Temporal Reasoning in Natural Language Inference

Siddharth Vashishtha
University of Rochester
svashis3@cs.rochester.edu

Adam Poliak
Barnard College
apoliak@barnard.edu

Yash Kumar Lal
Stony Brook University
yash.lal@stonybrook.edu

Benjamin Van Durme
Johns Hopkins University
vandurme@cs.jhu.edu

Aaron Steven White
University of Rochester
aaron.white@rochester.edu

Abstract

We introduce five new natural language inference (NLI) datasets focused on temporal reasoning. We recast four existing datasets annotated for *event duration*—how long an event lasts—and *event ordering*—how events are temporally arranged—into more than one million NLI examples. We use these datasets to investigate how well neural models trained on a popular NLI corpus capture these forms of temporal reasoning.

1 Introduction

The ability to reason about how events unfold in time is core to how humans structure their knowledge about the world (Casati and Varzi, 1996; Zacks and Tversky, 2001; Radvansky and Zacks, 2014), and modeling such temporal reasoning has been central to many classical AI approaches (McCarthy and Hayes, 1987; Kahn and Gorry, 1977; McDermott, 1982; Allen, 1984; Kowalski and Sergot, 1989; Pani and Bhattacharjee, 2001).

Natural language supports various forms of temporal reasoning, including reasoning about the chronology and duration of events, and many Natural Language Understanding (NLU) tasks and models have been employed for understanding and capturing different aspects of temporal reasoning (UzZaman et al., 2013; Llorens et al., 2015; Mostafazadeh et al., 2016; Reimers et al., 2016; Tourille et al., 2017; Ning et al., 2017, 2018a; Meng and Rumshisky, 2018; Ning et al., 2018b; Han et al., 2019; Naik et al., 2019; Vashishtha et al., 2019; Zhou et al., 2019, 2020). More broadly, the ability to perform temporal reasoning is important for understanding narratives (Nakhimovsky, 1987; Jung et al., 2011; Cheng et al., 2013), answering questions (Bruce, 1972; Khashabi, 2019; Ning et al., 2020), and summarizing events (Jung et al., 2011; Wang et al., 2018).

Order
▶ *We waited until 2:25 PM and then left.*
The waiting started before the leaving started.
▶ *Reggie said he will pay us soon.*
The paying ended before the saying started.
Duration
▶ *The greeter said there was about 15 mins waiting.*
The saying did take or will take shorter than an hour.
▶ *Randy , this is the issue I left you the voice mail on.*
The leaving did take or will take longer than a day.

Table 1: NLI sentence pairs from our recasted datasets. ▶ indicates the line is a context, and the following line is its corresponding hypothesis. Hypotheses in green indicate that the context entails the hypothesis; those in red indicate that it does not entail the hypothesis.

Given that temporal reasoning is integral to natural language understanding (NLU) and that Natural Language Inference (NLI) is a common framework for evaluating how well models capture semantic phenomena integral to NLU (Cooper et al., 1996; Dagan et al., 2006; White et al., 2017; Poliak et al., 2018), it is important to evaluate how well different classes of NLI models trained on common generic NLI datasets capture temporal reasoning.

We present five new NLI datasets recasted from four existing temporal reasoning datasets: (i) TempEval3 (TE3; UzZaman et al., 2013); (ii) TimeBank-Dense (TB-D; Chambers et al., 2014); (iii) Richer Event Description (RED; O'Gorman et al., 2016); and (iv) UDS-Time (UDS-T Vashishtha et al., 2019). Our new NLI datasets focus on two key aspects of temporal reasoning: (a) *temporal ordering* and (b) *event duration*.

We present strong baseline models for our temporal reasoning focused NLI datasets and also investigate the performance of common neural NLI models on these datasets. Our experiments demonstrate that common neural based NLI models trained on a popular dataset do not sufficiently capture temporal reasoning and require additional supervised training on datasets specific to temporal reasoning.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4070–4078

2 Motivation

A text often does not contain explicit mentions of how long events last or whether some events are contained within another. Consider (1).

(1) We waited until 2:25 pm and then left.

