Prior Knowledge-Guided Adversarial Training

Prior Knowledge-Guided Adversarial Training

Lis Kanashiro Pereira[1], **Fei Cheng**[2], **Wan Jou She**[3]
Masayuki Asahara[4], **Ichiro Kobayashi**[5]
[1]National Institute of Information and Communications Technology (NICT), Japan
[2]Kyoto University, Japan
[3]Kyoto Institute of Technology, Japan
[4]National Institute for Japanese Language and Linguistics (NINJAL), Japan
[5]Ochanomizu University, Japan
liskanashiro@nict.go.jp, feicheng@i.kyoto-u.ac.jp, wjs2004@kit.ac.jp
masayu-a@ninjal.ac.jp, koba@is.ocha.ac.jp

Abstract

We introduce a simple yet effective **P**rior **K**nowledge-**G**uided **ADV**ersarial Training (**PKG-ADV**) algorithm to improve adversarial training for natural language understanding. Our method simply utilizes task-specific label distribution to guide the training process. By prioritizing the use of prior knowledge of labels, we aim to generate more informative adversarial perturbations. We apply our model to several challenging temporal reasoning tasks. Our method enables a more reliable and controllable data training process than relying on randomized adversarial perturbation. Albeit simple, our method achieved significant improvements in these tasks. To facilitate further research, we will release the code and models.

1 Introduction

Class imbalance, a classification setting where one or multiple classes (minority classes) are considerably less frequent than others (majority classes), is a common yet challenging problem in natural language processing (NLP) (Henning et al., 2023). The uneven distribution of target categories often leads to lower performance for minority classes. Despite that, NLP research often overlooks the importance of incorporating methods for addressing it, and finding effective solutions remains an open research challenge (Henning et al., 2023). While deep learning models have been successful in various NLP tasks, they are sensitive to changes in the input data distribution.

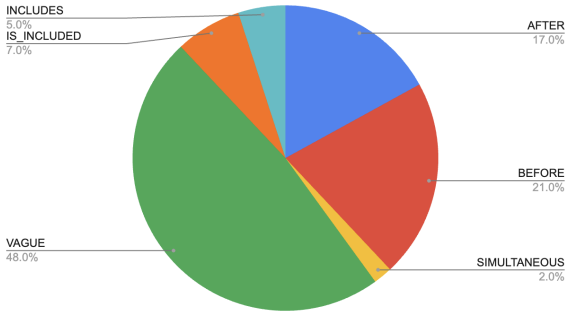
We explore the use of adversarial training techniques to enhance model performance on such scenarios. More specifically, our proposed approach incorporates the knowledge of task-specific label distribution into the adversarial training process. Typically, the perturbation direction is chosen to mislead the model to flip the current model prediction away from the correct label. However, this strategy might not be optimal because it does not make use of the knowledge of task-specific label distribution during the training process. We hypothesize that such information might indicate which category a model is more likely to misclassify as another category. We focus on temporal reasoning tasks. These tasks are essential for NLP, for timing events, for estimating their duration, frequencies, ordering, etc. Due to the nature of the task, classes are highly imbalanced, as shown in Figure 1 and Table 2. Even the performance of recent large language models (LLMs), such as ChatGPT, is still underperformed by a large margin by simpler and smaller models such as BERT and RoBERTa (Yuan et al., 2023; Chan et al., 2024), indicating the inherent challenge of temporal reasoning tasks. For instance, on the TB-Dense dataset (Cassidy et al., 2014), due to the high label imbalance, the model might misclassify the samples with the true label "VAGUE" as "BEFORE" or "AFTER", as these labels occur more often in the dataset, as shown in Figure 1. Our model, PKG-ADV , can intentionally attack those vulnerable categories and learn how to better distinguish each label class, improving the model performance.

Our experimental results show that, despite its simplicity, our proposed model outperforms standard fine-tuning and a strong adversarial training method on several challenging temporal reasoning tasks. Moreover, our model can outperform ChatGPT-based models with a large gap. Our findings contribute to the understanding and improvement of adversarial training in NLP and can help enhance model performance in scenarios with imbalanced classes, such as temporal reasoning tasks.

