Learning Fine-Grained Knowledge about Contingent Relations between Everyday Events

Elahe Rahimtoroghi, Ernesto Hernandez and Marilyn A Walker

Natural Language and Dialogue Systems Lab
Department of Computer Science
University of California Santa Cruz
Santa Cruz, CA 95064, USA
Introduction

Goal

- Capture common-sense knowledge about the fine-grained events of everyday experience
  - opening a fridge enabling preparing food
  - getting out of bed being triggered by an alarm going off

Contingency relation between events (Cause and Condition)
Much of the user-generated content on social media is provided by ordinary people telling stories about their daily lives.

Camping Trip
We packed all our things on the night before Thu (24 Jul) except for frozen food. We brought a lot of things along. We woke up early on Thu and JS started packing the frozen marininated food inside the small cooler... In the end, we decided the best place to set up the tent was the squarish ground that’s located on the right. Prior to setting up our tent, we placed a tarp on the ground. In this way, the underneath of the tent would be kept clean. After that, we set the tent up.

Storm
I don’t know if I would’ve been as calm as I was without the radio, as the hurricane made landfall in Galveston at 2:10AM on Saturday. As the wind blew, branches thudded on the roof or trees snapped, it was helpful to pinpoint the place... A tree fell on the garage roof, but it’s minor damage compared to what could’ve happened. We then started cleaning up, despite Sugar Land implementing a curfew until 2pm; I didn’t see any policemen enforcing this. Luckily my dad has a gas saw (as opposed to electric), so we helped cut up three of our neighbors’ trees. I did a lot of raking, and there’s so much debris in the garbage.

- Rich with common-sense knowledge about contingent relations between events
  - placing a tarp, setting up a tent
  - the hurricane made landfall, the wind blew, a tree fell
  - started cleaning up, cut up the trees, raking

This fine-grained knowledge is simply not found in previous work on narrative event collections.
Much of the previous work is not focused on a particular relation between events (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Manshadi et al., 2008; Nguyen et al., 2015; Balasubramanian et al., 2013; Pichotta and Mooney, 2014)

Main focus is on newswire

Evaluation criteria: narrative cloze test

Contingency

Personal stories

New evaluation method as well as previous work
Challenge: Personal stories provide both advantages and disadvantages

- Told in chronological order
- Temporal order between events is a strong cue to contingency
- Their structure is more similar to oral narrative (Labov and Waletzky, 1967; Labov, 1997) than to newswire
- Only about a third of the sentences in a personal narrative describe actions (Rahimtoroghi et al., 2014; Swanson et al., 2014)
- Novel methods are needed to find useful relationships between events

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
<th>Story Clause</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Orientation</td>
<td>Now, on with this week’s story...</td>
</tr>
<tr>
<td>2</td>
<td>Orientation</td>
<td>The last month has been hectic.</td>
</tr>
<tr>
<td>3</td>
<td>Orientation</td>
<td>Turbo charged.</td>
</tr>
<tr>
<td>4</td>
<td>Orientation</td>
<td>Lot’s of work because I was learning from Tim, my partner in crime.</td>
</tr>
<tr>
<td>5</td>
<td>Orientation</td>
<td>This hasn’t been helped by the intense pressure in town due to the political transition coming to an end.</td>
</tr>
<tr>
<td>6</td>
<td>Orientation</td>
<td>This week things started alright and on schedule.</td>
</tr>
<tr>
<td>7</td>
<td>Action</td>
<td>But I managed to get myself arrested by the traffic police (roulage) early last Wednesday.</td>
</tr>
<tr>
<td>8</td>
<td>Action</td>
<td>After yelling excessively at their outright corrupted methods</td>
</tr>
<tr>
<td>9</td>
<td>Action</td>
<td>and asking incessantly for what law I actually broke,</td>
</tr>
<tr>
<td>10</td>
<td>Action</td>
<td>they managed to bring me in at the police HQ.</td>
</tr>
<tr>
<td>11</td>
<td>Action</td>
<td>I was drawing too much of a curious crowd for the authorities.</td>
</tr>
<tr>
<td>12</td>
<td>Action</td>
<td>In about half an hour at police HQ I had charmed every one around.</td>
</tr>
<tr>
<td>13</td>
<td>Action</td>
<td>I had prepared my “gift” as they wished.</td>
</tr>
<tr>
<td>14</td>
<td>Evaluation</td>
<td>Decision withheld, they decided that I neednt to bother,</td>
</tr>
<tr>
<td>15</td>
<td>Evaluation</td>
<td>they liked me too much.</td>
</tr>
<tr>
<td>16</td>
<td>Evaluation</td>
<td>I should go free.</td>
</tr>
</tbody>
</table>
Event Representation and Extraction