Although (1) does not explicitly mention how long the waiting lasted, one can reasonably guess that it lasted somewhere between minutes to hours—definitely not months or years. Zhou et al. (2020) note that common sense inference is required to come to such conclusions about an event's duration and text might even contain *reporting biases* when highlighting rarities (Schubert, 2002; Van Durme, 2011; Zhang et al., 2017a; Tandon et al., 2018), potentially making it hard to learn using common language modeling-based methods.

Popular NLI datasets contain hypotheses which are elicited by humans (Bowman et al., 2015; Williams et al., 2018). Although the context sentences for these datasets come from multiple genres, the constructed hypotheses do not necessarily capture semantic phenomenon which are essential for any robust NLU inference system. Recent work has catered to the lack of such inference capabilities by focusing on semantic phenomenon such as paraphrastic inference and anaphora resolution (White et al., 2017), veridicality (Poliak et al., 2018; Ross and Pavlick, 2019), and various other implicatures and presuppositions (Jeretic et al., 2020).

Even though temporal reasoning is crucial for event understanding, no datasets focused on temporal reasoning exist in the NLI format. To fill this lacuna, we recast four existing datasets to create NLI pairs that explicitly require reasoning about event duration and chronological ordering. Table 1 shows examples from two of our recasted datasets.

3 Dataset Creation

We construct five new NLI datasets recast from four existing datasets that focus on two key aspects of temporal reasoning: (a) *temporal ordering* and (b) *event duration*. Across these datasets, we have more than a million NLI examples and we retain the training, development, and test splits from the original (for datasets in which such splits exist). Table 2 reports the total number of NLI pairs in each of our recast datasets.

Phenomenon	Dataset	# NLI Pairs
duration	UDS-Time	504,136
order	UDS-Time	562,944
order	TempEval3	27,240
order	TimeBank-Dense	11,910
order	RED	5,578

Table 2: Recast datasets statistics

3.1 Temporal Ordering

To generate hypotheses for our temporal ordering datasets, we create 8 templates which refer to the start-points and end-points of events in a pair of two events. The templates are shown in Table 3.

We recast 4 datasets: (i) TE3; (ii) TB-D; (iii) RED; and (iv) UDS-T. UDS-T directly annotates for the relation between start and end points of events in an event pair, making hypothesis generation with our templates straight-forward. In contrast, TE3, TB-D, and RED annotate event pairs for categorical temporal relations based on those proposed by Allen (1983). Using each category's definition, we map that category to a *template predicate*—a function from hypothesis templates to {*entailed*, *not-entailed*}—summarized in Table 3.

TE3, which comprises of the TimeBank (Pustejovsky et al., 2003) and AQUAINT (Graff) corpora, contains 13 *temporal links*: *before* (`B`), *ibefore* (`IB`), *after* (`A`), *iafter* (`IA`), *isincluded* (`II`), *includes* (`I`), *begins* (`BE`), *begun-by* (`BB`), *ends* (`E`), *ended-by* (`EB`), *during* (`D`), *simultaneous* (`S`), and *identity*.[1] Each of these relations unambiguously maps to a template predicate.

TB-D uses a reduced set of relations: *before* (`Bt`), *after* (`At`), *isincluded* (`II`), includes (`I`), *simultaneous* (`S`), and *vague* (the last of which we ignore); as does RED: *before* (`Br`), *begins-on* (`BO`), *ends-on* (`EO`), *contains* (`C`), and *simultaneous* (`S`). This reduction results in the categories being ambiguous with respect to certain hypothesis templates. For instance, for Template 3 (*X ended before Y started*) knowing that X is *before* (`Bt`, `Br`) Y in the TB-D and RED sets does not give enough information about the ending point for X because these relations are not defined to have a strict ending boundary—in contrast to *before* (`B`) in TE3. We thus exclude hypothesis templates for ambiguous TB-D or RED relations.

For RED, we collapse relations with the same prefix into a single relation, e.g *before/causes*, *before/precondition* is collapsed into `Br`. We ignore

[1] For our purposes, *identity* and *simultaneous* denote the same relation.

