STAT581: Final Project Explanation Regeneration from Knowledge Base

Tianzhi Cao (tc324),Abhishek Bhatt (ab2083),Harshini Bonam (sdb202)

December 17, 2020

Introduction

Modern language models are able to answer most questions accurately, but unable to explain reasoning behind their answers to the target user (interpretable inference). For example,

1. An AI tutoring system providing correct answers but unable to explain why they are correct, limits the student's ability to acquire a coherent grasp of the subject matter
2. A medical recommendation system that suggests a patient receive a particular surgery, but unable to explain why, presents challenges towards trusting that system

As shown in Figure 1, the problem we focus on is regenerating detailed gold explanations for standardized elementary science exam questions by selecting facts from a knowledge base of semi-structured tables.
Machine reading comprehension (MRC)[1], an area within Natural Language Understanding (NLU)[2], refers to the ability of intelligent systems to read and understand textual data. One of the key MRC tasks is Question answering (QA), that provides a natural way of testing a model's capability to comprehend textual information.

Problem Description

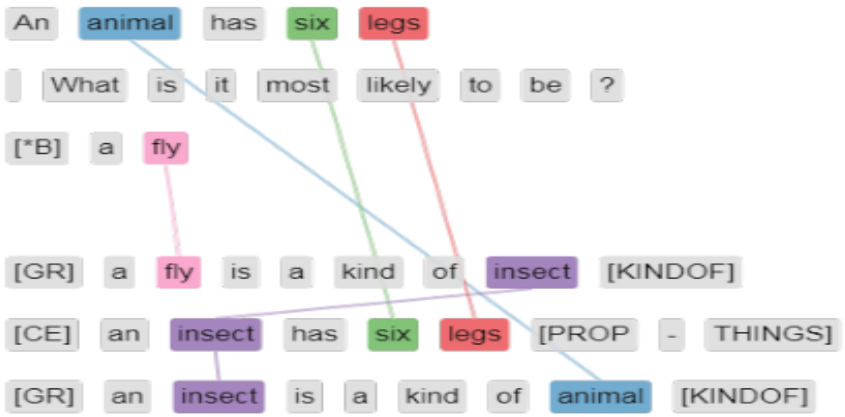

Figure 1: Explanation regeneration example

Approach

The approach leveraged to achieve a solution to this problem is the concept of multi-hop inference[3], which refers to combining more than one piece of information (say combining free-text sentences from the web, or combining linked facts from a structured knowledge base) to solve an inference task, such as question answering.

Figure 2: Multi-hop inference example

As shown in Figure 2, each explanation can be represented by an explanation graph, that combines facts together to form a detailed explanation for the reasoning required to answer a question. The facts include both core scientific facts (multi-hop reasoning) as well as detailed world knowledge (common-sense inference).

Dataset

The dataset used for this task is WorldTree explanation corpus[3] (Jansen et al., 2018). It includes approximately 2,200 standardized elementary science exam questions (3rd to 5th grade) drawn from the Aristo Reasoning Challenge (ARC) corpus (Clark et al., 2018). The questions in this corpus include detailed explanations for their answers, in the form of graphs of separate atomic facts that are connected together by having lexical overlap (i.e. shared words) with each other, and/or the question or answer text.

AnswerKey	explanation	question
B	73fa-1e22-26a8-1a7c|CENTRAL 5be6-58b4-ec52-40b2|GROUNDING e565-87e6-6f00-1598|CENTRAL d257-69b5-a1b9-f7e7|GROUNDING 0da9-2984-f126-bcad|GROUNDING 825b-a440-75bd-3001|CENTRAL e473-7297-61c1-56fa|GROUNDING af26-8083-b908-7d3b|GROUNDING dbe8-e776-f804-99a0|GROUNDING 395b-4954-311f-4749|CENTRAL	Which of the following best explains why the Sun appears to move across the sky every day? (A) The Sun rotates on its axis. (B) Earth rotates on its axis. (C) The Sun orbits around Earth. (D) Earth orbits around the Sun.

