Question Classification with Machine Learning

Jason Liang, YSP Student, *Westford Academy*
Stephanie Martinez, YSP Student, *Prospect Hill Academy Charter School*
Shuyang Cao, Computer Science, *Northeastern University*
Professor Lu Wang, Computer Science Department, *Northeastern University*

Young Scholars Program · Northeastern University Center for STEM Education · Northeastern Khoury College of Computer Sciences · NSF

Abstract

Currently, question taxonomies that are widely adopted in factoid question generation systems are too shallow or specific for open-ended questions. The purpose of this research project is to develop a more widely applicable question classification system. The system aims to identify the intention of the question asker, which is more meaningful in a diverse setting. To do so, a question taxonomy is developed that provides fine granularity for intentions while applicable to multiple domains. A question classification system is then created through feeding a model a myriad of question examples with their intentions already labeled. After testing the model on a held-out question set, the model reaches an accuracy of roughly 43.09%.

Background

Popular question taxonomies today are often based on interrogative words or are specific to a field. Using the former is typically too shallow for open-ended questions. Using the latter to classify a diverse text of questions often is not as useful. For example, Bloom's taxonomy, which is heavily involved in the student-teacher relationship in education, would not be as reliable when used to classify questions from a social media page or online forum. These factors create a need for a more widely applicable question taxonomy.

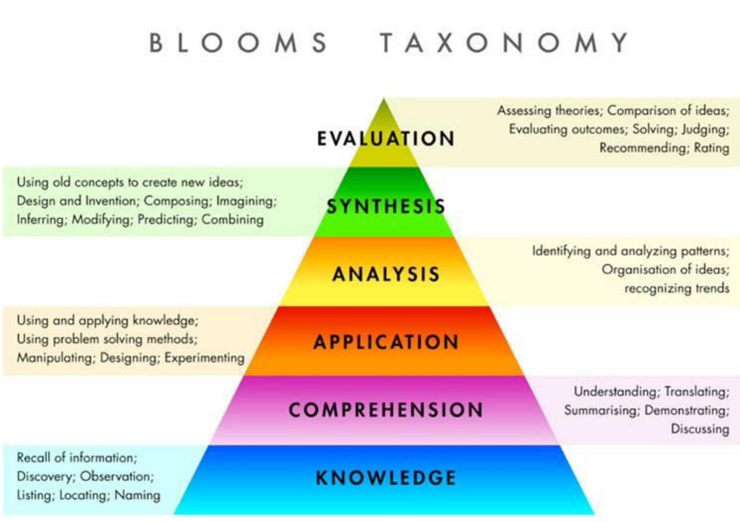


Fig. 1: Bloom's Taxonomy

Table 1: Graesser, Person, and Huber (1992)'s Question Categories

Question category	Abstract specification
1. Verification	Is X true or false? Did an event occur?
2. Disjunctive	Is X, Y, or Z the case?
3. Concept completion	Who? What? When? Where?
4. Feature specification	What qualitative properties does X have?
5. Quantification	How much? How many?
6. Definition questions	What does X mean?
7. Example questions	What is an example of a category?
8. Comparison	How is X similar to or different from Y?
9. Interpretation	What can be inferred from given data?
10. Causal antecedent	What state causally led to another state?
11. Causal consequence	What are the consequences of a state?
12. Goal orientation	What are the goals behind an agent action?
13. Procedural	What process allows an agent to reach a goal?
14. Enablement	What resource allows an agent to reach a goal?
15. Expectation	Why did some expected event not occur?
16. Judgmental	What value does the answerer give to an idea?
17. Assertion	A declarative statement that indicates the speaker does not understand an idea.
18. Request/Directive	The questioner wants the listener to perform some action.

Fig. 2: Original Taxonomy used

While state-of-the-art classification models can achieve impressive performance on many popular classification datasets, they require huge computing resources to train and evaluate. In this project, Fasttext, an open-source library released by Facebook, was used. It can run on computers with limited computing capability and yield satisfactory results on many text classification benchmarks. Fasttext employs multiple techniques, such as the vectorization of words and linear classification and allows for much customization of model parameters.

