Fine-Grained Temporal Relation Extraction

Siddharth Vashishtha
University of Rochester

Benjamin Van Durme
Johns Hopkins University

Aaron Steven White
University of Rochester

Data and code available at:

http://decomp.io

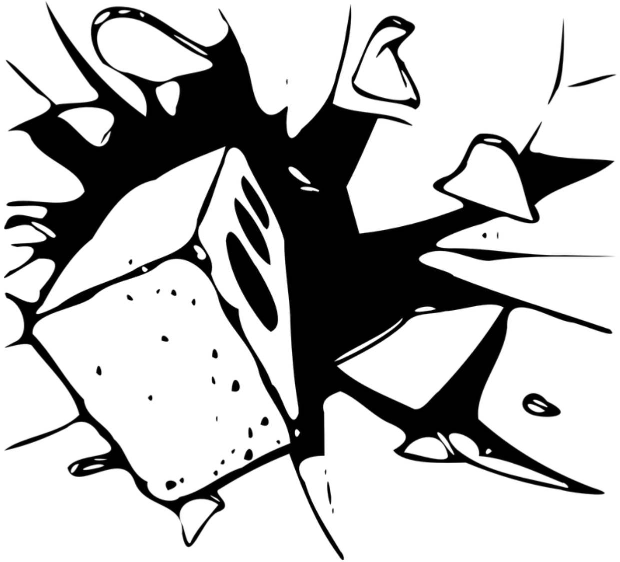
Overarching claim

Humans are good at extracting the chronology of events from linguistic input.

Overarching claim

Consider the narrative:

At 3pm, a boy **broke** his neighbor's window.



Overarching claim

Consider the narrative:

At 3pm, a boy **broke** his neighbor's window. He was **running away,** when the neighbor **rushed out** to **confront** him.

Overarching claim

Consider the narrative:

At 3pm, a boy **broke** his neighbor’s window. He was **running away,** when the neighbor **rushed out** to **confront** him. His parents were **called** but couldn’t **arrive** for two hours because they **were still at work.**

Overarching claim

Consider the narrative:

At 3pm, a boy **broke** his neighbor’s window. He was **running away**, when the neighbor **rushed out** to **confront** him. His parents were **called** but couldn’t **arrive** for two hours because they **were still at work.**

Each predicate denotes some event

A typical timeline of events

At 3pm, a boy **broke** his neighbor’s window.

break

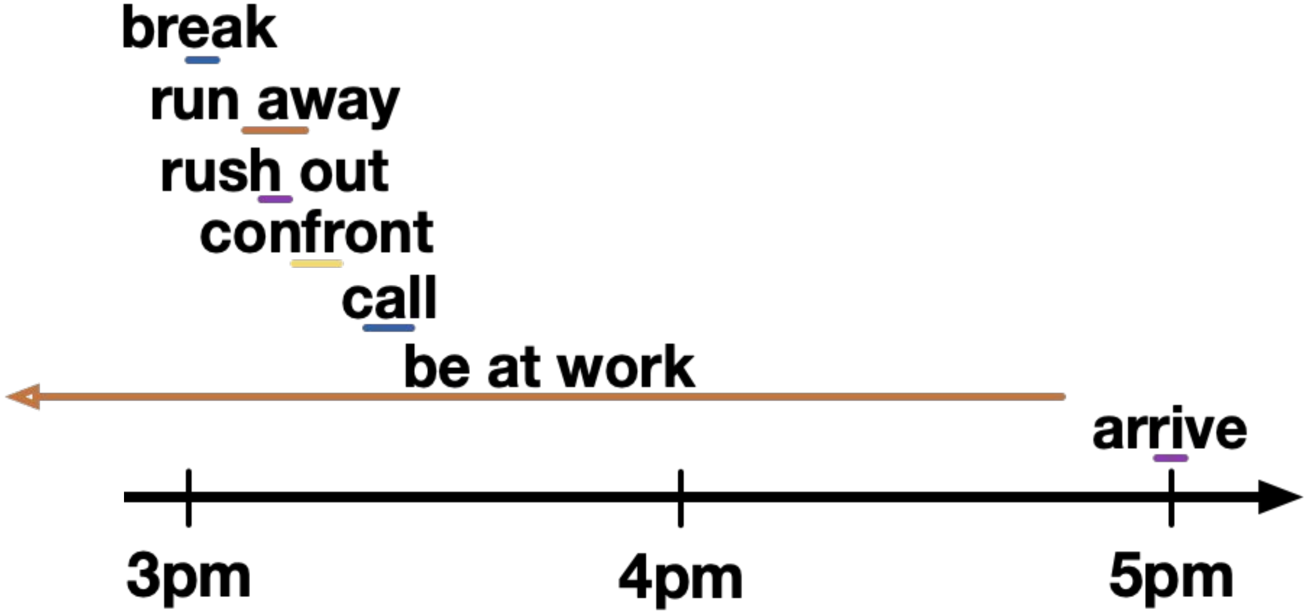
3pm

A typical timeline of events

At 3pm, a boy **broke** his neighbor’s window. He was **running away,** when the neighbor **rushed out** to **confront** him.

A typical timeline of events

At 3pm, a boy **broke** his neighbor's window. He was **running away,** when the neighbor **rushed out** to **confront** him. His parents were **called** but couldn't **arrive** for two hours because they **were still at work**.



Objective

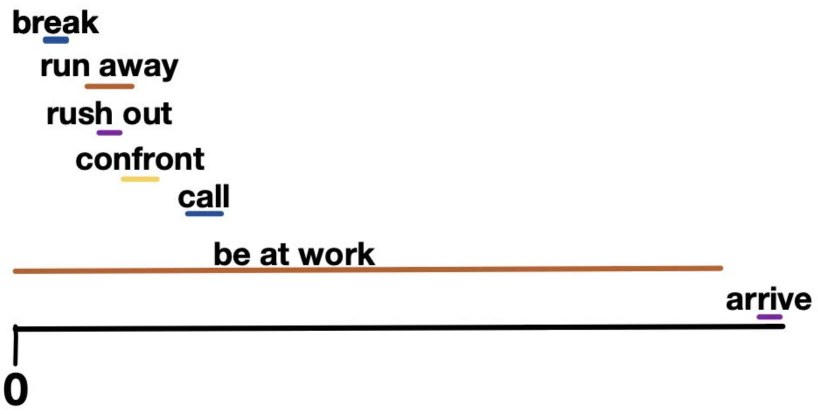
Input Document:

At 3pm, a boy **broke** his neighbor’s window. He was **running away,** when the neighbor **rushed out** to **confront** him. His parents were **called** but couldn’t **arrive** for two hours because they **were still at work**.

Objective

Input Document:

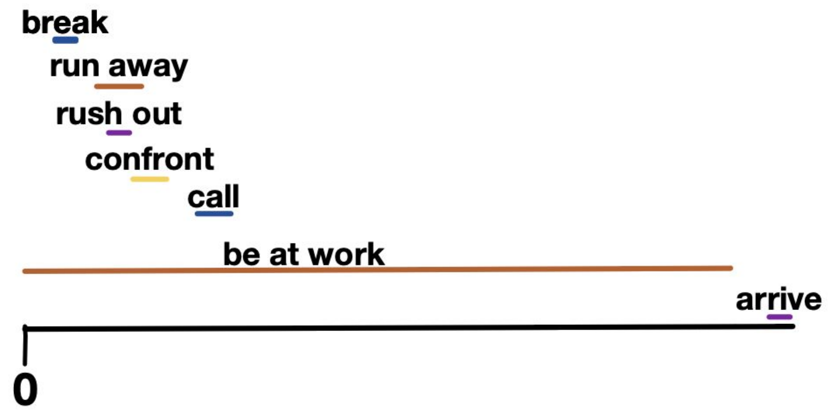
At 3pm, a boy **broke** his neighbor's window. He was **running away,** when the neighbor **rushed out** to **confront** him. His parents were **called** but couldn't **arrive** for two hours because they **were still at work**.



Objective

Input Document:

At 3pm, a boy **broke** his neighbor's window. He was **running away,** when the neighbor **rushed out** to **confront** him. His parents were **called** but couldn't **arrive** for two hours because they **were still at work**.



Two components are crucial:

1. **Relations** between events
2. **Durations** of individual events

Outline

Background

Methodology

Model

Results

Model Analysis

Conclusion

Background

Categorical Temporal Relations

A standard approach: Pairwise categorical temporal relation extraction based on **Allen Relations** (1983).

(Pustejovsky et al., 2003; Styler IV et al., 2014; Minard et al., 2016)

Categorical Temporal Relations

A standard approach: Pairwise categorical temporal relation extraction based on **Allen Relations** (1983).

Relation	Illustration	Interpretation
$X < Y$ $Y > X$	X Y	X takes place before Y
$X\,\mathbf{m}\,Y$ $Y\,\mathbf{mi}\,X$	X Y	X meets Y (*i* stands for *inverse*)
$X\,\mathbf{o}\,Y$ $Y\,\mathbf{oi}\,X$	X Y	X overlaps with Y
$X\,\mathbf{s}\,Y$ $Y\,\mathbf{si}\,X$	X Y	X starts Y
$X\,\mathbf{d}\,Y$ $Y\,\mathbf{di}\,X$	X Y	X during Y
$X\,\mathbf{f}\,Y$ $Y\,\mathbf{fi}\,X$	X Y	X finishes Y
$X = Y$	X Y	X is equal to Y

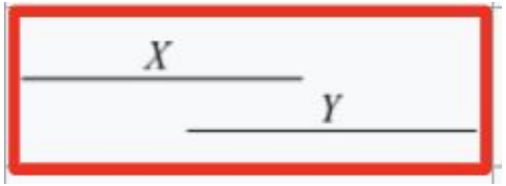
For example: X takes place before Y

Categorical Temporal Relations

A standard approach: Pairwise categorical temporal relation extraction based on **Allen Relations** (1983).

