استدلال در بحث کاوی

علی اصغر تقی‌زاده
دانشکده مهدسی کامپیوتر
دانشگاه علم و صنعت ایران
a_taghizadeh@comp.iust.ac.ir

چکیده

مباحثه به عنوان یکی از موضوعات مطرح شده در زبان‌شناسی تاریخچه‌ی بسیار طولانی‌ای دارد. هدف از بررسی مباحثه، تجزیه آن به عناصر تشکیل دهنده و تحلیل آن‌هاست و هدف از بحث کاوی انجام این تجزیه و تحلیل از طریق مدل‌های مبتنی بر هوش مصنوعی است. یکی از موضوعات چالش برانگیز اخیر در بحث کاوی درک استدلال می‌باشد. درک استدلال به این معنی است که مدل هوش مصنوعی بتواند روابط منطقی را به درستی تشخیص دهد و تصمیم منطقی اتخاذ کند. مسئله مطرح شده در این باره در بحث کاوی به این صورت است که ادعا و دلیل در ارتباط با یک موضوع آورده می‌شود و مدل باید بتواند فرضیه‌ی درست را از بین دو فرضیه‌ی ورودی تشخیص دهد. در این گزارش به بررسی استفاده یادگیری انتقالی از مجموعه دادگان ساخته شده برای یادگیری منطق، به منظور حل این مسئله پرداخته خواهد شد.

۱ مقدمه

مباحثه یکی از موضوعات مورد علاقه فلاسفه بوده و تاریخچه آن به دوران ارسطو برمی‌گردد. در تعریف مباحثه آمده است که:"مباحثه یک فعالیت زبانی، اجتماعی و عقلانی است که هدف آن متقاعد کردن در ارتباط با صحیح بودن یک نقطه نظر با استفاده از ترکیبی از عبارات به منظور توجیه کردن یا رد کردن عبارت گفته شده در نقطه نظر است."

علاوه بر فلسفه، مباحثه در علم زبان‌شناسی هم مورد توجه بوده است. در سالهای اخیر مدل تولمین[۱] به عنوان یکی از مدل‌های مطرح برای تجزیه‌ی مباحثه به عناصر سازنده‌ی آن مطرح شده است. این عناصر عبارت‌اند از ادعا[۲]، فرضیه[۳]، فرضیه‌های پنهان [۴]، توصیف‌کننده[۵]، رد کننده[۶] و پشتیبان[۷] است.

در سالهای اخیر مباحثه مورد توجه محققان پردازش زبان طبیعی نیز بوده است. به دلیل شباهات برخی از کارهای پیشین نمی‌توان زمان دقیقی برای مطرح شدن بحث کاوی در نظر گرفت. اما می‌توان کارهای پروفسور گورویخ[۸] را به عنوان یکی از مهم‌ترین فعالیت‌های اخیر در نظر گرفت که چارچوب بحث کاوی را شکل داده است.

از موضوعات تعیین شده در بحث‌کاوی می‌توان به تشخیص متن‌های دارای بحث، تشخیص عناصر بحث و روابط بین آن‌ها اشاره کرد. یکی از موضوعاتی که در SemEval ۲۰۱۸ مطرح شده است، بررسی استدلال در بحث‌کاوی است [۱]. استدلال که در سال اخیر بسیار مورد توجه محققان هوش مصنوعی بوده است، در بسیار از موضوعات می‌توان تعریف کرد که به عنوان نمونه به پرسش و پاسخ می‌توان اشاره کرد. منظور از استدلال این است که مدل شبکه‌ی عصبی بتواند در پاسخی که می‌دهد روابط

[1] Toulmin
[2] claim
[3] premise
[4] warrant
[5] qualifier
[6] rebuttal
[7] backing
[8] https://www.informatik.tu-darmstadt.de/ukp/ukp_home/staff_ukp/prof_dr_iryna_gurevych/index.en.jsp

منطقی‌ای که در ورودی داشته است را نقض نکند. برای مثال در جمله‌ی « علی از حسین بلند تر است و حسین از حسن بلند تر است » اگر از مدل پرسیده شود « علی از حسن بلندتر» باید مدل بتواند به عنوان جمله‌ی درست این جمله را تشخیص بدهد.

