Inferring Narrative Causality between Event Pairs in Films

Zhichao Hu 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, mawalker@ucsc.edu

Abstract

To understand narrative, humans draw inferences about the underlying relations between narrative events. Cognitive theories of narrative understanding define these inferences as four different types of causality, that include pairs of events A, B where A physically causes B (X drop, X break), to pairs of events where A causes emotional state B (Y saw X, Y felt fear). Previous work on learning narrative relations from text has either focused on "strict" physical causality, or has been vague about what relation is being learned. This paper learns pairs of causal events from a corpus of film scene descriptions which are action rich and tend to be told in chronological order. We show that event pairs induced using our methods are of high quality and are judged to have a stronger causal relation than event pairs from Rel-grams.

1 Introduction

Telling and understanding stories is a central part of human experience, and many types of human communication involve narrative structures. Theories of narrative posit that NARRATIVE CAUSALITY underlies human understanding of a narrative (Warren et al., 1979; Trabasso et al., 1989; Van den Broek, 1990). However previous computational work on narrative schemas, scripts or event schemas learn "collections of events that tend to co-occur" (Chambers and Jurafsky, 2008; Balasubramanian et al., 2013; Pichotta and Mooney, 2014), rather than causal relations between events (Rahimtoroghi et al., 2016). Another limitation of previous work is that it has mostly been applied to newswire, limiting what is learned to relations between newsworthy events, rather than everyday events (Rahimtoroghi et al., 2016; Hu et al., 2013; Beamer and Girju, 2009; Manshadi et al., 2008).

Our focus here is on NARRATIVE CAUSALITY (Trabasso et al., 1989; Van den Broek, 1990), the four different relations posited by narrative theories to underly narrative coherence:

- PHYSICAL: Event A physically causes event B to happen
- MOTIVATIONAL: Event A happens with B as a motivation
- PSYCHOLOGICAL: Event A brings about emotions (expressed in event B)
- ENABLING: Event A creates a state or condition for B to happen. A enables B.

Previous work on learning causal relations has primarily focused on physical causality (Riaz and Girju, 2010; Beamer and Girju, 2009), while our aim is to learn event pairs manifesting all types of narrative causality, and test their generality as a source of causal knowledge. We posit that film scene descriptions are a good resource for learning narrative causality because they are: (1) action rich; (2) about everyday events; and (3) told in temporal order, providing a primary cue to causality (Beamer and Girju, 2009; Hu et al., 2013).

Film scenes contain many descriptions encoding PHYSICAL CAUSALITY, e.g. in Fig. 1, Scene 1, Frodo grabs Pippin's sleeve, causing Pippin to spill his beer (*grab - spill*). Pippin then pushes Frodo away, causing Frodo to stumble backwards and fall to the floor (*push - stumble*, *stumble - fall*, and *push - fall*). But they also contain all other types of narrative causality: in Scene 2, Gandalf has to stoop, because he wants to avoid hitting his head on the low ceiling (*stoop - avoid*: MOTIVATIONAL). He then looks around, and enjoys the result of looking: the familiarity of Bag End (*look - enjoy*: PSYCHOLOGICAL). He turns, which causes

Proceedings of the SIGDIAL 2017 Conference, pages 342–351,
Saarbrücken, Germany, 15-17 August 2017.

#	Scene
1	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.
2	Bilbo leads Gandalf into Bag End... Cozy and cluttered with souvenirs of Bilbo's travels. Gandalf has to **stoop** to **avoid** hitting his head on the low ceiling. Bilbo hangs up Gandalf's hat on a peg and trots off down the hall. Bilbo disappears into the kitchen as Gandalf **looks** around.. **enjoying** the familiarity of Bag End... He **turns**, **knocking** his head on the light and then walking into the wooden beam. He groans.
3	Bilbo **pulls out** the ring... he **stares at** it in his palm. With all his will power, Bilbo **allows** the ring to slowly **slide off** his palm and drop to the floor. The tiny ring lands with a heavy thud on the wooden floor.
4	GANDALF... lying unconscious on a cold obsidian floor. He **wakes** to the sound of ripping and tearing ... **rising** onto his knees... lifting his head... Gandalf stands as the camera pulls back to reveal him stranded on the summit of Orthanc.