Relation	Illustration	Interpretation
$X < Y$ $Y > X$	X Y	X takes place before Y
$X\,\mathbf{m}\,Y$ $Y\,\mathbf{mi}\,X$	X Y	X meets Y (*i* stands for *inverse*)
$X\,\mathbf{o}\,Y$ $Y\,\mathbf{oi}\,X$	X Y	X overlaps with Y
$X\,\mathbf{s}\,Y$ $Y\,\mathbf{si}\,X$	X Y	X starts Y
$X\,\mathbf{d}\,Y$ $Y\,\mathbf{di}\,X$	X Y	X during Y
$X\,\mathbf{f}\,Y$ $Y\,\mathbf{fi}\,X$	X Y	X finishes Y
$X = Y$	X Y	X is equal to Y

For example: X overlaps with Y

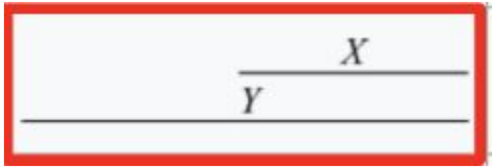


Categorical Temporal Relations

A standard approach: Pairwise categorical temporal relation extraction based on **Allen Relations** (1983).

Relation	Illustration	Interpretation
$X < Y$ $Y > X$	X Y	X takes place before Y
$X\ \mathbf{m}\ Y$ $Y\ \mathbf{mi}\ X$	X Y	X meets Y (*i* stands for *inverse*)
$X\ \mathbf{o}\ Y$ $Y\ \mathbf{oi}\ X$	X Y	X overlaps with Y
$X\ \mathbf{s}\ Y$ $Y\ \mathbf{si}\ X$	X Y	X starts Y
$X\ \mathbf{d}\ Y$ $Y\ \mathbf{di}\ X$	X Y	X during Y
$X\ \mathbf{f}\ Y$ $Y\ \mathbf{fi}\ X$	X Y	X finishes Y
$X = Y$	X Y	X is equal to Y

For example: X finishes Y



Corpora

- TimeBank corpus

(Pustejovsky et al., 2003)

Corpora

- TimeBank corpus
- TempEval tasks

(Verhagen et al., 2007, 2010; UzZaman et al., 2013)

Corpora

- TimeBank corpus
- TempEval tasks
- TimeBank-Dense

(Cassidy et al., 2014)

Corpora

- TimeBank corpus
- TempEval tasks
- TimeBank-Dense
- Richer Event Description (RED)

(O’Gorman et al., 2016)

Corpora

- TimeBank corpus
- TempEval tasks
- TimeBank-Dense
- Richer Event Description (RED)
- Hong et al. (2016)

Corpora

- TimeBank corpus
- TempEval tasks
- TimeBank-Dense
- Richer Event Description (RED)
- Hong et al. (2016)
- Grounded Annotation Framework (GAF)

(Fokkens et al., 2013)

Models

- Hand-tagged features with multinomial logistic regression and Support Vector Machines (SVM)

(Mani et al., 2006; Bethard, 2013; Lin et al., 2015)

Models

- Hand-tagged features with multinomial logistic regression and Support Vector Machines (SVM)
- Combined rule based and learning-based approaches

(D’Souza and Ng, 2013)

Models

- Hand-tagged features with multinomial logistic regression and Support Vector Machines (SVM)
- Combined rule based and learning-based approaches
- Sieve-based architectures— CAEVO and CATENA

(Chambers et al., 2014; Mirza and Tonelli, 2016)

Models

- Hand-tagged features with multinomial logistic regression and Support Vector Machines (SVM)
- Combined rule based and learning-based approaches
- Sieve-based architectures— CAEVO and CATENA
- Structured learning approaches

(Leeuwenberg and Moens, 2017 ; Ning et al., 2017)

Models

- Hand-tagged features with multinomial logistic regression and Support Vector Machines (SVM)
- Combined rule based and learning-based approaches
- Sieve-based architectures— CAEVO and CATENA
- Structured learning approaches
- Neural Network based approaches

(Tourille et al., 2017; Cheng and Miyao, 2017; Leeuwenberg and Moens, 2018, Dligach et al., 2017)

Models

- Hand-tagged features with multinomial logistic regression and Support Vector Machines (SVM)
- Combined rule based and learning-based approaches
- Sieve-based architectures— CAEVO and CATENA
- Structured learning approaches
- Neural Network based approaches
- Jointly modeling causal and temporal relations

(Ning et al., 2018)

Models

- Hand-tagged features with multinomial logistic regression and Support Vector Machines (SVM)
- Combined rule based and learning-based approaches
- Sieve-based architectures— CAEVO and CATENA
- Structured learning approaches
- Neural Network based approaches
- Jointly modeling causal and temporal relations
- Event durations from text

(Pan et al., 2007; Gusev et al., 2011; Williams and Katz, 2012)

Corpora Drawbacks

Corpora Drawbacks

- Event durations are not explicitly captured.

Corpora Drawbacks

- Event durations are not explicitly captured.

```
<TIMEX TYPE="TIME"> twelve o'clock noon </TIMEX>
```

```
<TIMEX TYPE="DATE"> fiscal 1989's fourth quarter </TIMEX>
```

Corpora Drawbacks

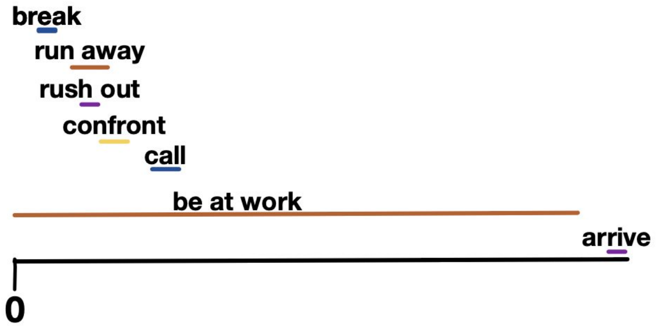
- Event durations are not explicitly captured.

- Experts are needed to annotate these datasets.



Corpora Drawbacks

- Event durations are not explicitly captured.
- Experts are needed to annotate these datasets.
- Event timelines are not directly captured and it is not trivial to create document timelines.



Corpora Drawbacks

- Event durations are not explicitly captured.
- Experts are needed to annotate these datasets.
- Event timelines are not directly captured and it is not trivial to create document timelines.

  However, approaches have been used to create relative timelines from the temporal relations

(Leeuwenberg and Moens, 2018)

Methodology

Representing Event Timelines

- A novel **Universal Decompositional Semantics** (UDS) framework for temporal relation representation that puts <u>event duration front and center.</u>

Representing Event Timelines

- A novel **Universal Decompositional Semantics** (UDS) framework for temporal relation representation that puts <u>event duration front and center.</u>

- We map the events or situations to a timeline represented in real numbers.

Representing Event Timelines

- A novel **Universal Decompositional Semantics** (UDS) framework for temporal relation representation that puts <u>event duration front and center.</u>

- We map the events or situations to a timeline represented in real numbers.

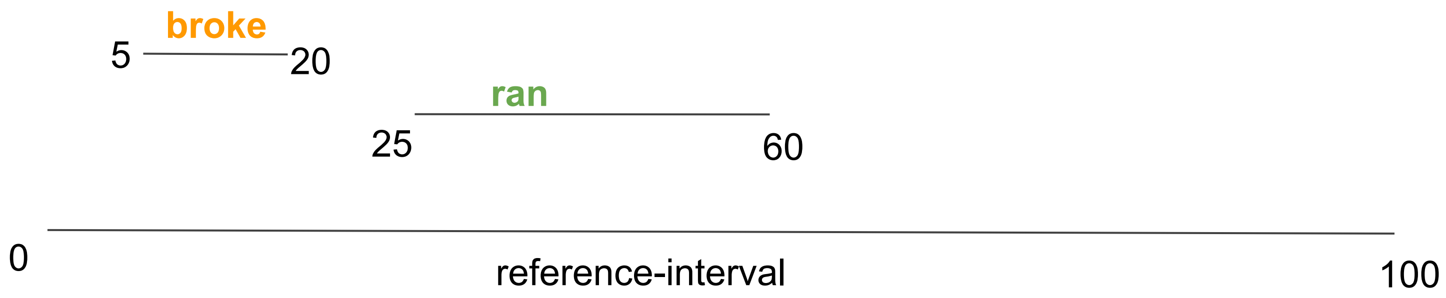
Sam broke the window and ran away.

Representing Event Timelines

- A novel **Universal Decompositional Semantics** (UDS) framework for temporal relation representation that puts <u>event duration front and center.</u>

- We map the events or situations to a timeline represented in real numbers.

Sam broke the window and ran away.



Protocol Design

- We ask questions about the chronology of events and the duration of each event
- Annotated example (next slide)

What to [1] feed my dog after gastroenteritis ? My dog has [2] been [2] sick [2] for about 3 days [2] now .

