Temporal Event Reasoning Using Multi-source Auxiliary Learning Objectives

Xin Dong[1](✉), Tanay Kumar Saha[2], Ke Zhang[2], Joel Tetreault[2], Alejandro Jaimes[2], and Gerard de Melo[1,3]

[1] Rutgers University, New Jersey, USA
xd48@rutgers.edu, gdm@demelo.org
[2] Dataminr Inc., New York, USA
{tsaha,kzhang,jtetreault,ajaimes}@dataminr.com
[3] Hasso Plattner Institute/University of Potsdam, Potsdam, Germany

Abstract. Temporal event reasoning is vital in modern information-driven applications operating on news articles, social media, financial reports, etc. Recent works train deep neural nets to infer temporal events and relations from text. We improve upon the state-of-the-art by proposing an approach that injects additional temporal knowledge into the pretrained model from two sources: (*i*) part-of-speech tagging and (*ii*) question constraints. Auxiliary learning objectives allow us to incorporate this temporal information into the training process. Our experiments show that these types of multi-source auxiliary learning objectives lead to better temporal reasoning. Our model improves over the state-of-the-art model on the TORQUE question answering benchmark by 1.1% and on the MATRES relation extraction benchmark by 2.8% in F1 score.

Keywords: Temporal event reasoning · Auxiliary learning · Question answering

1 Introduction

Temporal event reasoning is a crucial yet under-explored aspect of interpreting text in modern information systems, enabling people to infer the timeline of narrated events. Past work has often cast this as a Relation Extraction task [2,12,13] that involves predicting temporal relationships between two events mentioned in a given piece of text, such as Before or After. Another recently proposed task is that of reading comprehension about temporal relations [11]. Given an input text, the system answers temporal questions pertaining to some event. Compared with the aforementioned temporal relationship prediction task, the advantage of such a Question Answering (QA) problem formulation is that questions can encode a richer, more diverse range of complex temporal relationships

Work done as a Research Intern at Dataminr, Inc.

M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 102–110, 2022.
https://doi.org/10.1007/978-3-030-99739-7_12

and phenomena, such as overlap, uncertainty, negation, hypotheticals, and repetition, to name a few. For instance, we may ask a challenging question incorporating negation such as “What has not happened after investigators made good progress?”

Table 1. Excerpts from input passages with different verb POS tags.

Example	POS tag	Temporal information
People have **predicted** his demise so many times...	VBN: verb, past participle	event has happened
Security Council **passed** a resolution ...	VBD: verb, past tense	event happened

Table 2. Question Answering samples from TORQUE [11].

Passage: They were traveling in an up-armored high-mobility, multi-purpose, wheeled vehicle when this occurred. Those injured were evacuated by air to a nearby forward operating base for treatment.

Questions	**Answers**
What events have already finished?	traveling, occurred, evacuated
What will happen in the future?	No answer.
What events happened during their travel?	occurred, evacuated
What events have begun but has not finished?	treatment
What happened after it occurred?	evacuated, treatment
What happened before the injured were treated?	traveling, occurred, evacuated

Auxiliary learning is a common means of improving the performance on a primary task of interest [6,8,15]. In our work, we propose two auxiliary tasks to acquire better temporal reasoning abilities: (*i*) part-of-speech (POS) tagging, and (*ii*) question constraints. POS tagging as an auxiliary task is able to ensure a better understanding of tense-related information within a sentence. For example, as shown in Table 1, the word “predicted” in “People have predicted his demise so many times ...” is labeled as VBN (past participle), while “passed” is labeled as VBD (past tense) in “Security Council passed a resolution ...”. Being able to capture such distinctions enables the model to more accurately distinguish what happened from what *has* (perhaps more recently) happened.

The second auxiliary task, question constraints, can be viewed as a self-supervised task and is induced based on a temporal question answering dataset. As shown in Table 2, for a given text passage, the dataset provides a set of questions, and different questions tend to call for different answers. For example, the set of answers to “What events have already finished?” and “What will happen in the future?” should typically be disjoint. Hence, we explore the value of question constraint rules between pairs of questions for a passage. We induce such rules automatically based on their answer overlap, and subsequently enforce them by training the model with the auxiliary classification task of identifying the kind of answer overlap.

We propose a novel multi-source auxiliary learning objective that incorporates the two auxiliary tasks to improve the performance in two temporal event reasoning tasks. Our method achieves a new state-of-the-art performance on the TORQUE [11] dataset (QA setup), improving over previous work by 0.8 F1 points (absolute). Having fine-tuned the model on this QA setup so as to learn complex temporal cues, we further demonstrate the generalizability of our approach by showing that the fine-tuned encoder can then be further fine-tuned to improve the top performance on MATRES [12] (Relation Extraction setup) by 2.3 F1 points. Finally, we show that our approach is particularly performant in a low-resource setting, yielding absolute improvements of up to 19.5%.