	Hypothesis Template	Entailing Relations		
		TE3	TB-D	RED
1	X started before Y started	B,I,EB,IB,D	Bt,I	Br,C, EO?
2	X started before Y ended	B,I,II,S,IB,BB,BE,EB,D,E	Bt,I,II, At?	Br, C, EO, BO, S
3	X ended before Y started	B	?Bt?	Br?
4	X ended before Y ended	B,II,BE,IB,D	Bt,II	Br,BO?
5	Y started before X started	A,II,IA,E	At,II	EO?
6	Y started before X ended	A,I,II,S,IA,BB,BE,EB,D,E	At,I,II, Bt?	C, EO, BO, S, Br?
7	Y ended before X started	A	At?	-
8	Y ended before X ended	A,I,IA,BB	At,I	C, BO?

Table 3: Hypothesis templates for temporal ordering of events X and Y and the relations that entail those templates. If a relation does not entail a hypothesis template, then that template is mapped to *not-entailed* for that relation. A relation with a ? denotes that the relation cannot determine whether the template is *entailed* or *not-entailed*.

relations with *overlap* prefix as they do not have a clear boundary for start or end points of events.

3.2 Temporal Duration

To generate hypotheses for our temporal duration dataset, we create 18 hypothesis templates that refer to a range of likely durations for an event, based on two metatemplates: (i) *X did last or will last longer than* LOWER-BOUND and (ii) *X did last or will last shorter than* UPPER-BOUND, where LOWER-BOUND and UPPER-BOUND range over *a second*, *a minute*, *an hour*, *a day*, *a week*, *a month*, *a year*, *a decade*, and *a century*.[2]

We recast a single dataset—UDS-T—which contains annotations for the duration of an event drawn from the following 11 labels: *instantaneous*, *seconds*, *minutes*, *hours*, *days*, *weeks*, *months*, *years*, *decades*, *centuries*, and *forever*. For each event, we create two or four NLI pairs (depending upon the true label) to capture the duration information.

The *entailed* hypothesis of the NLI pair takes a range of duration values derived from the gold duration label for the given event. The lower limit of the range is one rank less than the gold label—e.g. for *minutes*, the LOWER-BOUND is *a second*—and the upper limit is one rank greater than the gold label—e.g. for *minutes*, the UPPER-BOUND is *an hour*. Two entailed hypotheses are then generated from these two limits, one corresponding to the lower limit—*longer than a second*, and the other corresponding to the upper limit—*shorter than an hour*. The corresponding *not-entailed* hypotheses are then generated by inverting the entailed hypothesis—e.g. for *minutes*: *shorter than a second* and *longer than an hour*. In cases, where the gold duration label is *instantaneous* or *forever*, only one *entailed* and one *not-entailed* pair in created.

3.3 Development and Test Splits

For the development and test set in UDS-T, there are three gold labels for each event-pair, so for the *entailed* hypothesis in these cases, we take the lower limit of duration range as one rank less than the lowest of the three gold labels and the upper limit as one higher than the highest of the three gold labels. For instance, if the three gold labels in the development set for an event are: *hours, weeks, months*, then the lower limit is *minutes* and the upper limit is *years*. The *entailed* and *not-entailed* hypothesis can then be generated using the same method described for the train set earlier.

TE3 does not have a development set, so we randomly sample documents from the train data and set it aside as development set. We use the same number of documents as that in the test set. Similarly, RED does not contain development and test splits, so we randomly sample 20% of the documents from train, evenly splitting them to create a development and a test set.

3.4 Grammatical Hypothesis Generation

We define rules to help generate hypotheses that are grammatical. We define our rules based on the Part-of-Speech (POS) tag of the events (predicates) in the context.

UDS-T contains gold POS tags, and the gold dependency trees for all contexts. So for any predicate which is tagged as a VERB in the context, we use its inflected form as a gerund in the hypothesis. For example, *'we **waited** until ...'* becomes *'the waiting started ...'*. Predicates with other POS tags in UDS-T occur with a copular construction, so we add the prefix *being* before the predicate to make it grammatical, for example, *'we're **happy***

[2]We use ranges of durations instead of a single gold label value as this gives us a more robust way of capturing durations, especially for cases where the true duration label is ambiguous in a given context as described in example (1).

