Wibbly Wobbly Timely Wimely Questions: Classifying Timely And Timeless Questions

Vatsal Agarwal [1] **Pauline Comising** [1] **Harish Kumar** [1] **Madhava Paliyam** [1] **Nikhil Pateel** [1]

Abstract

Classifying timely and timeless questions can be a useful task for natural language processing. We present a method for determining if a question is timely or timeless using a pretrained BERT model and the Google NQ dataset. Past revisions of the source Wikipedia page are retrieved and the answers are compared to see how they change over time. We show that we can find timely questions with greater than chance probability and timeless questions with high accuracy using our method.

1. Problem Overview

Question answering (QA) is a fundamental task in natural language processing. The objectives are to understand questions posed in natural language and retrieve relevant information within a database or text to answer them. The answers to some questions remain unchanged and static regardless of the time they are asked. We will refer to these questions as timeless questions. As examples, the answers to the questions "When was the Magna Carta signed" and "Who was the first person to walk on the moon" do not change over time. On the other hand, some questions will have different answers based on when they are asked, and we will refer to these as timely questions. To illustrate, the answer to "Who won the world cup" will change every four years, and the answer to "How many farms are there in Alaska" will vary depending on how many Alaskan farms exist at the asking time.

The purpose of our project is to build a pipeline that determines the "timeliness" of questions, as understanding whether a question is timely or timeless could be useful for accurate question answering. When an individual asks a timely question, they would expect the system to return the most recent answer. Detecting timeliness in queries is a nifty skill for voice assistant software and chatbots; if a question is understood to be timely, updated answers could be routinely checked to provide accurate results for possible future queries.

[1]Department of Computer Science, University of Maryland, Maryland, USA. Correspondence to: Nikhil Pateel <github.com/npateel/timelywimely>.

1.1. Related Work

There has been extensive work examining methods to improve temporal reasoning in natural language processing tasks such as question-answering. The 2002 TERQAS workshops spurred initial work in this area and aimed to answer temporally-based questions about events and entities in articles and lead to the development of TimeML, a set of rules for encoding temporal information in documents (Radev et al., 2002). (Breck et al., 2000) was one of the first to build a QA system which incorporated temporal information. (Saquete et al., 2009) explored more complex strategies to incorporate temporal information. Specifically, they proposed a method to first identify temporal questions and decompose them into simpler questions. They then obtain answers to this set of questions, and then recombine them following the original temporal relationships to generate the complete answer. Recently, the TEQUILA method was described by (Jia et al., 2018) as an extension of (Saquete et al., 2009). Thus, our work focuses on improving the first task of better identifying temporal questions.

2. Methodology

2.1. Dataset

We used Google's Natural Questions (NQ) dataset (Kwiatkowski et al., 2019). The dataset has been used for open-domain question answering research, and is "the first large publicly available dataset to pair real user queries with high quality annotations of answers in documents". It consists of 307,373 anonymized natural questions asked via the Google search engine, and each question is paired with the full HTML markup of a Wikipedia page. Human annotators have marked the location of long and short answers to each question. A model that is trained on the NQ dataset is required to read the full Wikipedia article, and locate the answers. From our analysis, we estimate that less than 10% of the questions were timely questions.

SQuAD (Stanford Question Answering Dataset) 2.0 (Ra-

jpurkar et al., 2018) is a similar dataset that we explored. However, it has some key differences; questions in SQuAD have been artificially generated (unlike the real user queries of NQ), and each question is paired with a text paragraph from a Wikipedia page (whereas NQ provides the full page text).

2.2. String Comparison Metrics

2.2.1. Levenshtein Distance

We use a minimum string edit distance (otherwise known as the Levenshtein Distance) to measure the similarity between two answers. Since this metric does not rely on tokenization, it works well with smaller strings. The lack of word boundaries for Levenshtein distance can also be a major issue. Words like "kitten" and "smitten" will have a very low edit distance even though they are linguistically very different. When combined with a tokenizer, one can sort tokens alphabetically and then compute the Levenshtein distance to get a metric that respects word boundaries slightly more. However, the underlying metric will still disregard the differences in tokens. It is for this reason that we also used a BLEU score.

2.2.2. BLEU

BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), our other answer difference metric, is a measurement of overlap between one or multiple reference sentence(s) and a candidate sentence. BLEU defines overlap as the percentage of candidate sentence N-grams that are present in a reference sentence. When there are multiple reference sentences, the highest score is returned. We calculated this metric using the NLTK (Loper & Bird, 2002) package for BLEU scores, specifically utilizing the `sentence_bleu` function. The smallest unit in each function argument is a 1-gram, often a word.

