

Inference of Fine-Grained Event Causality from Blogs and Films

Zhichao Hu, Elahe Rahimtoroghi and Marilyn A Walker

Natural Language and Dialogue Systems Lab

Department of Computer Science, University of California Santa Cruz

Santa Cruz, CA 95064, USA

zhu@soe.ucsc.edu, elahe@soe.ucsc.edu, mawalker@ucsc.edu

Abstract

Human understanding of narrative is mainly driven by reasoning about causal relations between events and thus recognizing them is a key capability for computational models of language understanding. Computational work in this area has approached this via two different routes: by focusing on acquiring a knowledge base of common causal relations between events, or by attempting to understand a particular story or macro-event, along with its storyline. In this position paper, we focus on knowledge acquisition approach and claim that newswire is a relatively poor source for learning fine-grained causal relations between everyday events. We describe experiments using an unsupervised method to learn causal relations between events in the narrative genres of first-person narratives and film scene descriptions. We show that our method learns fine-grained causal relations, judged by humans as likely to be causal over 80% of the time. We also demonstrate that the learned event pairs do not exist in publicly available event-pair datasets extracted from newswire.

1 Introduction

Computational models of language understanding must recognize narrative structure because many types of natural language texts are narratively structured, e.g. news, reviews, film scripts, conversations, and personal blogs (Polanyi, 1989; Jurafsky et al., 2014; Bell, 2005; Gordon et al., 2011a). Human understanding of narrative is driven by reasoning about causal relations between the events and states in the story (Ger-

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 marinatinated 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**.

Figure 1: Part of a blog story about camping

rig, 1993; Graesser et al., 1994; Lehnert, 1981; Goyal et al., 2010). Thus previous work has aimed to learn a knowledge base of semantic relations between events from text (Chklovski and Pantel, 2004; Gordon et al., 2011a; Chambers and Jurafsky, 2008; Balasubramanian et al., 2013; Pichotta and Mooney, 2014; Do et al., 2011), with the long-term aim of using this knowledge for understanding. Some of this work explicitly models causality; other work characterizes the semantic relations more loosely as “events that tend to co-occur”. Related work points out that causality is granular in nature, and that humans flexibly move back and forth between different levels of granularity of causal knowledge (Hobbs, 1985). Thus methods are needed to learn causal relations and reason about them at different levels of granularity (Mulkar-Mehta et al., 2011).

One limitation of prior work is that it has primarily focused on newswire, thus have only learned relations about newsworthy topics, and likely the most frequent, highly common (coarse-grained) news events. But news articles are not the only resource for learning about relations between events. Much of the content on social media in personal blogs is written by ordinary people about their daily lives (Burton et al., 2009), and these blogs contain a large variety of everyday events (Gordon et al., 2012). Film scene descriptions

are also action-rich and told in fine-grained detail (Beamer and Girju, 2009; Hu et al., 2013). Moreover, both of these genres typically report events in temporal order, which is a primary cue to causality. In this position paper, we claim that knowledge about fine-grained causal relations between everyday events is often not available in news, and can be better learned from other narrative genres.

For example, Figure 1 shows a part of a personal narrative written in a blog about a camping trip (Burton et al., 2009). The major event in this story is *camping*, which is contingent upon several finer-grained events, such as *packing things the night before*, *waking up in the morning*, *packing frozen food*, and later on at the campground, *placing a tarp* and *setting up the tent*. Similarly film scene descriptions, such as the one shown in Figure 2, typically contain fine-grained causality. In this scene from Lord of the Rings, *grabbing* leads to *spilling*, and *pushing* leads to *stumbling* and *falling*.

We show that unsupervised methods for modeling causality can learn fine-grained event relations from personal narratives and film scenes, even when the corpus is relatively small compared to those that have been used for newswire. We learn high-quality causal relations, with over 80% judged as causal by humans. We claim that these fine-grained causal relations are much closer in spirit to those motivating earlier work on scripts (Lehnert, 1981; Schank et al., 1977; Wilensky, 1982; de Jong, 1979), and we show that the causal knowledge we learn is not found in causal knowledge bases learned from news.

Section 2 first summarizes previous work on learning causal knowledge. We then present our experiments and results on modeling event causality in blogs and film scenes in Section 3. Conclusions and future directions are discussed in Section 4.