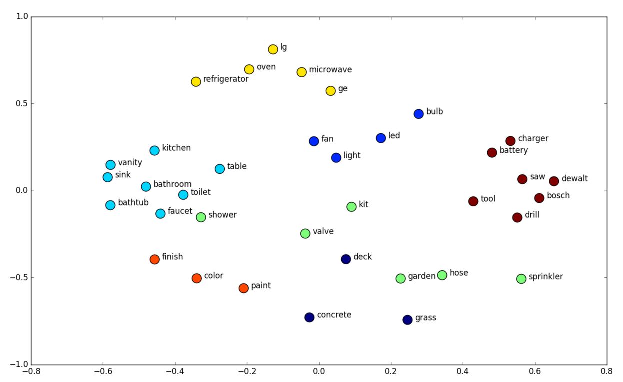


Fig.3: Example Word Vectors

Experimental Methods

In order to train the model, a dataset was needed. To start, existing questions formulated from recently published research articles were collected. An updated taxonomy more fit for diverse input was then created by generalizing some of the category definitions, combining two categories together, and removing very specific and infrequently seen categories. Using this personalized classification system, the previous dataset was reclassified. To broaden the dataset, questions from more mainstream question sites, specifically Yahoo and Reddit were also classified using the newly made taxonomy.

When comparing answers among group members, percent similarity and question type frequency distributions were created. These statistics gave a better picture of how ambiguous the new taxonomy was and patterns in the sentence structure of certain question types.

Fig. 5: Reddit Distributions

Questions with high agreement (at least 2 out of 3 members agree on the classification), were picked out from the dataset. ⅘ of these high agreement questions were used to train the model. The remaining ⅕ of the questions were used for evaluation of the model.

Specific techniques and parameters used while creating the model included a high learning rate, reading input in bigrams, and common preprocessing methods.

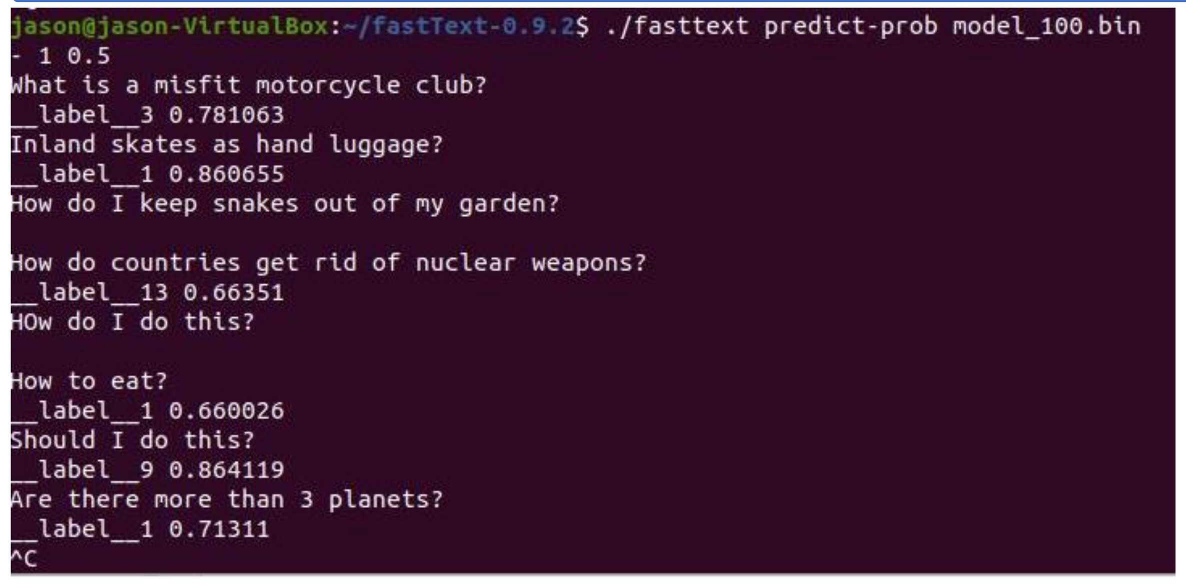
Results

Category	Definition	Examples
1. Verification	Is X true?	"Did Congress pass the law?"
2. Disjunctive	Is X or Y the case?	"Did Congress pass the law in 2016 or 2017?"
3-6. Concept question	Who? What? When? Where? What does X mean?	"What does the OCE do?" "What language do people speak in the Dominica republic?"
5. Extent	How much? How many? To what extent? How ADJ?	"How much does the appropriation offer for the plan?" "To what extent is the Renewable Fuel Standard accurate nationwide?"
7. Example	What is an example?	"What are some unintentional sources of cyber-based risks to federal systems?" "What are some examples to support or contradict this?"
8. Comparison	How is X similar to or different from Y?	"What is similar about the two proposals?" "How did the increase in grazing for BLM compare to that of FS?"
9-16. Judgmental	What can be inferred from the given data? What value does the answerer give to an idea?	"What does GAO think of the new measurement?"
10. Causal antecedent	What state causally led to another state?	"What has escalated the ongoing conflict in the Southeast?" "How are younger parents correlated with child poverty?"
11. Causal consequence	What are the consequences of a state?	"What are the negative consequences for the services if they do not evaluate their programs?" "How does the DATA Act affect OIGs?" "What would happen if employers violate the legislation?"
12. Goal orientation	What are the goals behind an agent's action?	"What is the purpose of this report?" "What was the motivation for the report written by the GAO?"
13-14. Procedural & Enablement	What process allows an agent to reach a goal? What resource allows an agent to reach a goal?	"What do insurers do now to avoid insuring individuals with higher risk?" "How might Russia gain considerable power in international affairs in the near future?" "What laws allow people to possess guns?"
15. Expectation	Why did some out-of-expectation events happen?	"Why does persecution against gang violence fail?" "How does FMCSA not address imprecise violation rates?"

