Finding Friends and Flipping Frenemies: Automatic Paraphrase Dataset Augmentation Using Graph Theory

Hannah Chen, Yangfeng Ji, David Evans
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
{yc4dx,yangfeng,evans}@virginia.edu

Abstract

Most NLP datasets are manually labeled, so suffer from inconsistent labeling or limited size. We propose methods for automatically improving datasets by viewing them as graphs with expected semantic properties. We construct a paraphrase graph from the provided sentence pair labels, and create an augmented dataset by directly inferring labels from the original sentence pairs using a transitivity property. We use structural balance theory to identify likely mislabelings in the graph, and flip their labels. We evaluate our methods on paraphrase models trained using these datasets starting from a pretrained BERT model, and find that the automatically-enhanced training sets result in more accurate models.

1 Introduction

Having high quality annotated data is crucial for training supervised machine learning models. However, producing large datasets with good labeling quality is expensive and labor intensive. Most NLP datasets rely on labels provided by human annotators with varying skills and limited training and expertise. The label instances are also often based on ambiguous definitions and guidelines.

To address this problem, we study automated techniques to improve datasets for training and testing. In particular, we focus on paraphrase identification task, which aims to determine whether two given sentences are semantically equivalent. The sentences and labels in a dataset can be viewed as nodes and edges of a graph. Moving from single labeled sentence pairs to a graph provides a better understanding of the sentence relations of the dataset, which can be exploited to infer additional edge labels. In particular, since paraphrases are an equality relation, we can perform a transitive closure on the graph to infer additional labels. In addition, we use the notion of balance (Harary, 1953) for signed graphs to identify conflicted relations. In the context of semantic relationships between pairs of sentences, any paraphrases of a given sentence cannot be a non-paraphrase of each other since they should all share an identical meaning.

Contributions. We show the benefits of representing sentence pair relations as a graph. We first construct a paraphrase graph with the original pairs and their relation labels from the Quora Question Pairs (QQP) dataset (Iyer et al., 2017) following the structure of a signed graph. With the graph structure and the transitivity of paraphrases, we can automatically infer new sentence pair relations directly from the original dataset (Section 3.1). In addition, we identify and correct likely mislabeled pairs based on violations of expected structural balance properties we expect a valid paraphrase graph to satisfy (Section 3.2). We found 90 seemingly mislabeled sentence pairs in the QQP dataset. We show that fine-tuning a BERT model on the augmented set improves its performance on both the original and augmented testing sets, decreasing the error rate from 10% to under 6% when testing on the augmented test set. We released the augmented QQP dataset and the implementation code. (https://github.com/hannahxchen/automatic-paraphrase-dataset-augmentation)

2 Representing Datasets as Graphs

A *signed graph* is a graph where each edge is labeled either positive or negative to indicate a relationship between the two connected nodes. For undirected graphs, this relationship is symmetric. A path is a set of connected edges with no repeated nodes, and a path with the last node connecting back to the first node forms a cycle. Given semantic interpretations of the edge labels, all paths in a signed graph should have certain properties.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4741–4751
November 16 - 20, 2020.

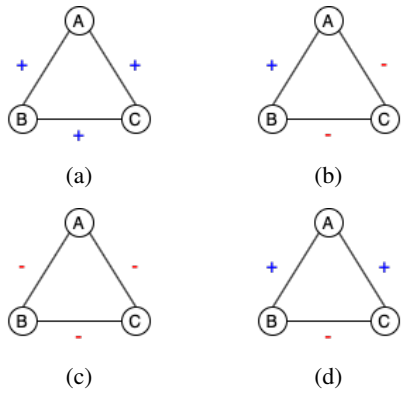


Figure 1: Four types of signed graphs of a triad. Signed graphs (a) and (b) are balanced; (c) is weakly balanced; (d) is imbalanced.