Figure 3: Questions example

ACTION.tsv	HISTORY.tsv	PROP-MAT-DURABILITY.tsv	PROP-REL-DIST.tsv	XIVORE.tsv
AFFECT.tsv	IFTHEN.tsv	PROP-MAT-OPACITY.tsv	PROP-RESOURCES-RENEWABLE.tsv	
ATTRIBUTE-VALUE-RANGE.tsv	INSTANCES.tsv	PROP-MAT-PURITY-MIXTURE.tsv	PROP-SOLUTION.tsv	
AFFORDANCES.tsv	KINDOF.tsv	PROP-ANIMAL-ATTRIB.tsv	PROP-SOLUBILITY.tsv	
CAUSE.tsv	LIFESPAN.tsv	PROP-ANIMAL-MAT-IM.tsv	PROP-STATESOFMATTER-SHAPEVOL.tsv	
CHANGE.tsv	LOCATIONS.tsv	PROP-ANIMAL-REPROD.tsv	PROP-STATESOFMATTER-TEMPS.tsv	
CHANGE-VEC.tsv	MADEOF.tsv	PROP-AVG-WEIGHT.tsv	PROP-THINGS.tsv	
CHEM-PERIODIC-TAB-FAM.tsv	MEASUREMENTS.tsv	PROP-CONDUCTIVITY.tsv	PROP-WARM-COLD-BLOODED.tsv	
CHEM-PERIODIC-TAB-LOC.tsv	NAMES.tsv	PROP-COUNTRY-HEMISPHERE.tsv	REQUIRES.tsv	
COMPARISON.tsv	OPPOSITES.tsv	PROP-DOMRECESS-TRAIT.tsv	SEQ-SPATIAL.tsv	
CONSUMERS-EATING.tsv	PARTOF.tsv	PROP-ENVIRONMENTATTRIB.tsv	SUBDIVISION-COUNTRY.tsv	
CONTAINS.tsv	PERCEPTIONS.tsv	PROP-FLEX-RIGIDITY.tsv	SUBDIVISION-GENERICSPATIAL.tsv	
CONVERSIONS.tsv	PREDATOR-PREY.tsv	PROP-GENERIC.tsv	SOURCEOF.tsv	
COUPLEDRELATIONSHIP.tsv	PROCESSROLES.tsv	PROP-HARDNESS.tsv	STAGE-IN-PROCESS.tsv	
DURATIONS.tsv	PROCESSSTAGES.tsv	PROP-INHERITEDLEARNED.tsv	SYNONYMY.tsv	
DURING.tsv	PROCESSSTAGES-ORDERS.tsv	PROP-INTENSIVE-EXTENSIVE.tsv	TRANSFER.tsv	
EXAMPLES.tsv	PROP-CHEM-ACIDITY.tsv	PROP-MAGNETISM.tsv	UNIT.tsv	
FORMEDBY.tsv	PROP-CHEM-CHARGE.tsv	PROP-ORBITAL-ROT.tsv	USEDFOR.tsv	
FREQUENCY.tsv	PROP-CHEM-REACT.tsv	PROP-QUANTITY-DATE.tsv	VEHICLE.tsv	
HABITAT.tsv	PROP-CHEM-ELEMSYMB.tsv	PROP-RECYCLABLE.tsv	WAVES.tsv	

Figure 4: Knowledge Base

As seen in Figure 3, the questions and answer keys are extracted in this format from questions.train.tsv (2207 questions) file. 81 categories of facts are extracted as knowledge base, from tsv files listed in Figure 4, that is used to generate an explanation for a given question and right answer choice.

Proposed Solution

[4]

Data Preparation

1. Facts : Each of the TSV files is read as a data frame without the headers. All columns except id and columns containing fact words/phrases are dropped. Then all columns containing fact words/phrases are concatenated to generate a single sentence. All these 81 data frames are finally merged together as a single data frame and exported as a CSV. This output forms the knowledge base in the next step below from which we extract explanation for a given question and correct answer string.

Listing 1: Data Preparation for Facts

```
library(tidyr)

# get all file names in tables folder
filenames <- list.files("tables", pattern="*.tsv", full.names=TRUE)
paste(filenames)

# read each file in tables folder as a dataframe
datalist = list()
for (i in 1:length(filenames)) {
    df <- read.delim(filenames[i], header=FALSE, sep="\t")

    # remove first row (header)
    df_without_header <- df[-1,]

    # remove last 3 columns except id
    df_trimmed <- df_without_header[, -c(length(df_without_header)-3,↩
        length(df_without_header)-2)]

    # concat all columns except last as one
    df_fact <- unite(df_trimmed[, -c(length(df_trimmed)-1)], 'fact', ↩
       colnames(df_trimmed[, -c(length(df_trimmed)-1)]), sep = " ", ↩
       remove = TRUE, na.rm = TRUE)