2 Related Work

Temporal Question Answering. Great strides have been made with new architectures and new self-supervised objectives to improve over vanilla BERT [3]. However, while models such as RoBERTa [10] and AlBERT [7] enable a better understanding of predicates and arguments for conventional QA tasks, our experiments show that they fail to yield substantial gains on temporal QA. Recently, Han et al. [5] presented a temporal-related language model with new self-supervised objectives for improved Temporal QA. In contrast to our approach, this method requires pre-defined event and temporal lexicons.

Temporal Relation Extraction. Compared with temporal QA, temporal relation (TempRel) extraction is widely studied in temporal event reasoning. Many TempRel datasets have been collected, such as TB-Dense [2], RED [13], and MATRES [12], and a variety of models target this task. For instance, Han et al. (2019) [4] present a joint event and temporal relation extraction model. Wang et al. (2020) [16] enforce logical constraints within and across temporal relations via differentiable learning objectives. Zhou et al. (2020) [18] incorporate probabilistic soft logic regularization and global inference.

Auxiliary Learning. There is a long history of research on multi-task learning [14], e.g., the Multi-Task Deep Neural Network (MT-DNN) seeks to learn representations across diverse natural language understanding tasks [9]. In auxiliary learning, there is a single primary task, and the role of the auxiliary tasks is to improve the performance and generalizability of this primary task. Trinh et al. (2018) [15] propose a method for better capturing long term dependencies in RNNs with an extra unsupervised auxiliary loss. Xu et al. (2021) [17] propose multi-task recurrent modular networks for any multi-task recurrent models.

3 Method

Following standard practice when training a deep network on multiple tasks [9], our model consists of a shared encoder and several task-specific classifiers on top of it. There is one such classifier for the primary task as well as two further ones for our proposed auxiliary tasks. This architecture allows the shared encoder to jointly learn from each of the tasks.

Shared Encoder. The encoder is from a pre-trained contextual representation model, denoted as $f_{\mathrm{se}}(\cdot;\theta_{\mathrm{se}})$. Given an input text sequence $\mathbf{s}$ consisting of T tokens $[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T]$, this encoder infers a contextual hidden representation $\mathbf{h_t} \in \mathbb{R}^d$ d of dimensionality d for each input token $\mathbf{x}_t$.

Primary Task. Our primary task-specific classification module $f_{\mathrm{p}}(\cdot;\theta_{\mathrm{p}})$ is responsible for the question answering task. It is applied for fine-tuning on top of the pre-trained model $f_{\mathrm{se}}(\cdot;\theta_{\mathrm{se}})$ and consists of a fully-connected layer with softmax activation to map $\mathbf{h_t} \in \mathbb{R}^d$ into $\mathbb{R}^{|\mathcal{Y}_p|}$. Here, $\mathcal{Y}_{\mathrm{p}}$ is defined as a set of binary output class labels denoting whether a given token is deemed a valid answer in response to the question.

Auxiliary Tasks. The model is additionally trained on two auxiliary tasks.

1. *POS tagging.* Our auxiliary POS tagging classification module $f_{\mathrm{pos}}(\cdot;\theta_{\mathrm{pos}})$ draws its input from the shared encoder $f_{\mathrm{se}}(\cdot;\theta_{\mathrm{se}})$. It then applies a linear mapping $\mathbf{h_t} \in \mathbb{R}^d$ into $\mathbb{R}^{|\mathcal{Y}_{\mathrm{pos}}|}$ followed by a softmax activation to predict a distribution over the set of POS tag classes $\mathcal{Y}_{\mathrm{pos}}$.
2. *Question Constraint Classification (Question CC).* For a given passage p from our primary QA task, we have a corresponding question set $Q = \{(q_i, a_i) \mid i \in \{1, ..., n\}, a_i \neq \emptyset\}$, where n is the number of questions and a_i is the answer set for question q_i. From this, we can obtain a set of question pairs $\mathcal{C} = \{\langle q_i, q_j \rangle \mid i < j; i, j \in \{1, ..., n\}\}$ and a set of answer pairs $\mathcal{A} = \{\langle a_i, a_j \rangle \mid i < j; i, j \in \{1, ..., n\}\}$. We consider the overlap of answers between two questions to acquire a constraint label for the question pair. In particular, the constraint label is chosen from a set of five relations $\mathcal{Y}_{qc} = \{$ Equal, Subset, Superset, Disjoint, Overlap$\}$, based on the corresponding conditions $(a_i = a_j)$, $(a_i \subset a_j)$, $(a_i \supset a_j)$, $(a_i \cap a_j = \emptyset)$, and $(a_i \cap a_j \neq \emptyset; a_i \cap a_j \neq a_i; a_i \cap a_j \neq a_j)$. To predict such labels, our model incorporates a question classification module $f_{\mathrm{qc}}(\cdot;\theta_{\mathrm{qc}})$ consisting of a fully-connected layer mapping $\mathbf{h_0} \in \mathbb{R}^d$ into $\mathbb{R}^{|\mathcal{Y}_{\mathrm{qc}}|}$ with softmax activation.