2.3. Pipeline

2.3.1. Answer Prediction

Retrieving answers for all revisions of a single question worked as follows. First, for each sample in a 20,000 question subset of the NQ train dataset, the question text, question HTML, Wikipedia article link, and all short answers were retrieved. We obtained the past revisions of the Wikipedia article for 2008, 2010, 2012, 2014, 2016, 2018, and current day by using the Wikimedia API. We chose this time period because we felt this represented a long enough time-span such that Wikipedia would have accurate information on most of the questions. The HTML was parsed using the Beautiful Soup library and the first 25 paragraphs of the article were extracted and fed to BERT along with the question.

The BERT-SQuAD model is a transformer model trained on the SQuAD dataset. BERT is a bidirectional transformer and has been proven to be successful at many natural language processing tasks such as machine translation and question answering (Devlin et al., 2018). We chose to use an implementation by (Kamal Raj, 2020) which we believed would be successful with the NQ dataset since the two are similar in the type of contextual information and questions that are available. By feeding each year's Wikipedia article revision and the question text to the model, we acquire an answer for each year.

2.3.2. Answer Filtering

After getting an answer for each year, we hypothesized that a timely question would have answers that change year to year, while a timeless question would have similar answers no matter which year was used as the context. However, before we could identify changes with the answers, we had to account for errors with the model's ability to answer questions. First, we removed any questions in which any of the answers returned a page that did not exist, or was redirected. Although a page that did not exist could indicate a timely question, we decided to remove those in case there was a problem with the Wikipedia link or the model.

After this initial filtering, our subset of 20,000 questions and answers was reduced to 10,183 samples. Then, using the string Levenshtein distance, we calculated the difference of each answer to the correct short answer provided in the NQ dataset. Some questions had multiple correct short answers, in which case we compared to the ground truth that had the closest match to the predicted answer. Any samples which had an incorrect answer using the predicted answer from the original HTML in the question were filtered out from our data. We defined an incorrect answer as having a string comparison score of at least 70. After removing these samples, we were left with a dataset of 3,408 samples to work with.

2.3.3. Answer Scoring

We defined the string distance score using the functions in the `fuzzywuzzy` library. This was an implementation of the Levenshtein distance that included two scores: token set score and token sort score. Token set score compared across the entire strings, while for the token sort score, the order of the tokens did not matter. We computed the average of both these scores for all the answers outputted from BERT compared with the ground truth answer. Next, for each question we computed the maximum string score and minimum string score and subtracted the two, with the intuition that if the range was large, then the answer would have changed a lot over the years.

Similarly, we also used BLEU scores to compare answers.

max - min	Question	GT Answers	Prediction: original	Avg Score: original	Avg Score: 2008-01-01 00:00:00	Predicted: 2008-01-01 00:00:00	Avg Score: 2010-01-01 00:00:00	Predicted: 2010-01-01 00:00:00	Avg Score: 2012-01-01 00:00:00	Predicted: 2012-01-01 00:00:00	Avg Score: 2014-01-01 00:00:00	Predicted: 2014-01-01 00:00:00
0.0	what is abby's real name from ncis	Pauley Perrette --	Valdosta State University Pauley Perrette	77.0	100.0	Pauley Perrette	100.0	Pauley Perrette	100.0	Pauley Perrette	100.0	Pauley Perrette

max - min	Question	GT Answers	Prediction: original	Avg Score: original	Avg Score: 2008-01-01 00:00:00	Predicted: 2008-01-01 00:00:00	Avg Score: 2010-01-01 00:00:00	Predicted: 2010-01-01 00:00:00	Avg Score: 2012-01-01 00:00:00	Predicted: 2012-01-01 00:00:00	Avg Score: 2014-01-01 00:00:00	Predicted: 2014-01-01 00:00:00
68.0	what is the population of the dalles oregon	13,620 --	13,620	100.0	50.0	12,156	32.0	12,156 at the 2000 census	32.0	12,156 at the 2000 census	100.0	13,620
78.0	who is going to host the 2024 olympics	Paris , France - -	Paris , France	100.0	26.0	Philippines	27.0	winning bid should be announced in 2017	21.0	bid should be announced in 2017 at the 129th IOC Session.	18.0	a city elected by the International Olympic Committee.