موضوعی که در ارتباط با استدلال در بحث کاوی تعیین شده است تشخیص فرضیه‌ی پنهان درست از نادرست است. در این موضوع ادعا و دلیل به همراه دو فرضیه‌ی که یکی از آن‌ها درست آورده شده است. مدل باید بتواند فرضیه‌ی درست را تشخیص دهد. اما مجموعه دادگان اولیه‌ای که برای این موضوع مطرح شده بود، توسط نایون و همکاران [۲] مورد بررسی قرار گرفت و مشخص شد که عملکرد خوب مدل‌ها در آن مجموعه داده به خاطر متعادل نبودن کلمات در بین نمونه‌ها هست. با ساختن مجموعه دادگان جدید چالش برانگیز از روی مجموعه دادگان ارائه شده عملکرد بسیاری از مدل‌ها کاهش چشمگیری پیدا کرده است .هدف از انجام این پروژه بهبود عملکرد بر روی این مجموعه دادگان جدید است.

در قسمت‌های بعدی ابتدا کارهای انجام شده در بحث کاوی مورد بررسی خواهند گرفت. سپس مدل پیشنهاد شده بررسی خواهد شد و بعد از آن نتایج به دست آمده ارائه خواهد شد. در نهایت به تحلیل کارهای انجام شده خواهیم پرداخت.

۲ کارهای مرتبط / پیش‌زمینه

در این بخش ابتدا به کارهای انجام شده در بحث کاوی خواهیم پرداخت و سپس کارهای انجام شده در زمینه استدلال و در نهایت دادگان چالشی را بررسی خواهیم کرد.

۱.۲ بحث کاوی

در SemEval 2018 مسئله‌ی درک استدلال در بحث کاوی که من بعد تحت عنوان ARCT از آن یاد می‌شود، به همراه یک مجموعه داده معرفی شد. یک نمونه از این مجموعه دادگان را در جدول ۱ مشاهده می‌کنید.

جدول ۱: یک نمونه از مجموعه دادگان ARCT

debateTitle	Do Home Schoolers Deserve a Tax Break?
debateInfo	Some conservatives want a federal credit for families who teach their children at home. What are its chances in the new Congress?
claim	Home schoolers do not deserve a tax break.
reason	facts over values religious emphasizes and education public undermines It co-exist. easily can two the though even
warrant 1	exist co easily so can two the
warrant 2	exist co hardly two the
correct warrant	2

در آزمایشات انجام شده برای این مجموعه دادگان معمولا ادعا و دلیل به همراه یکی از فرضیات به مدل داده می‌شود و مدل باید تشخیص دهد این فرضیه با ادعا و دلیل سازگار است یا نه. این نوع طرح‌بندی، مسئله بسیار شبیه به موضوع استنتاج در زبان طبیعی [۹] می‌کند. به همین دلیل بسیاری از راهبردهای موجود برای این موضوع برای ARCT هم کارگشا خواهد بود. چوی و همکاران [۳] از مدلی که تحت عنوان ESIM [۴] برای موضوع NLI مطرح شده بود استفاده کرده‌اند و به نوعی با تعریف شبکه در ساختار NLI از دادگان قبلی موجود برای این موضوع استفاده کرده و در واقع از یادگیری انتقالی بهره برده‌اند.

تیان و همکاران [۵] یک مدل شبکه‌ی عصبی انتها به انتها برای این مسئله ارائه داده‌اند که از مکانیزم توجه استفاده می‌کند. همچنین بوچن و همکاران [۶] از دانش موجود در گراف‌های دانش Fram-Net و Entity-Net استفاده کرده‌اند که هر کدام از روش‌های گفته شده توانسته‌اند موفق عمل کنند.

مدلی که تا به حال بهترین عملکرد برای ARCT داشته‌است مدل مبتنی برت بوده که به دقت ۷۳ درصد رسیده است [۲]. اما تحقیقات نایون و همکاران [۲] نشان داده است که این دقت بالا به علت توزیع نامتعادل ویژگی‌ها در بین کلاس‌ها هست که اگر این توزیع متعادل شود دقت مدل برت به ۵۳ درصد در بهترین حالت افت می‌کند.

