Improving Event Duration Prediction via Time-aware Pre-training

Zonglin Yang **Xinya Du** **Alexander Rush** **Claire Cardie**
Department of Computer Science
Cornell University
{zy223, arush}@cornell.edu
{xdu, cardie}@cs.cornell.edu

Abstract

End-to-end models in NLP rarely encode external world knowledge about length of time. We introduce two effective models for duration prediction, which incorporate external knowledge by reading temporal-related news sentences (time-aware pre-training). Specifically, one model predicts the *range/unit* where the duration value falls in (R-PRED); and the other predicts the *exact* duration value (E-PRED). Our best model – E-PRED, substantially outperforms previous work, and captures duration information more accurately than R-PRED. We also demonstrate our models are capable of duration prediction in the unsupervised setting, outperforming the baselines.

1 Introduction

Understanding duration of event expressed in text is a crucial task in NLP (Pustejovsky and Verhagen, 2009; Zhou et al., 2019). It facilitates downstream tasks such as story timeline construction (Ning et al., 2018; Leeuwenberg and Moens, 2019) and temporal question answering (Llorens et al., 2015). It is challenging to make accurate prediction mainly due to two reasons: (1) duration is not only associated with event word but also the context. For example, "watch a movie" takes around 2 hours, while "watch a bird fly" only takes about 10 seconds; (2) the *compositional nature* of events makes it difficult to train a learning-based system only based on hand annotated data (since it's hard to cover all the possible events). Thus, external knowledge and commonsense are needed to make further progress on the task.

However, most current approaches (Pan et al., 2011; Gusev et al., 2011; Vempala et al., 2018) focus on developing features and cannot utilize external textual knowledge. The only exception is the web count based method proposed by Gusev et al. (2011), which queries search engine with event word (e.g., "watch") and temporal units, and make predictions based on hitting times. However, this method achieves better performance when query only with the event word in the sentence, which means it does not enable contextualized understanding.

To benefit from the generalizability of learning-based methods and utilizing external temporal knowledge, we introduce a framework, which includes (1) a procedure for collecting duration-related news sentences, and automatic labeling the duration unit in it (Section 2.1); [1] (2) two effective end-to-end models that leverage external temporal knowledge via pre-training (Section 2.2). Specifically, our first model (R-PRED) predicts the most likely temporal unit/range for the event, with a classification output layer; and the other model (E-PRED) predicts the *exact* duration value, with a regression output layer. Our best model (E-PRED) achieves state-of-the-art performance on the TimeBank dataset and the McTACO duration prediction task. In addition, in the unsupervised setting, our model (E-PRED) trained with only collected web data outperforms the supervised BERT baseline by 10.24 F1 score and 9.68 Exact Match score on McTACO duration prediction task. We also provide detailed comparisons and analysis between the regression objective (E-PRED) and the classification objective (R-PRED).

2 Our Framework

2.1 Duration-relevant Sentences Collection and Automatic Labeling

We use multiple pattern-based extraction rules to collect duration-relevant sentences. To avoid the potential data sparsity problem, we extract them

[1] We'll release these weakly supervised duration-relevant sentences in https://github.com/ZonglinY/Improving-Event-Duration-Prediction-via-Time-aware-Pre-training.git

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3370–3378
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

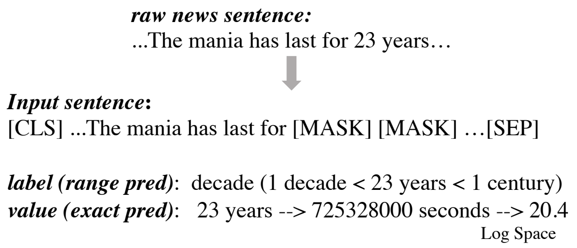
raw news sentence:
...The mania has last for 23 years…
Input sentence:
[CLS] ...The mania has last for [MASK] [MASK] …[SEP]
label (range pred): decade (1 decade < 23 years < 1 century)
value (exact pred): 23 years --> 725328000 seconds --> 20.4
Log Space

