Using a Bigram Event Model to Predict Causal Potential

Brandon Beamer and Roxana Girju

University of Illinois at Urbana-Champaign
{bbeamer,girju}@illinois.edu

Abstract. This paper addresses the problem of causal knowledge discovery. Using online screenplays, we generate a corpus of temporally ordered events. We then introduce a measure we call *causal potential* which is easily calculated with statistics gathered over the corpus and show that this measure is highly correlated with an event pair's tendency of encoding a causal relation. We suggest that causal potential can be used in systems whose task is to determine the existence of causality between temporally adjacent events, when critical context is either missing or unreliable. Moreover, we argue that our model should therefore be used as a baseline for standard supervised models which take into account contextual information.

1 Introduction

Automatic recognition and extraction of causal event sequences from text is a crucial task for many Computational Linguistics applications. It is a prerequisite in text coherence, entailment, question answering, and information retrieval (Goldman et al., 1999; Khoo et al., 2001; Girju, 2003). Put simply, it is a prerequisite to perform textual *reasoning*.

Whether two textual units (words, phrases or sentences) are in a causal relationship is largely dependent on context. But that is not to say that such pairs in general have no statistical tendencies when it comes to causality.

This paper describes a knowledge-poor unsupervised model relying on a statistical measure we call *causal potential*. Our focus is on event sequences as expressed by consecutive verbs in a discourse (event pairs). Event pairs with a high causal potential can be interpreted as being more likely to occur in causal contexts than events with low causal potential. This measure can then be used in more complex systems to gain causal intuitions in situations when context is scarce or unreliable. Therefore, we argue that our model should be used as a baseline by standard supervised models which take into account contextual information.

In this paper we evaluate our measure of causal potential and show that it correlates highly with human observances of causal events in text.

A. Gelbukh (Ed.): CICLing 2009, LNCS 5449, pp. 430–441, 2009.

2 Previous Work

Despite its importance to Computational Linguistics, the task of causal knowledge extraction has not been tackled much in this field. Many of the early works on this topic (Khoo et al., 2001; Girju, 2003) focused on predefined lexico-syntactic constructions encoding causal relations. Girju (2003), for example presents a supervised knowledge-intensive system which relies on "noun – verb – noun" constructions to identify new noun–noun pairs encoding cause-effect (e.g., *Tsunamis cause tidal waves*). The verb is identified from a set of 60 causal WordNet verbs such as *cause, lead to, provoke*. Her system obtains a precision score of 73.91% and she proves the importance of the system in the task of question answering.

Chang and Choi (2006) improve over Girju's approach by employing a dependency parser which identifies patterns of the type "NP_cause cue NP_effect". They use a Naive Bayes classifier based on cue phrase and lexical pair (NP_cause–NP_effect) probabilities which are learned from an unannotated corpus of examples. Chang and Choi (2006) start with the classifier proposed by Girju (2003) and then use expectation-maximization to bootstrap the final classifier. The expectation and maximization steps are repeated until no improvement is obtained. The reported performance is 80% F-measure.

In this paper we present a knowledge-poor unsupervised approach to the identification of causal relations between temporally adjacent events denoted by verbs. Unlike previous attempts, our system does not rely on any predefined lexico-syntactic pattern. Instead, events are identified only after part of speech tagging the text. In particular we introduce a measure called *causal potential* which relies on statistics gathered over a large unannotated corpus and show that this measure is highly correlated with human judgments.

3 Our Notion of Causality

As mentioned earlier, determining if two events are in a causal relationship is no simple matter. The task of constructing a definition of causality that is both rational and fits our linguistic intuitions still eludes many fields including Linguistics and Philosophy.

The challenge of defining causality has been pursued by philosophers for a long time. While they have not yet resolved the issue, numerous schools of thought have emerged from their efforts (Sosa and Tooley, 1993). Most relevant to our goal of language processing and understanding is a consistent set of annotation guidelines which capture our perception of causality as expressed by language. Thus, we are interested in causal theories which provide the annotator with a relatively objective test, allowing her to judge the causal relation between two events without relying on intuitions which will vary significantly from annotator to annotator. Additionally, the test should also be easy to perform mentally, without needing detailed philosophical knowledge about causality. After reviewing various causality theories in philosophical literature, two of them lend themselves as possibilities: *counterfactual* theories and *manipulation* theories.