```

```
    # trim leading and trailing spaces
    df_fact_tidy <- data.frame(lapply(df_fact, trimws), ↩
        stringsAsFactors = FALSE)

    # add the id column
    df_fact_tidy$id <- df_trimmed[, c(length(df_trimmed)-1)]

    datalist[[i]] <- df_fact_tidy
}

# merge all dataframes into one
all_facts = do.call(rbind, datalist)

# export output
write.table(all_facts, file = "all_facts.csv", row.names = FALSE, col↩
    .names = FALSE)
```

2. Questions : Each of the TSV files for questions is read as a dataframe. Then a function is called on each row to extract the correct answer key using the correct answer key. Another function extracts just the question string (without the answer keys) from each row. The question and correct answer strings are concatenated together as a single column. All other columns except merged ques ans and explantion are dropped. This final dataframe is exported as CSV file. This output is vectorized in the next step below and compared with knowledge base to extract most relevant facts based on cosine similarity.

Listing 2: Data Preparation for Questions

```

library(tidyr)
library(dplyr)
library(stringr)


# function to extract correct answer string from the question column ↩
    using the correct answer key
getCorrectAns <- function(quesString, key)
{
    keyString <- paste("\\(", key ,"\\)", sep="")
    spl1 <- strsplit(sub(keyString, "START", quesString), "START", ↩
        fixed = TRUE)
    res1 <- sapply(spl1, "[", 2)
    spl2 <- strsplit(sub("\\(", "END", res1), "END", fixed = TRUE)
    res2 <- sapply(spl2, "[", 1)
    res3 <- gsub("[()]", "", res2)
```

```
16      ans <- str_trim(res3)
17      return(ans)
18  }
19
20  # function to extract the question statement without the answer ↩
        choices from the question column
21  getQues <- function(quesString)
22  {
23      endchar <- paste("\\(", "A" ,"\\)", sep="")
24      spl <- strsplit(sub(endchar, "END", quesString), "END", fixed = ↩
            TRUE)
25      res <- sapply(spl, "[", 1)
26      ques <- str_trim(res)
27      return(ques)
28  }
29
30  # function to perform data preprocessing on a raw questions file
31  questionsPreprocessing <- function(split)
32  {
33      # read questions file
34      inputFile <- paste("questions/questions", ".", split, ".tsv", sep↩
            ="")
35      df <- read.delim(inputFile, header=TRUE, sep="\t")
36      df_raw <- select(df, 'question', 'AnswerKey')
37
38      # extract correct answer string
39      getCorrectAns_v <- Vectorize(getCorrectAns)
40      (df_raw$ans <- getCorrectAns_v(df_raw$question, df_raw$AnswerKey)↩
            )
41
42      # extract question statement
43      getQues_v <- Vectorize(getQues)
44      (df_raw$ques <- getQues_v(df_raw$question))
45
46      # concat question and correct answer strings
47      df_ques_ans <- select(df_raw, 'ques', 'ans')
48      df_out <- unite(df_ques_ans, 'merged_ques_ans', colnames(df_ques_↩
            ans), sep = " ", remove = TRUE, na.rm = TRUE)
49
50      # trim leading and trailing spaces
51      df_out_tidy <- data.frame(lapply(df_out, trimws), ↩
            stringsAsFactors = FALSE)
52
53      # add correct explanation column
54      df_out_tidy$explanation <- select(df, 'explanation')
55      print(df_out_tidy)
56
```

```
    # export output
    outputFile <- paste("questions", ".", split, ".csv", sep="")
    write.table(df_out_tidy, file = outputFile, row.names = FALSE, col.names = FALSE)
}

questionsPreprocessing("train")
questionsPreprocessing("dev")
questionsPreprocessing("test")
```

Similarity Calculation

The dataset of questions and facts are ready to be compared after they are extracted to csv/data frame. Our solution is split each questions/facts to words and build "dictionary" based on the words in questions. After that, we create a document-term-matrix, DTM, for questions and facts. Then we calculate the cosine similarity between each question and fact. Finally, we extract the top five relevant facts as the explain of this question.