2 Background and Related Work

Cognitive theories of narrative understanding define narrative coherence in terms of four different sources of causal inferences between events A and B (Trabasso and van den Broek, 1985; Warren et al., 1979; Trabasso et al., 1989; Van den Broek, 1990). (1) Physical: A physically causes event B. (2) Motivational: A happens with B as a motivation. (3) Psychological: A brings about emotions expressed by event B. (4) Enabling: A creates a

Pippin, sitting at the bar, chatting with Locals. Frodo leaps to his feet and pushes his way towards the bar. Frodo **grabs** Pippin’s sleeve, **spilling** his beer. Pippin **pushes** Frodo away...he **stumbles** backwards, and **falls** to the floor.

Figure 2: Film Scene from Lord of the Rings, Fantasy Genre

state or condition for B to happen.

There has been a great deal of interest in learning narrative relations or narrative schema in an unsupervised or weakly supervised manner from text. Here we focus on work where the resulting knowledge bases have been made publicly available, allowing us to compare the learned knowledge directly.

The VerbOcean project learned five different semantic relations between event types (verbs) from newswire, with the HAPPENS-BEFORE relation defined as “indicating that the two verbs refer to two temporally disjoint intervals or instances”. WordNet’s cause relation, between a causative and a resultative verb (as in buy::own) is tagged as an instance of HAPPENS-BEFORE in VerbOcean, consistent with the heuristic that temporal ordering is a major component of causality. Other examples of the HAPPENS-BEFORE relation in the VerbOcean knowledge base include marry::divorce, detain::prosecute, enroll::graduate, schedule::reschedule, and tie::untie (Chklovski and Pantel, 2004).

Balasubramanian et al. (2013) generate pairs of event relational tuples, called *Rel-grams*. The Rel-grams are publicly available through an online search interface¹. Rel-gram tuples are extracted using a co-occurrence statistical metric, Symmetric Conditional Probability (SCP), which combines Bigram probability in both directions as follows:

$$SCP(e_1, e_2) = P(e_2|e_1) \times P(e_1|e_2) \quad (1)$$

Their evaluation experiments directly compared the knowledge learned in Rel-grams to the previous work on narrative schemas (Chambers and Jurafsky, 2008, 2009), showing that they achieve better results, thus our work compares directly to the tuples available in Rel-grams.

Other work focuses more directly on learning **causal** or **contingency** relations between events.

¹<http://relgrams.cs.washington.edu:10000/relgrams>

Beamer and Girju (2009) introduced a distributional measure called *Causal Potential* to assess the likelihood of a causal relation holding between two events. This measure is based on Suppes’ probabilistic theory of causality (Suppes, 1970).

$$CP(e_1, e_2) = PMI(e_1, e_2) + \log \frac{P(e_1 \rightarrow e_2)}{P(e_2 \rightarrow e_1)} \quad (2)$$

where $PMI(e_1, e_2) = \log \frac{P(e_1, e_2)}{P(e_1)P(e_2)}$

where the arrow notation means ordered event pairs, i.e. event e_1 occurs before event e_2 . CP consists of two terms: the first is pair-wise mutual information (PMI) and the second is relative ordering of bigrams. PMI measures how often events occur as a pair (without considering their order); whereas relative ordering accounts for the order of the event pairs because temporal order is one of the strongest cues to causality (Beamer and Girju, 2009; Riaz and Girju, 2010, 2013). This work explicitly links their definitions to research using the Penn Discourse Treebank (PDTB) definition of CONTINGENCY.

Beamer and Girju (2009) applied the CP measure to 173 film scripts, resulting in a high correlation between human-judged causality and the CP measure. Their paper provides a list of 90 verb pairs, selected from the high, middle and low CP ranges in their learned causal pairs. We compare their 30 highest CP events with causal event pairs that we learn from film.

Riaz and Girju (2010) apply a similar measure to topic-sorted news stories about Hurricane Katrina and the Iraq War and present ranked causality relations between events for these topics, suggesting that topic-sorted corpora can produce better causal knowledge. Other work has also used CP to measure the contingency relation between two events, reporting better results than achieved with PMI or bigrams alone (Hu et al., 2013; Rahimtoroghi et al., 2016).

3 Methods and Evaluations

Our primary goal is simply to show that fine-grained causal relations can be learned from film scripts and blogs, and that these are not found in causal knowledge bases learned from newswire. In this section we describe our datasets and methods, and the present two evaluations. First, we evaluate whether the relations learned are causal

| Corpus | Number | Word Count |
|---------|--------|------------|
| Drama | 579 | 6,680,749 |
| Fantasy | 113 | 1,186,587 |
| Mystery | 107 | 1,346,496 |
| Camping | 1,062 | 2,207,458 |

Table 1: Number of documents and word count for each dataset

using human judgment HITs on Amazon Mechanical Turk. Second, we directly compare to event pair collections from other publicly available sources learned from news genre.