...' becomes *'the being happy started ...'*. We also attach three types of direct modifiers of the predicate in the context – adjectives, determiners, and negations – to make the reference of the predicate specific to the context in the hypothesis. For example, *'we're not **happy** ...'* becomes *'the not being happy started ...'*. For cases where the lemma of the event appears multiple times in the context, we attach the direct object modifier of the event to make the reference unambiguous in the context. For example, to refer to the highlighted predicate in the context – *'we **cleaned** the apartment .... and they cleaned the washroom ...'* – we use the hypothesis *'the cleaning the apartment started ...'*. We use the gold dependency trees of each context to obtain these modifiers of the predicate. We do not consider predicates with `AUX` and `DET` POS tags for our recasting.

TE3, TB-Dense, and RED do not have gold dependency trees. Hence, we process and tokenize each sentence in these corpora using `Stanza` (Qi et al., 2020) to predict the POS, lemma and dependency trees for all the sentences. To tag the copular predicates in these corpora, we use PredPatt (White et al., 2016; Zhang et al., 2017b) that uses the predicted dependency trees from Stanza. To get the inflection on each verb, we use the Universal Morphology corpora (UniMorph, Sylak-Glassman et al., 2015; Kirov et al., 2018) for English and back-off to `LemmInflect` for tokens not found in UniMorph.[3]

4 Dataset Validation

To assess whether the recast NLI pairs are correct, we conduct a validation experiment by randomly sampling 100 NLI pairs from the train split of each dataset. For each NLI pair, we ask the annotators to answer the question – *How likely is it that the second sentence is true if the first sentence is true?* We provide 5 options to choose from – *extremely likely, very likely, even chance, very unlikely, extremely unlikely*.

We recruited 48 annotators from Amazon Mechanical Turk to validate the sampled NLI pairs for each of our 5 recasted datasets. We selected only those annotators who passed an American native-speaker test with 90% or above accuracy. Each item in our validation task listed 10 NLI pairs.

If our recasting produces valid NLI pairs, we should see that *entailed* pairs receive higher likelihood judgments than *not-entailed* pairs, even when adjusting for the dataset the pair comes from, the annotator, the pair, and the list of pairs the annotator saw the pair in. To test this, we fit an ordinal mixed effects model to the likelihood responses given by annotators, with a fixed effect for the source of the NLI pair as well as random intercepts for annotator, pair, and list. We compare this model to a model that additionally includes a fixed effect for the entailment label associated with the pair by our recasting. We find a reliable positive effect of the label being *entailed* ($\chi^2(1) = 226.1, p < 0.001$), indicating our recasting method produces valid NLI pairs.

5 Experimental Setup

We use our recast datasets to explore how well different common classes of NLI models capture temporal reasoning. Specifically, we use three types of models: (i) neural bag of words (NBOW; Iyyer et al., 2015) (ii) InferSent (Conneau et al., 2017), and (iii) RoBERTa (Liu et al., 2019).[4] Our NBOW model represents contexts and hypotheses as an average of GloVe embeddings (Pennington et al., 2014). The concatenation of these representations is fed to a MLP with one hidden layer. The InferSent model encodes contexts and hypotheses independently with a BiLSTM and sentence representations are extracted using max-pooling. The concatenation of these sentences, their difference, and their element-wise product (Mou et al., 2016) are then fed to a MLP. For Roberta, we use a classification head on top of the pooled output of `roberta-large` to predict the labels.[5]

In our experiments, we train and test these models on each recast temporal dataset. For each model, we include a hypothesis-only baseline to evaluate how much the datasets test NLI as opposed to just the likely duration and order of events in general. Additionally, we train each model on Multi-genre NLI (MNLI, Williams et al., 2018) and test the model on our datasets to see if the model learns temporal reasoning from a generic NLI dataset that does not necessarily focus on temporal reasoning.

[3]Details about LemmInflect can be found at: https://github.com/bjascob/LemmInflect

[4]Code here: https://github.com/sidsvash26/temporal_nli

[5]We include implementation details in Appendix A.