Event: Verb Lemma (subj:Subject Lemma, dobj:Direct Object Lemma, prt:Particle)

<table>
<thead>
<tr>
<th>#</th>
<th>Sentence → Event Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>but it wasn’t at all frustrating putting up the tent and setting up the first night → put (dobj:tent, prt:up)</td>
</tr>
<tr>
<td>2</td>
<td>The next day we had oatmeal for breakfast → have (subj:PERSON, dobj:oatmeal)</td>
</tr>
<tr>
<td>3</td>
<td>by the time we reached the Lost River Valley Campground, it was already past 1 pm → reach (subj:PERSON, dobj:LOCATION)</td>
</tr>
<tr>
<td>4</td>
<td>then JS set up a shelter above the picnic table → set (subj:PERSON, dobj:shelter, prt:up)</td>
</tr>
<tr>
<td>5</td>
<td>once the rain stopped, we built a campfire using the firewoods → build (subj:PERSON, dobj:campfire)</td>
</tr>
</tbody>
</table>

- Multi-argument representation is richer, capable of capturing interactions between multiple events (Pichotta and Mooney, 2014)
- Event extraction
  - Stanford dependency parser
  - Stanford NER

Natural Language and Dialogue Systems UC Santa Cruz
Contributions

- Generate topic-sorted personal stories using bootstrapping

- Direct comparison of topic-specific data vs. general-domain stories
  - Learn more fine-grained and richer knowledge from topic-specific corpus
  - Even with less amount of data

- Two sets of experiments
  - Directly compare to previous work
  - Introduce new evaluation methods
Semi-Supervised Algorithm for Generating Topic-Specific Dataset

Corpus

Labeled data

AutoSlog-TS

870 more Camping Trip stories
971 more Storm stories

Event-patterns

NP-Prep-(NP): CAMPING-IN
(subj)-ActVB-Dobj: WENT-CAMPING

small set (~200-300) of stories on the topic

Camping: 299
Storm: 361
Causal Potential (Beamer and Girju 2009)

- Unsupervised distributional measure
- Tendency of an event pair to encode a causal relation
- Probability of occurring in a causal context

\[ CP(e_1, e_2) = \log \frac{P(e_2|e_1)}{P(e_2)} + \log \frac{P(e_1 \rightarrow e_2)}{P(e_2 \rightarrow e_1)} \]

- Calculate CP for every pair of adjacent events
  - Skip-2 bigram model
  - Two related events may often be separated by a non-event sentences
Evaluations

- Narrative cloze test
  - Sequence of narrative events in a document from which one event has been removed
  - Predict the missing event

- Unigram model results nearly as good as other complicated models (Pichotta and Mooney, 2014)
Automatic Two-Choice Test

- Automatically generated set of two-choice questions with the answers
  - Modeled after the COPA task (An Evaluation of Commonsense Causal Reasoning, Roemmele et al., 2011)
  - From held-out test sets for each dataset
- Each question consists of one event and two choices
  
  **Question event:** arrange (dobj:outdoor)
  
  Choice 1: help (dobj:trip)
  Choice 2: call (subj:PERSON)

- Predict which of the two choices is more likely to have a contingency relation with the event in the question
Comparison to Previous Work: Rel-gram Tuples (Balasubramanian et al., 2013)

- **Rel-grams**: Generate pairs of relational tuples of events
  - Use co-occurrence statistics based on Symmetric Conditional Probability
  - Publicly available through an online search interface
  - Outperform the previous work

\[
SCP(e_1, e_2) = P(e_2|e_1) \times P(e_1|e_2)
\]