3.1 Datasets

Topical coherence and similarity of events within the corpus used for learning event relations can be as important as the size of the corpus (Riaz and Girju, 2010; Rahimtoroghi et al., 2016). We use two datasets for learning causal event pairs: first-person narratives from blogs (Burton et al., 2009; Rahimtoroghi et al., 2016), and film scene descriptions (excluding dialogs because dialogs are not as action-rich) (Walker et al., 2012; Hu et al., 2013). Our experiment on blogs learns causal relations from a topic-sorted corpus of ~ 1000 camping stories. We also posit that the genre of a film may select for similar types of events. However genres can be defined broadly or narrowly, e.g. the Drama genre overlaps with many other genres. We thus compare two narrow film genres of Fantasy and Mystery with the Drama genre from an existing corpus (Walker et al., 2012; Hu et al., 2013). The raw numbers for each subcorpus are shown in Table 1. Note that Camping corpus consists of blog posts which are much shorter compared to movie scripts. Thus their word count is much smaller compared to films corpus despite the larger number of documents.

3.2 Methods

In the blogs, related event pairs are more frequently separated by utterances that provide state descriptions or affective reactions to events (Swanson et al., 2014). As a result, we use Causal Potential (CP) measure to assess the causal relation between events and apply skip-2 bigram method for modeling event pairs. But in film scenes, events are very densely distributed, thus related event pairs are often adjacent to one another and therefore nearby events are more likely to be causal. So, for event pairs extracted from

| Camping Event Pairs |
|---|
| person - pack up → person - go - home |
| person - wake up → person - pack up - backpack |
| person - eat - breakfast → person - pack up - camp-site |
| person - head → hike up |
| person - pack up - car → head out |
| Fantasy Event Pairs |
| person - slam - something → shut |
| send - something → fly - something |
| person - watch → something - disappear |
| person - pick up - something → carry - something |
| person - turn → face - person |
| Mystery Event Pairs |
| bind → gag |
| person - reach → touch - something |
| person - pull - something → reveal - something |
| person - look → confuse |
| person - come → rest |
| Drama Event Pairs |
| person - slam - something → shut |
| person - offer - something → something - decline |
| person - rummage → person - find - something |
| send - something → something - fly |
| send - something → sprawl |

Table 2: High-CP pairs from Camping, Fantasy and Mystery datasets

films we use a variant of CP measure, shown in Eq. 3, that accounts for different window sizes and punishes event pairs from larger window sizes (Riaz and Girju, 2010, 2013; Do et al., 2011; Pichotta and Mooney, 2014).

$$CP_{variant}(e_1, e_2) = \sum_{i=1}^{w_{max}} \frac{CP_i(e_1, e_2)}{i} \quad (3)$$

where w_{max} is the max window size (how many events after the current event are paired with the current event). $CP_i(e_1; e_2)$ is the CP score for event pair $e_1; e_2$ calculated using window size i .

3.3 Experiments and Results

We process the data in each dataset and calculate causal potential score for each extracted event pair, resulting in a rank-ordered list of causal event pairs. We evaluate the top 100 event pairs for camping, and the top 684 event pairs for films. We take a number of event pairs from each film genre (proportional to the number of films in that genre, see Table 1 and 3), then remove duplicate event pairs, which result in the 684 event pairs from film. Table 2 presents examples of learned high-

| Genre | # High-CP Pairs | % Causality |
|---------|-----------------|-------------|
| Drama | 655 | 82.6 |
| Fantasy | 127 | 90.7 |
| Mystery | 122 | 87.7 |

Table 3: Percentage of high-CP pairs labeled as causal by AMT worker, comparing with low-PC pairs, in film genres Drama, Fantasy and Mystery.

CP event pairs from each corpus. In our following Mechanical Turk experiments, Turkers have to pass qualification tests similar to the actual HITs to be able to participate in our task.

In a study on each genre of films, we compare high-CP pairs to a random sample of low-CP pairs on Mechanical Turk to see if pairs with high CP score more strongly encode causal relations than ones with low CP. For every event pair in the 684 high pairs, we randomly select a low pair in order to collect human judgments on Mechanical Turk. The task first defines events and event pairs, then gives examples of event pairs with causal relations. Turkers are asked to select the event pair that is more likely to manifest a causal relation. The results, summarized in Table 3, show that humans judge a large majority of the high-CP pairs to have a causal relation and the results vary by genre. The causality rate is achieved for more focused genres, Fantasy (90.7%) and Mystery (87.7%), despite their smaller size, and the lowest for Drama (82.6%). We believe this result is further evidence that topical coherence improves causal relation learning (Rahimtoroghi et al., 2016; Riaz and Girju, 2010).

