A Question Classification System Based on the Covid-Q Dataset

Christopher Liu

Eric Zhang

Yunyi Zhang

Abstract

In the entire last year, Internet users have asked thousands of questions about Covid-19 on different online platforms. In order to compile all these questions semantically for future use (such as building a question answering machine), we decide to design a question classification system. In this paper, we propose a fast and simple algorithm to classify the questions related to Covid-19. First we enumerate several pertinent features to train the classifiers specifically for the pandemic questions. Then we investigate quantitatively the contributions of different features on the training model. We also discuss our retrospection to the experimental result and propose some possible solutions for future improvement.

1 Introduction

The rampage of corona-virus has already raised people's concerns over their daily routines. Everyday, millions of worried Internet users toss numerous questions on different online platforms and forums, ranging from symptoms of the disease to the purchase of medical inventory. In order to process all these inquiries more efficiently, it is imperative to develop a question answering system tuned to Covid-19 related data. As the starting point, we have decided to develop a Covid-19 question classification system to categorize all the questions in the hopes that a latter developed answering machine can pinpoint the answers more accurately.

In this paper, we propose a simple and fast algorithm for classifying different types of questions related to Covid-19.

2 Related Work

In terms of a general question classification system, [Huang et al.][1] [2] have identified several essential features, such as question wh-word, head word and hypernyms. Different features bring different levels of improvements over the accuracy of the system: in the experiment with SVM, the feature set with wh-word and head word brings an accuracy high up to 92.6%, while the feature set with solely unigram brings an accuracy of merely 88.0%. Some features even have adverse effects over the efficiency of the system. For instance, bigram and trigram both noticeably lowers the accuracy compared (with the accuracy of 85.6 and 68.) to unigram (88.0). This is potentially because in question classification tasks, we don't really care about the relationship between words, unlike we do in POS tagging training process. Therefore, it is evident that the careful selection of the

features can greatly affect the performance of the system.

[Huang et al.][1] [2] also dig deeper to develop a series of question-dependent features. These features are constructed based on the specific expression of the questions. For example, the “how die” feature is used to describe all the questions which begin with a how and followed by a person's name or verb “die”. The benefit of such a feature is that it can bring further convenience for the system to correctly identify some other regular features such as head word and named entity. In the case of the “how die” feature, if a question is identified with such a feature, there is a high chance that the noun between “how" and “die” is the head word and even the name entity. Such question-dependent features can further improve the performance of the system.

[Metzler and Croft][3] demonstrate the limitation of some common dataset used by the classification system today: sometimes even the manually annotated question categories can be misleading as training data, given the ambiguity of the question themselves. For example, when facing questions such as “who developed potlatch?”, the annotators will hesitate to label it either as a question about “person” or a question about “nationality”. Due to lack of specific context, the annotators might unconsciously mislabel some of the questions, which ultimately leads to a flawed training dataset used by the system.

[Braun et al.][4] point out certain factors which might skew the experimental result: First of all, the performance of the system influences the domains of the data. Most question classification systems usually gain much higher accuracy when facing the concepts they are most familiar with, which depend on the service they target. Second, many Internet users might frame their questions in a way which is more understandable for the system but not for humans. With some less naturally expressed questions, the dataset can be biased in favor of the system performance. Considering the factors listed above, we should be more aware of the specificity and the inclusivity on the design of the algorithm.

3 Method and Implementation

Our methods and classification features mainly derive from the work of [Huang et al.][1] [2] , but with some added changes as well. We trained a maximum entropy classifier, using standard features like question word, head nouns and hypernyms. However, in respect to our covid-19 specific dataset, we selected other features more suited to differentiating the types of questions. This maximum entropy model was then run on a test of 668 real-world test questions as well as 237 author-generated questions from the COVID-Q question dataset.

Question Classification Features

We incorporated the following features to train the classifier, a modified selection of features from [Huang et al.][1] [2]’s classifier. Since the questions are often short (6 words or less), and the dataset being small, we determined that traditional sentence vectorization methods, such as bag of words and even TF-IDF, resulted in poor results for this type of problem. Therefore, we aimed to select the features that most precisely

determined their category classification, yet were the most similar between questions of the same category. We propose the following five features.