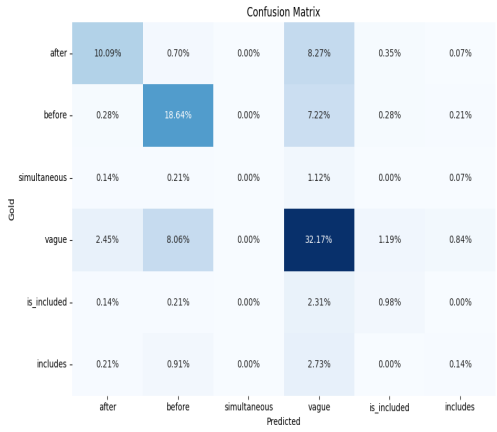
2 Adversarial Training for NLP

Standard training objectives seek to learn a function (a classifier) $f(x;\theta) : x \rightarrow C$, parametrized by θ, where C is the class label set. Given a training dataset D of input-output pairs (x, y)

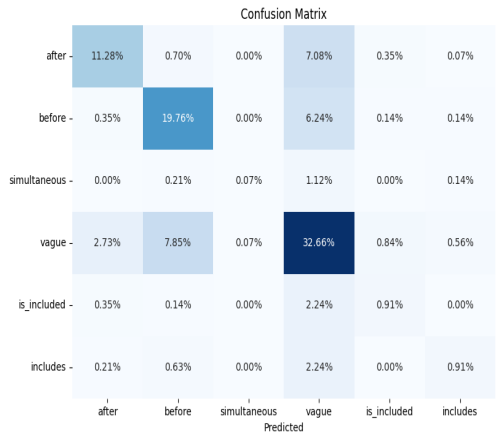
Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), pages 51–57



(a) Label Distribution (Train Set)



(b) RoBERTa with Standard Fine-Tuning



(c) RoBERTa with PKG-ADV Fine-Tuning

Figure 1: a) Label distribution from the TB-Dense (Cassidy et al., 2014) training dataset. b) Confusion matrix obtained after training on the RoBERTa_BASE model. c) Confusion matrix obtained after training on the RoBERTa_BASE model with the PKG-ADV algorithm. X-axis and Y-axis represent the predicted and gold labels, respectively.

and the loss function $l(.,.)$ (e.g., cross entropy), $f(x;\theta)$ is trained to minimize the empirical risk: $\min_\theta \mathbb{E}_{(x,y)\sim D}[l(f(x;\theta),y)]$. While this is effective in training a classifier, it usually suffers from overfitting and poor generalization to unseen cases. Recently, adversarial training has been proven effective in several tasks in nlp (Zhu et al., 2019; Jiang et al., 2019; Pereira et al., 2020). The standard approach is to add the adversarial perturbation to the embeddings. The input is augmented with a small perturbation that maximizes the adversarial loss:

$$\min_\theta \mathbb{E}_{(x,y)\sim D}[\max_\delta l(f(x+\delta;\theta),y)],$$

where the inner maximization can be solved by projected gradient descent (Madry et al., 2017). More recent approaches have explored adding the perturbation to other layers of the model (Pereira et al., 2021). Although these adversarial training algorithms substantially enhance model performance and generalization, such methods adopt *non-targeted* attacks, where the model prediction is not driven towards a specific incorrect label, i.e., such attacks lack a specific target. This might not be optimal, since many natural language processing (NLP) tasks are naturally imbalanced, as some labels occur much more frequently than others. In Figure 1, we illustrate this typical label imbalance scenario with the MATRES dataset (Ning et al., 2018a). Thus, there are consistently classes where the trained classifier may exhibit a higher error rate. This information can highlight the models' weaknesses. Our goal is to incorporate this prior knowledge to enhance model performance.