Figure 1: Film Scenes from Lord of the Rings

him to knock his head on the light (*turn - knock*: the weak causality of ENABLING).[1]

This paper learns causal pairs from a corpus of 955 films. Because previous work shows that more specific, detailed causal relations can be learned from topic-sorted corpora (Riaz and Girju, 2010; Rahimtoroghi et al., 2016), we explore differences in learning between genres of film, positing e.g. that horror films may feature very different types of events than comedies. We also test the quality of what is learned when we train on genre specific texts vs. the whole collection. Our results show that:

- human judges can distinguish between strong and weakly causal event pairs induced using our method (Section 3.1);
- our strongly causal event pairs are rated as more likely to be causal than those provided by the Rel-gram corpus (Balasubramanian et al., 2013) (Section 3.2);
- human judges can recognize different types of narrative causality (Section 3.3);
- using both whole-corpus and genre-specific methods yields similar results for quality, despite the smaller size of the genre-specific subcorpora. Moreover, the genre-specific method learns some event pairs that are different than whole corpus event-pairs, while still being high-quality. (Section 3.4);

We explain our method in Section 2, and then present experimental results in Section 3. We leave a more detailed discussion of related work until Section 4 when we can compare it more directly with our own.

2 Experimental Method

We estimate the likelihood of a narrative causality relation between events in film scenes.

2.1 Film Scenes & Pre-Processing.

We chose 11 genres with more than 100 films from a corpus of film scene descriptions (Walker et al., 2012; Hu et al., 2013),[2] resulting in 955 unique films. Film scripts were scraped from the IMSDb website, film dialogs and scene descriptions were then automatically separated. Films per genre range from 107 to 579. Films can belong to multiple genres, e.g. the scenes from *The Fellowship of the Ring* shown in Figure 1 would become part of the genres of Action, Adventure, and Fantasy. Each film's scene descriptions ranges from 2000 to 35000 words. Table 1 enumerates the sizes of each genre, illustrating the potential tradeoff between getting good probability estimates for event co-occurrence when the same events are repeated **within** a genre, vs. across the whole corpus. We use Stanford CoreNLP 3.5.2 to tokenize, lemmatize, POS tag, dependency parse and label named entities (Manning et al., 2014).

2.2 Compute Event Representations.

An event is defined as a verb lemma, as in previous work (Chambers and Jurafsky, 2008; Do et al., 2011; Riaz and Girju, 2010; Manshadi et al., 2008).

[1] Gandalf did not *turn* in order to *knock*, which would have been MOTIVATIONAL. Nor was it entailed that *turning* would cause *knocking*, which would have been PHYSICAL, because he clearly could have missed hitting his head if he had been more careful.

[2] From `https://nlds.soe.ucsc.edu/fc2`

Genre	# Films	Word Count	Example
Action	290	3,758,387	The Avengers
Adventure	166	2,115,247	Indiana Jones and the Temple of Doom
Comedy	347	3,434,612	All About Steve
Crime	201	2,342,324	The Italian Job
Drama	579	6,680,749	American Beauty
Fantasy	113	1,186,587	Lord of the Rings: Fellowship of the Ring
Horror	149	1,789,667	Scream
Mystery	107	1,346,496	Black Swan
Romance	192	2,022,305	Last Tango in Paris
Sci-Fi	155	1,964,856	I, Robot
Thriller	373	4,548,043	Ghost Rider

Table 1: Distribution of Films By Genre.

We extract events by keeping all tokens whose POS tags begin with VB: VB, VBD, VBG, VBN, VBP, and VBZ. This results in extracting deverbal nouns that implicitly evoke events, such as the events of *ripping* and *tearing* in Scene 4 of Figure 1. This definition also allows us to pick up *resultative clauses* along with the action that caused the result (Hovav and Levin, 2001; Goldberg and Jackendoff, 2004), e.g. in *He slammed the door shut*, both *slammed* and *shut* are picked up as verbs. We exclude light verbs e.g. *be, let, do, begin, have, start, try*, because they often only represent a meaningful event when combined with their complements.

We extract the subject (*nsubj, agent*), direct object (*dobj, nsubjpass*), indirect object (*iobj*) and particle of the verb (*compound:prt*), if any. In order to abstract and merge different arguments, we generalize the arguments to two types: *person* and *something*. We generalize an argument to *person* when: (1) its named entity type is PERSON; or (2) it is a pronoun (except "it"); or (3) it is a noun in WordNet with more than half of its Synsets having lexical filename noun.person, e.g. *doctor, soldier, waiter, man, woman*. Our narrative causal semantics would be more specific if we could generalize over other types of named entities as well, such as *location*. However Stanford NER identifiable named entities rarely occur in film data.