[1]feed

Range: 49 - 66

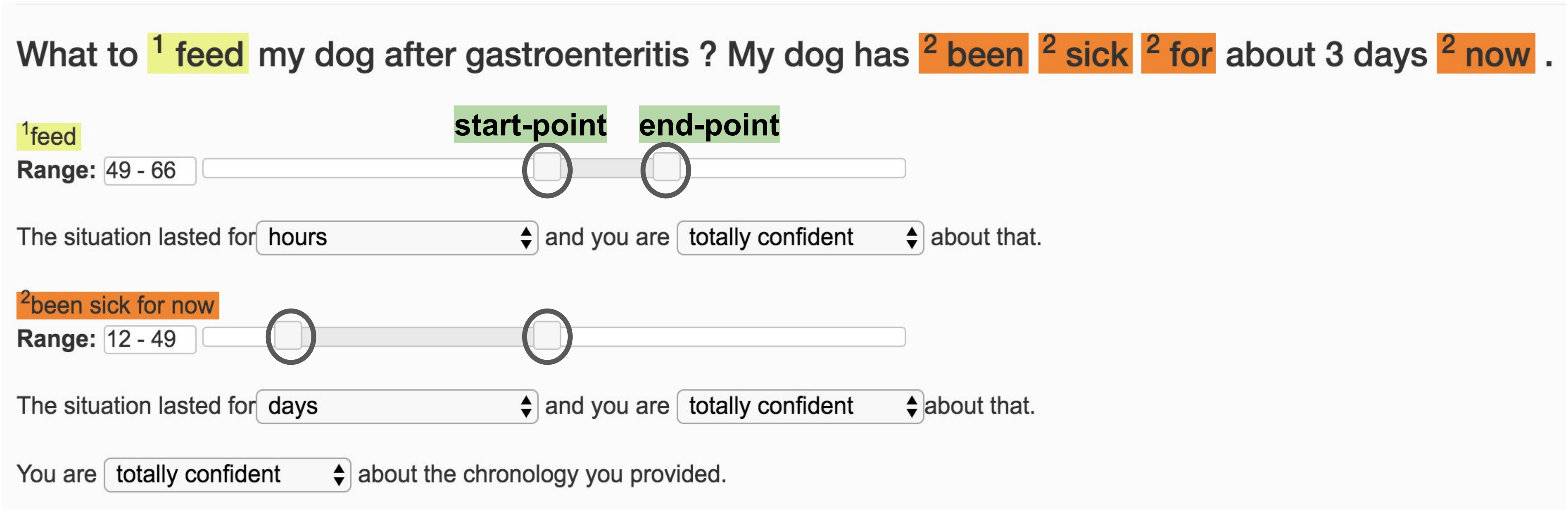
The situation lasted for hours and you are totally confident about that.

[2]been sick for now

Range: 12 - 49

The situation lasted for days and you are totally confident about that.

You are totally confident about the chronology you provided.


What to 1 feed my dog after gastroenteritis ? My dog has 2 been 2 sick 2 for about 3 days 2 now .
start-point
end-point
1feed
Range: 49 - 66
The situation lasted for hours and you are totally confident about that.
2been sick for now
Range: 12 - 49
The situation lasted for days and you are totally confident about that.
You are totally confident about the chronology you provided.

What to [1] feed my dog after gastroenteritis ? My dog has [2] been [2] sick [2] for about 3 days [2] now .

[1]feed

Range: 49 - 66

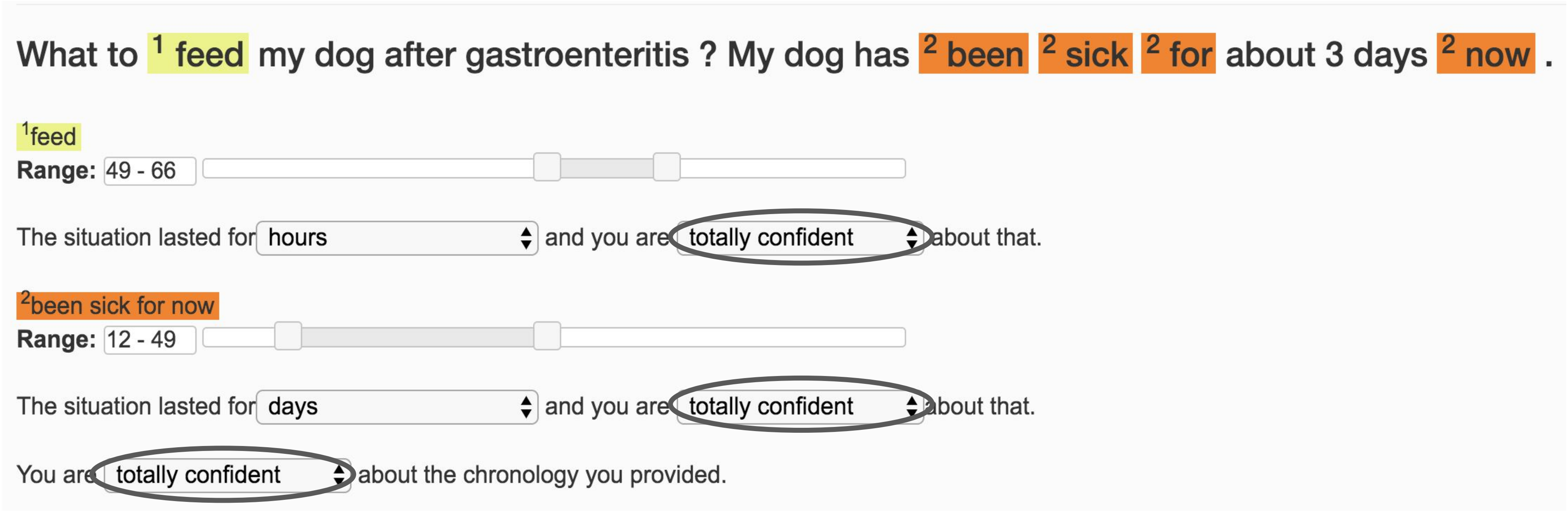
The situation lasted for hours and you are totally confident about that.

[2]been sick for now

Range: 12 - 49

The situation lasted for days and you are totally confident about that.

You are totally confident about the chronology you provided.


What to 1 feed my dog after gastroenteritis ? My dog has 2 been 2 sick 2 for about 3 days 2 now .
1feed
Range: 49 - 66
The situation lasted for hours and you are totally confident about that.
2been sick for now
Range: 12 - 49
The situation lasted for days and you are totally confident about that.
You are totally confident about the chronology you provided.

Data Collection

- We took English Web Treebank (EWT) from **Universal Dependencies (UD)** and designed a protocol to extract fine-grained temporal relations.

Data Collection

- We took English Web Treebank (EWT) from **Universal Dependencies (UD)** and designed a protocol to extract fine-grained temporal relations.

- Extracted predicates from UD-data using **PredPatt**

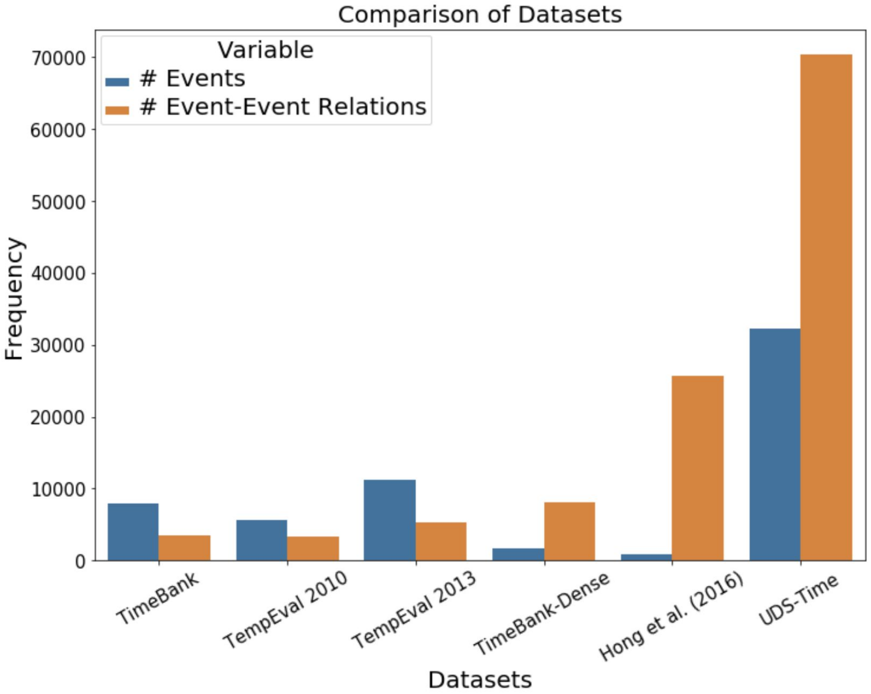
(White et al., 2016; Zhang et al., 2017)

Constructed Data

- We recruited 765 annotators from Amazon Mechanical Turk to annotate predicate pairs in groups of five. The resulting dataset is **UDS-Time.**

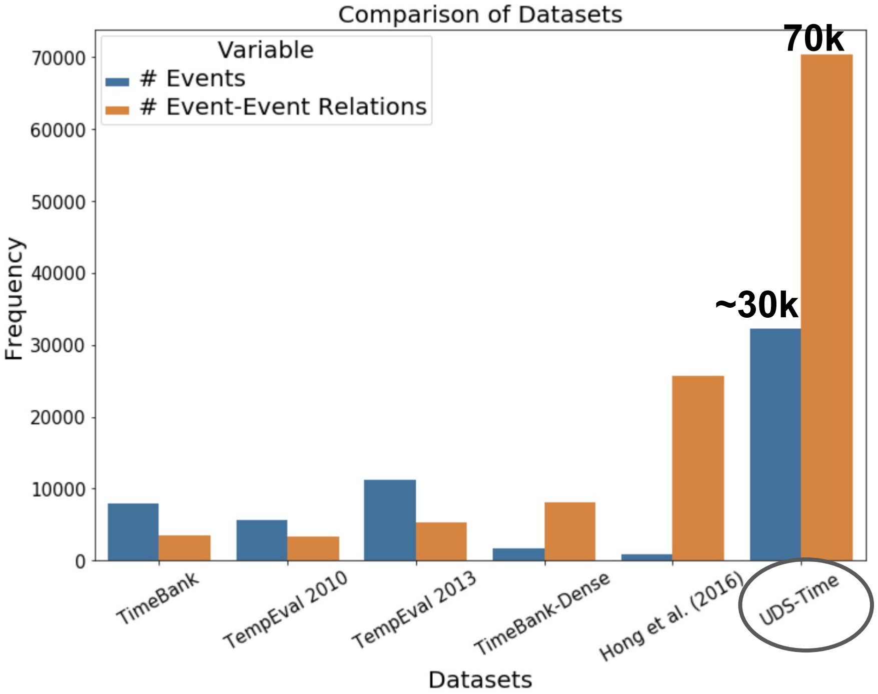
Constructed Data

- We recruited 765 annotators from Amazon Mechanical Turk to annotate predicate pairs in groups of five. The resulting dataset is **UDS-Time.**



