x_i}(v_i)}{N_{fq}} \qquad (۴)$$

که در آن N_{fq} تعداد سوالات پر شده در جدول، N_{fq} تعداد کل سوالات جدول، N_{cb} تعداد خانه‌هایی از جدول که محل تقاطع دو سوال بوده‌اند و هردو سوال پاسخ داده شده‌اند، N_b تعداد کل خانه‌های غیر سیاه جدول و $\sum p_{x_i}(v_i)$ مجموع احتمال ارائه شده برای پاسخ‌هایی که در جدول پر شده‌اند، می‌باشد.

$$h(x) = a * \frac{N_{fl}}{N_q} + b * \frac{N'_{cb}}{N_b} + c * \frac{\sum max(p_{x_i}(v_i))}{N_{fq}} \qquad (۵)$$

که در آن N_{fl} تعداد سوالات پاسخ داده نشده که مجموعه پاسخ‌های محتمل آن‌ها شامل حداقل یک عضو است، N'_{cb} تعداد خانه‌هایی از جدول

محل تقاطع دو سوال هستند و می‌توانند توسط سوال‌های پاسخ داده نشده پر شوند و $\sum \max(p(V_i))$ مجموع بیشینه احتمال ارائه شده در مجموعه پاسخ‌های محتمل برای سوالات پاسخ داده نشده، می‌باشد.

در فرمول تابع ارزیابی و مکاشفه‌ای مقادیر a، b و c ضرایبی هستند که میزان اهمیت هر بخش از تابع را مشخص می‌کنند و مقادیر در نظر گرفته شده برای آن‌ها که با آزمایش‌های تجربی بدست آمده به ترتیب ۱، ۴، و ۲ است.

در الگوریتم جستجو، سازگاری گره از قبل توسط پیمانه‌هایی که پاسخ‌های محتمل را تولید می‌کردند چک می‌شود. به این صورت که این پیمانه‌ها بر اساس قید طول کلمات که توسط جدول مشخص شده است، فقط کلمات با طول مناسب را نگه می‌دارند و سایر کلمات را از دامنه پاسخ‌های محتمل حذف می‌کنند. در مرحله انتخاب گره برای گسترش، در هر مرحله جدولی که مقدار تابع ارزیابی بیشتری داشته باشد برای گسترش و پر کردن سایر سوالات انتخاب می‌شود. سپس از میان سوالاتی که هنوز پر نشده‌اند، سوالی که بیشترین تعداد خانه‌های مشترک با سایرین را دارد انتخاب می‌شود تا پاسخ آن در جدول قرار گیرد و با توجه به اینکه یکی از سوالات در جدول قرار گرفته و قیود جدیدی در جدول اضافه شده‌اند، مجموعه پاسخ‌های احتمالی سایر سوالات مطابق این قیود محدود می‌شوند.

برای محدودتر کردن فضای جستجو، حلت‌های نامطلوب را هرس می‌کنیم. به این صورت که در هنگام جستجو همیشه بهترین جدولی که تا کنون یافته‌ایم یعنی جدولی که بیشترین مقدار تابع ارزیابی را داشته، نگه می‌داریم و اگر جدول‌هایی که در هنگام گسترش ایجاد می‌شوند، مقدار تابع مکاشفه‌ای آن‌ها کمتر از بهترین حلت باشد، این حلت‌ها را دور ریخته و هرس می‌شوند.

۳-٤- ارزیابی عملکرد پیمانه‌ها

برای ارزیابی عملکرد پیمانه‌ها از ۸۸ جدولی که به عنوان دادگان آزمون کنار گذاشته بودیم، استفاده می‌کنیم که این مجموعه داده حاوی ۸۱۰ ردیف داده متمایز است. برای هر کدام از پیمانه‌ها معیار فراخوانی[۱۳] را مطابق فرمول (۶) محاسبه می‌کنیم.

$$\text{فراخوانی} = \frac{\text{تعداد سوال‌هایی که پاسخ صحیح در لیست پاسخ‌های احتمالی وجود داشت}}{\text{تعداد کل سوالات}} \quad (۶)$$