For every event, we record the combinations of its arguments and particle for every instance. For example, the instance of event "pick" in sentence: *He picked it up... a pearl*, has combination *subj: person, dobj: something, iobj: none, particle: up*. We pick the combination with the highest frequency to represent the arguments and particle for each event.

2.3 Calculating Narrative Causality.

We use the Causal Potential (CP) measure in (1), shown to work well in previous work (Beamer and Girju, 2009; Hu et al., 2013):

$$CP(e_1, e_2) = PMI(e_1, e_2) + \log \frac{P(e_1 \rightarrow e_2)}{P(e_2 \rightarrow e_1)} \quad (1)$$

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

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).

We obtain the frequency of every event and event pair for each genre. Unseen event pairs are smoothed with frequency equal to 1. In this paper, the notion of window size indicates how many events after the current event are paired with the current event. We use window sizes 1, 2 and 3, and calculate narrative causality for each window size. In film scenes, events are very densely distributed, (see Figure 1), thus related event pairs are often adjacent to one another, but the discourse structure of film scenes, not surprisingly, also contain related events separated by other events (Grosz and Sidner, 1986; Mann and Thompson, 1987). For example, in Scene 3 of Figure 1, Bilbo pulling out the ring enables him to slide it off his palm later (*pull out - slide off*). Moreover, while related events are less

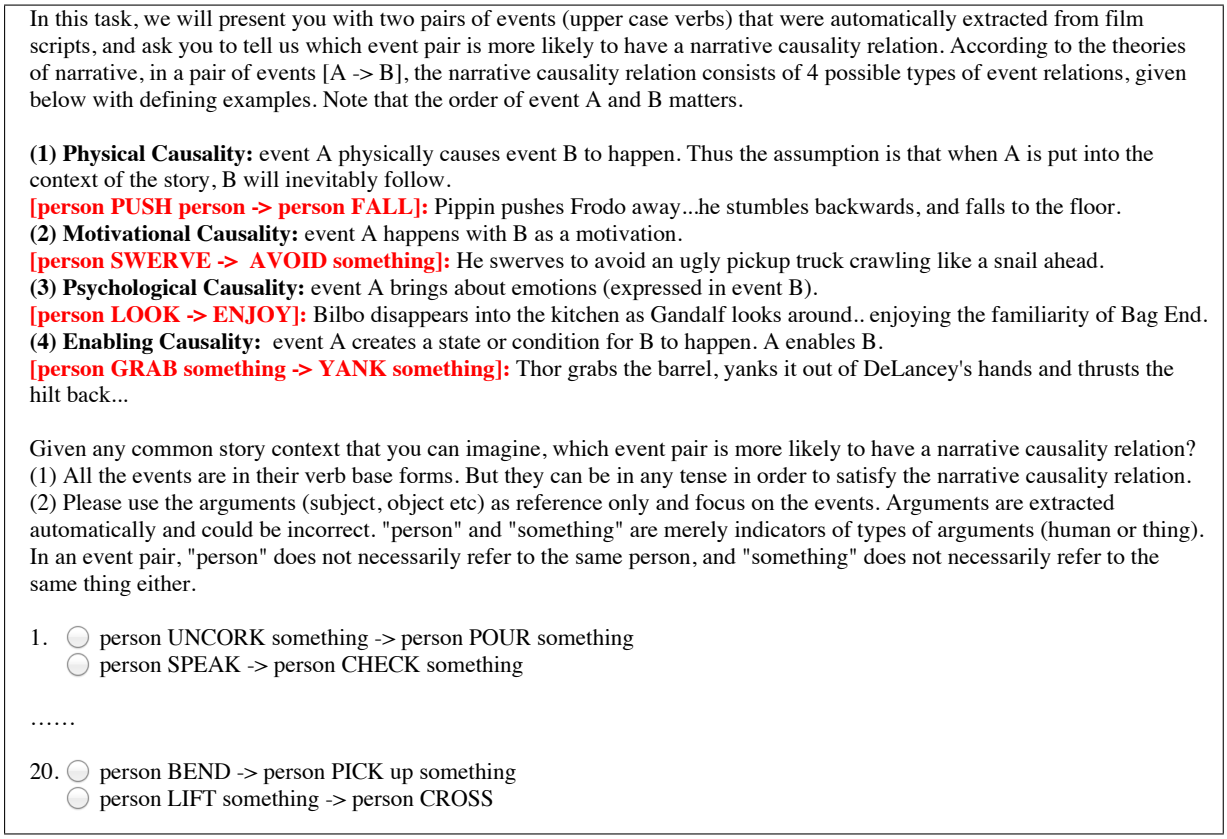

Figure 2: Instructions for the MT HIT.