Constructed Data

- We recruited 765 annotators from Amazon Mechanical Turk to annotate predicate pairs in groups of five. The resulting dataset is **UDS-Time.**

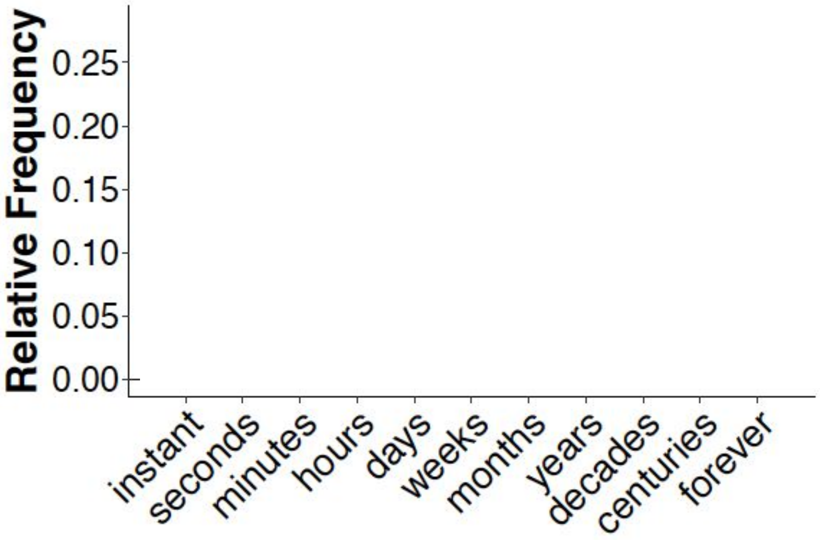


Data Distributions

Event Durations

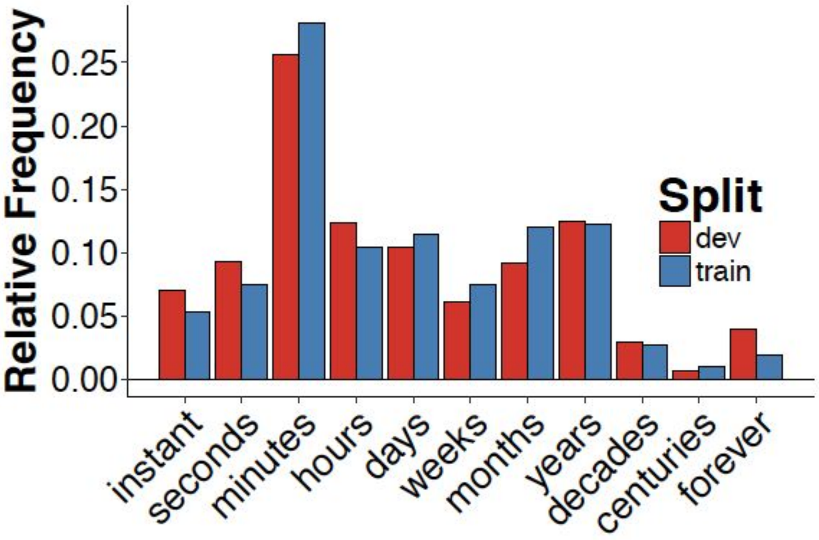
Data Distributions

Event Durations



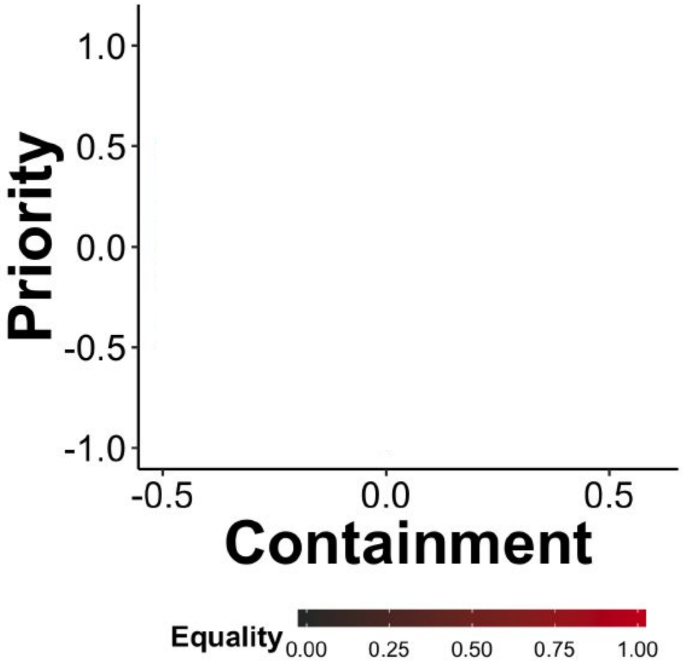
Data Distributions

Event Durations



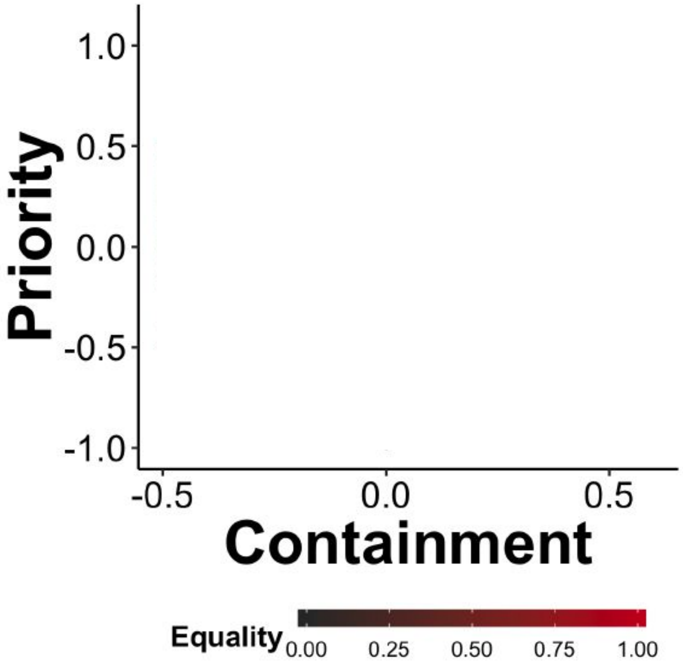
Data Distributions

Event Relations

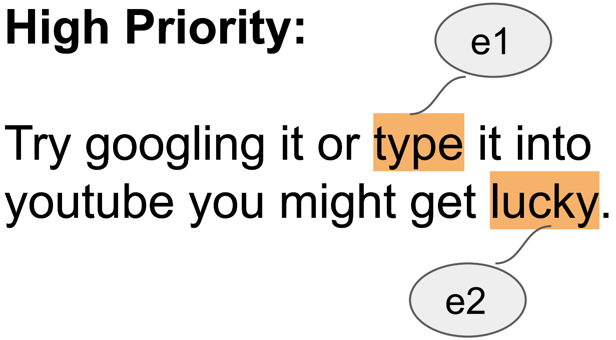


Data Distributions

Event Relations

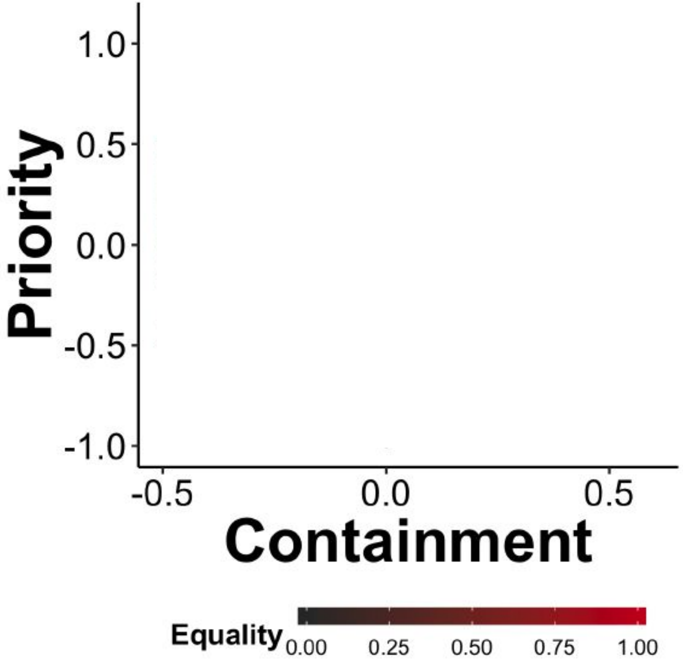


High Priority:



Data Distributions

Event Relations



High Containment:

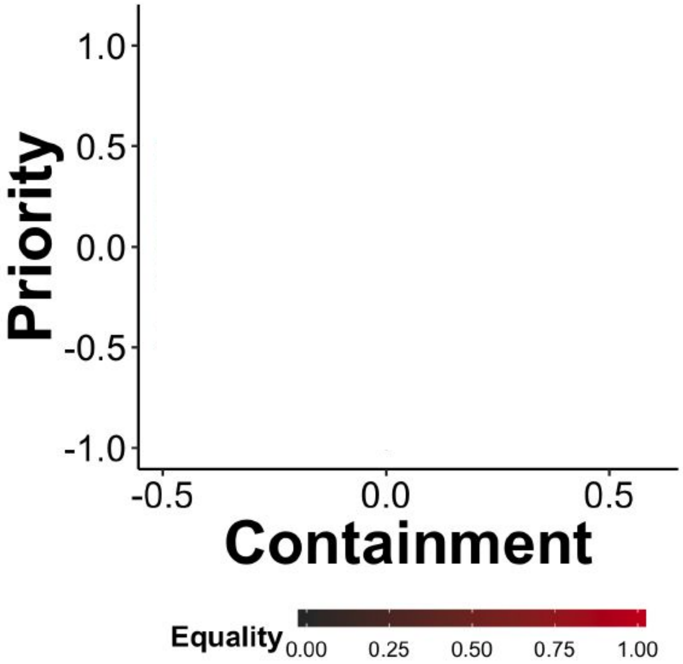
Both Tina and Vicky are excellent. I will definitely refer my friends and family.

e1

e2

Data Distributions

Event Relations



High Equality:

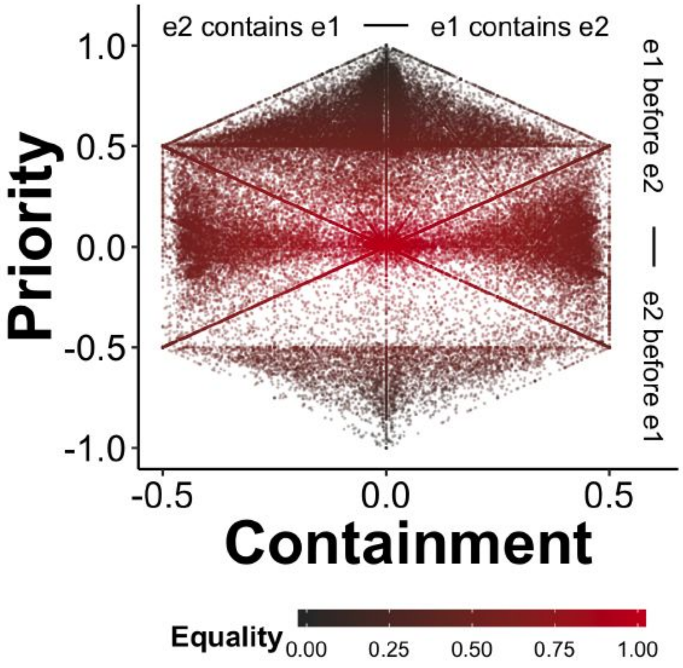
e1

I go Disco dancing and Cheerleading. It's fab!

e2

Data Distributions

Event Relations



Model

Goal

To model the **pairwise fine-grained temporal relations** and **durations** by attempting to automatically build featural representations of each predicate, its duration and its relation.

Model Architecture

1. **Event representation**

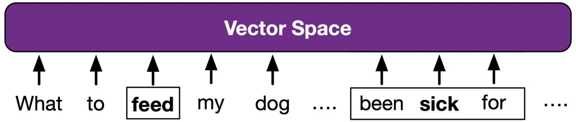
2. **Duration representation**

3. **Relation representation**

Model Architecture

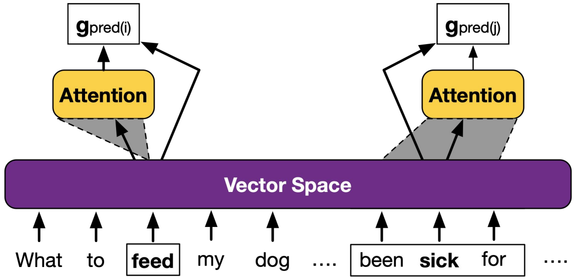
1. **Event representation**

What to feed my dog after gastroenteritis? My dog has been sick for about 3 days now.



Model Architecture

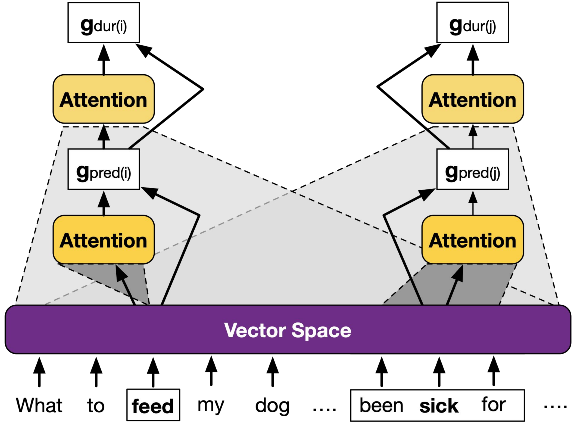
1. **Event representation**



What to feed my dog after gastroenteritis? My dog has been sick for about 3 days now.

Model Architecture

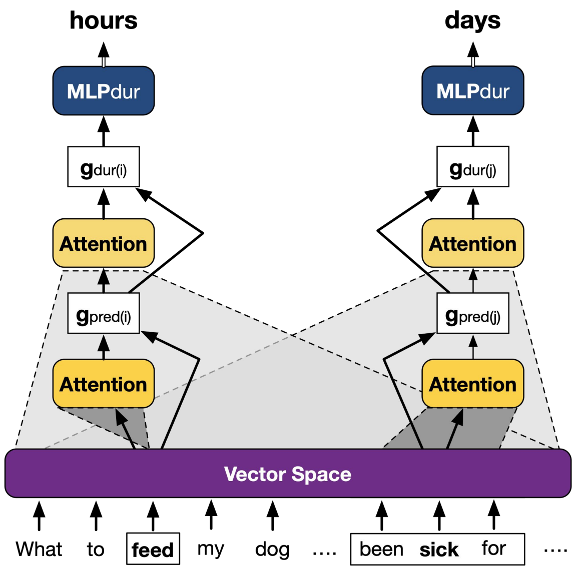
2. Duration representation



What to feed my dog after gastroenteritis? My dog has been sick for about 3 days now.

Model Architecture

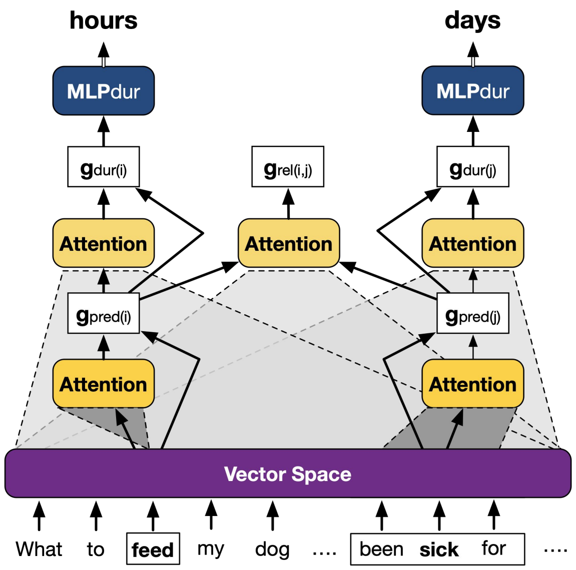
2. Duration representation



What to feed my dog after gastroenteritis? My dog has been sick for about 3 days now.

Model Architecture

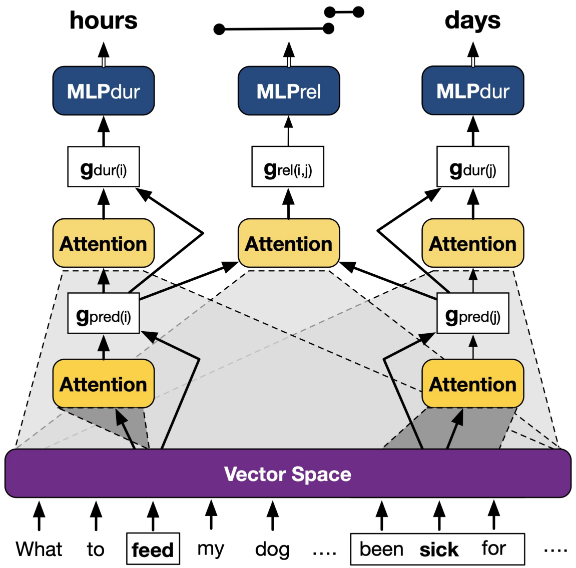
3. **Relation representation**



What to feed my dog after gastroenteritis? My dog has been sick for about 3 days now.

Model Architecture

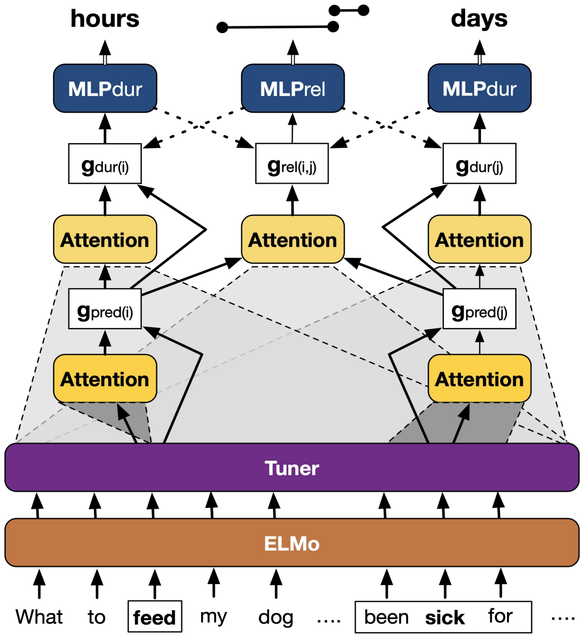
3. Relation representation



What to feed my dog after gastroenteritis? My dog has been sick for about 3 days now.