Model	UDS-duration	UDS-order	TempEval3	TimeBank-Dense	RED
Majority	50.00	54.52	54.57	50.54	52.51
MNLI Baseline					
NBOW	43.37	33.83	34.13	35.36	34.02
InferSent	49.04	49.58	47.57	47.84	47.64
RoBERTa	50.28	55.01	54.87	51.01	52.97
Hypothesis-Only					
NBOW	84.96	54.52	54.57	50.54	52.51
InferSent	91.18	71.96	63.22	68.29	63.47
RoBERTa	91.52	71.22	**63.25**	68.83	52.51
Context and Hypothesis					
NBOW	82.45	54.52	54.57	50.54	52.51
InferSent	92.65	73.22	62.2	68.29	63.47
RoBERTa	**94.51**	**80.17**	54.57	**94.60**	**80.59**

Table 4: Accuracies on the test set of our recast datasets as predicted by different settings of our models.

6 Results & Discussion

Table 4 shows the accuracy of different models on our recast temporal datasets. We report the majority baseline (MAJ) of always predicting the label that appeared the most in training. We see that the models trained on MNLI perform poorly on our recast datasets, even worse than MAJ baseline in many cases. This indicates that the models trained on MNLI do not learn representations well enough to infer temporal reasoning in our datasets.

The hypothesis-only models provide an interesting limitation of NBOW and InferSent. Both NBOW and InferSent hypothesis-only models are as good as, or even better, than the normal models across all datasets. RoBERTa, however, improves when given the context, across all datasets, with TempEval3 as the exception. This suggests that RoBERTa embeddings are better able to capture the semantics of the context than NBOW and InferSent. In fact, NBOW and InferSent may just predict the label based on information about lexical entities in the hypothesis.

Context in duration All three hypothesis-only models achieve high accuracy on the NLI dataset based on UDS-Duration. Even RoBERTa seems to fail to capture anything extra from the context. To analyze this anomaly, we create a hypothesis-template based majority baseline inferred from the UDS-Duration train data and find that it achieves an 80.2% accuracy on the test set. This indicates that the data is skewed for each template, which might be caused by the skewed *minutes* duration label in UDS-T (roughly 28% of the UDS-T train set contains *minutes* as the true duration label). This template based majority prediction is noteworthy as the models pretrained on MNLI fail to infer the correct labels even when the labels are skewed per template. The neural models see a 10% gain in accuracy over the template-sensitive majority, indicating that the models are learning the range of durations for different entities. Another possible reason that the context does not help much for duration is that events often have a modal distribution for a duration label, similar to the explanation for the recast NER data in Poliak et al. (2018)

7 Conclusion

To better capture temporal reasoning inference capabilities, we create a million NLI pairs recast from existing corpora in the literature that focus on two aspects of temporal reasoning – temporal duration and temporal order. We test existing models trained on MNLI on our datasets and find that a generic NLI model is not able to capture temporal reasoning. We show that training on our datasets can improve the performance of models in capturing temporal reasoning, and some aspects of temporal reasoning, specifically how long an event lasts, might be learned from lexical entities alone. We hope that our recast datasets push the research community to further explore how learning temporal reasoning could benefit other tasks.

Acknowledgments

We are grateful to the FACTS.lab at the University of Rochester as well as three anonymous reviewers for useful comments on this work. This research was supported by the University of Rochester, JHU HLTCOE, and DARPA KAIROS. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

James F Allen. 1984. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154.

Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Bertram C Bruce. 1972. A model for temporal references and its appliicationin a question answering program. *Artificial intelligence*.

Roberto Casati and Achille C. Varzi. 1996. *Events*. Dartmouth.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Yao Cheng, Peter Anick, Pengyu Hong, and Nianwen Xue. 2013. Temporal relation discovery between events and temporal expressions identified in clinical narrative. *Journal of biomedical informatics*, 46:S48–S53.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

David Graff. *The aquaint corpus of English news text:[content copyright] Portions© 1998-2000 New York Times, Inc.,© 1998-2000 Associated Press, Inc.,© 1996-2000 Xinhua News Service*. Linguistic Data Consortium.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. *arXiv preprint arXiv:2004.03066*.

Hyuckchul Jung, James Allen, Nate Blaylock, William de Beaumont, Lucian Galescu, and Mary Swift. 2011. Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 workshop*, pages 146–154.

Kenneth Kahn and G Anthony Gorry. 1977. Mechanizing temporal knowledge. *Artificial intelligence*, 9(1):87–108.