3.1 Counterfactual Tests of Causality

Counterfactual theories (see Menzies (2008) for an overview) examine causality via counterfactual statements. For example, the statement *Mary shooting John caused his death* has the counterfactual equivalent: *John would not have died (at that moment) had Mary not shot him.* Shibatani (1976) offers a rigorous counterfactual definition of causation:

Two events qualify as a causative situation if

(a) the relation between the two events is such that the speaker believes that the occurrence of one event, the 'caused event', has been realized at t_2, which is after t_1, the time of the 'causing event'; and if
(b) the relation between the causing and the caused event is such that the speaker believes that the occurrence of the caused event is wholly dependent on the occurrence of the causing event; the dependency of the two events here must be to the extent that it allows the speaker to entertain a counterfactual inference that the caused event would not have taken place at the particular time if the causing event had not taken place, provided that all else had remained the same.

Hence a counterfactual test for an annotator deciding if event A causes event B could be to ask herself the following questions/criteria:

(i) Did event A occur before (or simultaneously with) event B?
(ii) Is the occurrence of event B wholly dependent on the occurrence of event A?
(iii) Had event A not taken place, could one necessarily infer that event B would not have taken place?

If and only if a given situation satisfies all these constraints, the annotator would decide that the two events in question are causally related. A test like this satisfies our constraint that annotation tests should be easy, and should not require in-depth knowledge of works in philosophy and logic. However, it has a few problems. First, deciding if one event is "wholly dependent" on another is rather vague and possibly subjective. Second, assuming the annotator has a vague understanding of criterion (ii), criterion (iii) can produce false positives in cases where two events have a common cause. Consider the following example:

(1) *Mary shot John. John collapsed. John died.*

In this context, the statement "John's collapsing caused his death" is false, even though the counterfactual "Had John not collapsed, could one necessarily infer that John would not have died" is true. John collapsed because he was shot, hence had John not collapsed that would necessarily entail that he had not been shot, which would also entail that John would not have died. Having a good grasp on criterion (ii) can help an annotator avoid this pitfall, but *dependency* as it is used in the definition is not very well defined and this could lead annotators astray.

3.2 Manipulation Tests of Causality

Manipulation theories (see Woodward (2008) for an overview) examine causality via mental experiments where one manipulates one event and observes the behavior of another. A few examples of manipulation definitions of causality are:

(1) A causes B if control of A renders B controllable. A causal relation, then, is one that is invariant to interventions in A in the sense that if someone or something can alter the value of A the change in B follows in a predictable fashion. (Hoover, 1988)
(2) Z_1 is a cause of Z_2 is just a convenient way of saying that if you pick an action that controls Z_1, you will also have an action that controls Z_2. (Orcutt, 1952)

Most philosophers agree today that, insofar as the definition of causality is concerned, manipulation approaches are insufficient because it turns out to be quite impossible to discuss any notion of control in non-causal terms. Thus the definition becomes circular. This is not a problem for us though, as we are not seeking a rigorous philosophical definition of causality, but rather a relatively objective test for the purpose of linguistic annotation. In this capacity, manipulation theories turn out to be quite useful.

Thus, we can devise a new annotation test for causality. For example, an annotator deciding whether event A causes event B could ask herself the following questions instead. Answering *yes* to both would mean the two events are causally related:

(i) Does event A occur before (or simultaneously) with event B?
(ii) Keeping constant as many other states of affairs of the world in the given text context as possible, does modifying event A entail predictably modifying event B?

This annotation test is both simple to execute mentally and is relatively objective. Subjectivity would certainly arise in cases where the annotator is unaware of how certain things in the world work, but relying on the fact that many people share more or less the same baggage of commonsense knowledge this should not be a problem.

We word our annotation test in terms of the manipulation mindset because we believe the language is easier to understand. However, it is important to see that in practice both the counterfactual and the manipulation tests end up being mostly equivalent. The manipulation test instructs the annotator to modify event A and observe the behavior of B. The simplest and most extreme way to modify event A is to either add it to or remove it from the world. The outcome of adding the event is obvious since it corresponds to the situation described in the text. The only mental experiment the annotator ends up doing considers the outcome when the event is removed (i.e. the event does not happen). Hence, if the text describes event A preceeding event B, then since it is obvious that B happens when A happens given the context, the question is always "does B still happen when A does not?". This is a counterfactual test.

