دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

سیستم هوشمند پرسش و پاسخ مبتنی بر ویکی‌پدیا در حوزه‌ی پزشکی

غزاله خرادپور

دانشجوی کارشناسی رشتهٔ علوم کامپیوتر، دانشکدهٔ علوم کامپیوتر دانشگاه صنعتی امیرکبیر، تهران، ایران

استاد راهنما: دکتر اکبری

چکیده: در این پروژه، سعی بر آن بود که با وجود سیستم پرسش و پاسخ مبتنی بر مقاله‌های ویکی‌پدیای[1] غیرفارسی، داده‌هایی به زبان فارسی –چه به صورت اتوماتیک و چه به صورت دستی– جمع‌آوری شود و سپس به کمک سیستم‌های هوشمند موجود، سیستمی پیاده‌سازی کنیم که کاربران بتوانند با وارد کردن سؤالات پزشکی خود به زبان فارسی، پاسخی درخور همان سوال به همان زبان فارسی دریافت کنند.

کلمات کلیدی: سیستم پرسش و پاسخ، ویکی‌پدیا، پزشکی، دیتا، اتوماتیک، پردازش زبان طبیعی

۱ مقدمه

پرسش و پاسخ یک زمینهٔ تحقیقاتی در رشتهٔ علوم کامپیوتر و در حوزهٔ بازیابی اطلاعات و پردازش زبان طبیعی[2] است که هدف از آن طراحی سیستم‌هایی است که به‌طور خودکار به سوالات مطرح شده توسط انسان در ساختار زبان طبیعی پاسخ تولید می‌کند. پیاده‌سازی سیستم پرسش و پاسخ، معمولاً به‌صورت یک برنامهٔ رایانه‌ای است که پاسخ‌های خود را با پرس‌وجو از یک پایگاه دادهٔ ساخت‌یافته از دانش یا اطّلاعات، معمولاً یک پایگاه دانش ایجاد می‌کند. به طور معمول، سیستم‌های پرسش و پاسخ می‌توانند پاسخ‌ها را از یک مجموعهٔ بدون ساختار از اسناد زبان طبیعی دریافت کنند. از سری اسناد زبان طبیعی که برای دریافت جواب سؤالات از آن استفاده می‌شود، می‌توان به مجموعه‌ای محلّی از متون مرجع اسناد و صفحات وب سازمان داخلی، گزارش‌های خبری خبرگزاری‌ها، مجموعه‌ای از صفحات ویکی‌پدیا و زیرمجموعه‌ای از صفحات وب جهانی اشاره کرد. برای دریافت جواب‌ها از اسناد دو نوع منبع دامنه‌محدود و دامنه‌باز وجود دارد که هر کدام شرایطی دارند و انتخاب ما اسناد دامنه‌محدود بوده است به این معنی که با یک دامنهٔ خاصّی از موضوعات سروکار دارند و موضوع مورد نظر ما حوزهٔ پزشکی بوده است. در آینده آن را به دامنه‌باز تبدیل خواهیم کرد.

حال سیستم پرسش و پاسخ هوشمند در حوزهٔ پزشکی می‌تواند کمک بزرگی برای کاربران باشد از آن جا که پزشک و اطّلاعات پزشکی در اختیار تمام مردم نیست و حتّی اگر در دسترس هم باشد، ممکن است در یافتن جواب‌های خود دچار گمراهی شوند. افراد با حلّ مشکلات پزشکی سطحی خود، امکانات محدود پزشکی را در اختیار افرادی قرار می‌دهند که به آن نیاز شدیدی دارند. از طرفی جواب دادن به سؤالات پزشکی نیازمند دانستن دانش بسیار است و هر فردی در شرایط عادی آن‌ها را نمی‌داند و وقت زیادی ندارد تا منابع مختلف را مطالعه کند، این سیستم می‌تواند بسیار کمک‌کننده باشد. علاوه بر این پزشکان نیز می‌توانند صحّت تشخیص و دانسته‌های خود را از این طریق تثبیت کنند و دانش پزشکی و ارتباط با بیماران به سمت هوشمندتر شدن می‌رود. بنابراین سیستمی که بتواند به پرسش‌های کاربران پاسخ مناسب برگرداند، بسیار خدمت بزرگی است و برای آن‌ها کم‌هزینه است. تنها با وارد کردن سوالات خود به زبان طبیعی، به جواب متناسب می‌رسند.