۲.۲ استدلال

می‌توان گفت یکی از شگفتی‌های مدل‌های ترنسفورمر توانایی آن‌ها در یادگیری روابط منطقی زبان طبیعی است. طبق آزمایشات [۷] و همکاران مشخص شده است که مدل ترنسفرمر قادر است با دقت ۹۹ درصد به سوالات منطقی ساده درست پاسخ دهد.

[9] Natural Language Inference

نمونه‌های آزمایش‌های آن‌ها که به صورت ساختگی است شامل یک جهان بسته است که هر بار ترکیبی از این جهان ساخته می‌شود و چند سوال در ارتباط با آن جهان پرسیده می‌شود. یک نمونه از این مجموعه داده در تصویر ۱ قابل مشاهده است.

Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

If someone is young and round then they are kind.
Big people are rough.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

شکل ۱: نمونه‌ای از مجموعه دادگان آزمایشات کلارک و همکاران برای یادگیری روابط منطقی

ریچارد و همکاران [۸] نیز آزمایش مشابهی انجام داده‌اند. آن‌ها روابط منطقی را به چند دسته تحت عنوان بخش‌های معنایی تقسیم کرده‌اند که هر بخش بیانگر یک نوع رابطه‌ی منطقی است. برای مثال در یک بخش فقط نمونه‌هایی از روابط دودویی[۱۰] وجود دارد. چند نمونه از مجموعه دادگان ساخته شده توسط آن‌ها به همراه بخش‌های منطقی در تصویر قابل مشاهده است. دست‌آوردهای ریچارد و همکاران نیز این نکته را تایید می‌کند که ترنسفرمرها قادر به یادگیری روابط منطقی هست و به بیانی قابلیت استدلال را دارند. در بخش بعدی به ارائه‌ی راه حل بر اساس این قابلیت خواهیم پرداخت.

Fragments	**Example** (`premise,label,hypothesis`)	**Genre**	**Vocab. Size**	**# Pairs**	**Avg. Sen. Len.**
`Negation`	*Laurie has only visited Nephi, Marion has only visited Calistoga.* `CONTRADICTION` *Laurie didn't visit Calistoga*	Countries/Travel	3,581	5,000	20.8
`Boolean`	*Travis, Arthur, Henry and Dan have only visited Georgia* `ENTAILMENT` *Dan didn't visit Rwanda*	Countries/Travel	4,172	5,000	10.9
`Quantifier`	*Everyone has visited every place* `NEUTRAL` *Virgil didn't visit Barry*	Countries/Travel	3,414	5,000	9.6
`Counting`	*Nellie has visited Carrie, Billie, John, Mike, Thomas, Mark, .., and Arthur.* `ENTAILMENT` *Nellie has visited more than 10 people.*	Countries/Travel	3,879	5,000	14.0
`Conditionals`	*Francisco has visited Potsdam and if Francisco has visited Potsdam then Tyrone has visited Pampa* `ENTAILMENT` *Tyrone has visited Pampa.*	Countries/Travel	4,123	5,000	15.6
`Comparatives`	*John is taller than Gordon and Erik..., and Mitchell is as tall as John* `NEUTRAL` *Erik is taller than Gordon.*	People/Height	1,315	5,000	19.9
`Monotonicity`	*All black mammals saw exactly 5 stallions who danced* `ENTAILMENT` *A brown or black poodle saw exactly 5 stallions who danced*	Animals	119	10,000	9.38
`SNLI+MNLI`	*During calf roping a cowboy calls off his horse.* `CONTRADICTION` *A man ropes a calf successfully.*	Mixed	101,110	942,069	12.3

شکل ۲: نمونه‌ای از مجموعه دادگان آزمایشات ریچارد و همکاران برای یادگیری روابط منطقی

۳.۲ دادگان چالشی

تحقیقات نایون و همکاران [۲] نشان داده است که دقت بالای برت در ARCT به علت توزیع نامتعادل ویژگی‌ها در بین کلاس‌ها هست که اگر این توزیع متعادل شود دقت مدل برت به ۵۳ درصد در بهترین حالت افت می‌کند.