frequently separated (window size 3), we assume that unrelated events will be filtered out by their low probabilities. We thus define a *CPC* measure, shown in (2) that combines the frequencies across window size:

$$CPC(e_1, e_2) = \sum_{i=1}^{w_{max}} \frac{CP_i(e_1, e_2)}{i} \quad (2)$$

where w_{max} is the max window size. $CP_i(e_1, e_2)$ is the CP score for event pair e_1, e_2 calculated using window size i. The *CPC* measure combines frequencies across window sizes, but punishes event pairs from larger window sizes, thus assuming that nearby events are more likely to be causal.

3 Evaluation and Results

We posit that human judgments are the best way to evaluate the quality of the induced event pairs, as opposed to automatic measures such as Narrative Cloze, which assume that the event pairs in a particular instance of text can be used as held-out test data (Chambers and Jurafsky, 2008). Our first experiment tests whether event pairs with high *CPC* scores are more likely to have a narrative causality relation. Our second experiment compares pairs with high *CPC* scores with their corresponding top Rel-gram pairs. Our third experiment tests whether annotators can distinguish narrative causality types. Our final experiment compares the quality and type of causal pairs learned on a per genre basis, vs. those learned on the whole film corpus.

3.1 High vs. Low CPC Event Pairs

After processing all the data, we have a list of event pairs scored by *CPC*, and rank-ordered within each genre. Some of the genre specific event pairs seem to intuitively reflect their genre, however there are many learned pairs that are in overlap across genres. We select the top 3000 event pairs with high scores from all the genres ("high pairs"). The number of event pairs from a genre is proportional to the number of films in that genre. We also select the bottom 6000 event pairs with low scores from all the genres using similar method ("low pairs"). Since many pairs are duplicated across genre, the high pairs and low pairs are then de-duplicated (two event pairs are defined as equal if they have the same verbs in the same order). We keep the arguments with the highest frequencies. This result in 960 high pairs. If an event has no subject, "person" is added as

#	High CPC Pair	Low CPC Pair
1	[person] *clink* [smth] - [person] *drink* [smth]	[person] *strike* - [person] *give* [person] [smth]
2	[person] *beckon* - [person] *come*	[smth] *become* - [person] *hide*
3	[person] *bend* - [person] *pick* up [smth]	[person] *lift* [smth] - [person] *cross*
4	[person] *cough* - [person] *splutter*	[person] *force* - [smth] *show* [smth]
5	[person] *crane* - [person] *see* [smth]	[person] *fade* - [person] *allow* [person]

Table 2: Narratively Causal Pairs where all 5 annotators selected the High CPC pair.

subject, since most events have human agents.

For every event pair in the 960 high pairs, we randomly select a low pair in order to collect human judgments on Mechanical Turk. The task first introduces event and event pair definitions, then defines the four types of narrative causality with corresponding examples. Turkers are asked to select the event pair that is more likely to manifest a narrative causality relation. Each HIT consists of 20 judgements, and we collect 5 judgements per HIT. Because this task requires some care, Turkers had to be prequalified. The qualification test aims to test Turkers' understanding of narrative causality. It is similar to the task itself, but with more obvious choices, such as high CPC pair *open - reveal* vs low CPC pair *pay - fade*. Figure 2 shows a simplified version of the HIT instructions.[3]

Genre	# High Pairs	% Causality
Action	320	86.3
Adventure	171	86.6
Comedy	384	84.9
Crime	23	84.9
Drama	**665**	**82.6**
Fantasy	**127**	**90.7**
Horror	156	87.2
Mystery	122	87.7
Romance	215	86.0
Sci-Fi	158	88.0
Thriller	405	87.7

Table 3: Percentages of high pairs that receive majority vote results by genre.

The results show that humans judge the high pairs as more likely to have a narrative causality relation in 82.8% of items. Among those, all the items receive 3 or more votes for the high pairs. Overall, all five Turkers select the high CPC pairs in 51% of the items. The average pairwise Krippendorff's Alpha score is respectable at 0.56.