Figure 1: An Example of Automatic Labeling

from a relatively large corpus. In particular, we use articles in DeepMind Q&A dataset (Hermann et al., 2015) which contains approximately 287k documents extracted from CNN and Daily Mail news articles. To avoid introducing potential bias from a single pattern, we design multiple patterns for extraction. Specifically, if a sentence contains words or its variants as "for", "last", "spend", "take", "over", "duration", "period", and within certain number of words there exists a numerical value and a temporal unit (including second, minute, hour, day, week, month, year, decade) , then we consider the sentence as containing duration information and keep the sentence. Further, we design rules to filter sentences with certain patterns to avoid common misjudgements of the patterns to reach higher precision in retrieving sentences with duration information. More details are illustrated in Appendix A.2.

We apply rules to create the labels (Figure 1), specifically, given a candidate sentence, we extract the duration expression (23 years) which consists of a number and unit, then we normalize it to "second" space. We use the logarithm of the normalized value as label for E-PRED; and use the closest temporal unit as label for R-PRED model. Then for the sentence itself, we replace its duration expression with [MASK]s.

2.2 Models for Duration Prediction

The structure of E-PRED and R-PRED is shown in Figure 2. We first pass the input sentence through BERT (Devlin et al., 2019) to obtain contextualized embedding for the masked tokens, $\mathbf{x}_0$, $\mathbf{x}_1$, ..., $\mathbf{x}_k$. Then we add a linear layer on top of the BERT representations for prediction. We propose two variations – E-PRED (with a regression layer) predicts the exact duration value v;

$$\mathbf{v} = \mathbf{W}_e \sum_{i=0}^{k} \mathbf{x}_i$$

R-PRED (with a cross-entropy layer) predicts the range r.

$$\mathbf{r} = \text{softmax}(\mathbf{W}_r \sum_{i=0}^{k} \mathbf{x}_i)$$

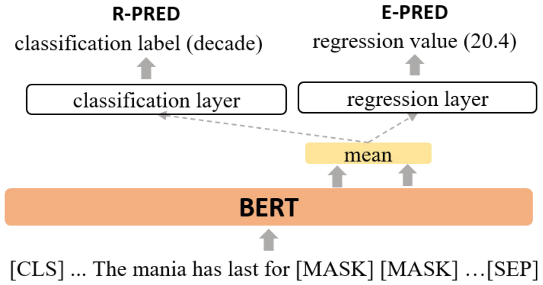
R-PRED
classification label (decade)
E-PRED
regression value (20.4)
classification layer
regression layer
mean
BERT
[CLS] ... The mania has last for [MASK] [MASK] …[SEP]

Figure 2: Models: R-PRED and E-PRED

3 Experiments and Analysis

3.1 Datasets and Evaluation Metrics

We evaluate our models on two duration-prediction benchmarks – TimeBank (Pan et al., 2011) and McTACO-duration (Zhou et al., 2019). **TimeBank**[2] annotates 48 non-Wall-Street-Journal articles (non-WSJ) and 10 WSJ articles. Specifically, it annotates duration for an event trigger (e.g., "watched") in the sentence (e.g., I watched a movie yesterday). Non-WSJ articles are splitted to generate train set and test set, and WSJ articles are used to generate testWSJ set, serving as an additional evaluation set. The Coarse-Grained task requires predicting whether the event takes less than a day or longer than a day; the Fine-Grained task requires predicting the most likely temporal unit (e.g., second, minute, hour, etc.). To transform the sentences into the input format of our models. We insert duration pattern (", lasting [MASK] [MASK], ") after event word and use the new sentence as the input sequence. For example, one sentence in TimeBank is "Philip Morris Cos, adopted a defense measure ...". Our method will convert it to "Philip Morris Cos, adopted, lasting [MASK] [MASK], a defense measure ...". Our strategy of directly adding duration pattern is possible to help pre-trained model to utilize learned intrinsic textual representation for duration prediction (Tamborrino et al., 2020).