max - min	Question	GT Answers	Prediction: original	Avg Score: original	Avg Score: 2008-01-01 00:00:00	Predicted: 2008-01-01 00:00:00	Avg Score: 2010-01-01 00:00:00	Predicted: 2010-01-01 00:00:00	Avg Score: 2012-01-01 00:00:00	Predicted: 2012-01-01 00:00:00	Avg Score: 2014-01-01 00:00:00	Predicted: 2014-01-01 00:00:00	Avg Score: 2016-01-01 00:00:00	Predicted: 2016-01-01 00:00:00
100.0	who is the original singer of i will always love you dolly parton	Dolly Parton --	Dolly Parton	100.0	74.0	country singer-songwriter Dolly Parton	100.0	Dolly Parton	100.0	Dolly Parton	0.0	1974	78.5	singer-songwriter Dolly Parton.

	max - min	Question	GT Answers	Prediction: original	Avg Score: original	Avg Score: 2008-01-01 00:00:00	Predicted: 2008-01-01 00:00:00	Avg Score: 2010-01-01 00:00:00	Predicted: 2010-01-01 00:00:00	Avg Score: 2012-01-01 00:00:00	Predicted: 2012-01-01 00:00:00	Avg Score: 2014-01-01 00:00:00	Predicted: 2014-01-01 00:00:00
314	100.0	Robert Pershing Wadlow - -	who was the tallest person that ever lived	82.5	Robert Wadlow - wikipedia Robert Pershing Wadlow	100.0	Robert Pershing Wadlow	100.0	Robert Pershing Wadlow	100.0	Robert Pershing Wadlow	100.0	Robert Pershing Wadlow

Figure 1. Levenshtein Examples (top is timeless, middle is timely, third is false positive, bottom is false negative).

In the implementation of BLEU scores that we used, we were able to input multiple reference sentences and obtain the best match with the test sentence. In addition, we could choose the N-grams that the BLEU score used. Since we used short answers from the NQ dataset and our predictions were only a few words long, we decided to use unigrams and bigrams.

We computed four BLEU scores for each year's predicted answer.

1. Bigram BLEU: Bigram BLEU score based on the predicted answer and the ground truth answer.
2. Unigram BLEU: Unigram BLEU score based on the predicted answer and the ground truth answer
3. Previous years only Bigram BLEU: Bigram BLEU score between the predicted answer and using all the answers from the years before it as the reference.
4. Previous years only Unigram BLEU: Unigram BLEU score between the predicted answer and using all the answers from the years before it as the reference.

The scores across the years were averaged to get four scores for each question. The maximum of these scores was used to qualify timely and timeless questions. Timely ones would have low maximum BLEU scores and timeless ones would have higher ones.

3. Results

3.1. Levenshtein Results

We first evaluated the answers using the aforementioned string distance score. Generally, questions with a small difference between the maximum and minimum Levenshtein distance were reliably timeless. Despite the rarity of timely questions in our dataset, our approach was still somewhat effective at identifying them. Many questions that had a larger Levenshtein range were found to be timely.

Examples of correctly identified timeless and timely questions are shown in Fig. 1. Depicted are timely questions, such as those associated with changing populations or future events. We can note how the model's answer changes over time. While changes such as difference in population

Maximum BLEU Avg	Max BA Choice	Overall FW Avg	Question	GT Answers	Prediction: original	Avg Score: original	Avg Score: 2008-01-01 00:00:00	Predicted: 2008-01-01 00:00:00	Avg Score: 2010-01-01 00:00:00	Predicted: 2010-01-01 00:00:00	Avg Score: 2012-01-01 00:00:00	Predicted: 2012-01-01 00:00:00	Avg Score: 2014-01-01 00:00:00
100.0	Avg BLEU: Uni-Original	100.0	name of the horse in steptoe and son	Hercules --	Hercules	100.0	100.0	Hercules.	100.0	Hercules.	100.0	Hercules.	100.0

Maximum BLEU Avg	Max BA Choice	Overall FW Avg	Question	GT Answers	Prediction: original	Avg Score: original	Avg Score: 2008-01-01 00:00:00	Predicted: 2008-01-01 00:00:00	Avg Score: 2010-01-01 00:00:00	Predicted: 2010-01-01 00:00:00	Avg Score: 2012-01-01 00:00:00	Predicted: 2012-01-01 00:00:00
0.000000	Avg BLEU: Uni-Original	61.714286	when was the last time the philadelphia flyers made the playoffs	2017 -- 18 --	2017 -- 18	100.0	57.0	1980-81	36.0	1997	71.0	2010–11
0.000000	Avg BLEU: Uni-Original	37.857143	how many starbucks stores are in the world	28,218 --	28,218	100.0	33.0	15,011	33.0	16,635	67.0	18,887