ما برای پیدا کردن جواب سؤالات پاسخ‌بلند خود از ویکی‌پدیا به عنوان منبع دایرةالمعارف‌گونه استفاده می‌کنیم. ویکی‌پدیا منبع قدرتمندی است که اطّلاعات به‌روزی دارد و هر روز بهتر می‌شود؛ چون توسّط عدّهٔ زیادی از مردم نوشته و ویرایش می‌شود، کم‌تر حاوی اطّلاعات غلط است و می‌توان سیستم‌های هوشمندی روی آن پیاده کرد.

---

[2] NLP

	question	answer	link	paragraph_text	text	title	categories
0	دلیل ایجاد اسهال چرب چیست؟	علت آن عدم جذب صحیح چربی‌ها در روده باریک است....	https://fa.wikipedia.org/wiki/%D8%A7%D8%B3%D9%...	اسهال به طور کلی، به عنوان افزایش وزن مدفوع (ب...	اسهال چرب (انگلیسی: Steatorrhea یا چربپیخال...	اسهال چرب	['رده:Infobox medical condition', 'رده:اختلالا...]
1	عوارض اسهال چرب چیست؟	اسهال چرب، سبب ایجاد اختلال شدید در فرآیند جذب...	https://fa.wikipedia.org/wiki/%D8%A7%D8%B3%D9%...	اسهال چرب با حجم زیاد و براق از عوارض اسهال‌ها...	اسهال چرب (انگلیسی: Steatorrhea یا چربپیخال...	اسهال چرب	['رده:Infobox medical condition', 'رده:اختلالا...]
2	درمان واریس مری چگونه انجام می‌شود؟	درمان با اتساع بالون یا تزریق واژوپرسین وریدی ...	https://fa.wikipedia.org/wiki/%D9%88%D8%A7%D8%...	درمان با اتساع بالون یا تزریق واژوپرسین ... وریدی	به) واریس مری (انگلیسی: Esophageal varices)، وا...	واریس مری	['رده:Infobox medical condition', 'رده:بیماری'\...]
3	خطر اصلی و عوارض واریس مری چیست؟	خطر اصلی واریس مری، پارگی عروق متسع و خونریزی ...	https://fa.wikipedia.org/wiki/%D9%88%D8%A7%D8%...	به) واریس مری (انگلیسی: Esophageal varices)، وا...	به) واریس مری (انگلیسی: Esophageal varices)، وا...	واریس مری	['رده:Infobox medical condition', 'رده:بیماری'\...]
4	عوارض جانبی میزوپرستول چیست؟	سردرد، تهوع، کرامپ رحمی، دل درد، اسهال، نفخ و ...	https://fa.wikipedia.org/wiki/%D9%85%DB%8C%D8%...	سردرد، تهوع، کرامپ رحمی، دل درد، اسهال، نفخ و ...	به) میزوپروستول (انگلیسی: Misoprostol) ترکیبی س...	میزوپرستول	['رده:بیماری'\u200cهای گوارشی', 'رده:داروهای ضر...
...	...	...	...	...	...	...	...

شکل ۱: نمونه‌ای از دادهٔ جمع‌آوری‌شده

۲ منابع داده

منابع اصلی ما داده‌های تمامی صفحات ویکی‌پدیا و داده‌های جمع‌آوری‌شده بودند که از داده‌های ویکی‌پدیا به عنوان منبعی برای استخراج جواب استفاده می‌کنیم و از داده‌های جمع‌آوری‌شده برای آموزش دادن مدلی که قادر است بهترین جواب برای سؤال پرسیده‌شده را برگرداند. چالش اصلی کار ما، پیدا کردن دادهٔ مناسب به زبان فارسی بود و از آن جا کار مشابهی قبلاً به زبان فارسی انجام نشده است و به طور کلّی در پردازش زبان طبیعی به فارسی پیشینهٔ خاصّی وجود ندارد، دادهٔ مناسبی هم در اختیار نداشتیم. چندین منبع برای این نوع مجموعه‌داده‌ها وجود داشتند که هم محدود بودند و هم ساختاری به شکل آن چه ما می‌خواستیم نداشتند. کاری که ما می‌خواستیم انجام دهیم، چون پردازش زبان طبیعی بود و باید از گونه‌ای از شبکهٔ عصبی بهره می‌گرفتیم، به دیتای بسیار زیادی برای آموزش نیاز داشتیم تا به هدفی که برای خود تعیین کرده بودیم، برسیم.

