# Temporal Relation Classification using Boolean Question Answering

**Omer Cohen**
Efi Arazi School of Computer Science
Reichman University, Israel
omeromy@gmail.com

**Kfir Bar**
School of Computer Science,
College of Management, Israel
kfirbarmail@gmail.com

## Abstract

Classifying temporal relations between a pair of events is crucial to natural language understanding and a well-known natural language processing task. Given a document and two event mentions, the task is aimed at finding which one started first. We propose an efficient approach for temporal relation classification (TRC) using a boolean question answering (QA) model which we fine-tune on questions that we carefully design based on the TRC annotation guidelines, thereby mimicking the way human annotators approach the task. Our new QA-based TRC model outperforms previous state-of-the-art results by 2.4%.

## 1 Introduction

Events in stories are not necessarily mentioned in a chronological order. The timeline of events is important for understanding the main narrative of a story as well as the correct order of actions. For example, the timeline may be used directly by clinicians looking for a convenient way to explore the disease course of their patients, or by algorithms to follow instructions in the right order, given as text, such as in cooking recipes. Building the timeline is done based on two main subtasks: (1) event extraction, that is, detecting the most important events in a given textual input, and (2) temporal relation classification (TRC), also known as temporal relation extraction, which is about putting two events, given as gold spans, in the right chronological order. For example, consider the following text: "Before you put the cake in the oven, say a little prayer." In the first subtask, known as *event extraction*, we would like to detect only the relevant events for our domain of interest. In this case, the words **put** and **say** are both verbs representing some relevant actions; therefore, we mark them as events. In the second subtask, TRC, we put every two events in a chronological order by classifying them using a closed set of temporal relations. In this case, the two events **put** and **say** should be assigned with the label *AFTER* indicating that **put** is happening after **say** in a chronological order.

In this study we focus on TRC, which is typically handled as a classification problem of two events provided along with the context in which they are mentioned. MATRES (Ning et al., 2018b) is one of the dominant datasets for TRC comprised of news documents manually annotated with temporal relation labels. The events are deterministically chosen to be all actions (mostly verbs) mentioned in the documents. Every pair of events $(n, m)$ are manually labeled with one of four labels: BEFORE ($n$ happened before $m$), AFTER ($n$ happened after $m$), EQUAL ($n$ and $m$ happened at the same time), and VAGUE (it is impossible to know which event happened before the other).

Traditional classification approaches have already been demonstrated for TRC. In this work, we get inspiration from a relatively new promising approach for solving natural language processing (NLP) tasks, in which the target algorithm is based on a reduction of the task to another problem. In our case, we solve the TRC problem using a model that handles the boolean question-answering (QA) task, which is about answering a Yes/No question given a passage used as a context. We decide to use boolean QA as our proxy problem due to the way the annotation work for building MATRES has been done. In the main annotation guidelines of MATRES (Ning et al., 2018b), the annotators are asked to assign a label to a pair of events $(n, m)$ by answering the two following questions: (1) Is it possible that the start time of $n$ is before the start time of $m$? and (2) Is it possible that the start time of $m$ is before the start time of $n$? There are four possible answer combinations, each is mapped to one label: (*yes*, *no*) ⇒ BEFORE, (*no*, *yes*) ⇒ AFTER, (*no*, *no*) ⇒ EQUAL, and (*yes*, *yes*) ⇒ VAGUE. Therefore, we transform an instance of TRC, composed of a pair of events and a document,

*Findings of the Association for Computational Linguistics: ACL 2023*, pages 1843–1852
July 9-14, 2023 ©2023 Association for Computational Linguistics

into a pair of Yes/No QA instances, one for each of the two questions, and then fine-tune a Yes/No QA model to answer them. The final prediction is made based on the combination of the Yes/No answers retrieved by the QA model.