	fact	id
1	a vehicle for something allows; enables that something ...	bb32-0bc0-3629-6bca
2	adding heat to something kills viruses; bacteria in that s...	1966-99de-7765-39de
3	some adult animals lay eggs	73df-7e6e-db00-ae55
4	a warm front is when warm air mass rises and passes ov...	f992-5698-76aa-c6de
5	an airplane flies at high altitudes, between 5000 and 30...	3ad2-6e55-7ae1-f182
6	the airplane was invented in 1903	5730-38d6-f120-68f8
7	Alexander Graham Bell invented the telephone in Boston	2c3e-5dac-b4e5-5bc8
8	amphibians hatch from eggs	37af-5fb1-9a80-3548
9	young amphibians breathe through gills	359d-9e24-9685-50ca
10	adult amphibians live on land	4e5c-f633-ddba-96d9
11	young amphibians live in water	ad33-c9d9-8287-cba0
12	young amphibians undergo metamorphosis	710f-aa1b-106c-0fef
13	an electric car uses less gasoline than a regular car	8948-37e5-94f2-ca16
14	panting is when an animal breathes quickly through the...	e7c9-77b1-0847-9902
15	exhaling is when an animal expels air from the lungs	ef31-4d23-6b78-5f80
16	shivering is when an animal creates heat by shaking to k...	3b42-d1e1-9cb2-7b6b
17	panting is when an animal hangs its tongue out of its m...	709d-6f18-874c-f16a
18	an animal knows how to do instinctive behaviors when i...	277a-fd34-e20b-0e05
19	hunting is when an animal kills another animal to eat th...	e1e4-0797-e545-e336
20	hunting is when a human kills an animal for food; recre...	1593-f053-a8da-abc1
21	some animals shed fur in warm weather	f1a3-386d-38d5-de6d
22	migration is when animals move to different locations in...	4d20-3294-ea14-e444
23	all animals breathe	e5c6-c007-ec45-6843
24	most animals avoid bad odors	9a8e-ea9a-3158-05d6

Showing 1 to 25 of 9,625 entries

Figure 5: All _facts data frame

	ques	explain.explanation
1	Which of the following best explains why the Sun appe...	73fa-1e22-26a8-1a7c|CENTRAL 5be6-58b4-ec52-40b2|GR...
2	Why are different stars seen during different seasons? E...	b018-4375-d391-cb98|CENTRAL 3dd5-2225-75cc-e350|LEX...
3	Stars are organized into patterns called constellations. ...	b018-4375-d391-cb98|CENTRAL db62-4c2e-6a9d-8b12|G...
4	How does the appearance of a constellation change dur...	db62-4c2e-6a9d-8b12|CENTRAL fc96-f0ff-1c18-337e|CENT...
5	Which event occurs on a daily cycle? The Sun rises and s...	a972-0cbb-3c14-b098|CENTRAL 9318-8c1b-0902-d588|LEX...
6	Which of the following statements best explains why st...	49f5-727d-92b5-03aa|CENTRAL 825b-a440-75bd-3001|CE...
7	The Sun appears to move across the sky each day, rising ...	73fa-1e22-26a8-1a7c|CENTRAL 5be6-58b4-ec52-40b2|CE...
8	One evening as it is getting dark, Alex sits on the front ...	73fa-1e22-26a8-1a7c|CENTRAL fae5-1f95-1138-9bc4|LEXG...
9	From Earth, the Sun appears brighter than any other sta...	4e12-62f2-0830-d602|CENTRAL a538-175f-9223-d117|CEN...
10	Many stars can be seen in the sky at night. Which state...	4e12-62f2-0830-d602|CENTRAL a538-175f-9223-d117|CEN...
11	Which statement correctly describes a relationship betw...	12b0-e9bb-c84e-6a42|CENTRAL 04ee-408f-6c7a-54fc|GRO...
12	Kevin is observing the sky on a clear night. With the una...	73aa-f84a-b42e-53b3|GROUNDING 1fb6-af63-74de-86fd|...
13	The measure of the amount of light received on Earth fr...	dbe8-e776-f804-99a0|GROUNDING 1bae-622f-4093-d6da...
14	At night, the Moon is the brightest object in the sky. Wh...	a423-40e8-3886-4df5|CENTRAL a538-175f-9223-d117|CEN...
15	The Big Bang theory states that the universe started as a...	cb78-4afe-a7f9-ca1c|CENTRAL
16	In New York State, the longest period of daylight occurs ...	4c40-fb3b-92a4-0c8d|ROLE 90a3-648e-1622-edc2|GROU...
17	The length of daylight changes as the seasons change d...	66df-2200-f730-17c3|CENTRAL e172-fe13-821d-196c|CEN...
18	Michael learned that the movement of Earth in the solar ...	7959-4428-d766-0494|CENTRAL e473-7297-61c1-56fa|GR...
19	In Alaska, there are fewer hours of daylight in the winte...	842e-1407-d27c-3e94|ROLE b695-f668-1dec-659c|CENTR...
20	How does the length of daylight in New York State chan...	ac73-12b7-56aa-d852|GROUNDING 3d48-ef4c-bf06-a0c5|...
21	Earth's rotation (turning on its axis) causes day and night.	d047-b416-20a3-5fca|CENTRAL 73fa-1e22-26a8-1a7c|CEN...
22	Which of these cycles takes 24 hours? Earth rotating on ...	d047-b416-20a3-5fca|CENTRAL 7bde-6b21-5c29-4e15|CE...
23	The number of daylight hours in New York State change...	e172-fe13-821d-196c|BACKGROUND 8a35-c04c-91c8-dfde...
24	In New York State, the shortest period of daylight occurs...	e288-5cf7-f60d-4e8d|ROLE 50a6-138c-0abe-6496|ROLE 5...