Daniel Khashabi. 2019. *Reasoning-Driven Question-Answering for Natural Language Understanding*. Ph.D. thesis, University of Pennsylvania.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: universal morphology. *arXiv preprint arXiv:1810.11101*.

Robert Kowalski and Marek Sergot. 1989. A logic-based calculus of events. In *Foundations of knowledge base management*, pages 23–55. Springer.

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 preprint arXiv:1907.11692*.

Hector Llorens, Nathanael Chambers, Naushad Uz-Zaman, 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.

John McCarthy and Patrick J Hayes. 1987. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in nonmonotonic reasoning*, pages 26–45.

Drew McDermott. 1982. A temporal logic for reasoning about processes and plans. *Cognitive science*, 6(2):101–155.

Yuanliang Meng and Anna Rumshisky. 2018. Context-aware neural model for temporal information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 527–536, Melbourne, Australia. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 130–136, Berlin, Germany. Association for Computational Linguistics.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Alexander Nakhimovsky. 1987. Temporal reasoning in natural language understanding: The temporal structure of the narrative. In *Proceedings of the Third Conference on European Chapter of the Association for Computational Linguistics*, EACL '87, page 262–269, USA. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037.

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)*, volume 1, pages 2278–2288.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. Torque: A reading comprehension dataset of temporal ordering questions. *arXiv preprint arXiv:2005.00242*.

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

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

AK Pani and GP Bhattacharjee. 2001. Temporal representation and reasoning in artificial intelligence: A review. *Mathematical and Computer Modelling*, 34(1-2):55–80.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. 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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Gabriel A Radvansky and Jeffrey M Zacks. 2014. *Event cognition*. Oxford University Press.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2195–2204.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? 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 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Lenhart Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of the second international conference on Human Language Technology Research*, pages 94–97. Morgan Kaufmann Publishers Inc.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680.

Niket Tandon, Bhavana Dalvi Mishra, Joel Grus, Wentau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. *arXiv preprint arXiv:1808.10012*.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2017. Neural architecture for temporal relation extraction: A bi-LSTM approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230, Vancouver, Canada. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, 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: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia. Association for Computational Linguistics.

Benjamin D Van Durme. 2011. Extracting implicit knowledge from text, proquest.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2018. Event phase oriented news summarization. *World Wide Web*, 21(4):1069–1092.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, TX. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Jeffrey M Zacks and Barbara Tversky. 2001. Event structure in perception and conception. *Psychological bulletin*, 127(1):3.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017a. Ordinal common-sense inference. *Transactions of the Association of Computational Linguistics*, 5(1):379–395.

Sheng Zhang, Rachel Rudinger, and Benjamin Van Durme. 2017b. An evaluation of predpatt and open ie via stage 1 semantic role labeling. In *IWCS 2017—12th International Conference on Computational Semantics (Short papers)*.

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. *arXiv preprint arXiv:2005.04304*.

A Model Implementation Details

For all of the experiments using Glove embeddings, we use 300-length dimensional embeddings. The MLP for the NBOW model has one hidden layer of 100 dimensions. The output from the hidden layer is fed to a logistic regression softmax classifier.

In InferSent, the encoders have one layer in each direction and we use Glove embeddings to initially represent the tokens. Sentence representations of length 2048 are extracted by max-pooling. The MLP has one hidden layer of 512 dimensions. We optimize the model using SGD. We set the initial learning rate to 0.1 and decay rate to 0.99 and we train over 20 epochs.

For Roberta, we use the `transformers` (Wolf et al., 2019) library from HuggingFace and use their `RobertaForSequenceClassification` class to implement our model. We use a mini-batch size of 16 trained over 2 GPUs with an Adam optimizer using 122 warmup steps and an initial learning rate of 2e-5 and a 0.1 weight decay. For UDS-T recast datasets we run the Roberta models for 2 epochs. For TE3, TBD, and RED we run the model for 10 epochs.

The MNLI dataset has three labels: *neutral*, *contradiction*, and *entailment*. For the MNLI Baseline models, we train the models to predict these three labels, but when we evaluate these models on our recast datasets, we follow common practice (Belinkov et al., 2019) by converting *neutral* and *contradiction* to the *not-entailed* label during test time.