4 The Corpus

In this section we present the text collection employed for this research along with details about the annotation guidelines.

4.1 Corpus Construction and Processing

Determining a causal relationship between two events A and B necessarily entails first establishing that A temporally precedes B, as it is impossible for a future event to cause one in the present or past. Since poor temporal judgments will most certainly hinder the performance of any causal prediction model, it is crucial to first establish an accurate temporal ordering of any event sequence on which causal predictions are to be made. And since the state of the art in automatic temporal ordering of narrative events has not yet achieved an adequate level of accuracy (Mani et al., 2006; Verhagen et al., 2007; Chambers and Jurafsky, 2008; Bethard and Martin, 2008), one of our goals was to find a large source of online text describing events in an already temporally ordered fashion. We achieved this goal by utilizing online screenplays.

Table 1. Portion of a prototypical screenplay

INT. NORTH KOREAN CONSULATE - HALLWAY - DAY

Joy opens the bathroom door. Sam is standing there, grinning.

JOY
There are six bathrooms in this house, Sam.

SAM
(fanning the air)
But only one with a smoking section.

She quickly closes the door behind her. Sam laughs.

Figure 1 shows a small portion of a prototypical screenplay. Screenplays can be broken up into two major components: action and dialog. As shown in the figure, action and dialog are explicitely separated via the format of the document; usually lines that contain different kinds of information (e.g. actions, dialog, scene breaks) have their own indentation and/or capitalization.

Screenplays are very useful for our purposes for two main reasons: (1) scene breaks are clearly marked, and thus it is easy to detect breaks in event sequences and (2), actions are consistently written in present tense. Thus, identifying the temporal order of textual events becomes trivial since the temporal order and event story order are usually equivalent.

Our corpus was generated from 173 screenplays downloaded from the internet[1]. Camera instructions, character dialog, and other non-action text was removed and scene breaks were used to separate the remaining action text. The

[1] The screenplays were extracted primarily from `joblo.com`.

resulting corpus consists of 2,554,364 word tokens describing textual actions with very confident temporal ordering. We then part-of-speech (POS) tagged (Roth and Zelenko, 1998) the words in the text collection thus acquired to identify and extract events (verbs).

4.2 Corpus Annotation

According to the definitions provided in Section 3 we generated a set of annotation guidelines. These guidelines state to annotate an example as causal if and only if the following conditions hold:

(i) Each event must be represented in text as one or more consecutive verbs.
(ii) Each event must refer to separate distinct events in context.
(iii) The events must occur either at the same time or in the order in which they were written.
(iv) When as much about the world as possible is held constant, manipulating the existence or manner of the first event must *predictably* and *unavoidably* manipulate the existence or manner of the second event.

We used two annotators who received the guidelines prior to annotation. Then, for each event pair they were provided with consecutive sentences containing the pair throughout the corpus and were asked to annotate each example as *Yes, No*, or *ERR*. *Yes/No* capture causal/non-causal relations. *ERR* captures instances which are not valid. By *valid instances*, we mean bigram instances which actually describe two unique events occurring in order. Possible reasons of invalid instances include: (a) the events do not occur in temporal sequence, (b) the two verbs do not describe distinct events, (c) there is a verbal event (that the POS tagger missed) written between the two marked events (i.e. they are not orthographically adjacent), (d) one or both of the marked events are mistagged as verbs by the POS tagger.

Examples of invalid instances are shown in situations (2) and (3) below.

(2) *John* **looked** *at Mary with sore eyes. They* **broke** *up five years ago but John always wanted a second chance.*

While such a case would be *extremely* rare in the corpus, the events are in past tense breaking the assumed temporal ordering of events. Hence the bigram instance *look→break* would not be considered when calculating the observed causal frequency of the bigram in general.

(3) *The bomb exploded,* **sending** *John* **flying**.

This is a much more common case in the corpus. This is an error however because the two highlighted verbs do not represent two distinct events – it is not the case that a *flying* event occurred just after a *sending* event. Hence this bigram instance of *send→fly* would not be considered during evaluation.

The following examples show event bigram instances from the corpus coupled with our annotation judgments and explanations.

(4) *It* **explodes** *with incredible force,* **sending** *dead bodies in all directions.* – **CAUSAL**

`Explanation:` The explosion occurs before the sending event. And if one could control whether the explosion happens, could also predictably control whether the sending of dead bodies happens.