Question words

Along with standard interrogative wh-words (who, what, where, when, why, how), we also selected additional pseudo-wh words. Since the COVID-Q dataset contained only questions, the some normally non-question words become selectable question words, given the context. For example, “are” is not inherently a question word (“We are human beings”), but since the COVID-Q dataset contains only questions, “are” becomes a question word (“Are bats carrying covid?”). In addition, we selected additional, more specific versions of wh-words that disproportionately appeared in the COVID-Q dataset, such as “how long.” These more specific question words more accurately determined the sentence’s question category—for example, “how long” was almost exclusively used by Speculation questions.

In a sentence with multiple questions (“Are pregnant women more susceptible to the covid virus and will it harm the fetus”) the dominant question word was selected through a priority ranking, determined by question specificity (i.e. “which” was only selected as a question word if “what” was not present in the sentence, and so on).

Question Word	Appearance Count
what	922
which	156
when	140
where	100
who	128
how long	42
are	397
will	342
is there	19
how	360
can	237
would	26
why	219

Table 1) The priority list as well as question frequency for each question word queried in our feature builder.

Normally, a priority list would bias the number of occurrences towards the top of the list. However, the presence of high numbers of occurrences relatively late in the priority list suggest that the relative diversity in question words indicates that this structuring was not a problem.

Root Word

We used the Stanford CoreNLP Dependency parser to analyze the syntactic structure of the sentences, using the sentence root word as a feature. This is also a common feature used in other NLP literatures.

Head Noun

After the root word, similar to [Huang et al.][1] [2], the head noun was included as a

feature. The head noun was found by the following simple algorithm: First, identify the syntactic root of the sentence as above. If the root is already a noun, that was selected as the head noun. If not, the first noun that was either tagged as the subject by the Stanford Parser or was the direct dependent of the root word was selected as the head noun.

Dependent

For many sentences, the head noun was "covid," due to the nature of the dataset. Therefore, the "dependent" was selected as a feature, or the main word that was governed by "covid." For example, in the sentence "Can covid live on surfaces", "Covid" is the head noun, while "surfaces" is the dependent. Because the dependent word often solely determines the category of the sentence (in this case, transmission), the dependent became a useful feature to include in the classifier.

Hypernyms

Due to our small dataset, we aimed to minimize the variance between questions similar in meaning. To that end, we included the direct WordNet hypernyms of both the head noun as well as the dependent.

Word Substitution

Much of the vocabulary in the Covid-19 dataset is relatively new. Covid, for example, is a word that only recently gained a new meaning to represent the pandemic. The Stanford CoreNLP parser, as a result, often mistakes words for others, such as "trump" for the verb "to triumph over," and "bill gates" as the two words "bill" and "gates."

So that the sentences could be analyzed correctly, we employed word substitution using a lookup table. The following replacements were used:

president trump	the president
trump	the president
bill gates	he
covid virus	virus
the covid	the virus
covid	the virus
the covid vaccine	the vaccine

Table 2) The word substitution lookup table. Similar to 1), the table is ordered by priority.

Training Model

The model we used was the python nltk MaxEnt classifier with IIS. We gained insights from [Huang et al.][1] [2] who uses the Maximum Entropy Model to train the classifier for questions as well. Maximum Entropy Model is a general machine learning model used in classification tasks. Since we are doing classification here as well, we use this model for our training. In investigating other classifiers such as an SVM classifier, we found they yielded similar but poorer results no matter with linear kernel or other kernels.

4 Result and Evaluation

We received the following accuracy results, broken down by features:

wh-word, head word, head noun	.4217
+non-wh question words	.4457
+hypernyms	.4550
+word replacement	.4956
+"covid" dependents	.5237

Table 3) The accuracy score table by different features

The results broken down by improvement in accuracy score:

wh-word, head word, head noun	
+non-wh question words	+.024
+hypernyms	+.0093
+word replacement	+.0406
+"covid" dependents	+.0281

Table 4) The improved accuracy score table by different features

The feature creating the most marked difference was the word replacement, followed by the inclusion of the dependent and non-wh question words. Including hypernyms resulted in a smaller increase to accuracy.

The categories with large amounts of miscategorizations (>.15) as seen in the confusion matrix (Appendix 2) are as follows: Societal Response as Societal Effects (11 / .15), Prevention as Individual Response (15 / .15), Economic Effects as Societal Effects (6 / .17), Having COVID as Transmission (7 / .25), and Comparison as Nomenclature (11 / .27).