McTACO is a multi-choice question answering dataset. **McTACO-duration**[3] is a subset of Mc-

[2] We use Gusev et al. (2011)'s split and obtain 1663/469/147 events in Train/Test/TestWSJ set respectively.

[3] In practice we collect context-question-answer triples that questions are about event duration and answers can be transformed to a duration value. We get 1060/2827 triples for dev/test set respectively (out of 1112/3032).

Model	Coarsed-Grained (Test)			Coarsed-Grained (TestWSJ)			Fine-Grained	
	<day F1	>day F1	Acc.	<day F1	>day F1	Acc.	Acc. (Test)	Acc. (TestWSJ)
Supervised Setting								
Majority class	-	76.90	62.47	-	76.99	62.58	59.28	52.38
Maximum Entropy (Pan et al., 2011)†	-	-	73.30	-	-	73.50	62.20	61.90
Maximum Entropy++ (Gusev et al., 2011)†	-	-	73.00	-	-	74.80	62.40	66.00
LSTM ensemble (Vempala et al., 2018)	64.29	82.69	76.69	73.20	87.78	83.21	-	-
TACOLM (Zhou et al., 2020)	80.58	88.88	85.86	**76.01**	**88.14**	**84.12**	-	-
R-PRED	**82.08**	87.72	85.43	70.15	81.12	76.87	82.09	76.19
w/o pre-training	80.94	86.19	84.01	73.46	79.93	77.32	80.38	**78.46**
E-PRED	80.63	**89.46**	**86.35**	70.67	85.39	80.50	**82.52**	**78.46**
w/o pre-training	78.73	88.16	84.79	73.50	86.21	81.86	80.34	77.02
Unsupervised Setting								
Majority	-	76.90	62.47	-	76.99	62.58	59.28	52.38
Web count, yesterday (Gusev et al., 2011)†	-	-	70.70	-	-	**74.80**	-	-
Web count, bucket (Gusev et al., 2011)†	-	-	72.40	-	-	73.50	66.50	**68.70**
R-PRED	63.19	80.39	74.41	5.19	66.36	50.34	69.72	43.54
E-PRED	60.14	82.52	**75.69**	2.86	69.64	53.74	**71.00**	41.50

Table 1: Performance on TimeBank. Results marked with † are reported in Gusev et al. (2011).

TACO whose questions are about event duration. Each data item includes a context sentence, a question, an answer (a duration expression) and a label indicating whether the answer is correct or not. To obtain the input sequence for our model, we convert the question to a statement using rule based method, and insert the same ", lasting [MASK] [MASK]." to the *end* of the statement sentence. For example, one question in McTACO-duration is "How long would they run through the fields?", our method will convert it to "they run through the fields, lasting [MASK] [MASK]." We then join the context sentence and newly obtained statement sentence as the input sequence.

We report F1 and accuracy for TimeBank Coarse-Grained task and accuracy for TimeBank Fine-Grained task. We report F1 and Exact Match (EM) for McTACO-duration.

3.2 Additional Dataset Details

In TimeBank Coarse-grained task, given an input event sentence, if prediction of E-PRED is smaller than 86400 seconds or prediction of R-PRED is "second" or "minute" or "hour", prediction will be "< day"; Otherwise prediction will be "> day". All models in TimeBank Fine-Grained task uses approximate agreement (Pan et al., 2011) during evaluation. In approximate agreement, temporal units are considered to match if they are the same temporal unit or adjacent ones. For example, "second" and "minute" match, but "minute" and "day" do not. It is proposed because human agreement on exact temporal unit is low (44.4%).

For McTACO-duration task, E-PRED uses $range$ as a hyper-parameter to define whether the answer is correct or not. Specifically, if the prediction of E-PRED is d, then only answers in $d \pm range$ in logarithmic second space are predicted as correct. We tune $range$ in development set. Here the $range$ we use is 3.0. R-PRED uses approximate agreement to predict correctness.