نحوه‌ی ساخت مجموعه دادگان چالش برانگیز به این صورت بوده است که در مجموعه دادگان اولیه با ادعا با یکی از فرضیات سازگار بوده است و اگر ادعا معکوس شود باید با فرضیه‌ی دیگر سازگار شود و مدل باید بتواند این موضوع را تشخیص داده و فرضیه‌ی درست را انتخاب کند. یک نمونه از این مجموعه دادگان در تصویر ۳ قابل مشاهده است.

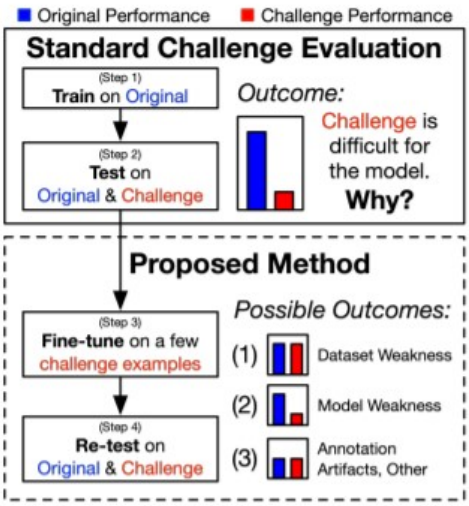
ساخت مجموعه دادگان چالشی فقط منحصر به بحث کاوی نمی‌شود و طبق گفته‌ی لیو و همکاران [۹] اخیرا چندین مجموعه دادگان چالشی ارائه شده است که مدل‌هایی که بر روی دادگان اصلی عملکرد خوبی داشته‌اند بر روی دادگان چالشی افت چشمگیری پیدا کرده‌اند. لیو و همکاران با بررسی این موضوع سعی داشته‌اند علت کاهش عملکرد را توجیه کنند. نتایج آن‌ها در تصویر ۴ قابل مشاهده است.

[10] boolean

Claim	Google is not a harmful monopoly
Reason	People can choose not to use Google
Warrant	Other search engines don't redirect to Google
Alternative	All other search engines redirect to Google

Reason (and since) **Warrant** $\rightarrow$ **Claim**
Reason (but since) **Alternative** $\rightarrow \neg$ **Claim**

شکل ۳: نمونه‌ای از مجموعه دادگان ARCT. فرضیه‌ی اول با ادعا سازگار است و فرضیه‌ی دوم با معکوس ادعا



شکل ۴: توضیح کلی درباره‌ی نحوه‌ی تشخیص علت کاهش عملکرد بر روی مجموعه دادگان چالشی طبق تحقیقات لیو و همکاران [۹]

لیو و همکاران[۹] پیشنهاد داده‌اند برای فهم علت کاهش عملکرد، ابتدا باید روی یک بخش یا کل مجموعه دادگان چالشی مدل قبلی آموزش ببیند و سپس دوباره بر روی بخش آزمون از روی آن مجموعه داده تست شود. در این صورت سه حالت پیش خواهد آمد: ۱- در صورتی که نتیجه بر روی مجموعه چالشی بهبود پیدا کرد نشان دهنده ضعیف بودن مجموعه دادگان اولیه است. ۲- در صورتی که همچنان عملکرد مدل ضعیف باشد نشان دهنده ضعیف بودن مدل است و ۳- در صورتی که عملکرد بر روی دادگان اولیه هم افت پیدا کرد نشان دهنده وجود اشکال در مجموعه دادگان است.

در بخش بعدی به بررسی مدل پیشنهاد شده خواهیم پرداخت.

۳ مدل پیشنهاد شده

با بررسی دادگان می‌توان به این نتیجه رسید که مدل در تشخیص تفاوت‌های بین یک متن و معکوس آن دچار اشتباه شده است. این امر شاید به علت کم بودن دادگان باشد (۱۲۱۰ نمونه آموزش). در همین راستا آزمایشات کلارک و همکاران [۷] جالب توجه به نظر می‌آید. چرا که در هر ترکیب منطقی از متن‌های تولید شده با کلمات محدود مدل قادر خواهد بود پاسخ درست تولید کند. به همین علت می‌توان با استفاده از یادگیری انتقالی و استفاده از دانش موجود در مجموعه دادگان کلارک و همکاران [۷] می‌توان قدرت استدلال مدل را بهبود داد.