۲-۱ ویکی‌پدیا

از ویکی‌پدیا به عنوان منبع دانش استفاده می‌شود. حال برای استخراج تمامی مقالات ویکی‌پدیا باید حدود ۵ میلیون صفحه استخراج شود و تمامی شکل‌ها و ساختارهای مربوطه از بین برود و فقط متن به تنهایی در نظر گرفته و ذخیره شود.

۲-۲ مجموعه‌داده‌های جمع‌آوری‌شده

برای این که بتوانیم مدلی آموزش دهیم تا نتیجهٔ نزدیک به مطلوب را به ما بدهد، نیاز به داده‌های بسیار زیادی از زبان فارسی داشتیم و از آن جا که مجموعه‌داده‌ها –خصوصاً در حوزهٔ پزشکی– بسیار کم بود، شروع به جمع‌آوری هزار ردیف دادهٔ فارسی –با این که باز هم مقدارش کم بود– کردیم. این دیتاها به صورت دستی از سایت ویکی‌پدیا استخراج شدند و ستون‌های آن شامل سؤال، جواب، متن پاراگراف (زمینه[۳])، تیتر مقاله و متن کل مقاله است. از متن پاراگراف و کلّ مقاله برای یاد دادن نحوهٔ استخراج جواب از زمینه به ماشین استفاده می‌کنیم. داده‌ها دارای ساختاری به شکلی زیر هستند:

- سؤال: سؤالی بر اساس مطلب موجود در ویکی‌پدیا طرح شده است.
- جواب: جواب سؤال طرح‌شده به صورت عینی از قسمتی از پاراگراف‌های یکی از مطالب ویکی‌پدیا نوشته شده است.
- زمینه: پاراگرافی که از آن جواب استخراج شده، ذخیره شده است. این ستون از داده‌ها برای آموزش مدل بسیار حائز اهمیّت است.
- تیتر: تیتری مطلبی از ویکی‌پدیا که سؤال و جواب از آن گرفته شده است.
- متن کل: متن تمام مقالهٔ مورد نظر در ویکی‌پدیا ذخیره شده است.
- رده‌ها: رده‌هایی که در قسمت پایینی صفحهٔ ویکی‌پدیا به صورت لیستی قرار داده شده‌اند، ذخیره شده است.

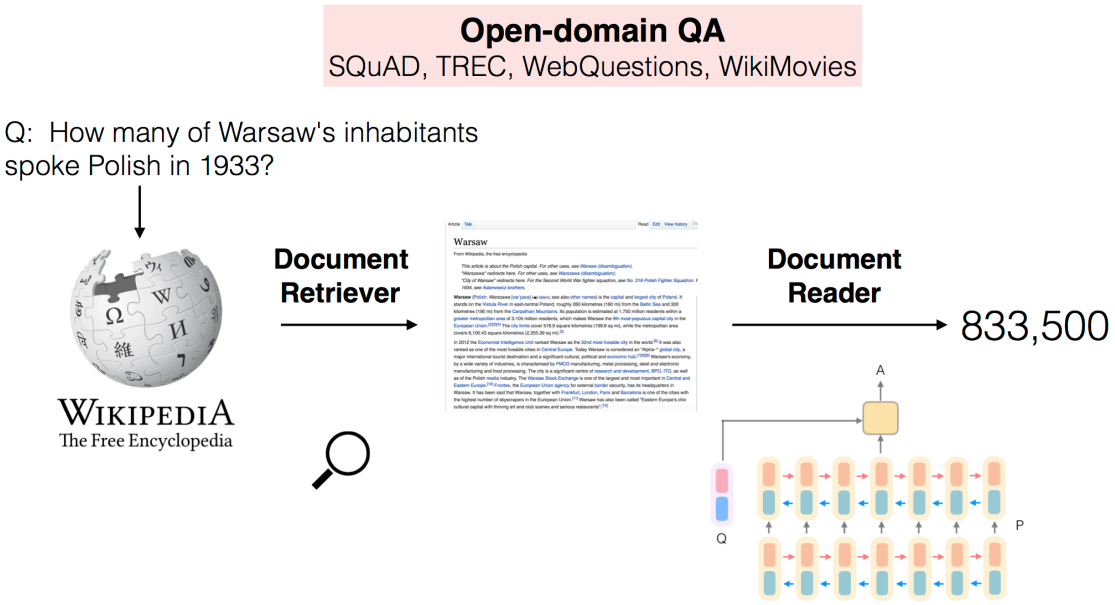
لازم به ذکر است که سه مورد انتهایی مجموعه‌داده‌ها به صورت اتوماتیک و با استفاده از کد استخراج شده است. نمونه‌ای داده‌ها در شکل ۱ آمده‌اند