Maximum BLEU Avg	Max BA Choice	Overall FW Avg	Question	GT Answers	Prediction: original	Avg Score: original	Avg Score: 2008-01-01 00:00:00	Predicted: 2008-01-01 00:00:00	Avg Score: 2010-01-01 00:00:00	Predicted: 2010-01-01 00:00:00	Avg Score: 2012-01-01 00:00:00	Predicted: 2012-01-01 00:00:00
5.952659	Avg BLEU: Uni-Previous	49.071429	when did the war of spanish succession end	1714 --	1712	75.0	7.5	the War of the Grand Alliance came to a close in 1697,	50.0	1697	50.0	1702

Figure 2. BLEU Examples (top is timeless, middle is timely, bottom is false positive).

are identified as expected, long, vague, or incorrect model outputs are still present. Both these differences in actual answer and model performance contribute to incorrect classification. Looking at the example timeless question, we see the model output remains the same over time, so our Levenshtein range is very small.

Despite its potential, we found that the model's ability to detect timely questions well was dependent on its ability to provide accurate answers. Oftentimes, the model would output an unreasonable answer for a single year, resulting in a higher Levenshtein distance between the ground truth. This caused false positives, depicted in Fig. 1. In order to tackle these issues, we explored methods to try and filter excessively long model outputs, but decided to utilize the BLEU score as an alternative metric to better capture semantic similarity.

3.2. BLEU Results

Using BLEU score metrics, identifying timeless vs timely questions was more successful. High BLEU scores corresponded to timeless questions and lower ones were indicative as timely questions. Similar to Levenshtein, the highly represented class of timeless questions were found reliably amongst those with the highest maximum BLEU score. An example is seen in Fig. 2 where the answers are non–changing. Timely questions found within the lowest scores showed changing answers and didn't have poor model output present. While inconsistent answers may have had a confounding effect on the Levenshtein approach, the BLEU score could ignore bad output and decrease the prevalence of false positives. Additionally, in the cases of false negatives, while there was poor model output present, there were also changes in answers seen in Fig. 2 where the year would change. These were incorrect yet reasonable answers.

In general, the concentration of timely questions within the bottom 5 percent of maximum average BLEU scores was much higher than that of its Levenshtein counterpart. This was explained by BLEU score's ability to be more robust to outliers and its use of "distance" from previous answers, not just the dataset's ground truth answers. The latter difference allows for each year's output to be directly compared to one another; Levenshtein only compared outputs with the ground truth. This allowed for a more intuitive approach to checking change over time. In being able to use previous years' answers as reference sentences to test candidate answers against, a single year's poor output does not impact any average BLEU score. Since the wibbly output is often one of many reference sentences, BLEU scores with previous–year reference sentences can calculate the score using more accurate outputs. So although poor model output was still the main reason for timeless questions amongst timely questions, the prevalence of these was much lower when using BLEU. Despite one wobbly output, the BLEU score could still remain high and accurately qualify a question as timeless.