(5) *Rudy* **picks** *up the phone and* **dials** *a number.* – **CAUSAL**
`Explanation:` The critical state of affairs to be aware of here is Rudy's implied intent to dial the number. Keeping that and everything else possible constant, one could control whether the dialing event happens by controlling whether the picking-up event happens.

(6) *As he* **says** *this, he* **holds up** *his right arm.* – **NOT CAUSAL**
`Explanation:` The saying and the holding up events happen simultaneously (which is fine), but independent of each other.

5 Model Description

We model the ordering of consecutive events by calculating bigram frequencies. From our corpus, we extracted 328,035 event instances defined over 4,368 unique events and 120,004 unique bigrams (pairs of adjacent verbs - in the same sentence or consecutive sentences). Using only statistics calculated over bigram and event frequencies, we then calculate a measure we call *Causal Potential*.

The causal potential of any two events is a measure which gauges how likely these two events are to be in a causal relationship without prior knowledge of any context. Causal potential ($\mathcal{C}$) is calculated via the following formula:

$$\mathcal{C}(e_1, e_2) = \log\left(\frac{P(e_2|e_1)}{P(e_2)}\right) + \log\left(\frac{P(e_1 \rightarrow e_2)}{P(e_2 \rightarrow e_1)}\right) \quad (1)$$

Our arrow notation ($\rightarrow$) denotes bigrams. Hence $e_i \rightarrow e_j$ denotes the event bigram (e_i, e_j) and $P(e_i \rightarrow e_j)$ is simply the frequency of occurrence of $e_i \rightarrow e_j$ in the corpus. To help avoid 0-probabilities, we adopt a very simple smoothing policy whereby non-existent bigram counts are assigned a frequency of 1.

There are two main intuitions behind our causal potential $\mathcal{C}$. The first term comes from the notion of probabilistic causation which defines it in terms of the causal event's occurrence increasing the probability of the result event (Suppies, 1970; Mellor, 1995). Thus the first term has high values when $P(e_2|e_1) > P(e_2)$ and has low values when $P(e_2|e_1) < P(e_2)$. The second term comes from the basic assumption that causes must preceed effects and thus if two events occur often in causal situations, they should also occur often in temporal order. Hence the second term has high values when they occur more often in order, and has lower values the more often they occur out of order. Satisfying both of these intuitions results in high values of $\mathcal{C}$ while lacking in one or both of them lowers the value of $\mathcal{C}$.

Our measure of causal potential has a range of $(-\infty, \infty)$. However since we smooth our frequency counts these types of results will never occur in practice (as shown in Figure 1).

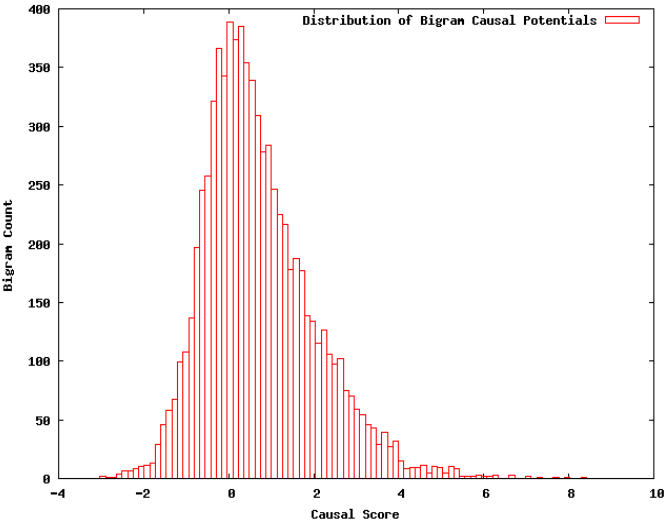

Fig. 1. Distribution of causal scores in screenplay corpus when $\mathcal{C}$ is calculated for each bigram in our corpus with frequency at least 5

6 Model Evaluation

Every consecutive event–event pair in the screenplay corpus was ranked by the system based on the causal potential model described above. Since our screenplay corpus is too large, we evaluated the causal potential on a sample of the corpus which was selected in the following way:

(a) All bigrams which had at least 5 instances were sorted by their scores ($\mathcal{C}$)
(b) We selected 90 bigrams from this sorted list: the top 30, the bottom 30, and 30 near the middle.
(c) For each of these bigrams we located all of its valid instances in the corpus.
(d) Each bigram instance was annotated in context according to the annotation test described in Section 3.2.
(e) From these annotations, Table 2 records the number of causal instances (column 2), the number of non-causal instances (column 3), and the number of invalid instances (column 4).
(f) *Observed causal frequency* (column 5) is simply column 2 divided by columns 2 and 3 added together.