3 Prior Knowledge Guided Adversarial Training

In our work, we propose to enhance the ALICE (Pereira et al., 2020) algorithm. ALICE is an adversarial training algorithm that combines the two approaches to estimate the perturbation δ: one that uses the label y (Zhu et al., 2019) and another that uses the model prediction $f(x;\theta)$, i.e., a "virtual" label (Miyato et al., 2018; Jiang et al., 2019). The first goal is to improve the robustness of our target label by preventing an increase in error for unperturbed inputs. The second goal is to enforce model smoothness, ensuring the model's output does not change significantly when a small perturbation is injected to the input. The formula of ALICE is shown below:

Algorithm 1 PKG-ADV : We explore incorporating the knowledge of task-specific label distribution into the adversarial training process. The two lines in blue color are the only changes from ALICE.

Input: T: the total number of iterations, $\mathcal{X} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$: the dataset, $f(x;\theta)$: the machine learning model parametrized by θ, σ^2: the variance of the random initialization of perturbation δ_1 and δ_2, δ_{1_r} and δ_{2_r}: the perturbations added to the embedding vector, ϵ: perturbation bound, K: the number of iterations for perturbation estimation, η: the step size for updating perturbation, τ: the global learning rate, α: the smoothing proportion of adversarial training in the augmented learning objective, Π: the projection operation and C: the classes.

1: **for** $t = 1, .., T$ **do**
2: **for** $(x, y) \in \mathcal{X}$ **do**
3: $\delta_1 \sim \mathcal{N}(0, \sigma^2 I)$
4: $\delta_2 \sim \mathcal{N}(0, \sigma^2 I)$
5: $y_t = sample(C_{\setminus y})$
6: **for** $m = 1, .., K$ **do**
7: $g_{adv} \leftarrow \nabla_{\delta_1} l(f(x + \delta_1; \theta), y_t)$
8: $\delta_1 \leftarrow \Pi_{\|\delta_1\|_\infty \leq \epsilon}(\delta_1 - \eta g_{adv})$
9: $g_{adv} \leftarrow \nabla_{\delta_2} l(f(x + \delta_2; \theta), f(x; \theta))$
10: $\delta_2 \leftarrow \Pi_{\|\delta_2\|_\infty \leq \epsilon}(\delta_2 + \eta g_{adv})$
11: **end for**
12: $g_\theta \leftarrow \nabla_\theta l(f(x + \delta_1; \theta), y)$ $+\alpha \nabla_\theta l(f(x + \delta_2; \theta), f(x; \theta))$
13: $\theta \leftarrow \theta - \tau g_\theta$
14: **end for**
15: **end for**

Output: θ

$$\min_\theta \mathbb{E}_{(x,y)\sim D}[\max_{\delta_1} l(f(x + \delta_1; \theta), y) + \alpha \max_{\delta_2} l(f(x + \delta_2; \theta), f(x; \theta))], \quad (1)$$

where δ_1 and δ_2 are two different perturbations, bounded by a general l_p norm ball, estimated by a fixed K steps of the gradient-based optimization approach and $p = \infty$. Effectively, the second term encourages smoothness in the input neighborhood, and α is a hyperparameter that controls the trade-off between standard errors and adversarial errors. ALICE has been originally applied for the commonsense reasoning task, however, it is a general algorithm that can be applied to other tasks as well.

We enhance ALICE by modifying the first term of Equation 1, to improve the robustness of our target label. PKG-ADV first samples a label from the class label set (excluding the correct label). This label class is sampled with a probability proportional to its frequency in the training dataset. Intuitively, we would like to focus training on prior knowledge at hand. This knowledge consists of the dataset label information, generated offline.

More specifically, PKG-ADV explicitly picks a target $y_t \neq y$ and tries to steer the model towards y_t. We accomplish this by sampling y_t from $C_{\setminus y} = C - \{y\}$ in proportion to the dataset label distribution. PKG-ADV can flexibly use different prior knowledge, i.e. the dataset label information, as shown in line 5. Then the adversarial sample is estimated by the opposite direction as in line 8. The two lines in blue color are the only changes from ALICE. At last, following Jiang et al. (2019) and Miyato et al. (2018), the adversarial regularizer is added to the standard training objective (e.g., cross-entropy between the correct label and prediction). The algorithm of PKG-ADV is depicted in Algorithm 1.