Model Architecture

Full Architecture



What to feed my dog after gastroenteritis? My dog has been sick for about 3 days now.

Results

Performance on UDS-Time (test set)

- We test 6 different variants of our model on the test set of UDS-Time

Performance on UDS-Time (test set)

- We test 6 different variants of our model on the test set of UDS-Time

Model			Duration			Relation		
Duration	Relation	Connection	ρ	rank diff.	R1	Absolute ρ	Relative ρ	R1
softmax	✓	-	32.63	1.86	8.59	77.91	68.00	2.82
binomial	✓	-	37.75	**1.75**	13.73	77.87	67.68	2.35
-	✓	Dur $\leftarrow$ Rel	22.65	3.08	-51.68	71.65	66.59	-6.09
binomial	-	Dur $\rightarrow$ Rel	36.52	1.76	13.17	77.58	66.36	0.85
binomial	✓	Dur $\rightarrow$ Rel	**38.38**	**1.75**	**13.85**	77.82	67.73	2.58
binomial	✓	Dur $\leftarrow$ Rel	38.12	1.75	13.68	**78.12**	**68.22**	**2.96**

Performance on UDS-Time (test set)

- We test 6 different variants of our model on the test set of UDS-Time

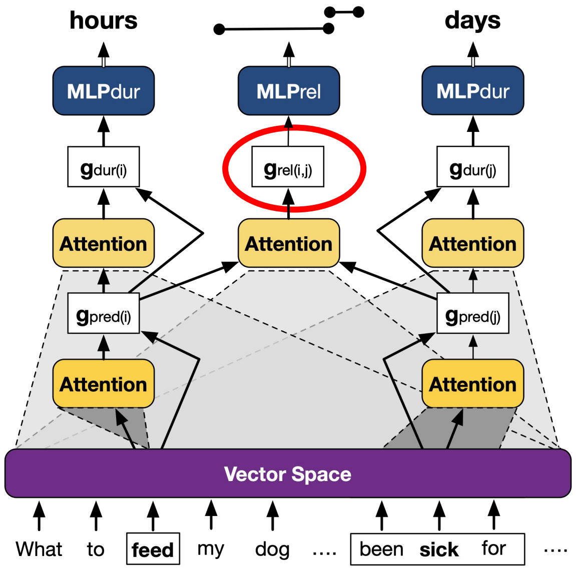
Model			Duration			Relation		
Duration	Relation	Connection	ρ	rank diff.	R1	Absolute ρ	Relative ρ	R1
softmax	✓	-	32.63	1.86	8.59	77.91	68.00	2.82
binomial	✓	-	37.75	**1.75**	13.73	77.87	67.68	2.35
-	✓	Dur $\leftarrow$ Rel	22.65	3.08	-51.68	71.65	66.59	-6.09
binomial	-	Dur $\rightarrow$ Rel	36.52	1.76	13.17	77.58	66.36	0.85
binomial	✓	Dur $\rightarrow$ Rel	**38.38**	**1.75**	**13.85**	77.82	67.73	2.58
binomial	✓	Dur $\leftarrow$ Rel	38.12	1.75	13.68	**78.12**	**68.22**	**2.96**

Performance on TimeBank-Dense

A transfer learning approach on TimeBank-Dense to predict **standard categorical temporal relations**

Performance on TimeBank-Dense

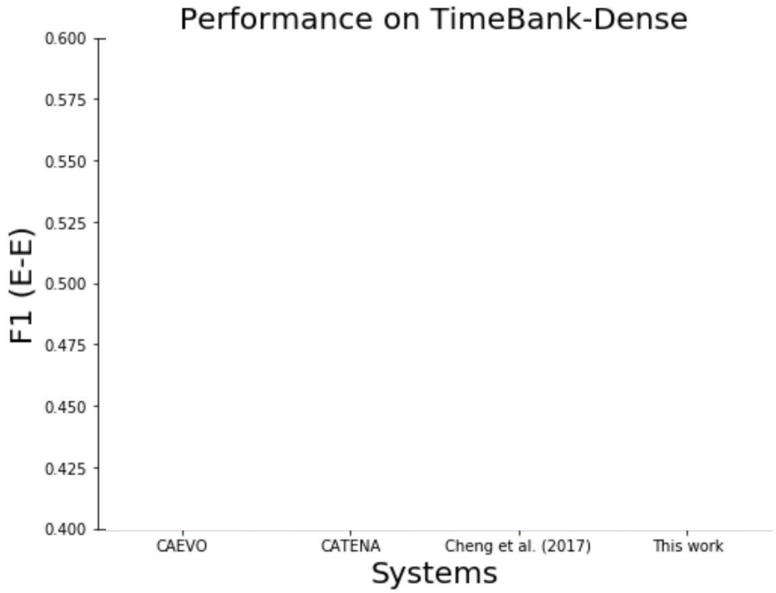
A transfer learning approach on TimeBank-Dense to predict **standard categorical temporal relations.**



Features

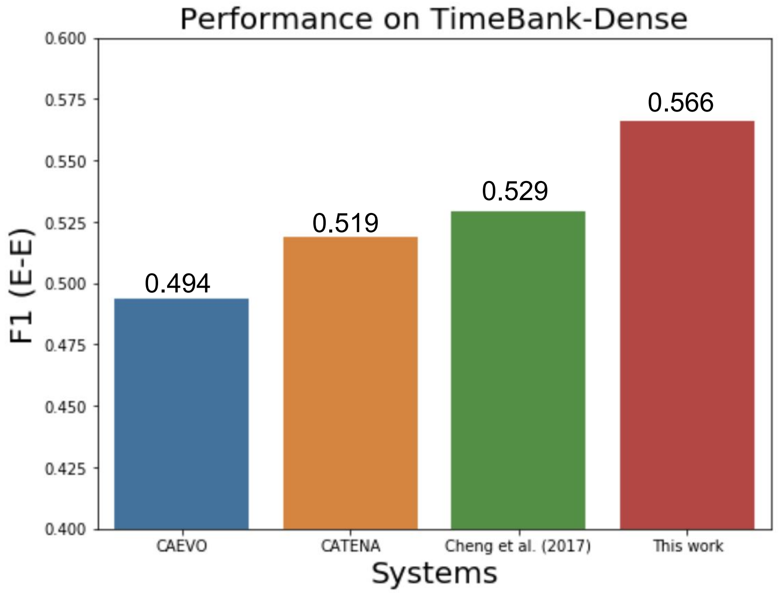
Performance on TimeBank-Dense

A transfer learning approach on TimeBank-Dense to predict **standard categorical temporal relations.**



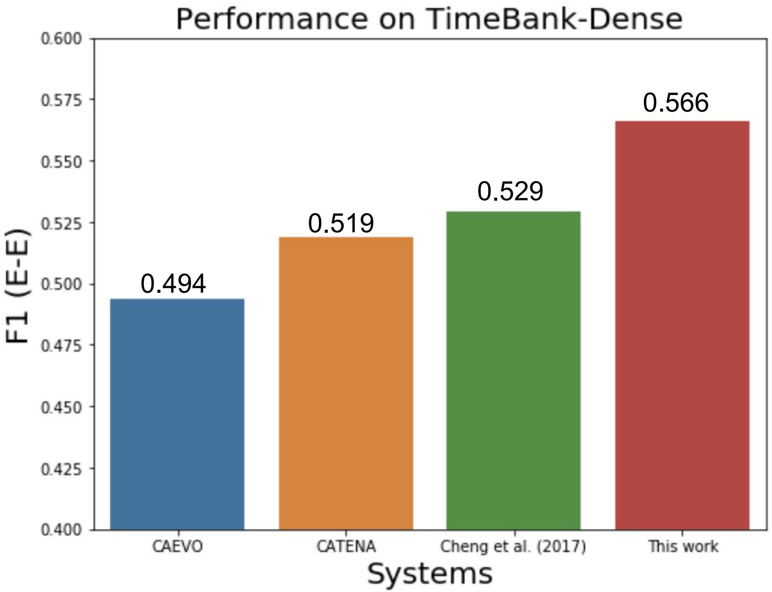
Performance on TimeBank-Dense

A transfer learning approach on TimeBank-Dense to predict **standard categorical temporal relations.**



Performance on TimeBank-Dense

A transfer learning approach on TimeBank-Dense to predict **standard categorical temporal relations.**



Our transfer learning approach beats most systems on TimeBank-Dense (**Event-Event Relations)**

Document Timelines

- A model to induce document timelines from the pairwise predictions

Document Timelines

- A model to induce document timelines from the pairwise predictions

- The Spearman correlation for timelines induced from our model and the timelines induced from the actual data:

    beginning point: 0.28

    duration: -0.097

Document Timelines

- A model to induce document timelines from the pairwise predictions

- The Spearman correlation for timelines induced from our model and the timelines induced from the actual data:

  beginning point: 0.28

  duration: -0.097

- The low correlation values suggest that even though the model is good at predicting pairwise predictions, it struggles to generate the entire document timeline

Model Analysis

Which words are attended to the most?

- We looked at the top 15 words in UDS-Time development set which have the highest mean duration-attention and relation-attention weights.

Which words are attended to the most? - Duration

- We looked at the top 15 words in UDS-Time development set which have the highest mean duration-attention and relation-attention weights.

Duration

Word	Attention (mean)	Rank (mean)	Freq
soldiers	0.911	1.28	69
months	0.844	1.38	264
Nothing	0.777	5.07	114
minutes	0.768	1.33	81
astronauts	0.756	1.37	81
hour	0.749	1.41	84
Palestinians	0.735	1.72	288
month	0.721	2.03	186
cartoonists	0.714	1.35	63
years	0.708	1.94	588
days	0.635	1.39	84
thoughts	0.592	2.90	60
us	0.557	2.09	483
week	0.531	2.23	558
advocates	0.517	2.30	105

Which words are attended to the most? - Duration

- We looked at the top 15 words in UDS-Time development set which have the highest mean duration-attention and relation-attention weights.