Table 3 shows positive and negative examples of sentences corresponding to event pairs in Table 2. The average inter-annotator agreement was 85%.

We restricted our evaluation to only bigrams with at least 5 instances to ensure that observed causal frequencies were calculated over a decent number of instances (more than 5 instances is a rare situation among bigrams).

Observed causal frequency is an estimate of the probability that two events will be in a causal relationship without prior knowledge of context. Table 2

Table 2. Bigrams' Causal Potentials vs. Observed Causal Frequencies. From left to right the columns are: *event bigram*, *number of times bigram instance was observed to represent a causal relationship*, *number of times bigram instance was observed to not represent a causal relationship*, *number of times bigram instances were extracted in error*, *calculated observed causal frequency (the second column divided by the sum of the second and third columns)*, *calculated causal potential* ($\mathcal{C}$).

Bigram	# Causal	# Non-causal	Error	Causal Freq.	$\mathcal{C}$
send → reel	0	0	11	N/A	8.358
offer → refuse	7	0	0	1.00	8.012
send → fly	0	0	39	N/A	7.799
send → sprawl	0	0	11	N/A	7.389
swerve → avoid	2	0	4	1.00	7.137
wear → tailor	0	0	6	N/A	6.725
round → bend	0	0	8	N/A	6.661
give → peck	0	0	6	N/A	6.368
send → crash	2	0	7	1.00	6.238
leave → strand	0	0	7	N/A	6.201
put → gear	0	0	10	N/A	6.075
explode → send	11	0	0	1.00	5.992
lean → kiss	40	1	0	0.98	5.965
pick → dial	32	1	0	0.97	5.946
try → lock	0	0	26	N/A	5.857
sound → echo	2	0	4	1.00	5.791
stumble → fall	19	0	1	1.00	5.721
open → reveal	88	0	2	1.00	5.636
swing → connect	7	0	0	1.00	5.564
unlock → open	14	1	3	0.93	5.468
cry → sob	6	0	0	1.00	5.437
sit → nurse	0	7	0	0.00	5.427
seize → drag	8	0	0	1.00	5.399
nod → satisfy	3	0	4	1.00	5.397
hit → send	13	1	1	0.93	5.393
hear → creak	0	0	7	N/A	5.368
scan → spot	6	0	0	1.00	5.364
kick → send	11	0	0	1.00	5.357
raise → aim	8	0	1	1.00	5.323
aim → fire	0	4	3	0.00	5.308
play → go	1	10	2	0.09	0.011
see → shove	1	6	0	0.14	0.010
face → see	10	5	34	0.67	0.010
sit → hear	0	29	0	0.00	0.010
dress → get	1	3	6	0.25	0.010
pull → look	7	15	2	0.32	0.009
roll → fall	2	2	2	0.50	0.009
get → appear	2	6	0	0.33	0.007
slap → look	2	5	3	0.29	0.007
roll → look	2	22	3	0.08	0.007
put → face	0	3	5	0.00	0.006
face → run	0	1	7	0.00	0.002
look → loom	0	8	0	0.00	-0.000
smile → move	0	8	0	0.00	-0.000
play → hear	0	7	2	0.00	-0.001
point → stand	0	8	4	0.00	-0.001
see → fire	2	5	5	0.29	-0.001
punch → look	1	4	2	0.20	-0.001
sit → remain	0	12	2	0.00	-0.003
fall → kick	1	4	1	0.20	-0.003
change → see	3	1	0	0.75	-0.003
go → jump	1	6	1	0.14	-0.005
hang → enter	0	9	1	0.00	-0.005
lead → walk	9	7	1	0.56	-0.005
pull → face	3	4	7	0.43	-0.007
say → hold	2	6	3	0.33	-0.008
pull → emerge	4	4	1	0.50	-0.008
turn → lie	3	13	2	0.19	-0.008
throw → look	10	18	11	0.36	-0.008
look → slap	3	5	2	0.38	-0.010
give → stand	0	7	4	0.00	-1.994
jump → look	1	4	2	0.20	-2.012
sit → fall	0	6	2	0.00	-2.014
shake → sit	0	7	0	0.00	-2.120
look → stop	1	12	3	0.08	-2.026
pick → sit	2	5	3	0.29	-2.036
stare → see	3	6	1	0.33	-2.040
take → enter	0	10	2	0.00	-2.076
look → lock	3	1	2	0.75	-2.102
listen → sit	0	6	0	0.00	-2.106
find → turn	3	4	4	0.43	-2.115
lead → come	2	4	0	0.33	-2.149
take → appear	1	4	1	0.20	-2.155
come → wait	7	2	1	0.78	-2.193
see → open	4	15	12	0.21	-2.217
move → nod	3	3	2	0.50	-2.243
reveal → pull	1	6	3	0.14	-2.253
open → wait	0	5	1	0.00	-2.319
lean → sit	0	7	0	0.00	-2.332
stand → reveal	1	7	2	0.13	-2.334
wear → see	3	4	4	0.43	-2.341
pass → walk	0	7	1	0.00	-2.367
know → look	4	4	11	0.50	-2.447
look → play	0	8	2	0.00	-2.517
turn → read	1	4	1	0.20	-2.556
open → watch	0	5	1	0.00	-2.570
wear → come	0	5	2	0.00	-2.691
stare → stand	0	6	1	0.00	-2.720
reach → walk	3	3	1	0.50	-2.902
enter → open	6	4	2	0.60	-2.948