## 2 Related Work

TRC has received increasing levels of attention in the past decade. There is a relatively long list of related shared tasks (Verhagen et al., 2007, 2010; Bethard et al., 2016; MacAvaney et al., 2017). Modern approaches for TRC use some sort of a neural network as a classifier. For example, Dligach et al. (2017) showed that a neural network that uses only words as input, performs better than the traditional models that process features which were manually created. A more modern approach for TRC is based on large pre-trained language models. Han et al. (2021) continued to pre-train a language model before fine-tuning it on TRC; Zhou et al. (2021) incorporated a global inference mechanism to tackle the problem at the document level; Han et al. (2019a) combined a recurrent neural network (RNN) over BERT (Devlin et al., 2019) embedding and a structured support vector machine (SSVM) classifier to make joint predictions; Ning et al. (2019) integrated BERT with a temporal commonsense knowledge base, and improved accuracy significantly by 10% over the previously known best result; and Han et al. (2019b) developed a multitask model for the two related subtasks, event extraction and TRC. Mathur et al. (2021) train a gated relational graph convolution network using rhetorical discourse features and temporal arguments from semantic role labels, in addition to some traditional syntactic features. Wang et al. (2022b) use a unified form of the document creation time to improve modeling and classification performance, and Wang et al. (2022a) improve the faithfulness of TRC extraction model. Zhang et al. (2021) built a syntactic graph constructed from one or two continuous sentences and combined it with a pre-trained language model. The best result so far has been reported recently by Zhou et al. (2022), who extract relational syntactic and semantic structures, and encode them using a graph neural network. In another recent work (Man et al., 2022), the authors introduce a novel method to better model long document-level contexts by detecting and encoding important sentences in the document. None of those studies use QA to address the TRC problem.

Our boolean QA-based approach continues to improve on Zhou et al.'s (2022) work, achieving a new stat-of-the-art result for TRC.

## 3 Datasets

We conduct experiments with two datasets. MATRES (Ning et al., 2018b) is a composition of three datasets (TIMEBANK, AQUAINT and PLATINUM) which were re-annotated following new guidelines. Following previous work, we use TIMEBANK and AQUAINT together as a training set and PLATINUM as a testing set. For validation and development we use a different dataset named TCR (Ning et al., 2018a), which has been used similarly in other works (Zhang et al., 2021). As mentioned above, MATRES has four labels: BEFORE, AFTER, EQUAL, and VAGUE. TimeBank-Dense (Cassidy et al., 2014), or TB-Dense in short, is the second dataset which we use in this work. TB-Dense has two additional labels: INCLUDES and IS-INCLUDED. Following common practices, we evaluate our models using the relaxed micro-average F1 score (i.e., for MATRES ignoring all mistakes on VAGUE instances during evaluation, and for TB-Dense completely removing VAGUE instances from the validation and testing sets). Overall, MATRES contains $12,736$ training instances, $837$ testing instances, and $2,600$ validation instances from TRC. TB-Dense contains $4,032$ training instances, $1,427$ testing instances, and $629$ validation instances. The label distributions is summarized under Appendix B.

## 4 Methodology

We design our problem as Yes/No question answering problem. Therefore, we fine-tune a pre-trained language model (PLM) by taking a Yes/No QA classification approach for which every instance is composed of a passage (text) and a question, provided along with a Yes/No answer. Our QA model is designed as a traditional classifier; the input is a concatenation of the passage and the question with a special separator token in between, and the output is a two-way label distribution vector. We use RoBERTa (Liu et al., 2019), which comes in two sizes, base and large; we use both.

An instance of TRC is composed of a document, two event spans, and a label. In order to use our QA model for TRC, we convert each such instance into two or three Yes/No QA instances, which we use for fine-tuning and testing. Each QA instance