5 Discussion and Future Work

Our accuracy was relatively low compared to classic open-domain question classifiers, which normally reach into 95% or greater. Based on our analysis, this is due to the combination of several factors.

Sample Set Size

One of the biggest roadblocks to our classifier is the limited size of the COVID-Q dataset. A statistics-based model such as ours would be more accurate given more data. In the future, we could potentially use bootstrap methods to upsample our data such that we have more data for our training.

Word Sense Disambiguation

Much of the critical vocabulary in this domain is new and has multiple meanings and ambiguous interpretations. For example, "covid" not only means the virus and/or its symptoms in questions such as "Can I catch covid from packages?" but also the overall pandemic lifestyle (lockdowns, etc.) in "when will covid be over?" Other questions that deal with nomenclature are particularly difficult, such as "Why Covid?"

Given that the highest boost to accuracy came from the word replacement table, we believe that the biggest roadblock to covid-related question classification is the ambiguous sense of covid vocabulary, with

words like “covid,” “the vaccine,” and “the pandemic” all containing multiple definitions even within the relatively small small sampleset. Since the words used in covid discourse are so new and so rapidly changing, the ability to disambiguate the vocabulary remains an important problem to solve.

We have made an attempt to mitigate these issues with a hard-coded word replacement table, but this fails to pick up the various senses of “covid” as a word, instead replacing them with a single word, “virus.” In the future, we would like to build a covid word sense disambiguation processor in order to more specifically parse the questions.

Word Sense Disambiguation

Many of the provided questions in the COVID-Q dataset could naturally fit into multiple categories based on context and intent. Slight (or even no) variations in question wording could place a question in different categories, which a statistical classifier would struggle with. For example, many Prevention questions deal with infection, disinfecting products, and techniques to reduce spread. Yet “Should I disinfect my groceries?”, marked as Prevention by our classifier, is categorized as Transmission in the dataset, because the motive for the question when put in context asks whether Covid can travel through food packaging. A larger dataset could help disambiguate context-sensitive questions by emphasizing the differences in subjects or root nouns.

Commonly Confused Categories

For the confusion matrix, it shows a relatively inconsistent number of correct predictions, with the best category being Testing (.67), and the worst Societal Response (.19). A factor contributing to this inconsistency is the small sample size and different size of classes (Transmission is by far the largest category), but we believe a factor in this inconsistency is that some categories are naturally more easily mistaken for others. For example, in the “transmission” questions, many questions start with “can”, as in “Can I get covid from my pet.” Thus, it’s easier to find patterns between these questions in the same class. On the other hand, Comparison questions were often mistaken for Nomenclature questions because they both often start with “why.” In these sentences:

Why is it covid? (Nomenclature)
Why is covid dangerous? (Comparison)

Our classifier classified both of these as Comparison. In the future, a feature more tuned to comparison words (-er, -est) would more easily pick up and prevent miscategorizations of this nature, but for other categories it will prove to be more difficult.

For example, it is difficult to distinguish between Prevention and Individual Response, as both deal with personal choice. Moreover, Societal Effects and Economic Effects both ask questions about the greater society, meaning the vocabulary is nearly identical (e.g. questions about “lockdowns” have both a societal and economic component, with categorization highly dependent on question motive).

In the future, we would like to

increase the accuracy of our model by including additional tagged questions, as well as potentially investigating using a combination of classifier types. In addition, as mentioned above, are interested in augmenting dictionaries and building statistical word-sense disambiguation classifiers for covid-related vocabulary.

Wh-Word Priority

The question word selection process, and more specifically the priority list, also resulted in some false positives for sentences with multiple question words. While nearly all questions contained only a single question word, our classifier potentially miscategorized compound questions. In the future, we can integrate our syntactic parser in this process, selecting a question word that was closest to the root of the sentence. While these compound questions would not make up a great number of answers, potentially complex question structures create a blind spot in our model that deserves investigation.

Training Model

We could also improve our classification performance by choosing a more advanced model. We know that the maximum entropy model is a standard ML model for classification, but in our case it's not performing well partially because it's overfitting our data. Besides upsampling our data, we could also try other advanced models including the SVM-TBL algorithm proposed by [Li et al.][5], which is an modified basic SVM model.