Structural Balance. Balance theory was proposed by Heider (1946) to study interpersonal relationships in social psychology. The idea was generalized to signed graphs by Harary (1953). A graph is said to be *balanced* if the product of the edge signs in every cycle is positive. There are only two types of conditions exist in a balanced signed graph: (1) all the nodes are connected with only positive edges, or (2) nodes can be divided into subsets such that nodes within each subset are connected with positive edges and nodes from different subsets are connected with negative edges. Figure 1 illustrates four possible sign combinations for a triad.

Paraphrase Graph. The definition of a paraphrase remains ambiguous and varies by task, but the most common definitions are similar to the one from Bhagat and Hovy (2013), which define paraphrases as sentences that convey the same meaning but are expressed in different forms. Since this notion is a symmetric relation, we can form an undirected signed graph by linking the sentence pairs from the paraphrase dataset with their annotated relations. Sentence pairs labeled as paraphrases are connected with positive edges; sentences labeled as non-paraphrases are connected with negative edges. A paraphrase cluster contains sentences connected with positive edges, and all sentences in the cluster should share the same meaning. Figure 2 shows how a paraphrase graph is constructed from selected labeled pairs in the QQP dataset.

3 Improving Datasets using Graphs

Typically, training sets for paraphrase identification are constructed by using annotations for sentence pairs provided by human annotators. Based on the semantics implied by the paraphrase and non-paraphrase labels, we can augment and correct the sentence-level paraphrase graph. Our method infers labels based on transitivity (Section 3.1), and identifies likely mislabelings based on expected graph consistency properties (Section 3.2).

3.1 Inferring New Labels (Finding Friends)

Since paraphrase is a reflexive, symmetric, and transitive relation, we can identify a set of semantically equivalent sentences if they are reachable by one another along the paraphrase links. We use Dijkstra's shortest path algorithm (Dijkstra, 1959) implemented by Networkx (Hagberg et al., 2008) to find paraphrase paths between nodes. Furthermore, we can infer additional non-paraphrase edges between nodes from two different paraphrase clusters if they are connected with one or more non-paraphrase links. Figure 2 illustrates how a paraphrase graph with inferred edge labels is constructed. For example, we can infer a positive link from node A to F, and a negative link from node A to C since A and D are paraphrases and C and D are non-paraphrases. By applying this method to the entire dataset, we expand the training set size for QQP by 60.7% (Section 4.1).

3.2 Fixing Mislabelings (Flipping Frenemies)

Based on the concept of structural balance for signed graphs, a balanced paraphrase graph can either have the entire sets of sentences being paraphrases of each other, or multiple subset groups of paraphrases with several sentences from different groups being connected with negative links. Our algorithm finds inconsistencies by identifying negative edges within a paraphrase cluster. Given the transitive relation of paraphrases, we correct the false negative links into positive. We found 88 mislabeled pairs in the QQP training set, and 2 pairs in the testing set. See Appendix A.1 for some examples, and A.3 for entire list of identified pairs.

For clusters with only negative edges like the triad in Figure 1c, even though the relation is imbalanced according to the definition, we are unable to determine whether there should be a pair of paraphrases in the graph without knowing the actual semantic meaning of the sentences. Therefore, we use the weaker form of structural balance to represent graphs with all negative edges. We only consider the negative links within a paraphrase cluster as potentially mislabeled relations.