3.3 Baselines

We compare to strong models in the literature. For TimeBank, **Majority Class** always select "month" as prediction ("week", "month" and "year" are all considered as match because of approximate agreement). In the supervised setting, **Maximum Entropy** (Pan et al., 2011) and **Maximum Entropy++** (Gusev et al., 2011) are two models which utilize hand-designed time-related features. Difference is that Maximum Entropy++ uses more features than Maximum Entropy. **LSTM ensemble** (Vempala et al., 2018) is an ensemble LSTM (Hochreiter and Schmidhuber, 1997) model which utilize word embeddings. **TACOLM** (Zhou et al., 2020) is a concurrent work to our methods that also utilize unlabeled data. It uses a transformer-based structure and is also pre-trained on automatically labeled temporal sentences. Different from our model, TACOLM focuses on classification model and providing better representation instead of directly generating predicted duration. Here TACOLM forms Coarse-Grained task as a sequence classification task and uses the embedding of the first token of transformer output to predict from "< day" or "> day".

For McTACO-duration, **BERT_QA** (Zhou et al., 2019) is the BERT sentence pair (question and answer) classification model trained with McTACO-duration; **BERT_QA full** is the same model trained with all of McTACO examples. **TACOLM** here shares the same structure with BERT_QA but uses transformer weights pre-trained on collected data. To be fair, train data for TACOLM is McTACO-duration, the same as R-PRED and E-PRED. For the unsupervised setting, for TimeBank, we compare to **Web count-yesterday** and **Web count-bucket** (Gusev et al., 2011). They are rule-based approaches which rely on search engine.

3.4 Results

Table 1 presents results for TimeBank. In the supervised setting E-PRED achieves the best performance in Coarse-Grained task ("Test set") and Fine-Grained task, while it receives a lower performance than TACOLM in Coarse-Grained task ("TestWSJ"). In addition, E-PRED achieves best performance in Test set in unsupervised setting while it receives lower performance in TestWSJ set. However, Test set has a similar distribution with train set, while TestWSJ's is different (from a different domain). Therefore, performance on Test set should be a more important indicator for comparison.

We attribute the possible limitation of our models in TimeBank (especially TestWSJ set) experiments to reporting bias, relatively limited number of automatically collected data and mismatch of our duration pattern and TimeBank style annotation. More details are explained in Section 3.5. TACOLM's better performance in Coarse-Grained task in TestWSJ set might caused by its more compatible input format with TimeBank (it marks each event word that has a duration annotation in collected data) and its larger number of collected data from more sources.

Table 2 presents result on McTACO-duration. In supervised setting, E-PRED achieves the best performance. This table indicates that pre-training for incorporating external textual knowledge is helpful for both R-PRED and E-PRED. Plus, E-PRED which is trained with *only* web collected data still outperforms BERT_QA by a large margin.

We observe that E-PRED and R-PRED does not receive much performance gain from task-specific training. We attribute it to the noise introduced during transforming the QA data to fit in our models' input-output format. Specifically, we use the average of all correct answers as duration value label. This process is not guaranteed to get the expected duration value for each input event sentence.

Model	F1	EM
Supervised setting		
BERT_QA	51.95	30.32
BERT_QA full	56.98	32.26
TACOLM (Zhou et al., 2020)	57.60	33.50
R-PRED	55.36	25.48
w/o pre-training	50.05	22.58
E-PRED	**63.63***	**39.68***
w/o pre-training	45.31	25.48
Unsupervised Setting		
R-PRED	54.14	25.16
E-PRED	**62.19**	**40.00**

Table 2: Performance on McTACO-duration. * indicates that the difference compared to BERT_QA is statistically significant ($p < 0.01$) using Bootstrap method (Berg-Kirkpatrick et al., 2012)

3.5 Analysis

E-PRED or R-PRED? We provide insights on why BERT with regression loss generally outperforms BERT with a classification loss.