[3]The full instructions provide more examples and background information.

Table 2 shows items where all 5 Turkers selected the high pair. For example, *clink - drink* in Row 1 could have either a MOTIVATIONAL or ENABLING narrative causality depending on the context, but the causal relation in either case is much clearer than with the low CPC pair *strike - give*. Row 2 and Row 5 *beckon - come* and *crane -see* both have ENABLING causality which is a weakly causal relation, but again more meaningful than their low CPC counterparts. In Row 3, it is clear that a person often *bends* with the motivation to *pick up* something. In row 4 a person *cough*s, PHYSICALLY causes him to *splutter* everywhere.

Table 3 shows majority vote results for percentages of high pairs that are considered to exhibit more narrative causality, sorted by genre. The results for all genres are good, ranging from ∼82% to ∼91%. Interestingly, Drama has the highest number of films with the lowest percentage of judged narrative causality, while Fantasy has the lowest number of films with the highest judged narrative causality. This may be because the Drama category is a catch-all (over half of the films are categorized this way suggesting that it has low coherence as a genre). The poor performance on Drama would then be consistent with previous work that shows that topical coherence (genre in this case) improves causal relation learning (Rahimtoroghi et al., 2016; Riaz and Girju, 2010). We will return to this point in Section 3.4.

3.2 CPC vs. Rel-gram Event Pairs

We then compare the narrative causality event pairs (high pairs) with event pairs from the Rel-grams corpus (Balasubramanian et al., 2012, 2013). Rel-grams (Relational n-grams) are pairs of open-domain relational tuples (T,T'). They are analogous to lexical n-grams, but is computed over relations rather than over words. For example, "A person who gets arrested is typically charged with some activity." yield the tuple: T = ([police] *arrest* [person]) and T' = ([person] *be charge with* [activity]).

#	Narrative Causality (CPC) Pairs	Rel-gram Pairs	CPC Vote #
1	[person] *clear* [smth] - [person] *reveal* [smth]	[person] *clear* [smth] - [person] *hit* [smth]	5
2	[person] *embrace* - [person] *kiss*	[person] *embrace* [person] - [person] *meet* [person]	5
3	[person] *empty* [something] -[person] *reload*	[person] *empty* [smth] - [person] *shoot* [person]	5
4	[person] *marry* [person] - [person] *think*	[person] *marry* [person] - [person] *die* [something]	5
5	[person] *stumble* - [smth] *fall*	[person] *stumble* upon [person] - [person] *take* [person]	5
6	[person] *gaze* - [smth] *drift*	[person] *gaze* at [person]- [person] *see* [person]	0
7	[person] *reveal* [smth] - [person] *sit*	[person] *reveal* [person] - [person] *see* [person]	0
8	[person] *watch* - [person] *appal*	[person] *watch* [person] - [person] *see* [person]	0

Table 4: Items where either CPC event pairs or Rel-gram event pairs were strongly preferred.

Over 1.8M news wire documents are used to build a database of Rel-grams co-occurence statistics.

Using a similar HIT template, we randomly sample 100 high CPC event pairs from the 960 high CPC pairs, where we ensure that each of the first events of the pairs are distinct. We use the publicly available search interface for Rel-grams[4] to find Rel-gram statement pairs that have the same first event. Modeling our own experimental setup we set the co-occurrence window to 5[5], and select the Rel-gram pair with the highest #50(FS) (frequency of first statement occurring before second statement within a window of 50).

To make Rel-gram event pairs similar to ours, we generalize their arguments to "person" and "something" manually. We keep the verb particle if any. For example, the Rel-gram pair "[person] *remain* in [location] - [person] *become* [leader]" is generalized to "[person] *remain* in [something] - [person] *become* [something]". It is possible that this disadvantages Rel-grams in some way, but our main focus is on the causality relation between verbs, which should not be affected. Moreover the two sets of event pairs cannot be compared without this generalization. The same 5 annotators participate in this 5 HITs (100 items).