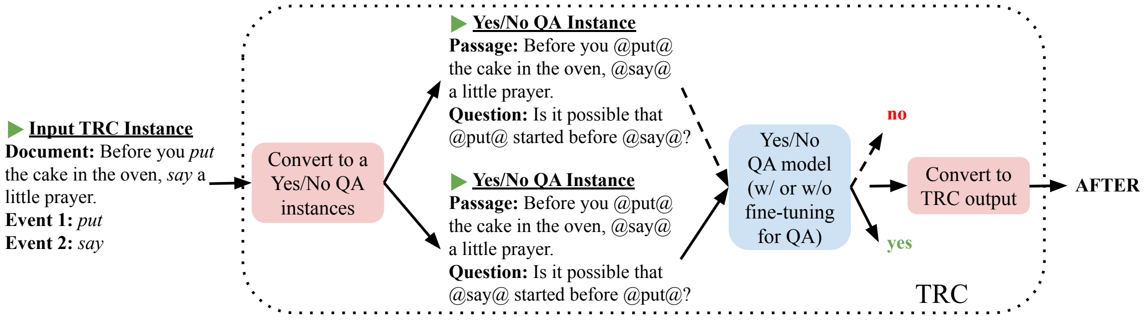
▶ Input TRC Instance
Document: Before you put the cake in the oven, say a little prayer.
Event 1: put
Event 2: say
Convert to a Yes/No QA instances
▶ Yes/No QA Instance
Passage: Before you @put@ the cake in the oven, @say@ a little prayer.
Question: Is it possible that @put@ started before @say@?
▶ Yes/No QA Instance
Passage: Before you @put@ the cake in the oven, @say@ a little prayer.
Question: Is it possible that @say@ started before @put@?
Yes/No QA model (w/ or w/o fine-tuning for QA)
no
yes
Convert to TRC output
AFTER
TRC

Figure 1: Using Yes/No QA for TRC. The input is a TRC instance from MATRES that we convert into two Yes/No QA instances, which we process with the QA model. The results are converted back into a TRC label using mapping rules (see the text and Table 1).

is composed of a passage and a question. Therefore, we cut the sentence from the input document, containing the spans of the two events, and use it as a passage. Sentence breaks are detected using full stops (e.g., a dot followed by a white space). The passage is paired with the Yes/No questions, generating multiple QA instances. MATRES uses a label set of size four, and TB-Dense has two additional labels: INCLUDES and IS-INCLUDED. Therefore, for MATRES we compose the following two question templates (<EVENT 1> and <EVENT 2> are used here as placeholders), inspired by the TRC annotation guidelines: **(1)** *Is it possible that <EVENT 1> started before <EVENT 2>*? and **(2)** *Is it possible that <EVENT 2> started before <EVENT 1>*? For TB-Dense, we add another question template: **(3)** *Is it possible that <EVENT 1> ended before <EVENT 2>*? We experiment with additional phrasing, as described in the following section. The answers to the questions are determined by the label of the TRC instance, using Table 1.

| Question Templates | | | MATRES | TB-Dense |
|---|---|---|---|---|
| 1 | 2 | 3 | | |
| *no* | *no* | <not used> | EQUAL | EQUAL |
| *yes* | *yes* | <not used> | VAGUE | VAGUE |
| *yes* | *no* | *yes* | BEFORE | BEFORE |
| *yes* | *no* | *no* | BEFORE | INCLUDES |
| *no* | *yes* | *yes* | AFTER | IS-INCLUDED |
| *no* | *yes* | *no* | AFTER | AFTER |

Table 1: Mapping of Yes/No answers to MATRES and TB-Dense labels. Question 3 is not used in MATRES. In TB-Dense, question 3 is not used only when the answers to questions 1 and 2 are either (*no*,*no*) or (*yes*,*yes*), respectively.

Each QA instance is processed independently during fine-tuning. At inference time we run the instances through the model and assign a TRC label based on the answers.

Naturally, a document may contain more events than the two relevant ones. Therefore, we use markers (Baldini Soares et al., 2019) in order to mark the two relevant events. Specifically, each relevant event is surrounded by the '@' character in both, the passage and the question. Figure 1 demonstrates how we process a MATRES instance.