Showing 1 to 25 of 2,207 entries

Figure 6: The training data frame

The initial data frame of facts and questions like graph below. There are 9625 facts and 2207 questions. Before all of that, step 0 is prepossessing data.

We build a function (Listing 3) to lower all capital character, remove non-alphanumeric symbols and collapse multiple spaces to avoid create too many similar words and meaningless words to "dictionary" (i.e. create "Sun" as "sun" as two words, create word " ").

Listing 3: Prepossessing Function

```
prep_fun=function(x){
  x = str_to_lower(x)
  # remove non-alphanumeric symbols
  x = str_replace_all(x, "[^[:alnum:]]", " ")
  # collapse multiple spaces
  str_replace_all(x, "\\s+", " ")
factrim=prep_fun(all_facts$fact)
quetrim=prep_fun(trainq$ques)
}
```

Step 1: is splitting word and build dictionary. First, we removed some "stop words" in dictionary. "Stop words" mean those words do not really have any meaning (i.e. "of', "this"). Due to our dataset are questions, we also include interrogative words. We also remove those words too frequently.

Listing 4: Tokenize dataset and create dictionary

```
#tokenize question/fact
quetok<-itoken(quetrim, preprocessor = tolower, tokenizer = word_tokenizer, progressbar = TRUE)
facttok<-itoken(factrim, preprocessor = tolower, tokenizer = word_tokenizer, progressbar = TRUE)
#make stop word list
stop_words = c("which", "why","for","its", "what", "where", "of", "as", "how", "is", "a", "that","this", "these","to", "he", "she", "are","in", "on","and","an","be","by","it")
vocab = create_vocabulary(quetok, stopwords = stop_words)
# Remove words too frequent
pruned_vocab = prune_vocabulary(vocab,term_count_min=1,doc_proportion_max=0.5)
vectorizer = vocab_vectorizer(pruned_vocab)
}
```

Step 2 is creating document term matrix. Text2vec library has a function to create it.

Listing 5: Create DTM for questions and facts

```
dtm_qes = create_dtm(quetok, vectorizer)
dtm_fact  = create_dtm(facttok, vectorizer)
```

Last step is creating cosine similarity matrix and extracting top 5 relevant facts as the explanation of questions. We build a table to save original questions and relevant explanations as csv file.

Listing 6: Build question and explanation table

```
simmat = sim2(dtm_qes, dtm_fact, method = "cosine", norm = "l2")#calculate cosine distance
lensim=nrow(simmat)#get length of questions
#get top five explains of each question
res<-data.frame()
i=1
while (i<=lensim){
  tcor=order(simmat[i,],decreasing = TRUE)[1:5]
  ans=paste(all_facts$fact[tcor],sep = ",")
  res<-rbind(res,c(trainq$ques[i],ans))
  i=i+1
}
#Write questions and explanation as table
write.table(res, file = "q&a.csv", row.names = FALSE, col.names = FALSE, sep=",")
```

An example question and explanation:

QuestionAnswer: Which of the following best explains why the Sun appears to move across the sky every day? Earth rotates on its axis.
Explanation: 1):the Earth rotating on its axis causes the Sun to appear to move across the sky during the day
2): the Earth rotates on its tilted axis
3):the Earth rotating on its axis causes stars; the Moon to appear to move across the sky at night
4):diurnal motion is when objects in the sky appear to move due to Earth's rotation on its axis
5):the moon rotates on its axis
As we see for this question, cosine similarity find the proper explanation successfully.