Auxiliary Learning Objectives. To inject the temporal knowledge into the primary QA training, we jointly learn the primary task along with the two auxiliary tasks. Hence, the overall loss function becomes

$$\mathcal{L} = \mathcal{L}_p + \lambda_1 \mathcal{L}_{\mathrm{pos}} + \lambda_2 \mathcal{L}_{\mathrm{qc}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{p}}, \mathcal{L}_{\mathrm{pos}}, \mathcal{L}_{\mathrm{qc}}$ are the QA loss, POS tagging loss, and question constraint classification loss, respectively, and λ_1, λ_2 are coefficients to control the influence of each auxiliary task loss term.

4 Experiments

4.1 Experimental Setup

Tasks and Datasets. For evaluation, we use TORQUE [11], a reading comprehension dataset of temporal ordering questions and answers. It provides 3.2k passages (∼50 tokens/passage), 24.9k events (7.9 events/passage), and 21.2k user-provided questions. For end-to-end training, the task is modeled as a binary classification problem that requires predicting for each token in the passage whether it is an answer. We also investigate pretraining on TORQUE to then improve on MATRES [12], a temporal relation (TempRel) extraction benchmark, consisting of 275 documents with entity relationships labeled as Before, After, Equal, or Vague. Regarding metrics, TORQUE is evaluated in terms of F1 score, Exact Match (EM), and Consistency (C). The latter is defined as the percentage of contrast groups for which a model's predictions have F1 $\leq 80\%$ for all questions in a group. The contrast groups provided by TORQUE consist of questions with contrasting changes to the temporal keywords, e.g., "What happened *after* the snow started?" versus "What happened *before* the snow started?". For MATRES, we report standard micro-averaged F1 scores.

Table 3. Hyper-parameter settings.

Parameter	TORQUE	MATRES
Max. sequence length	180	220
Batch size	12	10
Learning rate	1×10^{-5}	5×10^{-6}
# of training epochs	10	5
λ_1	0.001	–
λ_2	0.001	–

Model Details. For POS tagging as the auxiliary task, we invoke NLTK [1] to obtain POS tags on the TORQUE training set. The size of the POS tag inventory is 36. For question constraint classification, the number of question pairs extracted from the training set for the five labels defined in Sect. 3 are 4,307, 11,610, 6,181, 42,928, and 7,146, respectively. We adopt RoBERTa-Large [10] as the pre-trained encoder. To further evaluate the effectiveness of auxiliary learning, we use models fine-tuned on TORQUE first to evaluate on MATRES. We tune the hyper-parameters based on the respective development sets and list their values in Table 3. On TORQUE, as for the original baseline, we report average results over 3 random seeds, while on MATRES, we consider averages over 5 runs.

4.2 Results and Analysis

Table 4. Results from TORQUE experiments.

Method	F1	EM	C
RoBERTa-Large [11]	75.2	51.1	34.5
RoBERTa-Large			
+ Question CC	75.7	**51.3**	36.2
+ POS Tagging	75.8	50.7	35.6
+ POS Tagging + Question CC	**76.0**	51.2	**36.7**

TORQUE (Question Answering Setup). The current SOTA method on TORQUE is RoBERTa-Large [11]. Table 4 compares our approach against this baseline to evaluate the effectiveness of auxiliary learning. We first evaluate on RoBERTa-Large with either POS tagging or Question CC as the auxiliary task. Compared with RoBERTa-Large, we observe that adding Question CC improves the Consistency score, while POS tagging in particular improves the F1 score. This shows that our answer constraints lead to a better understanding of the differences between questions, while the POS tagging auxiliary task enables the model to better capture subtle differences. Our full method outperforms RoBERTa-Large across all three metrics, demonstrating that our multi-source auxiliary learning objective is effective for our primary QA task.

Table 5. Results on TORQUE with different ratios of training data.

Ratio	30%			50%			100%		
Method	F1	EM	C	F1	**EM**	**C**	**F1**	**EM**	**C**
RoBERTa-Large	57.3	37.9	20.1	73.3	46.3	32.0	75.2	51.1	34.5
Our approach	68.5	39.4	25.1	74.3	48.5	34.5	76.0	51.2	36.7
Improvement (%)	19.5%	4.0%	24.8%	1.4%	4.8%	7.8%	1.1%	0.2%	6.4%

Influence of Amount of Training Data for TORQUE. To assess the effectiveness of our method with limited amounts of training data on TORQUE, we compare our full multi-source auxiliary learning approach with RoBERTa-Large using different ratios of training data. As shown in Table 5, our method yields significant improvements over RoBERTa-Large in terms of F1 and C scores, especially with 30% of training data, which suggests that our auxiliary tasks are particularly fruitful when training data is scarce, although this also means that less supervision is available for POS tagging and question constraint induction Table 6.