4 Experiments

We compare PKG-ADV with ALICE (Pereira et al., 2020), a strong adversarial training baseline, and several state-of-the-art temporal reasoning models. We use the standard uncased RoBERTa$_{\text{BASE}}$ model (Liu et al., 2019b) as the text encoder, unless noted otherwise.

4.1 Datasets and Evaluation Metrics

We evaluated our model on the following tasks: temporal ordering prediction task, event duration prediction, and temporal commonsense reasoning. We used the following datasets, respectively: MATRES (Ning et al., 2018b), TimeML (Pan et al., 2006), MC-TACO (Ben Zhou and Roth, 2019), TB-Dense (Cassidy et al., 2014), and MAVEN-ERE (Wang et al., 2022). Details of each dataset are in Appendix A. We evaluate the performance of MATRES and MAVEN-ERE in terms of accuracy and F1-score, and TimeML in terms of accuracy. For the MC-TACO and TB-Dense datasets, we report F1 scores.

4.2 Implementation Details

Our model implementation is based on the MT-DNN framework (Liu et al., 2019a, 2020). We use RoBERTa$_{\text{BASE}}$ (Liu et al., 2019b) as the text encoder. RoBERTa remains a competitive pre-trained model for its size among NLP practitioners. We used ADAM (Kingma and Ba, 2014) as our optimizer with a learning rate in the range $\in \{9 \times 10^{-6}, 1 \times 10^{-5}\}$ and a batch size in the range $\in \{16, 32, 64\}$. The maximum number of epochs was set to 10. A linear learning rate decay schedule with warm-up over 0.1 was used unless stated otherwise. To avoid gradient exploding, we

Model	TimeML	MC-TACO	MATRES		TB-Dense	MAVEN-ERE	
	Acc	F1	Acc	F1	F1	Acc	F1
Standard (RoBERTa$_{\text{BASE}}$)	81.46	80.84	72.88	47.83	62.02	76.43	31.68
ALICE (RoBERTa$_{\text{BASE}}$)	83.15	82.59	71.57	47.02	63.49	77.06	31.27
Multi-Task (ALBERT-xxlarge) (Kimura et al., 2022)	81.10	80.30	**77.20**	-	-	-	-
ChatGPT (Bian et al., 2023)	-	46.79	-	-	-	-	-
ChatGPT_Prompt (Chan et al., 2024)	-	-	-	35.00	23.30	-	-
ChatGPT_PE (Chan et al., 2024)	-	-	-	27.00	47.90	-	-
ChatGPT_ICL (Chan et al., 2024)	-	-		25.00	44.90	-	-
PKG-ADV (RoBERTa$_{\text{BASE}}$)	**84.75**	**83.01**	73.00	**49.93**	**65.59**	**78.09**	**32.06**

Table 1: Test results. The best results are in **bold**. Standard denotes the standard fine-tuning procedure where we fine-tune RoBERTa on each task specific temporal reasoning dataset. PKG-ADV denotes our proposed models. Note that Standard, ALICE, and PKG-ADV models use RoBERTa$_{\text{BASE}}$ as the text encoder unless stated otherwise, and for a fair comparison, all these results are produced by ourselves.

clipped the gradient norm within 1. All the texts were tokenized using WordPiece and were chopped to spans no longer than 512 tokens. We also set the dropout rate of all the task-specific layers as 0.3. During adversarial training, we follow Jiang et al. (2019) and set the perturbation size to 1×10^{-5}, the step size to 1×10^{-3}, and to 1×10^{-5} the variance for initializing perturbation. We search the regularization weight α in $\{0.01, 0.1, 1\}$. We set the number of projected gradient steps to 1.

