IMapBooks - Automatic Question Response Grading

Jernej Janež
Fakulteta za računalništvo in informatiko
Ljubljana, Slovenija
jj5715@student.uni-lj.si

Rok Marinšek
Fakulteta za računalništvo in informatiko
Ljubljana, Slovenija
rm4153@student.uni-lj.si

Luka Podgoršek
Fakulteta za računalništvo in informatiko
Ljubljana, Slovenija
lp5796@student.uni-lj.si

***Abstract*—For the second assignment we had to solve an end-to-end natural language processing task. We have decided that we'll focus on automatic question response grading for IMapBooks. We got 12 different questions based on the Weightless story. We had to create three different models with different methods. Models and code can be accesed on github at https://github.com/marok39/onj-02-02.**

***Index Terms*—IMapBooks, NLP**

I. Introduction

In the second assignment we were given a dataset with 12 questions on which we had to create 3 different models. These models would then grade the response on a given question. Model "A" used only text to make inference and one correct answer. Model "B" used existing answers along with their scores and other story data. Model "C" used all the data from the dataset and additional resources.

A. The data

The dataset has 16 attributes *(ID, Time, User.ID, Book.ID, Page, State, Type.Code, Difficulty..1.5., Background.Knowledge.Needed..1.5., Graesser.Inference.Type, Question, Response, Glenn.s.rating, Amber.s.rating, Final.rating, Text.used.to.make.inference)*. For our purposes the last seven attributes are mostly used and with each model, more data is used for grading. The dataset has 12 questions and different number of answers for each question which can be seen in Table I.

TABLE I
Number of answers for each question

	Number of answers
Question 1	81
Question 2	81
Question 3	70
Question 4	76
Question 5	67
Question 6	69
Question 7	71
Question 8	69
Question 9	62
Question 10	65
Question 11	68
Question 12	71

Every answer to a question gets a final rating of either *0,0; 0,5* or *1,0*. In the Table II below we have gathered all the final ratings of all answers and sorted them accordingly. We can see that the worst answered question is the fifth one (*Why does Shiranna's father get sucked into a black hole in her nightmare?*) which got 26 incorrect, 10 semi-correct and 31 correct answers. On the other hand the best answered question was seventh one (*Why is every adult trying to shake her hand?*) which received only 1 incorrect and 70 correct answers. We also added another column (*Background knowledge*) to see if there is any correlation with the correctness of answers. There is some but nothing definitive.

TABLE II
Ratings for each question

Questions	Ratings			Background knowledge
	0,0	0,5	1,0	
How does Shiranna feel as the shuttle is taking off?	5	41	35	1
Why did her body want to float?	7	21	53	3
Why do Adam cheeks "flush"?	5	0	65	4
Why is the door locked?	18	26	32	1
Why does Shiranna's father get sucked into a black hole in her nightmare?	26	10	31	2
How do you think Shiranna's confidence has changed? What events caused this change?	9	13	47	1
Why is every adult trying to shake her hand?	1	0	70	1
What causes seasons to occur on Earth?	5	0	64	2
How does Shiranna feel about the zoo?	10	0	52	1
What is happening in the tunnel?	1	31	33	1
How does Adam's family structure differ from Shiranna's family structure?	0	5	63	1
Will the journey to Venus will be short or long? Why do you think so?	4	1	66	2

1) Inter-rater agreement: In the dataset we have Glenn's and Amber's ratings as well as the final one. We calculated the inter-rater agreement between Glenn and Amber and found out that out of all 850 answers they gave the same rating on 756 answers (88.94%). We also computed how their ratings

compared to the final rating of each answer and got these results: Glenn gave the same rating as the final one to 806 (94.82%) answers, whereas Amber gave 792 (93.18%).

II. Related work

Similar to our problem is the Conversational Question Answering (CoQA) dataset for measuring the ability of machines to participate in a question-answering style conversation [1]. In the paper they propose that a machine has to understand a text passage and answer a series of questions that appear in a conversation. Three goals were in mind, the first one was that CoQA would keep track of the conversation history - meaning questions can relate on the previous already asked questions, second one is that answers can be free-form text and not restricted to a contiguous span in a given passage and the third one is that CoQA can perform robustly across domains.

III. Methods

A. Model A

As previously mentioned, model A gets the least training data (inference text and one correct answer). This makes it difficult to decide for an effective method. As a basic baseline we have decided to use Word2Vec to determine a vector space for word embeddings using text used to make inference. The basic idea is to embed a correct answer in that vector space and check the cosine similarity between the correct answer and the answer we are predicting the score of. Our assumption is that the closer an answer is to the correct answer, the higher the probability of that example being correct as well.

B. TODO Model B

C. TODO Model C

IV. Results

A. Model A

Using the approach described above on the whole data set we achieved a F_{macro} score of 0.311 and F_{micro} score of 0.378. Given the very little data and very simple method we can be happy with the result. We noticed that our model labels responses with the score 0.5 overproportionally, especially compared to true labels. This could have multiple reasons. One of them could be because our approach works best if the differences between answers are extreme. Since they are not, most of the cosine similarity values are around 0.5, resulting in a poor distinction between classes.

B. TODO Model B

C. TODO Model C

V. Discussion

VI. Conclusion

References

[1] Reddy, Siva, Danqi Chen, and Christopher D. Manning. “CoQA: A conversational question answering challenge” arXiv preprint arXiv:1808.07042 (2018).