Table 6. Results on MATRES dataset.

Method	F1
Want et al. [16]	78.8
RoBERTa-Large	80.1
+ TORQUE	80.6
+ TORQUE (Question CC)	80.4
+ TORQUE (POS Tagging)	80.7
+ TORQUE (POS Tagging + Question CC)	**81.1**

MATRES (Relation Extraction Setup). As TORQUE provides more complex temporal information, we assess to what extent we can transfer the knowledge learned on it to the MATRES relation extraction task, so as to evaluate the generalizability of our auxiliary learning. As baselines, in addition to RoBERTa-Large, we consider Wang et al. [16], which incorporates temporal logic constraints among events into the training loss function. Our model is fine-tuned on TORQUE first and then further fine-tuned on MATRES. This outperforms the baselines, showing that MATRES can benefit from the auxiliary information provided by training on TORQUE first. In this regard, compared to versions with just one additional auxiliary task, our full auxiliary learning model proves the most effective at acquiring an understanding of temporal relationships.

5 Conclusion

We propose a method to inject additional temporal information with multi-source auxiliary learning objectives into pre-trained models for temporal event reasoning. In particular, we consider part-of-speech prediction and question answer constraint classification as additional objectives, and investigate how pretraining on question answering can benefit temporal relation extraction. Our experiments show that we achieve state-of-the-art results on TORQUE as well as on MATRES, and that our auxiliary learning method is particularly useful in low-resource settings.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
2. Chambers, N., Cassidy, T., McDowell, B., Bethard, S.: Dense event ordering with a multi-pass architecture. Trans. Assoc. Comput. Linguist. **2**, 273–284 (2014)

3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
4. Han, R., Ning, Q., Peng, N.: 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), pp. 434–444. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1041, https://www.aclweb.org/anthology/D19-1041
5. Han, R., Ren, X., Peng, N.: Deer: A data efficient language model for event temporal reasoning. arXiv preprint arXiv:2012.15283 (2020)
6. Jaderberg, M., et al.: Reinforcement learning with unsupervised auxiliary tasks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017, Conference Track Proceedings. OpenReview.net (2017). https://openreview.net/forum?id=SJ6yPD5xg
7. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: a lite bert for self-supervised learning of language representations. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=H1eA7AEtvS
8. Liu, S., Davison, A., Johns, E.: Self-supervised generalisation with meta auxiliary learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019). https://proceedings.neurips.cc/paper/2019/file/92262bf907af914b95a0fc33c3f33bf6-Paper.pdf
9. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4487–4496. Association for Computational Linguistics, Florence, Italy, July 2019. https://doi.org/10.18653/v1/P19-1441, https://www.aclweb.org/anthology/P19-1441
10. Liu, Y., et al.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
11. Ning, Q., Wu, H., Han, R., Peng, N., Gardner, M., Roth, D.: TORQUE: a reading comprehension dataset of temporal ordering questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1158–1172. Association for Computational Linguistics, Online, November 2020. https://doi.org/10.18653/v1/2020.emnlp-main.88, https://www.aclweb.org/anthology/2020.emnlp-main.88
12. Ning, Q., Wu, H., Roth, D.: 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), pp. 1318–1328. Association for Computational Linguistics, Melbourne, Australia, July 2018. https://doi.org/10.18653/v1/P18-1122, https://www.aclweb.org/anthology/P18-1122

13. O'Gorman, T., Wright-Bettner, K., Palmer, M.: Richer event description: integrating event coreference with temporal, causal and bridging annotation. In: Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016), pp. 47–56 (2016)
14. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
15. Trinh, T., Dai, A., Luong, T., Le, Q.: Learning longer-term dependencies in RNNS with auxiliary losses. In: International Conference on Machine Learning, pp. 4965–4974. PMLR (2018)
16. Wang, H., Chen, M., Zhang, H., Roth, D.: Joint constrained learning for event-event relation extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 696–706. Association for Computational Linguistics, Online, November 2020. https://doi.org/10.18653/v1/2020.emnlp-main.51, https://www.aclweb.org/anthology/2020.emnlp-main.51
17. Xu, D., et al.: Multi-task recurrent modular networks. In: AAAI, vol. 35, no. 12, pp. 10496–10504 (2021)
18. Zhou, Y., et al.: Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. arXiv e-prints pp. arXiv-2012 (2020)