۲-۳ مجموعه‌داده‌های ترجمه‌شده

از آن‌جا داده‌های جمع‌آوری شده به صورت دستی کافی نبودند، شروع به ترجمهٔ مجموعه‌دادهٔ SQuAD [۱] کردیم و با این که تقریباً ربط زیادی به حوزهٔ پزشکی نداشتند و API سیستم مترجم گوگل[۴] قادر به ترجمهٔ تمام و کمال مجموعه‌داده‌ها نبود، این داده‌ها هم بلااستفاده ماندند و در مراحل بعدتر و برای بهتر کردن مدل آموزش‌داده‌شده به ماشین از آن‌ها استفاده خواهیم کرد. علاوه بر خود مجموعه‌داده‌های این مجموعه، ستون‌هایی به آن‌ها اضافه کردیم که شامل ترجمه‌هایی به زبان فارسی هستند و سوال، جواب، تیتر مقاله و زمینه به زبان فارسی ترجمه شده‌اند.

[۳] context

[۴] Google Translate



شکل ۲: شیوهٔ کار سیستم DrQA

۳ مروری بر کارهای پیشین

برای سیستم پرسش و پاسخ به زبان فارسی، هیچ کاری انجام نشده است چون همان‌طور که قبلاً اشاره کردیم، مشکل کمبود داده و پیچیدگی‌های زبان فارسی چالش‌های زیادی را ایجاد می‌کند. به زبان‌های دیگر خصوصاً انگلیسی کارهای زیادی انجام شده است.

کارهای مشابه دیگری به عنوان مثال توسّط Ryu et al. [۲] انجام شده است که سیستم پرسش و پاسخ دامنه‌بازی طرّاحی کرده است که از ویکی‌پدیا به عنوان منبع مدل خود استفاده می‌کند و متن مقاله‌ای را با جواب‌های منطبق دیگر مانند جعبه‌های اطّلاعات [۵]، ساختار مقاله، ساختار دسته‌بندی و تعاریف ترکیب می‌کند. به طور مشابه Ahn et al. [۳] هم که از ویکی‌پدیا به عنوان منبع متن با ترکیب دیگر منابع استفاده می‌کند. علاوه بر آن، Buscaldi and Rosso [۴] از ویکی‌پدیا به عنوان منبع دانش خود استفاده می‌کند امّا به جای آن که جواب‌ها را از روی آن بیابد، از آن برای ارزیابی کردن جواب تولیدشده توسّط خود سیستم استفاده می‌کند و از دسته‌بندی‌های ویکی‌پدیا برای اطمینان از مجموعه الگوهایی که برای جواب پیش‌بینی شده، بهره می‌برد.

سیستم‌های پیشرفته‌تری از پرسش و پاسخ وجود دارند که از ویکی‌پدیا و وب استفاده می‌کنند؛ مانند QuASE، Microsoft's AskMSR [۵]، IBM's DeepQA و YodaQA [۶] که هر کدام از روش‌های خاصّ خود برای پیاده‌سازی این سیستم استفاده کرده‌اند.