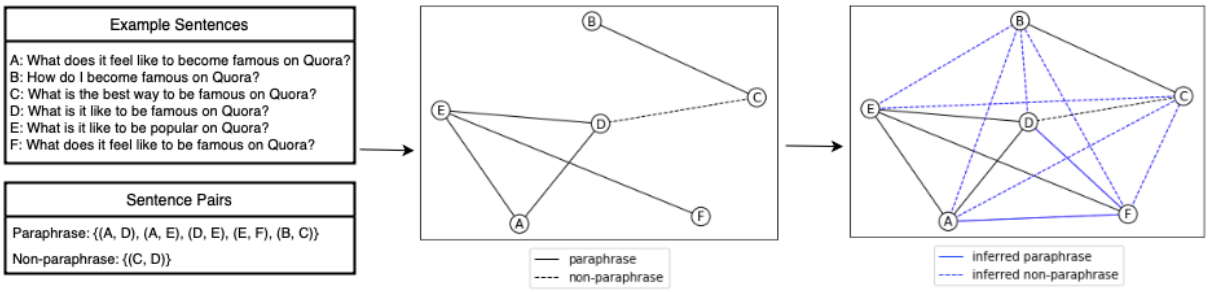


Figure 2: Example paraphrase graph constructed from example pairs from the Quora Question Pairs (QQP) dataset. The right most figure shows the inferred relations from the paraphrase graph.

4 Experiments

To understand the effectiveness and impact of our augmentation and correction methods, we compare the preforms of BERT models fine-tuned to the paraphrase identification task on the original QQP dataset and three datasets derived using the graph-based methods from the previous section.

4.1 Datasets

The Quora Question Pairs (QQP) dataset (Iyer et al., 2017) is based on questions extracted from Quora, where they aim to reduce the frequency of duplicated questions. Each pair is labeled as *duplicate* or *non-duplicate*. Duplicated questions are identified as having the same intent, meaning that they can be answered by the same answer. We consider the duplicate and non-duplicate labels comparable to paraphrase and non-paraphrase, and use the more familiar paraphrase terminology hereafter. This dataset is well suited to our approach since there are many sentences that appear in different pairs.

In addition to the original QQP dataset, we derived three additional datasets using the data augmentation and label correcting methods introduced in Section 3. Table 1 summarizes the four datasets.

Our inference method (Section 3.1) finds over 114,000 new paraphrase pairs and 137,000 non-paraphrase pairs across the dataset, expanding the training set by over 60%, and the testing set around 75%. The paraphrase ratio of the augmented training set remains similar as the original set. However, the ratio increases in the augmented testing set indicating the paraphrase clusters are sparser in the testing set. Our inconsistent label detection method (Section 3.2) detects 88 problematic labels in the training set and 2 problematic labels in the testing set. We flip the values of these labels in the Original-Flipped and Augmented-Flipped.[1]

[1]Other approaches would be worth exploring in future work such as removing the problematic pairs, manually inspecting them, and considering other labels involving sentences in problematic pairs as also likely to be problematic.

4.2 Model Training

We fine-tune the pretrained $\text{BERT}_{\text{BASE}}$ model on the four datasets with the default configuration (Devlin et al., 2019), and implement early stopping during training. We train the model on each dataset five times independently, and report the average accuracies and F1 scores in Table 2 and the detailed results with standard deviation in Appendix A.2.

4.3 Result Analysis

As shown in Table 2, the model trained on the Augmented-Flipped dataset has the best performance (both Accuracy and F1) on all testing datasets. The improvement in model accuracy on the Original dataset due to augmenting the training set is modest, but significant. The improvement increases when the flipped training sets are used, and is most substantial (reaching an error rate below 6%, compared to the original 10% error rate) when the testing is done using the Augmented testing set. According to the leaderboard of GLUE benchmark (Wang et al., 2018), an ALBERT based model (Lan et al., 2019) and ERNIE (Sun et al., 2019) are currently the top two models on QQP task with an accuracy of 91.0% and 90.9% on the original testing set. Comparing to these state-of-the-art models, we can reach a competitive performance with the simple data augmentation proposed in this work.