Firstly, we observe empirically that E-PRED generally outperforms R-PRED in TimeBank experiments. We attribute that E-PRED can catch more nuance information than R-PRED. For example, if the duration mentioned in the text is 40 min, then the generated label for R-PRED is "minute". While for E-PRED, the generated label is 40 minutes (1 min v.s. 40 min).

Secondly, E-PRED is more flexible and have a tunable range to predict the correctness (one of main reasons that E-PRED outperforms R-PRED largely in Table 2), while R-PRED can only use single bucket prediction or approximate agreement.

Effect of Time-aware Pre-training We observe that time-aware pre-training can lead to 5~18 F1 score improvement in McTACO-duration; while in TimeBank Coarse-Grained task, it can only lead to 1%~3% accuracy improvement in Test set, and causes around 1% accuracy drop in TestWSJ set.

We attribute the relatively limited effect of time-aware pre-training in TimeBank to reporting bias (Gordon and Van Durme, 2013) and data difference between McTACO-duration and TimeBank. Specifically, annotated events in McTACO-duration are

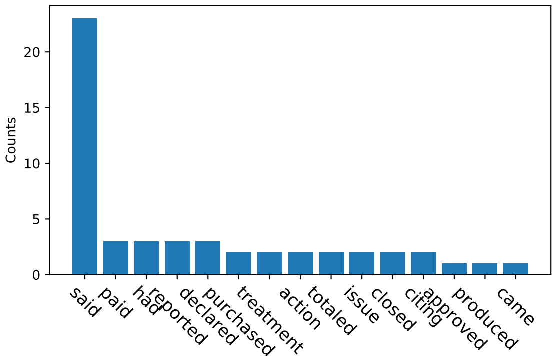

Figure 3: Times of event words that are predicted incorrectly by E-PRED in TimeBank TestWSJ set in unsupervised setting (only showing the 15 most frequent event words).

mainly description of concrete events, while annotated events in TimeBank are mainly abstract single words in the sentence. We consider that events in McTACO are more similar to events in our automatically collected data, while events in TimeBank are far less similar. Specifically, Figure 3 shows the most frequent single words annotated in TestWSJ that are predicted wrongly by E-PRED in unsupervised setting. We observe that event words in Figure 3 are mainly abstract and not durative, and people usually do not describe the duration of them in text (reporting bias). However, a larger collection of automatically collected data from different sources might alleviate this problem. More details on error analysis in TimeBank experiments can be found in Appendix A.4.

Another reason could be the mismatch of our designed duration pattern and TimeBank annotation style. Directly adding duration pattern after the annotated word might not comply with the sentences seen in pre-training data and might cause ambiguous reference of event.

Influence of Data Collection and Search Patterns We investigate how pre-training data collection affects the performance of our models. Table 3 shows performance of E-PRED in unsupervised setting pretrained w/ data collected with different methods. Specifically, we collect duration sentences from News or Wikipedia articles; sentences are collected by only the "for" pattern or "for|take|spend|last|lasting|duration|period" patterns (7 patterns). We find that E-PRED pre-trained with the three data collecting methods all achieves state-of-the-art performance in TimeBank Test (unsupervised setting) and get higher F1 score than BERT_QA supervised baseline. We find that pre-training with collected sentences can robustly increase our model's understanding of duration, and using more patterns for data collection is beneficial.

	TimeBank		McTACO-duration	
	Test	TestWSJ	F1	EM
Wiki (7 patterns)	70.15	**46.26**	57.34	36.77
News (only "for")	67.80	43.54	58.89	36.77
News (7 patterns)	**71.00**	41.50	**62.19**	**40.00**

Table 3: Effect of Data Collection and Search Patterns.