Fig. 6: Updated Taxonomy

Accuracy of the model after validation was approximately 43.09%. When classifying randomly, the accuracy was approximately 19.68%. Thus, with training, the model's accuracy increased by 23.41%. With a larger data set, its accuracy most likely would increase more.

Results Continued



```
jason@jason-VirtualBox:~/fastText-0.9.2$ ./fasttext predict-prob model_100.bin
- 1 0.5
What is a misfit motorcycle club?
__label__3 0.781063
Inland skates as hand luggage?
__label__1 0.860655
How do I keep snakes out of my garden?

How do countries get rid of nuclear weapons?
__label__13 0.66351
HOw do I do this?

How to eat?
__label__1 0.660026
Should I do this?
__label__9 0.864119
Are there more than 3 planets?
__label__1 0.71311
^C
```

Figure 7: Sample Questions with Results

On the right of the labels are probability values. The most likely label is outputted by the model. Oftentimes, if the question inputted has grammatical errors or is too complex (I.e. multiple conjunctions, unique terms), the model will output incorrect responses.

Conclusion

Compared with common classification tasks such as sentiment classification, where common models have above 60% accuracy, identifying the intention of the question asker is more difficult. Better performance can be expected if augmenting with a larger labeled dataset or if using state-of-the-art models that are pre-trained with large amounts of natural language data.

For the future, a major goal is to train the model to generate questions of each type in the taxonomy. With more time and resources to collect and classify questions, the model would have a larger data set to learn from. The model would ideally learn how to create templates based off of data sets. By observing patterns in each question type and creating its own templates, the model would be able to generate questions of any category from any source material when requested by the user.

In addition, a goal is to improve the model to be able to handle more complex inputs and grammatical errors.

References

[Example of Word Vectors]. (n.d.). https://devopedia.org/images/article/220/3225.1569667846.png
Olney, A. M., Graesser, A. C., & Person, N. K. (2012). Question Generation from Concept Maps. Dialogue and Discourse, 75–99. https://pdfs.semanticscholar.org/84b5/33b04ca5958664fbcc003be43852c59df1b4.pdf
Dalton, J. & Smith, D., (1986). Extending Children's Special Abilities: Strategies for primary classrooms (pp. 36-37). http://www.mandela.ac.za/cyberhunts/bloom.htm
Agarwal, M., Shah, R., & Mannem, P. (2011). Automatic Question Generation using Discourse Cues. Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, 1–9. https://3.basecamp.com/3761917/buckets/17375195/uploads/2735558915
reddit. (n.d.). Reddit. https://www.reddit.com/
Homepage. (n.d.). Yahoo Answers. https://answers.yahoo.com/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAANQ97CfRc1d0UFMjPT6oSVKJTjSckrCn191qJlw7CtT-neOG4utLP9iMOKPVrXTlAWJ1dWOAmsUSHy18mHj97ZWXKdQnecjpkVCSBGI7jk1bmWZc3MAojv62RbQxUyTvVBhGVqlUMK2JM8dcg19DUgQR3gYEjkF1g1X72-PoSG9m
OxfordSparks. (2017, January 11). What is Machine Learning? [Video]. YouTube. https://www.youtube.com/watch?v=f_uwKZIAeM0
[Robot Learning]. (n.d.). Artificial Intelligence: Preparing Students for the Future with AI. https://www.gettingsmart.com/wp-content/uploads/2019/09/4f06b7b2-753e-48b3-bb8f-777ca2653210_30212411048_2a1d7200e2_b.jpg

Acknowledgements

Department of Computer Science
Professor Lu Wang - Assistant Professor
Shuyang Cao - Ph.D Candidate for Computer Science
Department of STEM
Claire Duggan - Director of Programs and Operations
Natasha Zaarour, Salima Amiji, Nicholas Fuchs - YSP Coordinators

This work was supported by the National Science Foundation
Grant #