4.3 Main Results

We present our results in Table 1. We compare our model, PKG-ADV , with ALICE and other temporal reasoning models. Overall, the adversarial methods, ALICE and PKG-ADV , were able to outperform the standard fine-tuning approach (Standard) and the other baselines, without using any additional knowledge source, and without using any additional datasets other than the target task datasets. Overall, PKG-ADV was able to outperform the other baselines. Kimura et al. (2022) trains an ALBERT XXLarge v2 model using multi-task learning with several additional temporal datasets. Note that ALBERT XXLarge v2 is around 2x larger than the RoBERTa_BASE model. Except on the MATRES dataset, our PKG-ADV model trained on RoBERTa_BASE can outperform their model, without using any additional dataset. Bian et al. (2023) and Chan et al. (2024) use zero-shot inference and designs prompt templates for different datasets in the ChatGPT and ChatGPT_Prompt baselines. In the ChatGPT_PE baseline, Chan et al. (2024) manually designed a more sophisticated prompt template based on the expert understanding. The ChatGPT_ICL baseline refers to the in-context learning approach (Brown et al., 2020), where a number of input-output exemplars for the prompt were manually selected. We observe that still there is a considerable gap between these models and that of supervised methods. Chan et al. (2024) hypothesizes that the poor performance of ChatGPT might be attributed to inadequate human feedback during the model's training process on temporal features.

5 Conclusion

We have presented a Prior Knowledge Guided Adversarial Training (PKG-ADV) algorithm to improve adversarial training for natural language understanding. Albeit simple and drawn from a simple observation (label imbalance, common in most nlp tasks), incorporating task-specific label distribution into the training process for generating better adversarial perturbations has not yet been explored in the literature. PKG-ADV overall shows superior performance compared to standard fine-tuning, strong adversarial training baselines, and ChatGPT-based baselines. PKG-ADV can be applied to other language models as well by incorporating label distribution information. Other types of knowledge, such as annotator agreement data, might help further enhance the performance, and we leave this for future work.

6 Limitations and Ethical Statement

Although our method is task, model, and language-agnostic, we have conducted experiments only on English classification benchmarks, and using only the RoBERTa model. We focus on sentence-level tasks at this time. Although we focused on temporal reasoning tasks, our model can be generalized to other tasks as well. We plan to expand the scope of the experiments in the future. In our work, we have

only used publicly available datasets in our experiments, ensuring that there are no privacy concerns or violations.

Acknowledgments

We thank the reviewers for their helpful feedback. This work has been supported by the project KAKENHI ID: 21K17802.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Qiang Ning Ben Zhou, Daniel Khashabi and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *EMNLP*.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Mayuko Kimura, Lis Kanashiro Pereira, and Ichiro Kobayashi. 2022. Toward building a language model for understanding temporal commonsense. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 17–24, Online. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020. The microsoft toolkit of multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:2002.07972*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Qiang Ning, Hao Wu, and Dan Roth. 2018a. A multi-axis annotation scheme for event temporal relations. In *arXiv preprint arXiv:1804.07828*.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. In *ACL*.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R Hobbs. 2006. Extending timeml with typical durations of events. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 38–45.

Lis Pereira, Xiaodong Liu, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi. 2020. Adversarial training for commonsense inference. *arXiv preprint arXiv:2005.08156*.

Lis Kanashiro Pereira, Yuki Taya, and Ichiro Kobayashi. 2021. Multi-layer random perturbation training for improving model generalization efficiently. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 303–310, Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, et al. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764*.

A Evaluation Datasets

TimeML (Pan et al., 2006): This task involves predicting whether a given event has a duration longer or shorter than a day.

TB-Dense (Cassidy et al., 2014): TB-Dense is a public benchmark for temporal relation extraction (TRE). It was annotated from TimeBank (Pustejovsky et al., 2003) and TempEval (UzZaman et al., 2013). Given a passage and two event points, the task is to classify the relations between events into one of 6 types: BEFORE, AFTER, SIMULTANEOUS, VAGUE, IS_INCLUDED, and INCLUDES. An example of a sentence with two events, e1 and e2 (in bold) that hold the SIMULTANEOUS relation is shown below:

> Nobody **(e1:hurried)** her up. No one **(e2:held)** her back.