4 Additional Related Work

For **supervised** duration prediction, Pan et al. (2011) annotates duration length of a subset of events in TimeBank (Pustejovsky et al., 2003). New features and learning based models are proposed for TimeBank (Pan et al., 2011; Gusev et al., 2011; Samardzic and Merlo, 2016; Vempala et al., 2018). In particular, aspectual (Vendler, 1957; Smith, 2013) features have been proved to be useful. Concurrent to our work, Zhou et al. (2020) also utilize unlabeled data. Different from our work, they focus on temporal commonsense *acquisition* in a more general setting (for frequency, typical time, duration, etc.) and the models predict the discrete temporal unit, while we propose two models (classification and regression-based). In addition, they focus on providing better representation instead of directly generating duration prediction. For the **unsupervised** setting, Williams and Katz (2012); Elazar et al. (2019) use rule-based method on web data and generate collections of mapping from verb/event pattern to numeric duration value. Kozareva and Hovy (2011); Gusev et al. (2011) develop queries for search engines and utilize the returned snippets / hitting times to make prediction.

5 Conclusion

We propose a framework for leveraging free-form textual knowledge into neural models for duration prediction. Our best model (E-PRED) achieves state-of-the-art performance in various tasks. In addition, our model trained only with externally-obtained weakly supervised news data outperforms supervised BERT_QA baseline by a large margin. We also find that model trained with exact duration value seems to better capture duration nuance of event, and has more tunable range that is more flexible to make prediction for quantitative attributes such as duration.

Acknowledgments

We thank the anonymous reviewers for suggestions and Ben Zhou for running experiments.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005.

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.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Andrey Gusev, Nathanael Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. 2011. Using query patterns to learn the duration of events. In *Proceedings of the ninth international conference on computational semantics*, pages 145–154. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zornitsa Kozareva and Eduard Hovy. 2011. Learning temporal information for states and events. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 424–429. IEEE.

Artuur Leeuwenberg and Marie-Francine Moens. 2019. A survey on temporal reasoning for temporal information extraction from text. *Journal of Artificial Intelligence Research*, 66:341–380.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. Semeval-2015 task 5: Qa tempeval-evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multiaxis 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.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R. Hobbs. 2011. Annotating and learning event durations in text. *Computational Linguistics*, 37(4):727–752.

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.

James Pustejovsky and Marc Verhagen. 2009. Semeval-2010 task 13: evaluating events, time expressions, and temporal relations (tempeval-2). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 112–116.

Tanja Samardzic and Paola Merlo. 2016. Aspect-based learning of event duration using parallel corpora. *Essays in Lexical Semantics and Computational Lexicography–In Honor of Adam Kilgarriff, Springer Series Text, Speech, and Language Technology*.

Carlota S Smith. 2013. *The parameter of aspect*, volume 43. Springer Science & Business Media.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pretraining is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online. Association for Computational Linguistics.

Alakananda Vempala, Eduardo Blanco, and Alexis Palmer. 2018. Determining event durations: Models and error analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 164–168, New Orleans, Louisiana. Association for Computational Linguistics.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, pages 143–160.

Jennifer Williams and Graham Katz. 2012. Extracting and modeling durations for habits and events from twitter. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers), pages 223–227, Jeju Island, Korea. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

A Appendices

A.1 Hyper-Parameters

For pre-training BERT model with collected cheap supervised data, we use the same hyper parameters for time aware R-PRED and E-PRED:

- learning rate: 5e-5
- train batch size: 16
- optimizer: BertAdam (optimizer warmup proportion: 0.1)
- loss: mean square error loss (for E-PRED); cross entropy loss (for R-PRED)

For fine-tuning R-PRED or E-PRED with McTACO-duration or TimeBank data or fine-tuning BERT with McTACO-duration or TimeBank data, the hyper-parameter we use is:

- learning rate: 2e-5
- train batch size: 32
- optimizer: BertAdam (optimizer warmup proportion: 0.1)
- loss: mean square error loss (for E-PRED); cross entropy loss (for R-PRED)

A.2 Duration Data Collecting Method

We firstly use regular expression pattern to retrieve sentences that match with the pattern, then we use filter patter to filter out sentences that match with filter out pattern.