که یکی از بزرگ‌ترین و مشابه‌ترین کارها به کار ما، پروژه‌ای به نام DrQA[۷] است و کاری را انجام داده است که برنامه داشتیم انجام دهیم. بدنهٔ این پروژه به گونه‌ای است که با وارد کردن دادهٔ فارسی می‌توان مدلی را آموزش داد و از نتیجهٔ آن استفاده کرد. شکل کلّی این پروژه در شکل ۲ نشان داده شده است.

۳-۱ DrQA چیست؟

ک سیستم درک مطلب از طریق خواندن است که از ویکی‌پدیا به عنوان تنها منبع خود برای یافتن جواب استفاده می‌کند به طوری که انگار شخصی از روی دایرةالمعارف جواب سؤالات را می‌دهد. البتّه این سیستم به گونه‌ای طرّاحی شده است که می‌توان هر منبعی را به عنوان منبع پیدا کردن جواب سؤالات برای آن انتخاب کرد و قابلیّت مقیاس‌پذیری دارد. این سیستم پس از دریافت سؤال، به دنبال جوابی برای آن سؤال می‌گردد. تمرکز این پروژه بر روی سؤال و جواب‌های طولانی (نه یک کلمه‌ای) است و در این مسیر با چالش‌هایی هم‌چون پیدا کردن مقاله‌های مرتبط و پیدا کردن جواب از روی آن‌ها دست و پنجه نرم می‌کند. این پروژه از ویکی‌پدیای انگلیسی –از آن جا که بسیار غنی و باجزییات است– به عنوان منبع جواب‌های خود استفاده می‌کند.

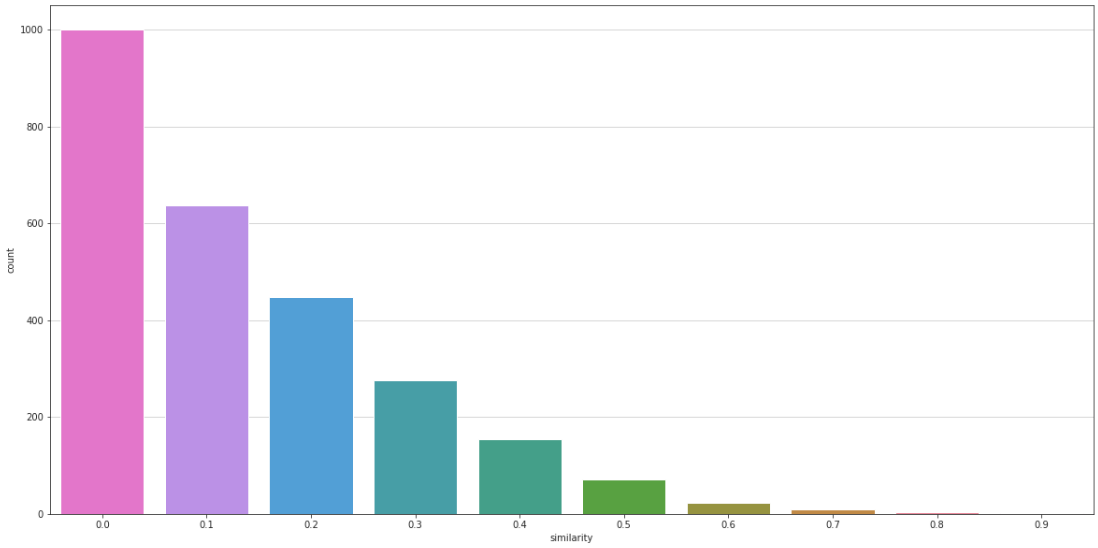
DrQA از دو قسمت اصلی تشکیل شده است که قسمت اوّل آن تمام مقالات ذخیره‌شده را می‌خواند و از بین آن ۵ مقاله‌ای پتانسیل دارا بودن جواب را در خود دارند، بر می‌گرداند و قسمت دوّم آن همان هستهٔ اصلی است که مدل روی آن آموزش دیده است.

این پروژه از شبکهٔ عصبی مکرر چندلایه[۶] که در حالت پیش‌فرض روی مجموعه‌داده‌های SQuAD آموزش داده شده است، به عنوان پردازش‌کنندهٔ مطالب استفاده می‌کند و در نهایت جواب‌های منتخب را با امتیاز حساب شده توسط مدل ارائه می‌دهد و بهترین آن را به عنوان جواب برمی‌گرداند.