Duration

Word	Attention (mean)	Rank (mean)	Freq
soldiers	0.911	1.28	69
months	0.844	1.38	264
Nothing	0.777	5.07	114
minutes	0.768	1.33	81
astronauts	0.756	1.37	81
hour	0.749	1.41	84
Palestinians	0.735	1.72	288
month	0.721	2.03	186
cartoonists	0.714	1.35	63
years	0.708	1.94	588
days	0.635	1.39	84
thoughts	0.592	2.90	60
us	0.557	2.09	483
week	0.531	2.23	558
advocates	0.517	2.30	105

- Words that denote some **time period** (months, minutes, hour etc.) have the highest mean duration attention-weights.

Which words are attended to the most? - Relation

- We looked at the top 15 words in UDS-Time development set which have the highest mean duration-attention and relation-attention weights.

Relation

Word	Attention (mean)	Rank (mean)	Freq
occupied	0.685	1.33	54
massive	0.522	2.71	66
social	0.510	1.68	57
general	0.410	3.52	168
few	0.394	3.07	474
mathematical	0.393	7.66	132
are	0.387	3.47	4415
comes	0.339	2.39	51
or	0.326	3.50	3137
and	0.307	4.86	17615
emerge	0.305	2.67	54
filed	0.303	7.14	66
s	0.298	4.03	1152
were	0.282	3.49	1308
gets	0.239	7.36	228

Which words are attended to the most? - Relation

- We looked at the top 15 words in UDS-Time development set which have the highest mean duration-attention and relation-attention weights.

Relation

Word	Attention (mean)	Rank (mean)	Freq
occupied	0.685	1.33	54
massive	0.522	2.71	66
social	0.510	1.68	57
general	0.410	3.52	168
few	0.394	3.07	474
mathematical	0.393	7.66	132
are	0.387	3.47	4415
comes	0.339	2.39	51
or	0.326	3.50	3137
and	0.307	4.86	17615
emerge	0.305	2.67	54
filed	0.303	7.14	66
s	0.298	4.03	1152
were	0.282	3.49	1308
gets	0.239	7.36	228

- Words that are either **coordinators** (such as *or* and *and*), or bearers of **tense information** - i.e. lexical verbs and auxiliaries, have the highest mean relation attention weights

Which words are attended to the most? - Relation

- We looked at the top 15 words in UDS-Time development set which have the highest mean duration-attention and relation-attention weights.

Relation

Word	Attention (mean)	Rank (mean)	Freq
occupied	0.685	1.33	54
massive	0.522	2.71	66
social	0.510	1.68	57
general	0.410	3.52	168
few	0.394	3.07	474
mathematical	0.393	7.66	132
are	0.387	3.47	4415
comes	0.339	2.39	51
or	0.326	3.50	3137
and	0.307	4.86	17615
emerge	0.305	2.67	54
filed	0.303	7.14	66
s	0.298	4.03	1152
were	0.282	3.49	1308
gets	0.239	7.36	228

- Words that are either **coordinators** (such as *or* and *and*), or bearers of **tense information** - i.e. lexical verbs and auxiliaries, have the highest mean relation attention weights

Conclusion

Introduction

- Overarching question: How do humans extract chronology of events?

Introduction

- Overarching question: How do humans extract chronology of events?

Background

- A standard approach in previous corpora: Categorical temporal relations

Introduction

- Overarching question: How do humans extract chronology of events?

Background

- A standard approach in previous corpora: Categorical temporal relations
- Limitations: no duration information, hard to annotate, lacking fine-grained relation distinctions

Introduction

- Overarching question: How do humans extract chronology of events?

Background

- A standard approach in previous corpora: Categorical temporal relations
- Limitations: no duration information, hard to annotate, lacking fine-grained relation distinctions

Methodology: A new approach

Introduction

- Overarching question: How do humans extract chronology of events?

Background

- A standard approach in previous corpora: Categorical temporal relations
- Limitations: no duration information, hard to annotate, lacking fine-grained relation distinctions

Methodology: A new approach

- Mapping events to timelines represented in real number

Introduction

- Overarching question: How do humans extract chronology of events?

Background

- A standard approach in previous corpora: Categorical temporal relations
- Limitations: no duration information, hard to annotate, lacking fine-grained relation distinctions

Methodology: A new approach

- Mapping events to timelines represented in real number
- Explicitly annotating event durations

Introduction

- Overarching question: How do humans extract chronology of events?

Background

- A standard approach in previous corpora: Categorical temporal relations
- Limitations: no duration information, hard to annotate, lacking fine-grained relation distinctions

Methodology: A new approach

- Mapping events to timelines represented in real number
- Explicitly annotating event durations
- Construction of a new dataset: UDS-Time

Model

Model

- Vector representation of events, event-duration, fine-grained temporal relations

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

- High correlation (~77%) for start-points and end-points in pairwise event relations

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

- High correlation (~77%) for start-points and end-points in pairwise event relations
- Reasonable duration rank-difference of 1.75 by the best model

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

- High correlation (~77%) for start-points and end-points in pairwise event relations
- Reasonable duration rank-difference of 1.75 by the best model
- Competitive performance on TimeBank-Dense Event-Event Relations

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

- High correlation (~77%) for start-points and end-points in pairwise event relations
- Reasonable duration rank-difference of 1.75 by the best model
- Competitive performance on TimeBank-Dense Event-Event Relations
- Low correlation between induced document timelines from actual annotations and predicted values

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

- High correlation (~77%) for start-points and end-points in pairwise event relations
- Reasonable duration rank-difference of 1.75 by the best model
- Competitive performance on TimeBank-Dense Event-Event Relations
- Low correlation between induced document timelines from actual annotations and predicted values

Model Analysis

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

- High correlation (~77%) for start-points and end-points in pairwise event relations
- Reasonable duration rank-difference of 1.75 by the best model
- Competitive performance on TimeBank-Dense Event-Event Relations
- Low correlation between induced document timelines from actual annotations and predicted values

Model Analysis

- Most attended words for duration-attention are words which denote some time-span such as *month, minutes, year, week* etc.

Model

- Vector representation of events, event-duration, fine-grained temporal relations
- Neural Network architecture with linguistically motivated self-attention mechanism

Results

- High correlation (~77%) for start-points and end-points in pairwise event relations
- Reasonable duration rank-difference of 1.75 by the best model
- Competitive performance on TimeBank-Dense Event-Event Relations
- Low correlation between induced document timelines from actual annotations and predicted values

Model Analysis

- Most attended words for duration-attention are words which denote some time-span such as *month, minutes, year, week* etc.
- Most attended word for relation-attention are either coordinators (*or, and*) or words containing tense information (*present tense, past tense*)

THANK YOU!

Data and code available at:

http://decomp.io

References

- Marvin Minsky. 1975. A framework for representing knowledge. The Psychology of Computer Vision
- Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In Proceedings of the 4th International Joint Conference on Artificial Intelligence-Volume 1, pages 151–157. Morgan Kaufmann Publishers Inc.
- Leslie Lamport. 1978. Time, clocks, and the ordering of events in a distributed system. Communications of the ACM, 21(7):558–565.
- James F Allen and Patrick J Hayes. 1985. A commonsense theory of time. In Proceedings of the 9th International Joint Conference on Artificial Intelligence-Volume 1, pages 528–531. Morgan Kaufmann Publishers Inc.
- Chung Hee Hwang and Lenhart K Schubert. 1994. Interpreting ense, aspect and time adverbials: A compositional, unified approach. In Temporal Logic, pages 238–264. Springer.
- 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.
- William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. Transactions of the Association for Computational Linguistics, 2:143.
- 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.
- Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In Proceedings of the 4th International Workshop on Semantic Evaluations, pages 75–80. Association for Computational Linguistics
- Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 57– 62. Association for Computational Linguistics.
- 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), volume 2, pages 1–9.

References

- 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), volume 2, pages 501–506.
- 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.
- Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. 2016. Building a cross-document event-event relation corpus. In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with Association for Computational Linguistics 2016 (LAW-X 2016), pages 1–6.
- Antske Fokkens, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. 2013. GAF: A grounded annotation framework for events. In Workshop on Events: Definition, Detection, Coreference, and Representation, pages 11–20.
- Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 753–760. Association for Computational Linguistics.
- Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), volume 2, pages 10–14.
- Chen Lin, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2015. Multilayered temporal modeling for the clinical domain. Journal of the American Medical Informatics Association, 23(2):387–395.
- Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations with rich linguistic knowledge. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 918–927.
- 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.
- Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 64–75.

References

- Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. 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.
- Julien Tourille, Olivier Ferret, Aurelie Neveol, 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), volume 2, pages 224–230.
- Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional lstm over dependency paths. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 1–6
- Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1237–1246.
- Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, volume 2, pages 746–751.
- Feng Pan, Rutu Mulkar-Mehta, and Jerry R Hobbs. 2007. Modeling and learning vague event durations for temporal reasoning. In Proceedings of the 22nd National Conference on Artificial Intelligence. Volume 2, pages 1659–1662. AAAI Press.
- 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.
- 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: Short Papers-Volume 2, pages 223–227. Association for Computational Linguistics.
- 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), volume 2, pages 501–506.
- Sheng Zhang, Rachel Rudinger, and Benjamin Van Durme. 2017. An evaluation of predpatt and open ie via stage 1 semantic role labeling. In IWCS 201712th International Conference on Computational Semantics (Short papers).

References

- 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: Short Papers-Volume 2, pages 223–227. Association for Computational Linguistics.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60.
- 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 EMNLP.

Appendices

Appendix A

Pivot-Predicate

- Adjacent sentences in a document were concatenated together to be able to capture inter-sentential temporal relations.