MATRES (Ning et al., 2018b): This dataset was annotated from TimeBank (Pustejovsky et al., 2003), AQUAINT, and Platinum documents. The task involves predicting the temporal relation between a pair of input events in a span of text. It originally contains 13,577 pairs of events annotated with a temporal relation (BEFORE, AFTER, EQUAL, VAGUE). The relations named EQUAL and VAGUE are equivalent to SIMULTANEOUS and NONE in TB-Dense. An example of a sentence with two events, e1 and e2 (in bold) that hold the BEFORE relation is shown below:

> At one point , when it **(e1:became)** clear controllers could not contact the plane, someone **(e2:said)** a prayer.

MC-TACO (Ben Zhou and Roth, 2019): This task entirely focuses on temporal commonsense reasoning. It considers five temporal properties, (1) duration (how long an event takes), (2) temporal ordering (typical order of events), (3) typical time (when an event occurs), (4) frequency (how often an event occurs), and (5) stationarity (whether a state is maintained for a very long time or indefinitely). It contains 13k tuples, each consisting of a sentence, a question, and a candidate answer, that should be judged as plausible or not. An example from the dataset is below. The correct answer is in **bold**.

> *Paragraph*: Carl Laemmle, head of Universal Studios, gave Einstein a tour of his studio and introduced him to Chaplin.
>
> *Question*: How long did the tour last?
>
> a) 9 hours
>
> b) **45 minutes**
>
> c) 15 days
>
> d) 5 seconds

MAVEN-ERE (Wang et al., 2022): This is a unified large-scale human-annotated event relation extraction dataset. It was annotated at the document level from Wikipedia and FrameNet (Baker et al., 1998), for four tasks: event coreference, temporal, causal, and subevent relations. In our work, we focus on the sentence level temporal event pair relations. Given a passage and two event points, the task is to classify the relations between

Dataset	#Train	#Test	#Label	Label Distribution		Metrics
				Train	Test	
TimeML	1,248	1,003	2	yes: 789, no: 459	yes: 610, no: 393	Accuracy
MC-TACO	3,783	9,442	2	yes: 1,229, no: 2,554	yes: 3,198, no: 6,244	F1-Score
TB-Dense	4,177	1,426	6	VAGUE: 2015, BEFORE: 885, AFTER: 730, IS_INCLUDED: 275, INCLUDES: 209, SIMULTANEOUS: 63	VAGUE: 638, BEFORE: 380, AFTER: 278, IS_INCLUDED: 52, INCLUDES: 57, SIMULTANEOUS: 22	F1-Score
MATRES	12,740	837	4	BEFORE: 6,425, AFTER: 1,416, VAGUE: 4,481, OVERLAP: 418	BEFORE: 427, AFTER: 271, VAGUE: 30, OVERLAP: 109	Accuracy & F1-score
MAVEN-ERE	44,586	10,488	6	BEFORE: 35273, CONTAINS: 5204, SIMULTANEOUS: 2392, OVERLAP: 1605, BEGINS-ON: 58, ENDS-ON: 54	BEFORE: 8092, CONTAINS: 1426, SIMULTANEOUS: 609, OVERLAP: 346, BEGINS-ON: 6, ENDS-ON: 9	Accuracy & F1-score

Table 2: Summary of the English evaluation datasets.

events into one of 6 types: BEFORE, SIMULTANEOUS, CONTAINS, OVERLAP, ENDS-ON, and BEGINS-ON. Despite its larger size, the authors highlight that the label distribution in the dataset is severely unbalanced, but decided to keep the unbalanced distribution so that the dataset reflects the real-world data distribution (Wang et al., 2022). An example of a sentence with two events, e1 and e2 (in bold) that hold the BEFORE relation is shown below:

> It (**e1: turned**) again to 270 then began an abnormal (**e2: descent**).