The results show that humans judge the CPC pairs to be more likely to manifest a narrative causality relation 81% of the time. The average pairwise Krippendorff's Alpha score of all Turkers is 0.482. Table 4 shows items where all Turkers judge the CPC pairs as more likely to be causally related. For example, in Row 1 to *clear* seems more likely to enable something being *revealed*, instead of causing a person to *hit* something. In Row 2, even though *embrace* and *kiss* might only have an ENABLING narrative causality relation, the reversed causality between *embrace* and *meet* in the Rel-gram pair is based on symmetric conditional probability (SCP) rather than explicit causal modeling. SCP combines Bigram probability in both directions as follows:

$$SCP(e_1, e_2) = P(e_2|e_1) \times P(e_1|e_2) \quad (3)$$

In Row 4, *marrying* someone might just possibly enable one to think about something, but could hardly enable/cause someone to die. In Row 5 *stumble* physically causes one to *fall*, while it is more difficult to see the causal relation between *stumbling on* someone and then a person *taking* another person (somewhere).

Narrative Causality Type	Count	Example Pair
Physical	13	*fire - blast*
Motivational	29	*bend - retrieve*
Psychological	9	*look - astonish*
Enabling	28	*lean - whisper*

Table 5: Distribution of narrative causality types .

3.3 Narrative Causality Types

Although theories of narrative posit four different types of narrative causality, previous work has not conducted reliability studies with non-experts such as Turkers. Here we explore whether humans can distinguish narrative causality types, by asking Turkers to decide which relation holds between an event pair. The instructions contain descriptions of narrative causality types and the strength of these relations (from strong to weak: PHYSICAL, MOTIVATIONAL, PSYCHOLOGICAL and ENABLING (Trabasso et al., 1989)). Because the stronger types of narrative causality could also be considered ENABLING, Turkers are instructed to choose the strongest narrative causality that could be applied to the event pair.

[4] http://relgrams.cs.stonybrook.edu/

[5] The search interface does not support a window size of 3, thus we chose 5 as it's the closest window size larger than 1.

Fantasy	CPC	Action	CPC
[person] *slam* [smth] - *shut*	4.95	[person] *huff* - [person] *puff*	5.57
send [smth] - [smth] *fly*	4.89	*bind* - *gag*	5.50
[person] *watch* - [smth] *disappear*	4.87	[smth] *swerve* - *avoid* [smth]	5.21
[person] *turn* - *face* [person]	4.83	[person] *bend* - [person] *pick* up [smth]	5.01
[person] *pull* [smth] - *reveal* [smth]	4.70	*send* [smth] - [smth] *tumble*	4.85
[person] *pick* up [smth] - *carry* [smth]	4.54	*send* [smth] - *sprawl*	4.83
[person] *reach* - [person] *pull* [smth]	4.42	[person] *slam* [smth] - *shut*	4.79
Sci-Fi	**CPC**	**Thriller**	**CPC**
[person] *bend* - [person] *pick* up [smth]	4.88	*bind* - *gag*	5.66
follow - [person] *gaze*	4.83	[smth] *swerve* - *avoid* [smth]	5.37
[person] *grab* [smth] - [person] *yank* [smth]	4.83	[person] *rummage* - [person] *find* [smth]	5.05
send [smth] - [smth] *fly*	4.81	[person] *inhale* - peroson *exhale*	5.04
[person] *slam* [smth] - *shut*	4.78	[person] *slam* [smth] - *shut*	5.00
[person] *grab* [smth] - [person] *drag* [person]	4.77	*send* [smth] - [smth] *fly*	4.97
[person] *reach* - *touch* [smth]	4.67	[person] *reach* - [person] *produce* [smth]	4.81

Table 6: Event pairs with Highest CPC scores from Fantasy, Action, Sci-Fi and Thriller genres.

We select 100 pairs randomly from the high CPC pairs of the 479 questions that had the highest Turker agreement. Among all 100 questions, 79% of the items receive a majority vote result (3 or more Turkers selecting the same answer). The distribution of narrative causality types of the 79 items is shown in Table 5. Interestingly, films are full of motivational causality, which often reflect action sequences where protagonist pursue particular narratively relevant goals (Rapp and Gerrig, 2006, 2002).

3.4 Genre Specific Causality

Previous work suggests that topical coherence and similarity of events within the corpus used for learning causal/contingent event relations might be as important as the size of the corpus (Riaz and Girju, 2010; Rahimtoroghi et al., 2016). In other words, smaller corpora filtered by topic or genre might be more useful than large undifferentiated sets (Riloff, 1996), although obviously very large corpora that are topic or genre sorted could be even more useful. We therefore test whether separating films by genre yields higher quality event pairs than a method that combines all films, irrespective of genre. We assume that the very notion of a film genre defines a set of films with similar types of events.