Observations and Analysis

In this dataset, each question has a correct explanation in the form of facts ids. Using that as ground truth, we compute the percentage of questions in the dataset that cannot find any fact relevant to its correct answer explanation (e.g. the ground truth explanation includes facts 1,2,3,6,9 and our approach finds and explanation with facts 4,5,7,8,10). This is observed for 30.26% of all the questions that we run the solution for. This implies that there are around 70% questions, for which our implementation can find at least one relevant fact for its correct answer. For comparison, if we randomly choose 5 random facts, only for 0.4% questions can we get at least one relevant fact.
On an average, for a given question, our solution can find 0.9954 relevant facts. The observation that out of 5 highest cosine similarity facts, only 1 is relevant to the explanation on average does

not ensure a good accuracy. Following are some reasons that we believe are a cause for this:

1. Some irrelevant word in the ques+correct ans string might repeat multiple times. For example:
   QuestionAnswer: Why are different stars seen during different seasons? Earth changes position in its orbit during different seasons..
   Explanation generated: 1):different seasons occur during different times of the year
   2): different substances are different in density
   3): different classes of rocks are formed by different methods
   4): different containers usually are different in shape
   5): the particles are different in arrangement in different substances
   In this example, "different" is not a significant word that should contribute to the explanation. But the cosine similarity measure finds it "important" because it repeats three times, hence the inaccurate observation.

2. Some questions need "explanations" of explanation. For example:
   QuestionAnswer: In New York State, the longest period of daylight occurs during which month? June
   Explanation generated:1)New York; New York State is a state located in the United States of America
   2)June is a kind of month
   More relevant facts that should be incorporated in explanation:
   United States of America is located in the Northern Hemisphere.
   One possible approach to solve this issue is to run our solution recursively, such that we use the first explanation as the question+ans string and rerun the explanation generation. Though this might be a slower implementation, we can find an "explanation's explanation" in this way.

3. Cosine similarity is more likely to find short facts where there is just one prominent word or phrase. For example:
   If the question have the word moon, we always find a fact even if it is irrelevant:
   "the moon is a kind of moon"
   It is very short (after delete stop words, the length of this sentence is 3: "moon kind moon") and contain "moon" two times. Even though this fact is "useless", the cosine similarity will add it every time just because ratio of "moon" is very high in this facts.

Setup Instructions

Following steps should be followed to run this solution:
1. Download 581Project Final file, make sure your R working directory is ./581Project Final
2. Run data_preparation_questions.R and data_preparation_facts.R. These two R script will generate two file: all_facts.csv and questions.rds (data preparation)
3. Run get five.R. After run each line of this file, it will show the result of the percent that the question cannot find any correct correspond facts by our algorithm as "not find rate"; the question cannot find any correct correspond facts by random 5 facts as "random not find rate"; and the average facts found by each question as "average each question find". It also will generate a csv file which names qa.csv. This csv record all the questions in column one and five facts found by our algorithm in column 2-6.

Conclusion and Future Work

We observe that for around 70% of the questions, our solution can find at least one relevant fact that explains its correct answer. There is room to improve our solution further using learning to rank approaches which can correctly rank the extracted facts. This project was a good introduction to the field of Natural Language Processing in R for us. We intend to build upon our learning from this project, to get better results in the future by using deep learning approaches.

References

[1] Machine reading comprehension. [Online]. Available: https://futuretodayinstitute.com/trend/artificial-intelligence/machine-reading-comprehension-mrc/

[2] L. Models. [Online]. Available: https://en.wikipedia.org/wiki/Language_model

[3] P. Jansen. (2020) Worldtree corpus (v2.1) of explanation graphs and inference patterns supporting multi-hop inference. [Online]. Available: http://cognitiveai.org/explanationbank/

[4] M. Z. S. D. A. M. Rajarshi Das, Ameya Godbole. (2019) Chains-of-reasoning at textgraphs 2019 shared task: Reasoning over chains of facts for explainable multi-hop inference. [Online]. Available: https://www.aclweb.org/anthology/D19-5313/