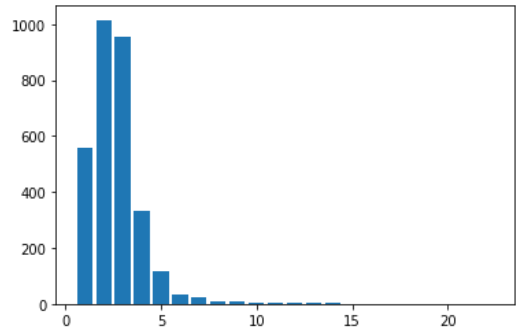
نتیجه ارزیابی پیمانه‌ها در جدول (۱) قابل مشاهده است.

در شکل (۵) نمودار فراوانی تعداد کلمات شرح‌ها در مجموعه دادگان آزمون و در شکل (۶) نمودار فراخوانی تمام پیمانه‌ها در کنار هم بر حسب تعداد کلمات موجود در شرح قابل مشاهده است.

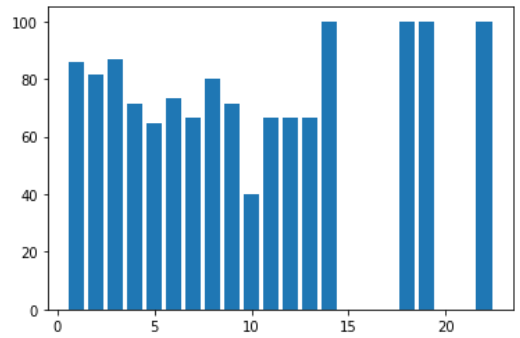
همچنین در شکل (۷) نمودار فراوانی طول پاسخ در مجموعه دادگان آزمون و در شکل (۸) نمودار فراخوانی پیمانه‌ها بر حسب طول پاسخ قابل مشاهده است. همان طور که در این نمودارها مشخص است با افزایش طول پاسخ، فراخوانی کاهش می‌یابد زیرا پاسخ‌هایی که طول بیشتری دارند معمولا مرکب از دو یا چند کلمه هستند و پیدا کردن آن‌ها سخت‌تر است.

جدول (۱) : نتیجه ارزیابی پیمانه‌ها

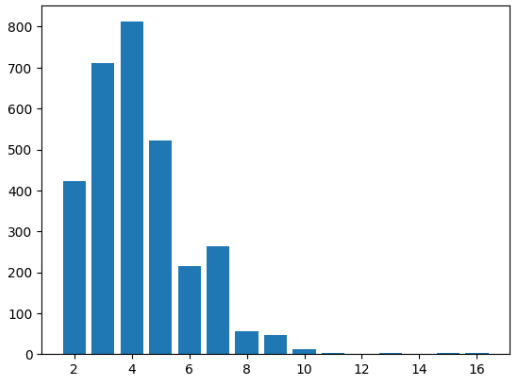
نام پیمانه	فراخوانی
استخراج کلمات مترادف و متضاد (فارس‌نت)	٪۷۰.۸۱
استخراج کلمات مترادف و متضاد (فارسی‌یار)	٪۷۲.۵۳
بازیابی اطلاعات از ویکی‌پدیا	٪۲۰.۹۰
جستجوی الستیک روی دادگان جداول	٪۷۳.۵۲
شباهت‌یابی معنایی روی دادگان جداول	٪۷۳.۴۰
پرامپتینگ و مدل زبانی	٪۱۰.۳۲
بازیابی اطلاعات از فرهنگ لغات	٪۷.۷۱
تمام پیمانه‌ها در کنار هم	٪۸۲.۱۹



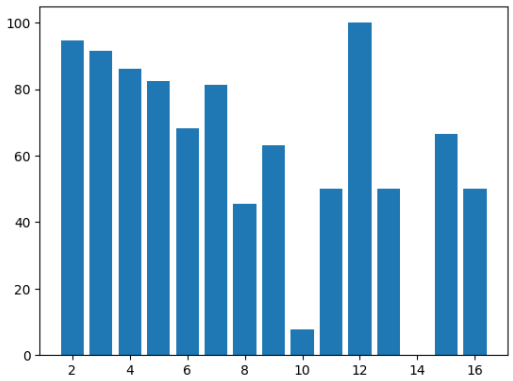
شکل (۵): نمودار فراوانی تعداد کلمات شرح در مجموعه دادگان آزمون