We first compute a list of *CPC* scores using films from all genres and take 960 event pairs with highest scores. Comparing the 960 event pairs from all films with the 960 pairs from merging genres described in Section 3.1, we find that 728 pairs overlap between the two sets. Thus with the smaller genre-specific corpora we learn more than 70% of the same causal pairs. The results shown in Table 3 suggest furthermore that the genre-specific pairs are high quality.

However, it is still possible that the 232 pairs from each set that are not in overlap vary in quality from the 728 pairs that are in overlap. We therefore pick 100 random pairs from each set, match the pairs randomly to form items, and repeat the event pairs comparison HIT with these pairs. The results suggest that there are no differences between the two methods as far as quality: in 48 of the 100 questions, pairs from genre-separated method have Turkers' majority vote, vs. in 52 of the 100 questions pairs from combined genres have the majority vote.

Moreover we obtain **more** high-quality, reliable narrative causality relations using both methods, and we learn some genre-specific causal relations that we do not learn on the whole corpus. Table 8 shows the the overlap in learned pairs amongst the top 30 CPC pairs in five of the most distinct genres (genres with highest percentages in Table 3: Fantasy, Sci-Fi, Horror, Mystery and Thriller) vs. all films (All). Mystery has the smallest overlap with All, followed by Fantasy and Sci-Fi.

To illustrate some of the differences, Table 6 shows event pairs with the highest *CPC* scores in Fantasy, Action, Sci-Fi and Thriller genres. Table 7 shows event pairs unique to each genre within its top 30 CPC pairs.

We also compare our 960 pairs from merging genres described in Section 3.1 with 200 event pairs extracted from camping and storm personal blog stories in Rahimtoroghi et al. (2016). The only pairs that overlap are: *sit* - *eat*, *play* - *sing*, illustrating again that causal relations learned are not as dependent on the size of the corpus, as they are on its topical and event-based coherence. Since most previous work on narrative schemas, scripts,

Genre	Event pairs
Fantasy	*struggle - get, reveal - stand, see - stand, get - marry, sit - sip, nod - head, make - break, spin - face, take - bite, watch - disappear, pick - carry*
Sci-Fi	*hear - echo, see - come, look - alarm, widen - see, head - stop, clear - reveal, sit - study, look - puzzle, peek - see*
Horror	*listen - hear, stare - fascinate, hear - muffle, slow - stop, peel - reveal, reach - yank, reach - handle, grab - handle*
Mystery	*slip - fall, dig - pull, walk - reach, look - confuse, sit - eat, knock - open, look - horrify, stop - look, sit - look, seem - lose*
Thriller	*look - wonder, raise - fire, poise - strike, sit - hunch, rape - murder*
All	*sit - leg, whoop - holler, huff - puff, disappear - reappear, cease - exist, dive - swim, spur - gallop, offer - decline, contain - omit, hoot - holler, pay - heed*

Table 7: Event pairs unique to Fantasy, Sci-Fi, Horror, Mystery, Thriller genres and all films.

Genre	All	Thr	Mys	Hor	Sci
Fan	8	9	13	15	14
Sci	8	12	14	18	
Hor	10	14	14		
Mys	7	12			
Thr	18				

Table 8: Overlap in learned pairs among the most distinct genres (Fantasy, Sci-Fi, Horror, Mystery and Thriller) vs. all films (All).

event schemas or rel-grams has only been applied to one large corpus of newswire (Gigaword corpus), these methods have only learned relations about newsworthy topics, and even then, perhaps only the most frequent, highly common news events. In contrast, both our approach and that of Rahimtoroghi et al. (2016) learn fine-grained causal relations that underly narratives, which we believe are more in the spirit of Schank's original motivation for scripts (Lehnert, 1981; Schank et al., 1977; Wilensky, 1982; de Jong, 1979).

4 Related Work

Hu et al. (2013) tested four methods for inducing pairs of adjacent events with contingency/causality relations from film scenes, including Causal Potential, Pointwise Mutual Information, Bigram Model and Protagonist-based Model. Rahimtoroghi et al. (2016) also used a modified version of the the CP measure, adjusted to account for the discourse structure of personal narratives in blogs. Here we use a much larger set of films and apply different techniques and a detailed evaluation. Our learned causal pairs and supporting film data are available for download [6].