- Two experiments:
  - Content of the learned event knowledge
  - Method: one of the baselines on our data
Baselines

- **Event-Unigram**
  - Produce a distribution of normalized frequencies for events

- **Event-Bigram**
  - Bigram probability of every pair of adjacent events using skip-2 bigram model

- **Event-SCP**
  - Symmetric Conditional Probability between event tuples (Balasubramanian et al., 2013)
Datasets

- General-domain dataset
  - Train (4,000 stories)
  - Held-out test (200 stories)

- Topic-specific dataset

<table>
<thead>
<tr>
<th>Topic</th>
<th>Dataset</th>
<th># Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camping Trip</td>
<td>Hand-labeled held-out test</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>Hand-labeled train (Train-HL)</td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>Train-HL + Bootstrap (Train-HL-BS)</td>
<td>1,062</td>
</tr>
<tr>
<td>Storm</td>
<td>Hand-labeled held-out test</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Hand-labeled train (Train-HL)</td>
<td>263</td>
</tr>
<tr>
<td></td>
<td>Train-HL + Bootstrap (Train-HL-BS)</td>
<td>1,234</td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Model</th>
<th>Train Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camping Trip</td>
<td>Event-Unigram</td>
<td>Train-HL-BS</td>
<td>0.507</td>
</tr>
<tr>
<td></td>
<td>Event-Bigram</td>
<td>Train-HL-BS</td>
<td>0.510</td>
</tr>
<tr>
<td></td>
<td>Event-SCP</td>
<td>Train-HL-BS</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td>Causal Potential</td>
<td>Train-HL</td>
<td>0.631</td>
</tr>
<tr>
<td></td>
<td>Causal Potential</td>
<td>Train-HL-BS</td>
<td>0.685</td>
</tr>
<tr>
<td>Storm</td>
<td>Event-Unigram</td>
<td>Train-HL-BS</td>
<td>0.510</td>
</tr>
<tr>
<td></td>
<td>Event-Bigram</td>
<td>Train-HL-BS</td>
<td>0.523</td>
</tr>
<tr>
<td></td>
<td>Event-SCP</td>
<td>Train-HL-BS</td>
<td>0.516</td>
</tr>
<tr>
<td></td>
<td>Causal Potential</td>
<td>Train-HL</td>
<td>0.711</td>
</tr>
<tr>
<td></td>
<td>Causal Potential</td>
<td>Train-HL-BS</td>
<td>0.887</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event-Unigram</td>
<td>0.478</td>
</tr>
<tr>
<td>Event-Bigram</td>
<td>0.481</td>
</tr>
<tr>
<td>Event-SCP (Rel-gram)</td>
<td>0.477</td>
</tr>
<tr>
<td>Causal Potential</td>
<td>0.510</td>
</tr>
</tbody>
</table>

General-Domain Stories

- CP results stronger than all the baselines
- Results on topic-specific dataset is significantly stronger than general-domain narratives
- More training data collected by bootstrapping improves the accuracy
Compare Camping Trip Event Pairs against the Rel-gram tuples

- Find tuples relevant to Camping Trip
  - Used our top 10 indicative event-patterns, generated and ranked during the bootstrapping
  - Apply filtering and ranking
  - Evaluate top N = 100

<table>
<thead>
<tr>
<th>[person]</th>
<th>go to</th>
<th>camp</th>
<th>[&lt;&lt;]</th>
<th>[person]</th>
<th>work with</th>
<th>[person]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[person]</td>
<td>go to</td>
<td>camp</td>
<td>[&lt;&lt;]</td>
<td>[person]</td>
<td>go with</td>
<td>[organization]</td>
</tr>
<tr>
<td>[person]</td>
<td>go to</td>
<td>camp</td>
<td>[&lt;&lt;]</td>
<td>[person]</td>
<td>be director of</td>
<td>[organization]</td>
</tr>
<tr>
<td>[person]</td>
<td>go to</td>
<td>camp</td>
<td>[&lt;&lt;]</td>
<td>[person]</td>
<td>lose</td>
<td>[person]</td>
</tr>
<tr>
<td>X:[person]</td>
<td>go to</td>
<td>camp</td>
<td>[&lt;&lt;]</td>
<td>X:[person]</td>
<td>go with</td>
<td>[person]</td>
</tr>
<tr>
<td>X:he</td>
<td>go to</td>
<td>camp</td>
<td>[&lt;&lt;]</td>
<td>X:he</td>
<td>go in</td>
<td>[time_unit]</td>
</tr>
<tr>
<td>[person]</td>
<td>go to</td>
<td>camp</td>
<td>[&lt;&lt;]</td>
<td>[person]</td>
<td>leave</td>
<td>[person]</td>
</tr>
</tbody>
</table>
Evaluation on Mechanical Turk