The models that trained on the Original set has a small performance drop when tested on the Augmented testing set. Since this testing set has a much higher paraphrase ratio, it means that the original model is better at predicting non-paraphrases than paraphrases. It fails to give correct predictions on the augmented paraphrase pairs. This also shows the benefit of augmenting the sentence pairs by rep-

Dataset	Training Set Size		Testing Set Size		Paraphrase Ratio (%)	
	Paraphrase	Non-paraphrase	Paraphrase	Non-paraphrase	Training	Testing
Original	134,378	229,468	14,885	25,545	36.93	36.82
Original-Flipped	134,446	229,380	14,886	25,544	36.96	36.82
Augmented	220,890	363,986	42,570	28,164	37.77	60.18
Augmented-Flipped	220,978	363,898	42,572	28,162	37.78	60.19

Table 1: Training and testing size and percentage of paraphrase pairs for each dataset. *Dataset*-Flipped denotes the dataset with the problematic labels flipped.

	Testing Set:	Original		Original-Flipped		Augmented		Augmented-Flipped	
	Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Training Set	Original	90.35	87.05	90.09	86.72	89.72	91.50	89.72	91.50
	Original-Flipped	90.15	86.78	90.16	86.80	93.47	94.59	93.46	94.58
	Augmented	90.61	87.48	90.61	87.48	93.89	94.95	93.87	94.94
	Augmented-Flipped	**90.96**	**88.01**	**90.95**	**88.00**	**94.21**	**95.23**	**94.19**	**95.22**

Table 2: Model performance evaluated on the four datasets. Both accuracy and F1 score are scaled by 100.

resenting sentence pair relations as a graph, which helps us generate more paraphrase pairs for training and improve model accuracy on paraphrases.

Since there are only two mislabeled sentence pairs in the testing set (and 88 in the training set), it is unsurprising that the impact of flipping the inconsistent labels is small. Still, in all cases we observe the models trained with the flipped training sets have higher accuracy than those trained on the corresponding dataset with the problematic labels. Interestingly, we find that the model trained on Original-Flipped reaches a similar performance as the model trained on Augmented, when tested on the Augmented and Augmented-Flipped testing sets. This shows the benefits of correcting the labels identified as problematic.

5 Related Work

The most closely related work from Shakeel et al. (2020) also applies paraphrase graphs to generate additional paraphrase and non-paraphrase pairs. Similar to our method, they infer non-paraphrase pairs from sentences within different paraphrase groups, and use transitivity to find new paraphrase pairs. Different from our work, they generate additional paraphrase pairs by pairing sentences to themselves, and reversing the order of each sentence pair. Other than using structural balance, their method can only identify conflicted labels between pairs of sentences. In addition, they only apply data augmentation on the training sets and evaluate their models directly on the original testing sets. We infer additional data and identify conflicts for both training and testing, which illustrates the full potential of our data augmentation method.

Besides, Chen et al. (2012) propose a graph-based method to improve the quality of paraphrase generation. They represent phrases as nodes and translation similarities as edges from a bilingual parallel corpus, and infer paraphrases with the pivot based method, which finds phrases with the same translation. However, this method can only infer new paraphrases within a path length of two. Homma et al. (2017) use a simpler approach by generating new paraphrase pairs with the reflexive and symmetric property of paraphrases with no graph involved. The non-paraphrase pairs are sentences randomly selected from two different pairs, which can not be guaranteed to have a correct relation.

6 Conclusion

In this paper, we show the benefit of representing datasets as graphs. We develop methods based on graph theory to automatically expand a paraphrase dataset and improve labeling consistency. Our experiments show an improvement on the Augmented-Flipped testing set after correcting the conflicted labels in the Original training set, and the combination of the two methods produce a model that gives the best performance across all testing sets.

References

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Mei-Hua Chen, Shi-Ting Huang, Chung-Chi Huang, Hsien-Chin Liou, and Jason S. Chang. 2012. PREFER: Using a graph-based approach to generate paraphrases for language learning. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

E. W. Dijkstra. 1959. A note on two problems in connexion with graphs. *Numer. Math.*, 1(1):269–271.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*.

Frank Harary. 1953. On the notion of balance of a signed graph. *Michigan Math. J.*, 2(2):143–146.

Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112. PMID: 21010780.

Yukiko Homma, Stuart Sy, and Christopher Yeh. 2017. Detecting duplicate questions with deep learning.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

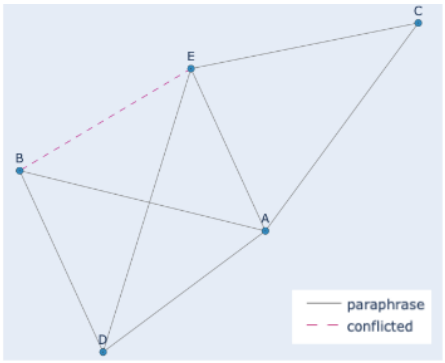
Muhammad Shakeel, Asim Karim, and Imdadullah Khan. 2020. A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. *Information Processing and Management*, 57:102204.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

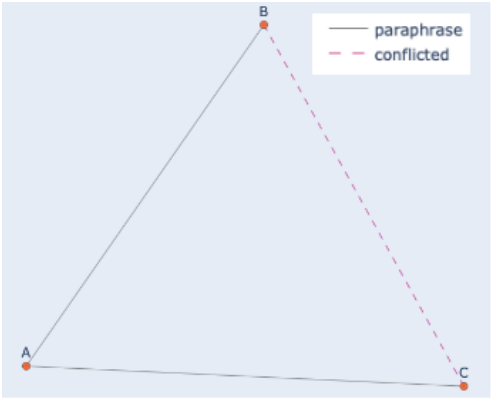
A Appendix

A.1 Mislabeling Examples



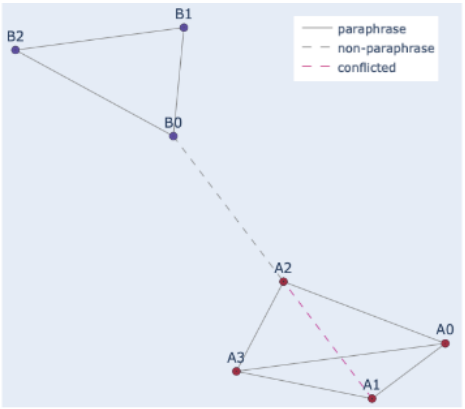
Sentences
A: What's the one thing you think about when you wake up?
B: What is the first thing you think when you wake up in morning?
C: What's the first thing you do when you wake up every morning?
D: What's the first thing you do after you wake up?
E: What is the first thing you do when you wake up?

(a)



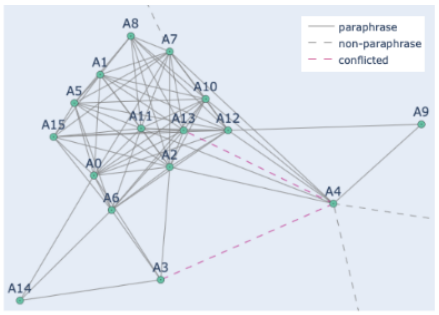
Sentences
A: What is the colour of sun?
B: What is the colour of the Sun?
C: What is the color of the sun?

(b)



Sentences
A0: How do I recover deleted text messages on my iPhone?
A1: How can I recover the deleted messages from my iPhone 5?
A2: How do you recover deleted messages from an iPhone?
A3: How can I recover deleted text message in iPhone?
B0: How can I recover deleted text messages?
B1: Can I retrieve deleted text messages?
B2: How do I retrieve deleted text messages?

(c)



Sentences
A0: What are some ways to delete my Yahoo Mail account permanently?
A1: How do you permantley delete a Yahoo e-mail account?
A2: What is the safest way to delete my Yahoo account?
A3: What is the best way to disable your Yahoo account permanently?
A4: How do you disable a Yahoo account?
A5: What are some ways to delete my Yahoo Mail account?
A6: How do you delete your Yahoo account permanently?
A7: How do you delete your Yahoo email account?
A8: How can you terminate your Yahoo e-mail account?
A9: How do I deactivate a Yahoo! account?
A10: How can you delete your Yahoo mail account?
A11: How can can I delete my yahoo email account?
A12: How can you delete your Yahoo account?
A13: How do you permanently delete your Yahoo account?
A14: How can I deactivate my yahoo account permanently?
A15: How can you delete your Yahoo Mail ID?