شکل (٦): نمودار فراخوانی پیمانه‌ها برحسب تعداد کلمات شرح



شکل (۷): نمودار فراوانی طول کلمه پاسخ در مجموعه دادگان آزمون



شکل (۸): نمودار فراخوانی پیمانه‌ها برحسب طول کلمه پاسخ

۳-۵- ارزیابی عملکرد نرم افزار

برای ارزیابی عملکرد کلی نرم افزار حل جدول، از همان مجموعه داده آزمون که در بخش قبل توضیح داده شد استفاده کردیم و دو معیار فراخوانی و دقت[۱۴] را مطابق فرمول‌های (۷) و (۸) محاسبه کردیم.

$$\text{فراخوانی} = \frac{\text{تعداد خانه‌هایی از جدول که به درستی مقداردهی شده‌اند}}{\text{تعداد کل خانه‌های جدول}} \quad (۶)$$

$$\text{دقت} = \frac{\text{تعداد خانه‌هایی از جدول که به درستی مقداردهی شده‌اند}}{\text{تعداد خانه‌های پر شده}} \quad (۷)$$

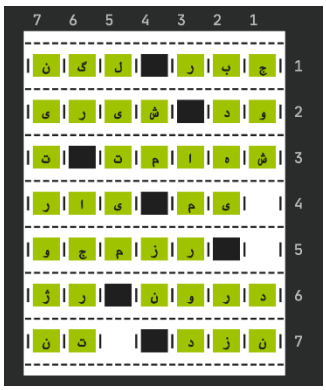
پس از ارزیابی‌های انجام شده مشخص شد که نرم افزار ما به صورت میانگین، فراخوانی **۶۸.۸۶** درصد و دقت **۸۰.۲۲** درصد در حل جدول دارد.

به عنوان نمونه عملکرد نرم افزار برروی جدول شکل (۱) در شکل (۹) قابل مشاهده است.

۴- نتیجه گیری

حل کردن جدول کلمات متقاطع به کمک هوش مصنوعی مسئله‌ای چالش برانگیز و پیچیده است که نیازمند اطلاعات و داده در زمینه‌های مختلف است. برای اینکه بتوانیم یک نرم افزار حل کننده جدول کلمات با دقت خیلی بالا داشته باشیم، نیازمند مجموعه داده‌ای در ابعاد یک موتور جستجو هستیم تا بتوانیم در زمینه‌های مختلف اطلاعات داشته باشیم و همچنین باید بتوانیم به شکل مناسبی از این مجموعه داده اطلاعات مورد نظر را بازیابی کنیم که داشتن این حجم از داده و بازیابی مناسب آن نیازمند زمان، منابع و هزینه بسیار زیادی است.

به همین دلیل در این پروژه ما بر روی بخش‌های کوچک‌تر و هدفمندتری از داده متمرکز شدیم تا بتوانیم نیازمندی خود را برطرف کنیم. به همین منظور در این پروژه از مجموعه دادگان لغت‌نامه دهخدا، مجموعه مقالات فارسی ویکی‌پدیا، مجموعه جدول‌های سایت جدول‌یاب، فارس‌نت و فارسی‌یار استفاده کردیم تا بتوانیم تا حد قابل قبولی منبع دانش مورد نیاز برای حل جداول را شکل دهیم و سپس با استفاده از تکنیک‌های بازیابی اطلاعات و مدل‌های زبانی به دریافت اطلاعات از این مجموعه دانش پرداختیم. سپس با استفاده از الگوریتم



شکل (۹) : نمونه‌ای از عملکرد نرم افزار در حل جدول

جستجو با ارضای قیود نسبت به حل جدول و پیدا کردن پاسخ مناسب برای هر کدام از سوالات اقدام کردیم. و در نهایت به فراخوانی ۶۸.۸۶ درصد و دقت ۸۰.۲۲ درصد دست پیدا کردیم.