از طرفی ممکن است علت عملکرد پایین برت در مجموعه دادگان چالشی به خاطر این موضوع باشد که برت یک دانش توزیع شده از زبان را یاد می‌گیرد و ممکن است استفاده از گراف‌های دانش در برت کمک کننده باشد. از جمله روش‌هایی که به ترکیب گراف دانش و برت پرداخته‌اند می‌توان به KnowBert [۱۰] اشاره کرد که علاوه بر توابع هزینه اصلی برت یک تابع هزینه نیز به آن اضافه می‌شود تا دانش موجود در گراف‌های دانش نیز در برت تعبیه شود. با اینکه روش‌های دیگری نیز برای ترکیب گراف دانش با برت وجود داشت اما فقط آزمایش با KnowBert به زمان تحویل پروژه رسید.

در بخش بعدی نتایج حاصل از آزمایشات را بررسی خواهیم کرد.

۴ نتایج

ابتدا آزمایش‌هایی برای تکرار نتایج در مجموعه دادگان ARCT انجام شد. اطلاعات این مجموعه دادگان در جدول ۲ قابل مشاهده است.

جدول ۲: تعداد نمونه‌های موجود در مجموعه دادگان ARCT

آموزش	اعتبار	آزمون
۱۲۱۰	۳۱۶	۴۴

سه مدل مختلف از ترنسفرمرها شامل برت پایه[11]، آلبرت نسخه ۲[12] و دیستیل برت[13] برای آزمایشات انجام شد. دو مدل آخر ذکر شده به خاطر محدودیت منابع انتخاب شدند چرا که در مقایسه با برت پایه پارامترهای کمتری دارند. نتایج حاصل از اجرای آزمایشات به همراه بهترین نتیجه موجود (SOTA) در جدول ۳ قابل مشاهده است.

جدول ۳: نتایج آزمایشات بر روی مجموعه دادگان ARCT. اعداد بیانگر صحت هستند.

SOTA	albert	distilbert	bert
۷۳%	۵۰%	۷۳%	۶۵%

همانطور که مشاهده می‌شود عملکرد آلبرت از بقیه مدل‌ها ضعیف‌تر بوده است. علت آن را می‌توان در ظرفیت پایین این مدل نسبت به بقیه در نظر گرفت.

با اینکه بهترین مدل هم از نوع برت پایه بوده است اما عملکرد آزمایش انجام شده ضعیفتر بوده که علت آن را می‌توان در محدودیت منابع برای آموزش برت جستجو کرد.

همچنین مدل دیستیل برت همانند بهترین مدل موجود عمل کرده است.

اما طبق گفته‌ی لیو و همکاران [۹] باید مدل را بر روی مجموعه جدید نیز آموزش داد و نتایج را بر روی مجموعه آزمون آن ارزیابی کرد. نتایج آزمایشات انجام شده در جدول ۴ قابل مشاهده است.

جدول ۴: نتایج آزمایشات بر روی مجموعه دادگان چالشی. اعداد بیانگر صحت هستند.

SOTA	albert	distilbert	bert
۵۳%	۵۰%	۵۵%	۴۹%

همانطور که مشاهده می‌شود نتایج بهبود نداشته است و طبق گفته‌ی لیو و همکاران[۹] احتمال می‌رود مشکل از مدل باشد یا به عبارتی دیگر مدل ظرفیت کافی برای یادگیری مجموعه دادگان چالشی نداشته باشد.

با توجه به اینکه عملکرد مدل دیستیل برت نسبت به بقیه بهتر بوده است از این مدل برای آزمایش عملکرد از طریق یادگیری انتقالی از مجموعه دادگان استدلال [۷] استفاده شد. نتیجه این آزمایش در جدول قابل مشاهده است.