(d)

Figure 3: Graphs with Inconsistent Labelings.

A.2 Evaluation Results

Training Set		Testing Set			
		Original	Original-Flipped	Augmented	Augmented-Flipped
Original	Acc	90.35 ±0.14	90.09 ±0.11	89.72 ±0.36	89.72 ±0.37
	F1	87.05 ±0.33	86.72 ±0.30	91.50 ±0.37	91.50 ±0.38
	Recall	88.13 ±1.55	87.92 ±1.53	91.99 ±1.34	91.99 ±1.33
Original-Flipped	Acc	90.15 ±0.06	90.16 ±0.02	93.47 ±0.12	93.46 ±0.12
	F1	86.78 ±0.24	86.80 ±0.18	94.59 ±0.11	94.58 ±0.11
	Recall	87.92 ±1.22	87.96 ±1.13	94.80 ±0.42	94.79 ±0.43
Augmented	Acc	90.61 ±0.19	90.61 ±0.21	93.89 ±0.23	93.87 ±0.26
	F1	87.48 ±0.24	87.48 ±0.28	94.95 ±0.19	94.94 ±0.21
	Recall	89.07 ±0.42	89.04 ±0.53	95.48 ±0.22	95.47 ±0.23
Augmented-Flipped	Acc	90.96 ±0.11	90.95 ±0.12	94.21 ±0.04	94.19 ±0.04
	F1	88.01 ±0.11	88.00 ±0.12	95.23 ±0.03	95.22 ±0.03
	Recall	90.14 ±0.40	90.08 ±0.39	96.05 ±0.18	96.04 ±0.17

Table 3: Model performance evaluated on the four datasets. All metrics reported are scaled by 100. Standard deviations are calculated from training models five different times on the same training set.

A.3 Sentence Pairs with Conflicted Relation

This section shows all the sentence pairs we identified with conflicted relation in the QQP dataset. All the pairs are originally labeled as non-paraphrase, but are reachable by each other along the paraphrase links in the graph.

No.	Sentence Pair
1	What is the Sahara, and how do the average temperatures there compare to other deserts? What is the Sahara, and how do the average temperatures there compare to the ones in the Thar Desert?
2	What is the Sahara, and how do the average temperatures there compare to other deserts? What is the Sahara, and how do the average temperatures there compare to the ones in the Gibson Desert?
3	”What should I do if Quora marks my question as ””Needs Improvement””?” Why do all of my questions get markers for needing improvement no matter how carefully I word them?
4	Why is Saltwater Taffy candy imported in Italy? Why is salt water taffy candy either imported or unknown outside the USA?
5	Why is Saltwater taffy candy imported in Germany? Why is salt water taffy candy either imported or unknown outside the USA?
6	Why is Saltwater Taffy candy imported in Portugal? Why is salt water taffy candy either imported or unknown outside the USA?
7	How can I know who unfollowed me on Instagram? How do I know if someone unfollowed me on Instagram?
8	How do I see who viewed any video on instagram? How do l see who viewed my videos on Instagram?
9	I forgot my Facebook password and email password. How can I log into Facebook? I have forgotten my Facebook password and email password also?
10	How do I delete my own question from Quora? Why can’t you delete your own questions on Quora?