Do et al. (2011) used a minimally supervised approach, based on focused distributional similarity methods and discourse connectives, to identify causality relations between events in PDTB in context (both verbs and nouns) (Prasad et al., 2008). They present a detailed formula for calculating contingency/causality that takes into account several different kinds of argument overlap between adjacent events. However they do not provide any evidence that all the components of this formula actually contribute to their results.

Gordon et al. (2011) used event ngrams and discourse cues to learn causal relations from first person stories posted on weblogs and evaluated them with respect to the COPA SEM-EVAL task. Other related work learns likely sequences of temporally ordered events but does not explicitly model CAUSALITY (Chambers and Jurafsky, 2009; Balasubramanian et al., 2013; Manshadi et al., 2008).

Work on VerbOcean (Chklovski and Pantel, 2004) use lexical patterns to learn semantic verb relations of similarity, strength, antonymy, enablement and happens-before relations. Balasubramanian et al. (2013) use symmetric probability to learn semantically typed relational triples (actor, relation, actor), which they call Rel-grams (relational n-grams), and show that their schemas outperform previous work (Chambers and Jurafsky, 2009). We thus compared our event pairs with Rel-grams, showing that humans are more likely to perceive narrative causality in our event pairs.

5 Discussion and Future Work

We present an unsupervised model based on Causal Potential (Beamer and Girju, 2009) to induce event pairs with narrative causality relations from film scenes in 11 genres. Results from four human evaluations show that narrative causality event pairs induced using our method are of high quality, and are perceived as more causally related than corresponding Rel-grams. We show that humans can identify different types of narrative causality, but we leave automatic identification of these to future work. We also show that inducing narrative causality event

[6] https://nlds.soe.ucsc.edu/narrativecausality

pairs using both whole-corpus and genre-specific methods yields similar results for quality, despite the smaller size of the genre-specific subcorpora. Moreover, the genre-specific method learns high quality event pairs that are different than whole corpus event-pairs.

We are looking into applying and evaluating our CPC method to other genre and topic sorted datasets such as books and personal blogs. We want to expand our set of event pairs with narrative causality relations, which could potentially aid text understanding, information extraction, question answering, and content summarization. We also aim to explore features for narrative causality type classification. Information such as event A physically causes event B, or event C enables event D could further help aforementioned applications.

Acknowledgements

This research was supported by Nuance Foundation Grant SC-14-74, NSF #IIS-1302668-002, NSF CISE CreativeIT #IIS-1002921 and NSF CISE RI EAGER #IIS-1044693.

References

Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni, et al. 2012. Rel-grams: a probabilistic model of relations in text. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, pages 101–105.

Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni Mausam, and Oren Etzioni. 2013b. 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.

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.

Adele E Goldberg and Ray Jackendoff. 2004. The english resultative as a family of constructions. *Language* pages 532–568.

Andrew Gordon, Cosmin Bejan, and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. In *Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics* 12:175–204.

Malka Rappaport Hovav and Beth Levin. 2001. An event structure account of english resultatives. *Language* pages 766–797.

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.

Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive Science* 5(4):293–331.

W.C. Mann and S.A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. Technical Report RS-87-190, USC/Information Sciences Institute.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*. pages 55–60.

Mehdi Manshadi, Reid Swanson, and Andrew S Gordon. 2008. Learning a probabilistic model of event sequences from internet weblog stories. In *Proceedings of the 21st FLAIRS Conference*.

Karl Pichotta and Raymond J Mooney. 2014. Statistical script learning with multi-argument events. *EACL 2014* page 220.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. pages 2961–2968.

Elahe Rahimtoroghi, Ernesto Hernandez, and Marilyn A. Walker. 2016a. Learning fine-grained knowledge about contingent relations between everyday events. In *Proceedings of SIGDIAL 2016*. pages 350–359.

D.N. Rapp and R.J. Gerrig. 2002. Readers' reality-driven and plot-driven analyses in narrative comprehension. *Memory & Cognition* 30(5):779.

D.N. Rapp and R.J. Gerrig. 2006. Predilections for narrative outcomes: The impact of story contexts and reader preferences. *Journal of Memory and Language* 54(1):54–67.

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.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence*. pages 1044–1049.

R Schank, Robert Abelson, and Roger C Schank. 1977. *Scripts Plans Goals*. Lea.

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*.