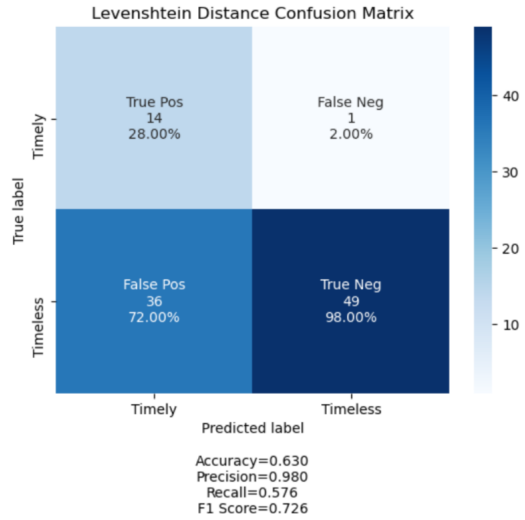

Figure 3. Confusion Matrix for Levenshtein Approach

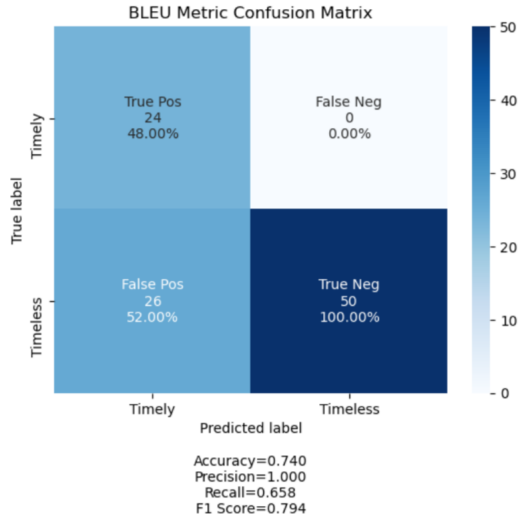

Figure 4. Confusion Matrix for BLEU Approach

3.3. Quantitative Analysis

To further investigate the efficacy of the two different metrics, we selected two sets of fifty questions that were filtered as timeless and timely. We then manually tallied the number of correctly classified questions for each set and visualized the results via confusion matrices along with their respective metrics in Fig. 3 and Fig. 4. Generally, we can observe that using the BLEU metric as a filter substantially reduces the number of timeless questions falsely predicted as timely, thereby increasing the accuracy and recall compared to the Levenshtein approach. Since we expected roughly 10% of questions to be timely, this means that for both of our models, we were able to detect timeliness better than chance. However, the proportion of questions with a low BLEU score is much smaller than the proportion of questions with a high Levenshtein range. This suggests that our current methodology may result in a higher number of false negatives with BLEU approach.

4. Limitations

While the BLEU score results are promising, there are some significant limitations with this methodology. The BLEU score approach filters out a large number of questions, giving a very high false negative rate. As for the Levenshtein methodology, we found that our timely detection rate was tightly linked with how well our model behaved. In general, we would expect that our model would have performed significantly better had it been trained on the NQ dataset beforehand. This would give our string metrics much more reliable data and improve the quality of our results. In addition, we had to filter out the majority of the dataset before even running our string comparisons largely due to issues with Wikimedia's API. We ended up filtering roughly 10,000 out of the 20,000 total questions because of these issues. An additional 7,000 were filtered from poor model performance.

Even with a perfect pre–trained model, we still run into some key limitations with our methodology. The first is that we assume that the answer to any question in the dataset can be found within the first 25 paragraphs. An analysis done by Google for the NQ dataset found that roughly 70% of questions can be found in a paragraph tag, but that leaves roughly 30% of questions that have answers in a table (which we do not add into our model inputs) (Kwiatkowski et al., 2019). Finally, as noted in the Results section, if a question's answer has not changed in the past 10 years, then our methodology will produce a false negative.

5. Future Work

With a longer project timeline, we would have liked to retrain a BERT model ourselves on the NQ dataset to get a more accurate model. In addition, once we have a more accurate set of timely and timeless questions, we would train a language model to classify questions based on their timeliness.

6. Contributions

6.1. Vatsal Agarwal

Vatsal helped with dataset processing and analysis, as well as the presentation and paper work.

6.2. Pauline Comising

Pauline ran the Google Trends baseline, helped with Wikimedia API integration, implemented BLEU score calculations, and helped with presentation and paper writing.

6.3. Harish Kumar

Harish ran the Google Trends baseline. He also helped with presentation and paper writing.

6.4. Madhava Paliyam

Madhava ran the BERT-SQUAD model on the train and development datasets. He implemented the comparisons using the `fuzzywuzzy` package. He also helped with the analysis.

6.5. Nikhil Pateel

Nikhil worked on the Google Trends baseline, Wikimedia API integration, as well as presentation and paper preparation.

Acknowledgements

We would like to thank Professors Jordan Boyd-Graber and Phillip Resnik for guidance on our project, as well as UMD for providing computational resources to make this project possible.

References

Breck, E., Burger, J., Ferro, L., Greiff, W., Light, M., Mani, I., and Rennie, J. Another sys called qanda. In *TREC*, 2000.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., and Weikum, G. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1807–1810, 2018.

Kamal Raj, Abubakar Abid, R. Bert-squad. https://github.com/kamalkraj/BERT-SQuAD, 2020.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Loper, E. and Bird, S. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pp. 63–70, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. URL https://doi.org/10.3115/1118108.1118117.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Radev, D., Sundheim, B., Ferro, L., Saurí, R., See, A., and Pustejovsky, J. Using timeml in question answering, 2002.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018. URL http://arxiv.org/abs/1806.03822.

Saquete, E., Vicedo, J. L., Martínez-Barco, P., Munoz, R., and Llorens, H. Enhancing qa systems with complex temporal question processing capabilities. *Journal of Artificial Intelligence Research*, 35:775–811, 2009.