Reference

1.Zhiheng Huang, Marcus Thint, Asli Celikyilmaz. "*Investigation of Question Classifier in Question Answering*" , Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, August 2009.

2. Zhiheng Huang, Marcus Thint, Zengchang Qin. "*Question Classification using Head Words and their Hypernyms*" Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, October 2008.

3. Donald Metzler, Bruce Croft. "*Analysis of Statistical Question Classification for Fact-Based Questions*" Information Retrieval, January 2005.

4. Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, Manfred Langen. " *Evaluating Natural Language Understanding Services for Conversational Question Answering Systems*" Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, August 2017.

5. LI Xin, HUANG Xuan-Jing, WU Li-de. *"Question Classification using Multiple Classifiers"* Proceedings of the Fifth Workshop on Asian Language Resources, January 2005

Appendix

1. Confusion matrix (Raw Count)

	SP	TN	NO	RE	SR	SE	OR	PR	TT	TE	CO	EE	SY	HC	IR
Speculation	12	1	2	1	1	2	0	3	2	2	2	4	0	0	1
Transmission	10	77	9	8	7	5	5	23	6	10	5	11	8	14	12
Nomenclature	1	0	14	0	1	1	0	2	1	2	1	1	3	0	0
Reporting	2	1	1	24	6	1	3	1	2	4	3	4	0	3	4
Societal Response	9	2	5	3	14	11	1	1	4	1	5	6	2	6	5
Societal Effects	9	5	8	4	12	19	0	5	3	5	3	6	2	6	5
Origin	6	2	1	1	1	5	23	1	5	0	3	4	2	0	0
Prevention	1	11	1	1	3	5	1	37	4	5	2	7	3	5	15
Treatment	2	5	2	5	2	2	3	4	12	2	0	3	0	4	5
Testing	2	1	1	2	0	2	0	3	1	28	1	0	0	0	1
Comparison	1	1	11	1	0	5	0	3	1	1	16	0	0	0	1
Economic Effects	2	1	2	1	0	6	0	1	2	1	2	12	0	4	1
Symptoms	1	1	2	1	0	0	1	0	1	0	0	2	5	2	2
Having Covid	1	7	3	1	0	3	0	1	0	0	2	1	0	7	2
Individual Response	0	5	1	2	0	2	0	4	2	3	0	4	2	2	10

2. Confusion Matrix (Percentage)

	SP	TN	NO	RE	SR	SE	OR	PR	TT	TE	CO	EE	SY	HC	IR
Speculation	.36	.03	.06	.03	.03	.06	.00	.09	.06	.06	.06	.12	.00	.00	.03
Transmission	.05	.37	.04	.04	.03	.02	.02	.11	.03	.05	.02	.05	.04	.07	.06
Nomenclature	.04	.00	.52	.00	.04	.04	.00	.07	.04	.07	.04	.04	.11	.00	.00
Reporting	.03	.02	.02	.41	.10	.02	.05	.02	.03	.07	.05	.07	.00	.05	.07
Societal Response	.12	.03	.07	.04	.19	.15	.01	.01	.05	.01	.07	.08	.03	.08	.07
Societal Effects	.10	.05	.09	.04	.13	.21	.00	.05	.03	.05	.03	.07	.02	.07	.05
Origin	.11	.04	.02	.02	.02	.09	.43	.02	.09	.00	.06	.07	.04	.00	.00
Prevention	.01	.11	.01	.01	.03	.05	.01	.37	.04	.05	.02	.07	.03	.05	.15
Treatment	.04	.10	.04	.10	.04	.04	.06	.08	.24	.04	.00	.06	.00	.08	.10
Testing	.05	.02	.02	.05	.00	.05	.00	.07	.02	.67	.02	.00	.00	.00	.02
Comparison	.02	.02	.27	.02	.00	.12	.00	.07	.02	.02	.39	.00	.00	.00	.02
Economic Effects	.06	.03	.06	.03	.00	.17	.00	.03	.06	.03	.06	.34	.00	.11	.03
Symptoms	.06	.06	.11	.06	.00	.00	.06	.00	.06	.00	.00	.11	.28	.11	.11
Having Covid	.04	.25	.11	.04	.00	.11	.00	.04	.00	.00	.07	.04	.00	.25	.07
Individual Response	.00	.14	.03	.05	.00	.05	.00	.11	.05	.08	.00	.11	.05	.05	.27