In our second evaluation method, we compare the learned CP event pairs to the existing causal knowledge collections. First, we compare our results to the Rel-grams data (learned from newswire) (Balasubramanian et al., 2013). For event pairs from films, we randomly sample 100 high-CP event pairs ensuring that each of the first events of the pairs are distinct. We use the publicly available search interface for Rel-grams to find tuples with the same first event for direct comparison of content of the learned knowledge. We set the co-occurrence window to 5, and select the Rel-gram tuples with the highest # 50 (FS) (frequency of first statement occurring before second statement within a window of 50) to choose high-quality tuples. We evaluate the extracted Rel-gram tuples using the same Mechani-

cal Turk HIT described above. Table 4 shows Mechanical Turk evaluation results for our method on films vs. Rel-grams: in 81% questions, humans judge the high-CP pairs to be more likely to manifest a causal relation. We believe this is because the fine-grained event pairs we learn do not exist in the Rel-gram collections and thus the Rel-gram tuples that matched our first events are not highly coherent, despite the filtering we applied.

| Dataset | Film | Rel-gram Tuples |
|-------------------------------|------|-----------------|
| Percentage of causal relation | 81 % | 19 % |

Table 4: Percentage of pairs judged as causal by AMT workers. Film vs. Rel-Grams.

For event pairs from camping blogs, we evaluate all 100 high-CP pairs in a Mechanical Turk study where Turkers are asked to choose whether an event pair has causal relation or not. We also evaluate Rel-gram tuples using the same task. However, Rel-grams are not sorted by topic. To find tuples relevant to Camping Trip, we use our top 10 indicative events and extracted all the Rel-gram tuples that included at least one event corresponding to one of the Camping indicative events, e.g. *go camp*. We remove any tuple with frequency less than 25 and sort the rest by the total symmetrical conditional probability. The evaluation results presented in Table 5 show that 82% of the blog pairs were labeled as causal, where as only 42% of the Rel-gram pairs were labeled as causal. We argue that this is mainly due to the limitations of the newswire data which does not contain the fine-grained everyday events that we have extracted from our corpus.

| Dataset | Camping | Rel-gram Tuples |
|-------------------------------|---------|-----------------|
| Percentage of causal relation | 82 % | 42 % |

Table 5: Percentage of pairs judged as causal by AMT workers. Camping blogs vs. Rel-Grams.

Next, we compare our results to the event pairs in VerbOcean (learned from newswire) with the HAPPENS-BEFORE relation (Chklovski and Pantel, 2004). We use all 6497 event pairs from VerbOcean, comparing with our 684 event pairs from films and 100 event pairs from camping blogs with high CP scores. Our result shows that there are 12 event pairs that exist in both VerbOcean and

films, e.g. *turn - leave* and *slow - stop*, and there is only one event pair that exist in both VerbOcean and camping blogs: *pack - leave*. This confirms that most causal relations learned from other narrative genres do not exist in the currently available knowledge bases extracted from newswire. A number of event pairs from these collections share the first event, e.g. *dig - find* and *scan - spot* from films vs. *dig - repair* and *scan - upload* from VerbOcean; *drive - park* and *pick - eat* from blogs vs. *drive - drag* and *pick - plunk* from VerbOcean.

Finally, we compare our high-CP pairs learned from film to the high-CP event pairs from Beamer and Girju (2009), learned from only 173 films. There is no public release of Beamer and Girju’s event pairs, thus we take the 29 event pairs with high CP score presented in the paper. A total of 14 of their 29 pairs are also in our top 684 film pairs. These include pairs such as *swerve - avoid*, *leave - stand* and *unlock - open*. However on our larger genre-sorted corpus we also learn pairs such as *grab - haul*, *scratch - claw* and *saddle- mount* that do not exist in their collection.

4 Conclusions and Future Work

Causality is often granular in nature with major events related to the occurrence of finer-grained events. In this position paper, we argue that the focus on newswire has inhibited attempts to learn fine-grained causal relations between everyday events, and that other narrative genres better support such learning. We use unsupervised methods to extract fine-grained causal event relations from films and blog posts about camping.