## 5 Experiments and Results

Table 2 summarizes our evaluation results on MATRES and TB-Dense, using the two sizes of RoBERTa. We compare our results with two baseline models, and some previous work. We experiment with three variations for the questions (only for the two MATRES-related questions; for TB-Dense we only use the best out of the three),[1] as reported in the three first rows of Table 2:

**QV1**: *<EVENT1> before <EVENT2>?*
**QV2**: *Is it possible that the start time of <EVENT1> is before the start time of <EVENT2>?*
**QV3**: *Is it possible that <EVENT1> started before <EVENT2>?*

We fine-tune our models for the duration of five epochs and evaluate them on the validation set every epoch; we use the best checkpoint as the output model. We run every experiment three times using different seeds and report on the averaged accuracy and standard deviation on the testing set.[2] The MATRES model with the best question variation (QV3) has been further processed with two additional procedures: Perturbation and fine-tuning with BoolQ.

**Perturbation.** To achieve better model generalization, we perturb the instances of the training

[1] Each question template has a symmetric question (omitted for lack of space).

[2] For more information please refer to Appendix A.

| **Model** | **MATRES** | | **TB-Dense** | |
|---|---|---|---|---|
| | **Base PLM** | **Large PLM** | **Base PLM** | **Large PLM** |
| **Ours** | | | | |
| Our-Model (QV1) | 84.7±0.7 | 85.2±0.6 | - | - |
| Our-Model (QV2) | 85.1±0.8 | 85.9±1.1 | - | - |
| Our-Model (QV3) | 85.4±0.6 | 86.3±0.7 | 72.9±0.5 | 73.21±0.6 |
| Our-Model (QV3) + AUG | **86.4**±0.5 | **87.7**±0.6 | **73.8**±0.7 | **74.34**±0.7 |
| Our-Model (QV3) + AUG + BoolQ | **86.4**±0.6 | 87.5±0.5 | - | - |
| **Baselines** | | | | |
| Standard QA (QV1) | 73.1±0.7 | 74.6±0.6 | 61.3±0.7 | 62.2±0.5 |
| Standard QA (QV2) | 71.1±0.6 | 72.5±0.7 | 60.1±0.6 | 61.3±0.6 |
| Sentence Classification | 70.2±0.7 | 70.9±1.1 | 58.4±0.4 | 59.7±0.6 |
| **Others** | | | | |
| Structrued Joint Model (Han et al., 2019b) | 75.5 | - | 64.5 | - |
| ECONET (Han et al., 2021) | - | 79.3 | - | 66.8 |
| (Zhang et al., 2021) | 79.3 | 80.3 | 66.7 | 67.1 |
| (Wang et al., 2020) | - | 78.8 | - | - |
| TIMERS (Mathur et al., 2021) | 82.3 | - | 67.8 | - |
| SCS-EERE (Man et al., 2022) | 83.4 | - | - | - |
| Faithfulness (Wang et al., 2022a) | 82.7 | - | - | - |
| DTRE (Wang et al., 2022b) | - | - | 72.3 | - |
| RSGT (Zhou et al., 2022) | 84.0 | - | - | - |

Table 2: Comparing micro-average F1 scores on MATRES and TB-Dense reported individually for the two sizes of the underlying PLM.

set, using nlpaug,[3] a data augmentation library for text. We employ the optical-character recognition (OCR) error simulation, using the default argument values, which replaces about 30% of the characters (except the characters of the events) with random letters or digits considered as common OCR mistakes (e.g., *l* vs. 1). We modify the original training instances in place; therefore, we do not increase the size of the training set. In Table 2 we refer to this procedure as AUG. It adds about 1% to F1 in the base model, and a slightly higher percentage in the large model, on both datasets.

**BoolQ.** Before fine-tuning on MATRES, we fine-tune the model on the BoolQ dataset (Clark et al., 2019) in which every instance is composed of a passage (text) and a question, provided along with a Yes/No answer. Overall, BoolQ has 9,427 training instances, which we use for fine-tuning. In Table 2 we refer to this procedure as BoolQ. As reported, this step does not improve performance. Therefore, we did not use it for TB-Dense.