Table 4: Sentence pairs with conflicted relation in the QQP training set. (continued)

No.	Sentence Pair
11	Can skipping increase my height? How can you increase your height?
12	Height: How would a 14 year old increase his height? How can you increase your height?
13	How can I increase my followers count in Quora? How do you become famous and receive a lot of views, upvotes, and followers on Quora?
14	How do I know whether someone really loves me? How do you know when someone really loves you?
15	What are some interesting facts that I should know? What is the most interesting fact that I know, but you don't, but I shouldn't?
16	What are some interesting facts that I should know? What is the most interesting fact that you know and I don't, but I should?
17	Can history repeat itself? What proves that history repeats itself?
18	How can I prepare for SSC CGL at home? Which is the best way to prepare for SSC CGL at home or by ourselves without coaching?
19	What are the best places to visit in Kerala? Why should I visit Kerala?
20	When do people usually get married? Why do people get married?
21	Why do people insist on driving slowly in the left (passing) lane? Why do people prefer driving in the overtaking lane?
22	How can I track my cheating partners devices? How do I track my cheating partner?
23	How do you disable a Yahoo account? How do you permanently delete your Yahoo account?
24	How do you disable a Yahoo account? What is the best way to disable your Yahoo account permanently?
25	What is digital marketing?What are the basics of Digital Marketing?What is digital marketing strategy?Who needs digital marketing services? What is exactly is digital marketing?
26	What are the safety precautions on handling firearms proposed by the NRA in the entire U.S. including it's territories and possessions? What are the safety precautions on handling shotguns proposed by the NRA in Mississippi?
27	Can you get pregnant the day before your period starts? When can women get pregnant in the menstrual cycle?
28	How do you find your life's purpose? What should the purpose of your life?
29	How do you find your life's purpose? What actually is the purpose of life?
30	What are the best programming languages to learn today? What is the best programming language for a beginner?
31	How do I prepare for IAS exam? What is best way to crack the Civil Services Exam?
32	How do I prepare for civil services? How do I start preparation for civil services exams?
33	Is Spotify not available in India? Why Spotify is not available in India?

Table 4: Sentence pairs with conflicted relation in the QQP training set. (continued)

No.	Sentence Pair
34	How much weight will I lose in a year if I exercise a lot and stop eating and drinking? If I start to exercise a lot, stop eating and drinking, will I lose a significant amount of weight in a year?
35	Is pro wrestling fake? Wwe is real fight?
36	Does a superfluid dark matter which ripples when Galaxy clusters collide and waves in a double slit experiment relate GR and QM? Is a sea of massive gravitons what ripples when galaxy clusters collide and is it what waves in a double slit experiment?
37	How can I make a girl fall in love with me? I haven't even kissed a girl. How can I make a girl fall in love with me?
38	Can I raise seed funding by writing a compelling business proposal alone? What is the best way to fund an idea?
39	How do I quickly and efficiently learn a new language? How do you learn to speak a foreign language?
40	Who is the most beautiful main female character in Game of Thrones? Who is the most beautiful woman on Game of Thrones?
41	Can hamsters eat carrot? Why or why not? Can hamsters eat honey? Why or why not?
42	How can I get rid of cellulite on my stomach? How do I get rid of cellulite on my butt?
43	Has anyone seen a genuine UFO? What was it like? Have you ever seen UFO?
44	How can I write a essay? How should one write an essay on myself?
45	How did India benefit from Indus water treaty with Pakistan? Why did India sign the Indus Water Treaty?
46	Who has been the most important person in your life and why? Who is the most important person in your life?
47	What app allows you to listen to music without WiFi or Internet? What music app is free without wifi connection?
48	How can I get free iTunes gift cards online? What's the best way to legally get free iTunes gift cards?
49	How can I learn to play chess? How do I play chess?
50	Can a pair of positive and negative energy virtual particles pop into existence from the vacuum? Do virtual particles and energy in vacuum really exist? How do we know if they exist if we can't observe them? Where are they?
51	Do you have a tattoo that you regret getting? Do you regret getting a tattoo?
52	What are some of the wittiest answers on Quora? What's the best answer in Quora?
53	Could the U.S. take over the world? How could the U.S. take over the world?
54	Can we store the energy from lightining? Why can't we harness lightning energy?
55	What is the first thing you do when you wake up? What is the first thing you think when you wake up in morning?