We show that more than 80% of the relations we learn are evaluated as causal, and that topical coherence plays an important role in modeling event relations. We also show that the causal knowledge we learn from other narrative genres does not exist in current event collections induced from newswire. We plan to expand our genre-specific experiments on the films corpus in future, as well as using other narrative datasets, like restaurant reviews, to extract fine-grained causal knowledge about events.

References

Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *EMNLP*. pages 1721–1731.

- Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *Computational Linguistics and Intelligent Text Processing*, Springer, pages 430–441.
- Allan Bell. 2005. News stories as narratives. *The Language of Time: A Reader* page 397.
- Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.
- Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. *Proceedings of ACL-08: HLT* pages 789–797.
- Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 602–610.
- Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *EMNLP*, volume 4, pages 33–40.
- G. F. de Jong. 1979. *Skimming Stories in Real Time: An Experiment in Integrated Understanding*. Ph.D. thesis, Computer Science Department, Yale University.
- Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 294–303.
- R.J. Gerrig. 1993. *Experiencing narrative worlds: On the psychological activities of reading*. Yale Univ Pr.
- Andrew Gordon, Cosmin Bejan, and Kenji Sagae. 2011a. Commonsense causal reasoning using millions of personal stories. In *Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*.
- Andrew S Gordon, Cosmin Adrian Bejan, and Kenji Sagae. 2011b. Commonsense causal reasoning using millions of personal stories. In *AAAI*.
- Andrew S Gordon, Christopher Wienberg, and Sara Owsley Sood. 2012. Different strokes of different folks: Searching for health narratives in weblogs. In *Privacy, Security, Risk and Trust (PASAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*. IEEE, pages 490–495.
- Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86.
- Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review* 101(3):371.
- Jerry R Hobbs. 1985. Granularity. In *Proceedings of the 9th international joint conference on Artificial intelligence-Volume 1*. Morgan Kaufmann Publishers Inc., pages 432–435.
- Zhichao Hu, Elahe Rahimtoroghi, Larissa Munishkina, Reid Swanson, and Marilyn A Walker. 2013. Unsupervised induction of contingent event pairs from film scenes. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 370–379.
- Dan Jurafsky, Victor Chahuneau, Bryan R Routledge, and Noah A Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* 19(4).
- Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive Science* 5(4):293–331.
- Rutu Mulkar-Mehta, Christopher Welty, Jerry R Hoobs, and Eduard Hovy. 2011. Using granularity concepts for discovering causal relations. In *Proceedings of the FLAIRS conference*.
- Karl Pichotta and Raymond J Mooney. 2014. Statistical script learning with multi-argument events. *EACL 2014* page 220.
- Livia Polanyi. 1989. *Telling the American Story: A Structural and Cultural Analysis of Conversational Storytelling*. MIT Press.
- Elahe Rahimtoroghi, Ernesto Hernandez, and Marilyn A. Walker. 2016. Learning fine-grained knowledge about contingent relations between everyday events. In *Proceedings of SIGDIAL 2016*, pages 350–359.
- Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*. IEEE, pages 361–368.
- Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. Citeseer.
- R Schank, Robert Abelson, and Roger C Schank. 1977. *Scripts Plans Goals*. Lea.
- Patrick Suppes. 1970. *A probabilistic theory of causality*. North-Holland Publishing Company Amsterdam.
- Reid Swanson, Elahe Rahimtoroghi, Thomas Corcoran, and Marilyn A Walker. 2014. Identifying narrative clause types in personal stories. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

- Tom Trabasso and Paul van den Broek. 1985. Causal thinking and the representation of narrative events. *Journal of Memory and Language* 24:612–630.
- Tom Trabasso, Paul Van den Broek, and So Young Suh. 1989. Logical necessity and transitivity of causal relations in stories. *Discourse processes* 12(1):1–25.
- Paul Van den Broek. 1990. The causal inference maker: Towards a process model of inference generation in text comprehension. *Comprehension processes in reading* pages 423–445.
- Marilyn Walker, Grace Lin, and Jennifer Sawyer. 2012. An annotated corpus of film dialogue for learning and characterizing character style. In *Language Resources and Evaluation Conference, LREC2012*.
- William H Warren, David W Nicholas, and Tom Trabasso. 1979. Event chains and inferences in understanding narratives. *New directions in discourse processing* 2:23–52.
- Robert Wilensky. 1982. Points: A theory of the structure of stories in memory. In Wendy G. Lehnert and Martin H. Ringle, editors, *Strategies for Natural Language Processing*.