Table 3. Positive and negative examples of event pairs corresponding to those in Table 2

Examples	Annotation
offer→refuse	
A member of the crew enters carrying a tray on which there is a half-filled glass of liquor, which Sandro takes and <offers> to Anna. Anna positively <refuses> it, and the sailor leaves as Sandro sets the glass down on a shelf.	Yes
explode→send	
Sykes is just past the cars when they <explode> – <sending> hoods and door pane and glass flying in all directions.	Yes
hit→send	
Bits of metal fall, <hit> the fan and <are sent> clanging off into space	No
hit→send	
Brody recoils in horror as the beast rushes past, he spins the wheel and <hits> the throttle, <sending> the launch hard to port.	Yes
pull→look	
Carmen <pulls> Johnny to a stop, <looks> him in the eye.	Yes
pull→look	
A bus <pulls> up to the bus stop. The black woman <looks> down at her watch.	No
lead→walk	
Hilary <leads> the line of vets toward the large anti-Vietnam war rally. The group of vets <walk> as Forrest tries to take another picture.	Yes
lead→walk	
The soldier <leads> the German Prisoner away. Maximus and Marcus <continue walking> in silence for a beat.	No
jump→look	
She <jumps> through the narrow opening as Han and Chewbacca <look> on in amazement.	Yes
jump→look	
Startled, Sid and Beth <jump> back. They <look> at each other and laugh.	No
see→open	
She <can't> see anything. She <throws open> the closet door.	Yes
see→open	
As Epps holds there she <sees> the vent <opening>.	No

shows that bigrams with high causal potential very often have high observed causal frequencies, while bigrams with low causal potential scores very often have low observed causal frequencies.

We used Spearman's rank correlation coefficient to measure the degree of correlation between the observed causal frequency and the calculated causal potential. Spearman's rank correlation coefficient is defined as:

$$\rho = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}} \tag{2}$$

where x_i and y_i are rankings of two lists. In our case x_i is the ranking of causal frequencies and y_i is the ranking of causal potentials. Spearman's rank correlation coefficient has a range of $[-1, 1]$. A coefficient of -1 corresponds to the two lists being perfectly uncorrelated (one is the reverse sort of the other), a coefficient of 1 corresponds to perfect correlation (the rankings of both lists are identical), and a coefficient of 0 for rankings being completely independent. The Spearman rank correlation coefficient between observed causal frequency and our measure of causal potential is $\rho = 0.497$.

7 Discussion

Our results show that our notion of causal potential is highly correlated with observed causal frequency; Spearman's rank correlation coefficient verifies that the observed ranking and the ranking predicted by our measure of causal potential are positively correlated. Bigrams which score very high have very high observed causal frequencies and those which score very low have very low observed causal frequencies.