- Document Retriever: تمام داده‌ها و مقالات ویکی‌پدیا

[۵] infoboxes

[۶] multi-layer recurrent neural network machine comprehension



شکل ۳: شباهت بین سؤال‌ها و جواب‌های یک ردیف در دادهٔ جمع‌آوری‌شده

به صورت ماتریس‌های وزن‌داده‌شده TF-IDF ذخیره می‌شوند. پس از آن که سؤال دریافت شد، توسط ادغام TF-IDF و n-grams –حتّی سریع‌تر و با دقّت بهتر از موتور جست‌وجوی خود ویکی‌پدیا– مقالات مرتبط را پیدا می‌کند.

- Document Reader: پس از کدگذاری پاراگراف‌ها (با استفاده از الگوریتم LSTM[۷])، استفاده از شبکهٔ عصبی مکرّر چندلایه کدگذاری سؤال و کدگذاری متن سؤال، جواب منتخب را از بین پاراگراف‌های مقالات منتخب مرحلهٔ قبل حدس می‌زند.

۴ پیش‌پردازش‌های انجام‌شده روی داده‌ها

بخش اعظم این پروژه جمع‌آوری داده‌ها بوده است که برای درک بهتر ساختارها و اجزای آن یک سری پیش‌پردازش‌ها انجام داده‌ایم تا بدانیم با چه جنس داده و با چه توزیعی از کلمات مواجه هستیم. یکی از کارهای اصلی‌ای که برای پیش‌پردازش داده‌ها در فرآیند پردازش زبان طبیعی انجام می‌شود، حذف کلمات کلیدی [۸] است که انجام دادن آن در این پروژه بی‌معنی است. تنها کلمات غیرفارسی را از آن حذف کردیم و الگوهای زبانی را در آن بهبود بخشیدیم. دو نمونه از آمارهایی که در مورد مجموعه‌داده‌ها رسم کرده‌ایم، در شکل ۴ و ۳ آمده‌اند.

برای این که ارتباط بین سؤالات و جواب‌های هر ردیف از داده‌ها را بدانیم، با استفاده از روش TF-IDF متن سؤال‌ها و جواب‌ها به بردار تبدیل کردیم و سپس با روش Cosine Similarity شباهت بین سؤال‌ها و جواب‌های هر ردیف را بررسی کردیم. سؤال‌ها و جواب‌ها تا حدی خوبی به هم شبیه بودند و ارتباط بین‌شان مشهود بود از آن جا امتیاز شباهتشان با روش مذکور، به طور متوسط روی ۰/۳ بود و این عدد نسبتاً خوبی است.

شکل ۴: ابر کلمات استفاده شده در متن‌های مجموعه‌داده‌های جمع‌آوری‌شده

۵ نتیجه

این پروژه با بیشتر شدن داده‌ها، به امتیاز بیشتری خواهد رسید و هدف ما این است که این پروژه را در زمینه‌های دیگری به غیر از پزشکی نیز گسترش دهیم و در مراحل بعدتر به الگوریتم‌های موجود در این زمینه بهبود بخشیم.

---

[۷] long short term memory

[۸] stopwords

مراجع

[1] K. L. P. L. Pranav Rajpurkar, Jian Zhang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” 2017.

[2] H.-K. K. Pum-Mo Ryu, Myung-Gil Jang, “Open domain question answering using Wikipedia-based knowledge model,” 2014.

[3] G. M. K. M. M. d. R. S. S. David Ahn, Valentin Jijkoun, “Using Wikipedia at the TREC QA Track.,” 2004.

[4] P. R. Davide Buscaldi, “Mining knowledge from Wikipedia for the question answering task,” 2006.

[5] M. B. Eric Brill, Susan Dumais, “An Analysis of the AskMSR Question-Answering System,” 2002.

[6] J. □. Petr Baudiš, “Modeling of the Question Answering Task in the YodaQA System,” 2015.

[7] J. W. A. B. Danqi Chen, Adam Fisch, “Reading Wikipedia to Answer Open-Domain Questions,” 2017.