Temporal Reasoning in Natural Language Inference

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

Adam Poliak
Barnard College
apolia@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 Rumlshisky, 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).

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.

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.

<table>
<thead>
<tr>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>► We waited until 2:25 PM and then left.</td>
</tr>
<tr>
<td>The waiting started before the leaving started.</td>
</tr>
<tr>
<td>► Reggie said he will pay us soon.</td>
</tr>
<tr>
<td>The paying ended before the saying started.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>► The greeter said there was about 15 mins waiting.</td>
</tr>
<tr>
<td>The saying did take or will take shorter than an hour.</td>
</tr>
<tr>
<td>► Randy, this is the issue I left you the voice mail on.</td>
</tr>
<tr>
<td>The leaving did take or will take longer than a day.</td>
</tr>
</tbody>
</table>

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4070–4078
November 16 - 20, 2020. ©2020 Association for Computational Linguistics
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.

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Dataset</th>
<th># NLI Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>duration</td>
<td>UDS-Time</td>
<td>504,136</td>
</tr>
<tr>
<td>order</td>
<td>UDS-Time</td>
<td>562,944</td>
</tr>
<tr>
<td>order</td>
<td>TempEval3</td>
<td>27,240</td>
</tr>
<tr>
<td>order</td>
<td>TimeBank-Dense</td>
<td>11,910</td>
</tr>
<tr>
<td>order</td>
<td>RED</td>
<td>5,578</td>
</tr>
</tbody>
</table>

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), begins (BE), begun-by (BB), ends (E), ended-by (EB), during (D), simultaneous (S), and identity. Each of these relations unambiguously maps to a template predicate.

TB-D uses a reduced set of relations: before (Bt), after (At), isincluded (II), 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

---

1For our purposes, identity and simultaneous denote the same relation.
To generate hypotheses for our temporal duration

<table>
<thead>
<tr>
<th>Hypothesis Template</th>
<th>TE3</th>
<th>Entailing Relations</th>
<th>RED</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 X started before Y started</td>
<td>B, I, E, B, I, D</td>
<td>Br, C, E O?</td>
<td></td>
</tr>
<tr>
<td>3 X ended before Y started</td>
<td>B</td>
<td>Br</td>
<td></td>
</tr>
<tr>
<td>4 X ended before Y ended</td>
<td>B, I, B E, I B, D</td>
<td>Br</td>
<td></td>
</tr>
<tr>
<td>5 Y started before X started</td>
<td>A, I I, A E</td>
<td>E O?</td>
<td></td>
</tr>
<tr>
<td>7 Y ended before X started</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>8 Y ended before X ended</td>
<td>A, I, I A, B B</td>
<td>C, B O?</td>
<td></td>
</tr>
</tbody>
</table>

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
...’ 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.

---

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

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

5We include implementation details in Appendix A.
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.

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

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

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, 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.

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


A Model Implementation Details

For all of the experiments using Glove embeddings, we use 300-length dimensional embeddings. The MLP for the NBOV 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.