**Baseline Algorithms.** To assess the contribution of our Yes/No QA design, we define two baseline algorithms. The first baseline is a traditional multi-class QA model, which is given with the same passage as in our original Yes/No QA model, paired with only one question that takes one of the labels as an answer. We experiment with two question variations:

**QV1**: *What is the chronological order of the two marked events: <EVENT 1> and <EVENT 2>?*
**QV2**: *Is <EVENT 1> happening before, after or at the same time as <EVENT 2>?*

The second baseline is a simple multiclass sentence-classification RoBERTa model, which receives as input for this model comprises only the passage, and the output is one of the labels from the dataset. As seen in Table 2, our models outperform the baselines and previous work, introducing a new state-of-the-art result for TRC on both datasets.[4]

## 6 Conclusions

We proposed a novel approach for TRC using a pre-trained language model fine-tuned for a Yes/No QA classification task. Our model was fine-tuned to answer questions which were originally designed to support decision making during the annotation process. We believe we have demonstrated the potential of this method to leverage the Yes/No QA design to break down the prediction process into a set of Yes/No questions; our approach outperforms existing methods, achieving a new state-of-the-art result for TRC on two datasets. There is a potential practical limitation to this work, which is related to time complexity and speed performance. Since every instance is transformed into multiple QA instances, it may take a relatively long time to process a document.

[3] https://nlpaug.readthedocs.io

[4] Qualitative analysis is provided in Appendix C.

## Limitations

There are two primary limitations of the system presented in this work. First, each set of questions we use for training the QA model is designed specifically for the dataset we trained our model on. While we provide a set of questions for each of the two common TRC datasets, we believe that training the model on other datasets may require rewrite of the questions. Second, as mentioned in the previous section, every TRC instance is converted into multiple QA instances which we then process individually. This may increase the overall inference time and pose a practical limitation which needs to be carefully considered.

## Acknowledgements

This research was supported by the Ministry of Science and Technology, Israel.

## References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, Valencia, Spain. Association for Computational Linguistics.

Rujun Han, I Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, Nanyun Peng, et al. 2019a. Deep structured neural network for event temporal relation extraction. *arXiv preprint arXiv:1909.10094*.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019b. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. ECONET: Effective continual pretraining of language models for event temporal reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029, Vancouver, Canada. Association for Computational Linguistics.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11058–11066.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: Document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533, Online. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 696–706. Association for Computational Linguistics.

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob R Gardner, Muhao Chen, and Dan Roth. 2022a. Extracting or guessing? improving faithfulness of event temporal relation extraction. *arXiv preprint arXiv:2210.04992*.

Liang Wang, Peifeng Li, and Sheng Xu. 2022b. DCT-centered temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2087–2097.

Shuaicheng Zhang, Lifu Huang, and Qiang Ning. 2021. Extracting temporal event relation with syntactic-guided temporal graph transformer. *arXiv preprint arXiv:2104.09570*.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022. RSGT: Relational structure guided temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2001–2010.

Yichao Zhou, Yu Yan, Rujun Han, J. Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *AAAI*.

## A Technical Details

All our models are trained with the same learning rate value of 0.00001 and a batch size value of 20. We use `Pytorch`'s distributed-data-parallel (DDP) mechanism with `SyncBatchNorm` over two GALAX GeForce RTX™ 3090 GPUs. Fine-tuning our QA model on the MATRES training set takes us about 25 minutes, and 13 minutes on TB-Dense.

## B Label Distribution

We summarize the label distributions of MATRES and TB-Dense in Tables 3 and 4, respectively.

| **Label** | **Train** | **Val.** | **Test** |
|---|---|---|---|
| VAGUE | 12.0 | 0.0 | 3.8 |
| EQUAL | 3.5 | 0.3 | 13.5 |
| BEFORE | 50.7 | 67.2 | 50.6 |
| AFTER | 33.8 | 32.5 | 32.1 |

Table 3: Label distribution (%) in MATRES.

| **Label** | **Train** | **Val.** | **Test** |
|---|---|---|---|
| VAGUE | 48.4 | 39.3 | 43.3 |
| EQUAL | 2.9 | 2.9 | 2.6 |
| BEFORE | 20.2 | 24.6 | 26 |
| AFTER | 16.9 | 27.4 | 19.3 |
| INCLUDES | 5.1 | 2.7 | 4.3 |
| IS-INCLUDED | 6.5 | 3.1 | 4.5 |

Table 4: Label distribution (%) in TB-Dense.