جدول ۵: نتیجه یادگیری انتقالی از مجموعه دادگان استدلال[۷] و آزمون آن بر روی مجموعه دادگان چالشی

distilbert
۵۰%

برای انجام یادگیری انتقالی هر دو مجموعه‌ی چالشی و استدلال به طور همزمان برای آموزش استفاده شدند. نتیجه‌ی حاصله بیانگر این است که یادگیری منطق نیز نتوانسته به بهود نتیجه کمک کند.

در نهایت به منظور استفاده از دانش خارجی از مدل KnowBert نیز استفاده شد. به علت محدودیت زمان، تنها آزمایشی که انجام شد استفاده از KnowBert به عنوان استخراج کننده ویژگی بوده است. به این معنی که صرفا نمونه‌ها به این شبکه داده شدند و از خروجی شبکه برای رده‌بندی توسط یک شبکه پیش خور استفاده شد بدون اینکه هیچ تنظیمی بر روی وزن‌های مدل انجام شود. نتایج در جدول ۶ قابل مشاهده است.

[11] bert-base-cased
[12] albert-base-v2
[13] distilbert-base-cased

جدول ۶: نتیجه استفاده از knowBert بر روی مجموعه دادگان چالشی

distilbert
۵۰%

تمامی مدل‌های برت استفاده شده با استفاده از کتابخانه هاگینگ فیس[۱۴] بوده است. همچنین برای مدل اخیر از کدهای ارائه شده توسط نویسنده[۱۵] استفاده شده است. تمامی کدهای پیاده سازی شده در گیت‌لب [۱۶] قابل دسترسی است.

۵ تحلیل

همانطور که مشخص شد هیچ یک از راهکارهای ارائه شده که شامل یادگیری انتقالی از منطق و استفاده از برت آموزش دیده بر روی گراف دانش بودند، نتوانستند عملکرد را بهبود دهند. این به معنای این است که این مجموعه دادگان بسیار چالشی می‌باشد.

از جمله کارهای دیگری که می‌توان انجام داد استفاده از مدل‌های دیگری برای تعبیه گراف دانش است که اخیرا ارائه شده. همچنین استفاده از مدل‌های تنظیم شده بر روی مجموعه دادگان‌های پرسش و پاسخ نیز می‌تواند مفید واقع شود.

منابع

[1] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein, "SemEval-2018 task 12: The argument reasoning comprehension task," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, (New Orleans, Louisiana), pp. 763–772, Association for Computational Linguistics, June 2018.

[2] T. Niven and H.-Y. Kao, "Probing neural network comprehension of natural language arguments," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 4658–4664, Association for Computational Linguistics, July 2019.

[3] H. Choi and H. Lee, "Gist at semeval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 773–777, 2018.

[4] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced LSTM for natural language inference," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 1657–1668, Association for Computational Linguistics, July 2017.

[5] J. Tian, M. Lan, and Y. Wu, "ECNU at SemEval-2018 task 12: An end-to-end attention-based neural network for the argument reasoning comprehension task," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, (New Orleans, Louisiana), pp. 1094–1098, Association for Computational Linguistics, June 2018.

[6] T. Botschen, D. Sorokin, and I. Gurevych, "Frame- and entity-based knowledge for common-sense argumentative reasoning," in *Proceedings of the 5th Workshop on Argument Mining*, (Brussels, Belgium), pp. 90–96, Association for Computational Linguistics, Nov. 2018.

[7] P. Clark, O. Tafjord, and K. Richardson, "Transformers as soft reasoners over language," *arXiv preprint arXiv:2002.05867*, 2020.

[8] K. Richardson, H. Hu, L. S. Moss, and A. Sabharwal, "Probing natural language inference models through semantic fragments.," in *AAAI*, pp. 8713–8721, 2020.

[9] N. F. Liu, R. Schwartz, and N. A. Smith, "Inoculation by fine-tuning: A method for analyzing challenge datasets," *arXiv preprint arXiv:1904.02668*, 2019.

[10] M. E. Peters, M. Neumann, R. L. Logan IV, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, "Knowledge enhanced contextual word representations," *arXiv preprint arXiv:1909.04164*, 2019.

[14] https://huggingface.co/

[15] https://github.com/allenai/kb

[16] https://gitlab.com/tagizaade/bert-argument-reasoner