برای کارهای آتی می‌توان بهبودهایی در بخش‌های مختلف نرم افزار ایجاد کرد تا عملکرد بهتری داشته باشد. یکی از این بهبودها استفاده از مجموعه دادگان وسیع‌تر و بزرگ‌تر است. مورد دیگری که قابل بهبود است، مدل‌های زبانی استفاده شده هستند که می‌توان با افزایش داده‌ها[۱۵] و استفاده ترکیبی از مجموعه داده‌ها عملکرد آن‌ها را بهتر کرد. همچنین می‌توان به مرحله ادغام پاسخ‌ها یک مرحله یادگیری اضافه کرد تا بتوان لیست پاسخ‌های محتمل نهایی را به شکل مناسبی رتبه بندی کرد[14]. علاوه بر این موارد میتوان بهبودهایی در الگوریتم جستجو، مانند یافتن توابع ارزیابی و مکاشفه‌ای مناسب‌تر و همچنین هرس بهتر فضای جستجو، ارائه کرد.

مراجع

[۱] جدول‌یاب، حل جدول آنلاین، مرداد ۱۳۹۶، https://www.jadvalyab.ir

[۲] دهخدا، علی‌اکبر، **لغت‌نامه دهخدا**، انتشارات دانشگاه تهران، ۱۳۷۷، (نسخه برخط در آدرس اینترنتی https://dehkhoda.ut.ac.ir)

[3] Ernandes, Marco, Angelini, Giovanni, Gori, Marco, *"WebCrow: a web-based system for crossword solving."*, AAAI'05: Proceedings of the 20th national conference on Artificial intelligence, Vol. 3, 2005

[4] Ginsberg, Matthew L, *"Dr. Fill: Crosswords and an implemented solver for singly weighted CSPs"*, Journal of Artificial Intelligence Research, Vol. 42, Issue 1, 2011

[5] Littman, Michael L., Keim, Greg A., Shazeer, Noam, *"A probabilistic approach to solving crossword puzzles."* Artificial Intelligence, Vol. 134, Issue 1-2, 2002

[6] Wallace E., Tomlin N., Xu A., Yang K., Pathak E., Ginsberg M., Klein D., *"Automated Crossword Solving"*, Annual Meeting of the Association for Computational Linguistics, 2022

[7] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., Monshizadeh, M., Assi, S. M., *"Semi automatic development of farsnet; the persian wordnet,"*, Proceedings of 5th global WordNet conference, Mumbai, India, 2010, vol. 29

[8] Khezry, Behrouz, Asgarian, Ehsan, *"FarsiYar: Persian Text mining and text processing tools"*, 2018, https://text-mining.ir

[9] Fellbaum Ch., *"WordNet: An Electronic Lexical Database."*, MIT Press, 1998
[10] McCandless M., Hatcher E., Gospod-netic O., *"Lucene in Action, Second Edition: Covers Apache Lucene 3.0"*, Manning Publications Co., Greenwich, CT, USA, 2010
[11] Reimers, Nils, Gurevych, Iryna, *"Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks"*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019
[12] Devlin J., Chang M., Lee K., Toutanova K.,*"BERT: Pre-training of deep bidirectional transformers for language understanding."* In NAACL, 2019
[13] Farahani, M., Gharachorloo, M., Farahani, M., Manthouri, M.,*"ParsBERT: Transformer-based Model for Persian Language Understanding"*, Neural Processing Letters, Vol. 53, Issue 6, 2021
[14] Barlacchi G., Nicosia M., Moschitti A., *"Learning to Rank Answer Candidates for Automatic Resolution of Crossword Puzzles"*, Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 2014.

زیرنویس

[1]module
[2]question answering model
[3] loopy belief propagation
[4]Depth First Search (DFS)
[5]heuristic
[6]crawler
[7]elasticsearch
[8]query
[9]language model
[10]semantic search
[11]finetune
[12] Constraint Satisfaction Problem (CSP)
[13]recall
[14]precision
[15]Data augmentation