## C Qualitative Analysis

Table 5 lists some examples from MATRES. The first column contains the passage in which we highlight the two relevant events. The second and third columns show the answers given by the fine-tuned boolean QA model, following by the forth and fifth columns which provide the corresponding model's label and the gold label, as assigned by the annotators. Finally, the last column provides indication for whether the model was right or wrong.

Some examples are relatively simple, while other are more challenging. For instance, Example 3 was manually assigned with EQUAL, indicating that none of the actions **found** and **floating** had started before the other. However, our QA model might be right about the second question, answering *yes*, since one may assume that the pigs were *floating* even before they were *found*.

Example 5 shows the difficulty in putting two events in a chronological order, when one of them did not really happen. This difficulty is addressed by the creators of MATRES by introducing the concept of *multi-axis modeling* to separate the story into different temporal axes, which allows the annotators to ignore some pairs of events that do not align chronologically.

| | Passage+Events | Ans. 1 | Ans. 2 | Prediction | Gold | Correct? |
|---|---|---|---|---|---|---|
| 1 | President Barack Obama arrived in refugee-flooded Jordan on Friday after scoring a diplomatic coup just before leaving Israel when Prime Minister Benjamin Netanyahu **apologized** to Turkey for a 2010 commando raid that **killed** nine activists on a Turkish vessel in a Gaza-bound flotilla. | No | Yes | AFTER | AFTER | Yes |
| 2 | The FAA on Friday **announced** it will **close** 149 regional airport control towers because of forced spending cuts – sparing 40 others that the FAA had been expected to shutter. | Yes | No | BEFORE | BEFORE | Yes |
| 3 | China's state leadership transition has taken place this month against an ominous backdrop. More than 16,000 dead pigs have been **found** **floating** in rivers that provide drinking water to Shanghai. | Yes | Yes | VAGUE | EQUAL | No |
| 4 | China's state leadership transition has taken place this month against an ominous backdrop. More than 16,000 dead pigs have been **found** floating in rivers that provide drinking water to Shanghai. A haze akin to volcanic fumes **cloaked** the capital, causing convulsive coughing and obscuring the portrait of Mao Zedong on the gate to the Forbidden City. | Yes | No | BEFORE | AFTER | No |
| 5 | Before the arrival of Keep, which Google launched this week, there was no default note-taking app for Android. It was a glaring hole, considering that Apple's iPhone has built-in Notes and Reminders apps that can be powered by Siri. Instead of **settling** for a bare bones app to fill the void, the search giant **took** things one step further. | Yes | No | BEFORE | AFTER | No |
| 6 | Former President Nicolas Sarkozy was **informed** Thursday that he would face a formal investigation into whether he **abused** the frailty of Liliane Bettencourt, 90, the heiress to the L'Oreal fortune and France's richest woman, to get funds for his 2007 presidential campaign. | No | Yes | AFTER | AFTER | Yes |

Table 5: Examples from MATRES, provided along with predictions given by our model.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

□ A1. Did you describe the limitations of your work?
*Left blank.*

□ A2. Did you discuss any potential risks of your work?
*Left blank.*

□ A3. Do the abstract and introduction summarize the paper's main claims?
*Left blank.*

□ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B □ Did you use or create scientific artifacts?

*Left blank.*

□ B1. Did you cite the creators of artifacts you used?
*Left blank.*

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Left blank.*

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Left blank.*

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Left blank.*

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Left blank.*

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Left blank.*

### C □ Did you run computational experiments?

*Left blank.*

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Left blank.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Left blank.*

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Left blank.*

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Left blank.*

**D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Left blank.*

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Left blank.*

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Left blank.*

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Left blank.*

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Left blank.*