Regular expression pattern: “(?:duration|period|for|last|lasting|spend |spent|over|take|took|taken)[∧,.!?;]*\d+ (?:second|minute|hour|day|week|month|year|decade)”

Filter pattern:

- if the matched sub-sentence contains “at” or “age” or “every” or “next” or “more than” or “per”
- if the matched sub-sentence match with “(?:first|second|third|fourth|fifth|sixth|seventh |eighth|ninth) time”
- if the matched sentence matches with “|d+ secondary”
- if the matched sentence matches with “(?:second|minute|hour|day|week|month|year |decade)[s]? old”

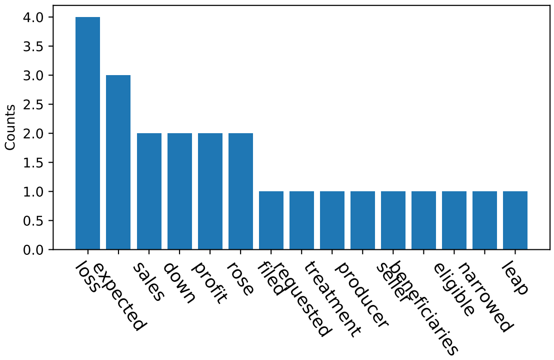

Figure 4: Times of event words that are predicted correctly in TimeBank TestWSJ set in unsupervised setting (only shows most frequent 15 event words)

A.3 Additional Details on Processing TimeBank and McTACO Data

Each annotated event trigger word in TimeBank are labeled with two duration values, max duration and min duration. We use the arithmetic mean of the two values to generate labels.

For TimeBank Fine-grained task, we use 7 temporal units as all possible labels (same setting with previous work (Gusev et al., 2011) (Pan et al., 2011)), including “second”, “minute”, “hour”, “day”, “week”, “month”, “year”. For R-PRED in McTACO task, we use 8 temporal units instead (adding “decade”)

A.4 Details on Correctly and Incorrectly Predicted Event Words in TimeBank Experiment

As shown in Figure 4, Figure 5 and Figure 6, we observe that correctly predicted words are generally more concrete and more possible to be described duration in text, which supports our analysis on reporting bias.

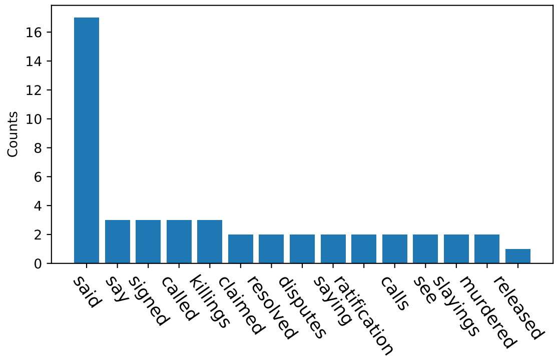

Figure 5: Times of event words that are predicted incorrectly in TimeBank Test set in unsupervised setting (only shows most frequent 15 event words)

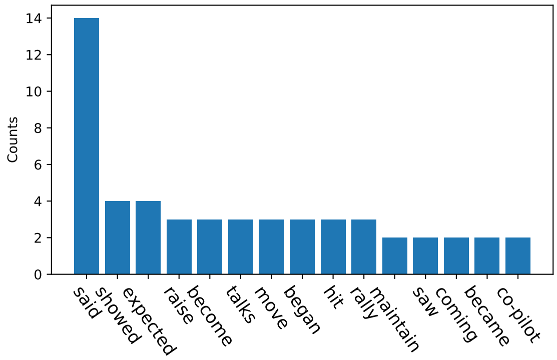

Figure 6: Times of event words that are predicted correctly in TimeBank Test set in unsupervised setting (only shows most frequent 15 event words)