- New method for evaluating topic-specific contingent event pairs
- Rate each pair
  0: The events are not contingent
  1: The events are contingent but not relevant to the specified topic
  2: The events are contingent and somewhat relevant to the specified topic
  3: The events are contingent and strongly relevant to the specified topic

- More readable representation for annotators:

  Subject - Verb Particle - Direct Object
  pack (subj:PERSON, dobj:car, prt: up) → person - pack up - car
## Rel-gram Evaluation Results

<table>
<thead>
<tr>
<th>Label</th>
<th>Rel-gram Tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contingent &amp; Strongly Relevant</td>
<td>7 %</td>
</tr>
<tr>
<td>Contingent &amp; Somewhat Relevant</td>
<td>0 %</td>
</tr>
<tr>
<td>Contingent &amp; Not Relevant</td>
<td>35 %</td>
</tr>
<tr>
<td>Total Contingent</td>
<td>42 %</td>
</tr>
</tbody>
</table>

Label >2: Contingent & strongly topic-relevant
Label = 2: Contingent & somewhat topic-relevant
1 ≤ Label < 2: Contingent & not topic-relevant
Label < 1: Not contingent
Two filtering methods
- Selected the frequent pairs for each topic and removed the ones that occur less than 5 times
- Used the indicative event-patterns for each topic and extracted the pairs that at least included one of these patterns

Rank by Causal Potential scores to identify the highly contingent ones
- Evaluated the top $N = 100$ pairs on Mechanical Turk task
We identify contingent event pairs that are highly indicative of a particular topic. We hypothesize that these event pairs serve as building blocks of coherent knowledge about events. Fig. 2 shows some examples of event pairs with high CP scores extracted from General-Domain stories.

To ensure that the Amazon Mechanical Turk annotations are reliable, we designed a scale of 0-3 as follows:

- 0: The events are not contingent.
- 1: The events are contingent and somewhat relevant to the specified topic.
- 2: The events are contingent and strongly relevant to the specified topic.
- 3: Both events strongly correspond to the specified topic.

The evaluations and results are presented in Table 6 and Table 7 respectively.

<table>
<thead>
<tr>
<th>Label</th>
<th>Camping</th>
<th>Storm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contingent &amp; Strongly Relevant</td>
<td>44 %</td>
<td>33 %</td>
</tr>
<tr>
<td>Contingent &amp; Somewhat Relevant</td>
<td>8 %</td>
<td>20 %</td>
</tr>
<tr>
<td>Contingent &amp; Not Relevant</td>
<td>30 %</td>
<td>24 %</td>
</tr>
<tr>
<td>Total Contingent</td>
<td>82 %</td>
<td>77 %</td>
</tr>
</tbody>
</table>

- Inter-annotator reliability
  - average kappa = 0.73 (substantial agreement)
Examples of Event Pairs

**Topic-Specific Dataset**

- climb → person - find - rock
- person - pack up - car → head out
- wind - blow - transformer → power - go out
- tree - fall - eave → crush
- hit - location → evacuate - person

**General-Domain Dataset**

- person - go → go down - trail
- person - find - fellow → go back
- person - see - gun → see - police
- person - go → person - walk down
Conclusions

- Learned new type of knowledge
  - Common-sense knowledge about everyday events focused on contingency relation
- Data collection
  - Semi-supervised bootstrapping approach create topic-sorted dataset
- New evaluation methods
  - Two-choice test and Mechanical Turk task
- Results
  - On topic-specific dataset is significantly stronger than general-domain
  - Method used on the news genre do not work as well on personal stories
  - Fine-grained relations we learn are not found in existing event collections
Thank you!