- Considering all possible event-pairs is infeasible. Hence, we design the following heuristic to select the pivot predicate from a sentence:

  We find the root-predicate of the sentence and if it governs a CCOMP, CSUBJ, or XCOMP, we follow that dependency to the next predicate until we find a predicate that doesn't govern a CCOMP, CSUBJ, or XCOMP.

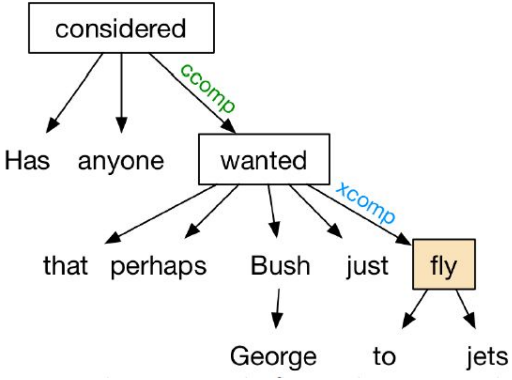
Appendix A

Pivot-Predicate

- Adjacent sentences in a document were concatenated together to be able to capture inter-sentential temporal relations.

- Considering all possible event-pairs is infeasible. Hence, we design the following heuristic to select the pivot predicate from a sentence:

  We find the root-predicate of the sentence and if it governs a CCOMP, CSUBJ, or XCOMP, we follow that dependency to the next predicate until we find a predicate that doesn't govern a CCOMP, CSUBJ, or XCOMP.



Sentence:
"Has anyone considered that perhaps George Bush just wanted to fly jets?"

Fig3: An example of our heuristic to find the pivot predicate

Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

- Time completion (< 60 seconds)

Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

- Time completion (< 60 seconds)
- Same slider positions in all annotations

Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

- Time completion (< 60 seconds)
- Same slider positions in all annotations
- Same duration values in all annotations

Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

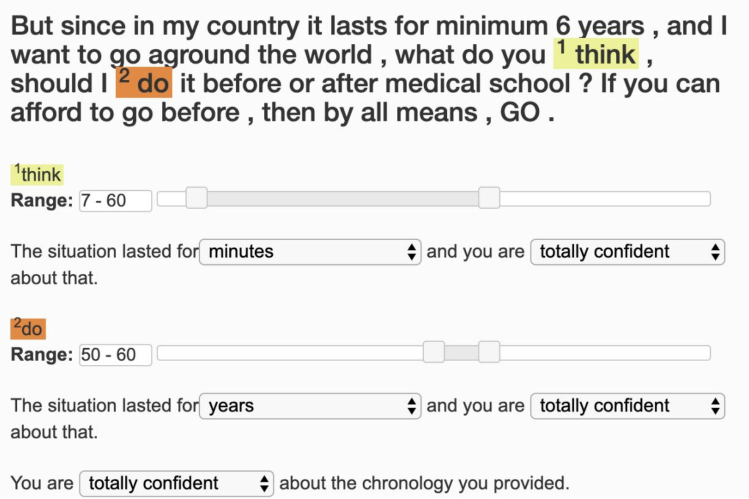
- Time completion (< 60 seconds)
- Same slider positions in all annotations
- Same duration values in all annotations
- Inconsistency between slider and duration values

Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

- Time completion (< 60 seconds)
- Same slider positions in all annotations
- Same duration values in all annotations
- Inconsistency between slider and duration values

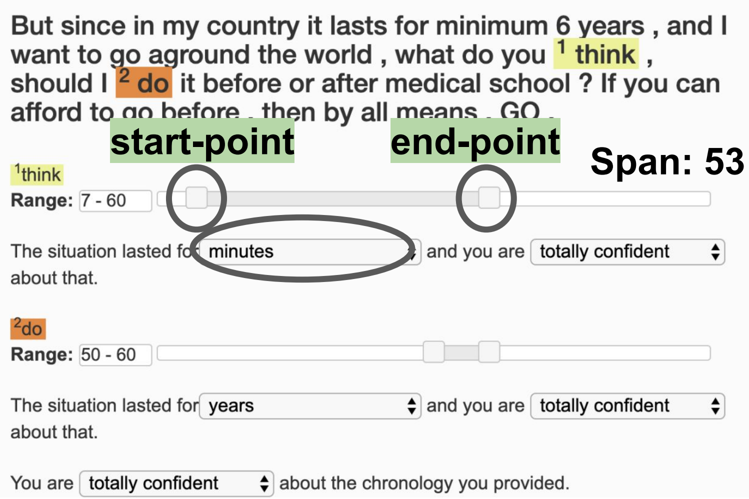


Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

- Time completion (< 60 seconds)
- Same slider positions in all annotations
- Same duration values in all annotations
- Inconsistency between slider and duration values

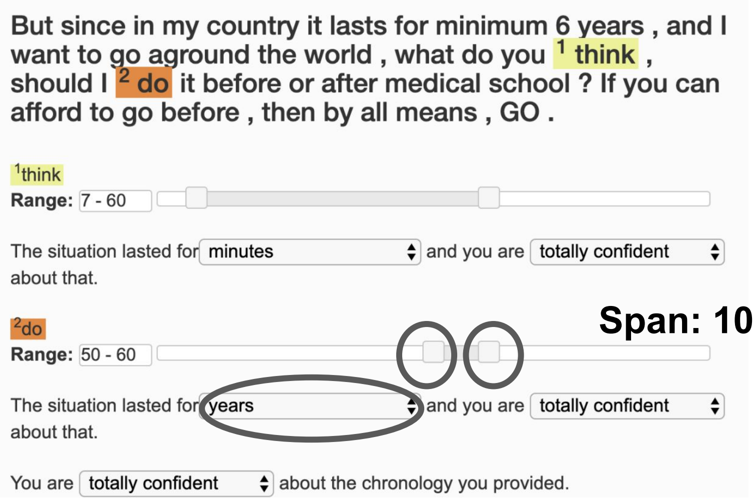


Appendix B

Rejecting Annotations

Multiple checks to detect potentially bad annotations:

- Time completion (< 60 seconds)
- Same slider positions in all annotations
- Same duration values in all annotations
- Inconsistency between slider and duration values



Appendix C

Inter-annotator Agreement

- 765 annotators from Amazon Mechanical Turk
- **Train set**: 1 annotation per predicate-pair
- **Dev, and Test set**: 3 annotations per predicate-pair

Appendix C

Inter-annotator Agreement

- 765 annotators from Amazon Mechanical Turk
- **Train set**: 1 annotation per predicate-pair
- **Dev, and Test set**: 3 annotations per predicate-pair

Relations:
Average Spearman Rank correlation between slider positions: **0.665** (95% CI=[0.661, 0.669])

Appendix C

Inter-annotator Agreement

- 765 annotators from Amazon Mechanical Turk
- **Train set**: 1 annotation per predicate-pair
- **Dev, and Test set**: 3 annotations per predicate-pair

Relations:
Average Spearman Rank correlation between slider positions: **0.665** (95% CI=[0.661, 0.669])

Durations:
Average Absolute difference in Duration rank: **2.24** scale points (95% CI=[2.21, 2.25])

- Heavy positive skew (γ_1 = 1.16, 95% CI=[1.15, 1.18])
- Modal rank difference is 1 (25.3% of the response pairs), with rank difference 0 as the next most likely (24.6%) and rank difference 2 as a distant third (15.4%).

Appendix D

Normalization

- Annotated Slider positions are normalized
- Absolute slider positions are meaningless
- Relative chronology preserved

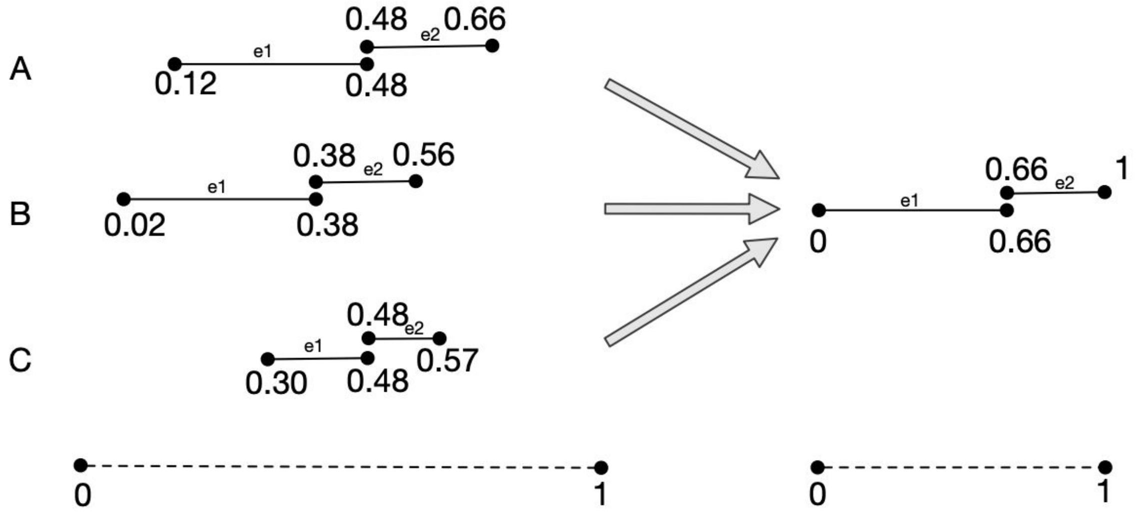


Fig: Normalization of slider values (a toy example with three annotators -- A, B, and C)

Appendix F

Further Analysis on Relations

- We rotate the predicted slider positions in the relation space as shown in Data Distribution and compare it with the rotated space of actual slider positions

- We obtain Spearman correlations of :
  0.19 for PRIORITY,
  0.23 for CONTAINMENT, and
  0.17 for EQUALITY