Table 4: Sentence pairs with conflicted relation in the QQP training set. (continued)

No.	Sentence Pair
56	What superpower would you like to have? What superpower would you not want, and why?
57	How does the HP OfficeJet 4620 Airprint compare to the HP DesignJet Z3200 24-in PostScript Photo Printer? How does the HP OfficeJet 4620 Airprint compare to the HP DesignJet Z5200 44-in Photo Printer?
58	Do you ever shut up? On your planet, do people ever shut up?
59	How can I recover the deleted messages from my iPhone 5? How do you recover deleted messages from an iPhone?
60	What is better? Quora or Yahoo Answers.? Why is Quora more popular than Yahoo answers?
61	What is the color of the sun? What is the colour of the Sun?
62	How do I engage in chat with a girl? How does one chat with girls?
63	How can you be more attractive? What can I do to make myself more attractive?
64	Why didn't Philippine ask the UN backed courts to rule, instead of the black-box operated and private Arbitration to rule for the SCS? Why didn't the Philippines ask UN backed courts, instead of Permanent Court of Arbitration over the water territorial disputes with China?
65	How do I ask a question intelligently? How do I ask a smart and intelligent question?
66	How do I stop negative self-talk and judging people? How do you stop negative self-talk?
67	What are the best rock bands? What is the best Rock band ever?
68	What are the best reasons for doing an MBA? Why should I do an MBA?
69	What are your favorite music and video clips? What's your favorite music video?
70	What are some small, but hot topics for mechanical engineering for PPT? What are the best easy and understandable topics for ppt for mechanical engineering?
71	What is the most happiest moment in life? What were the most happiest moments in your life?
72	What are the happiest moments in your life? What is the most happiest moment in life?
73	Do you think there is life on Mars? Is there life on Mars?
74	Why do black people have such intolerance towards LGBTQ people? Why is there hatred towards LGBTQ people?
75	How do I avoid heat in the stomach that are causing pimples? What are the best ways to avoid pimples?
76	Are there better alternatives to the current education system? What can be a feasible education system which is better than the present education system?
77	How do I prevent rape in Delhi? Is their any way to prevent rapes in Delhi?

Table 4: Sentence pairs with conflicted relation in the QQP training set. (continued)

No.	Sentence Pair
78	What are some good hotels in varanasi? Which hotels are best to stay in Varanasi for a family trip?
79	How do I find a mentor to help me in life? What is the best way to find mentors in life?
80	How will GST affect my taxation as a common man? How will the GST affect the common man in India?
81	How do I trademark a word I made up? How do you trademark a word that you made up?
82	What is it in soil that makes it smell so good after the first rain? Where does the smell of rain come from and why does it happen?
83	What is the maximum number of times you can fold a piece of paper? Why is it that a piece of square paper of any length can be equally folded only 8 times maximum?
84	How does the law of independent assortment compare to the law of segregation? What is the difference between Medel's law of segregation and law of independent assortment?
85	What is the second largest planet in the solar system, and how does its composition compare to Jupiter? What is the second largest planet in the solar system, and how does its composition compare to the other planets?
86	Which one is your favourite novel and why? Who is your favourite novelist and novel?
87	What are the benefits of water fasting? What are the health benefits of fasting?
88	What is the best online game? What is the best online games site?

Table 4: Sentence pairs with conflicted relation in the QQP training set.

No.	Sentence Pair
1	How cold can the Gobi Desert get, and how do its average temperatures compare to the ones in other deserts? How cold can the Gobi Desert get, and how do its average temperatures compare to the ones in the Gibson Desert?
2	Do running increase your height? How can you increase your height?

Table 5: Sentence pairs with conflicted relation in the QQP testing set.