Due to aforementioned issues of validity, some of the bigrams are not very reliable. These mostly correspond to bigrams in Table 2 which have a very low number of occurrences or a very high relative number of errors. This low occurrence rate/high error rate is the result of the system's simple approach to event extraction; in some cases, bigrams were not counted as they did not satisfy criterion (i) of our causal annotation test: the two events must occur in temporal sequence. While the nature of the corpus ensures event precedence most of the time, simple identification of part-of-speech tags is not enough to ensure event sequence. For example, some events occurring at the top of the list of causal potentials were: *send* → *sprawl* (c.f. *sends him sprawling*), and *realize/understand/believe* → *happen* (c.f. *realize/understand/believe what's happening*). The problem here is that simple part-of-speech patterns cannot capture the syntactic structure of the phrases and thus falsely extract numerous cases of subordinate clauses. The solution is to recognize structure in the corpus and extract accordingly. Such a solution is tractable and will be implemented in future work.

Another problem with our event extraction technique lies in the shortcomings of part-of-speech taggers. For example, here are a few bigrams extracted which also have a high causal potential: *wear* → *tailor* (c.f. *wears a tailored jacket*), and *leave* → *strand* (c.f. *leaves him stranded*). These are cases where a verb past participle is acting as a noun or adverb modifier. Similar problems can arise with verb gerund forms. In the future we will improve our tag patterns to account for the different verb forms instead of treating them all alike.

8 Conclusion

In this paper we described a knowledge-poor unsupervised causal event model which relies on a statistical measure we call *causal potential*. Causal potential can be easily calculated with simple statistics gathered from a corpus of temporally ordered event pairs. We have empirically shown that event pairs with a high causal potential are more likely to occur in causal contexts than events with a low causal potential and that events with a low causal potential are likely to not occur in causal contexts. This behavior lends our measure of causal potential to be used in cases where context is either absent or unreliable, to gain intuitions regarding the likelihood of two events to be causally related. Moreover, we argue that our model should therefore be used as a baseline for standard supervised models which take into account contextual information.

References

Bethard, S., Martin, J.: Learning semantic links from a corpus of parallel temporal and causal relations. In: The Human Language Technology (HLT) Conference, Short Papers, pp. 177–180 (2008)

Chambers, N., Jurafsky, D.: Unsupervised learning of narrative event chains. In: Human Language Technology Conference, pp. 789–797 (2008)

Chang, D.-S., Choi, K.-S.: Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Information Processing and Management 42(3), 662–678 (2006)

Girju, R.: Automatic detection of causal relations for question answering. In: The 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond (2003)

Goldman, S., Graesser, A., Broek, P.: Narrative comprehension, causality, and coherence: Essays in Honor of Tom Trabasso. Erlbaum Associates, Mahwah (1999)

Hoover, K.: The New Classical Macroeconomics. Basil Blackwell, Oxford (1988)

Khoo, C., Myaeng, S., Oddy, R.: Using cause-effect relations in text to improve information retrieval precision. Information Processing and Management 37, 119–145 (2001)

Mani, I., Verhagen, M., Wellner, B., Lee, C.M., Pustejovsky, J.: Machine learning of temporal relations. In: The Association for Computer Linguistics (ACL) Conference (2006)

Mellor, D.H.: The Facts of Causation. Routledge (1995)

Menzies, P.: Counterfactual theories of causation. The Online Stanford Encyclopedia of Philosophy (2008)

Orcutt, G.: Actions, consequences, and causal relations. Review of Economies and Statistics 34, 305–313 (1952)

Roth, D., Zelenko, D.: Part of speech tagging using a network of linear separators. In: The Association for Computational Linguistics Conference (1998)

Shibatani, M.: The grammar of causative constructions: A conspectus. In: Shibatani, M. (ed.) Syntax and Semantics. The Grammar of Causative Constructions, vol. 6. Academic Press, London (1976)

Sosa, E., Tooley, M. (eds.): Causation (Oxford Readings in Philosophy). Oxford University Press, Oxford (1993)

Suppies, P.: A Probabilistic Theory of Causality. North-Holland Publishing Company, Amsterdam (1970)

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: Semeval-2007 Task 15: Tempeval temporal relation identification. In: the Fourth International Workshop on Semantic Evaluations (SemEval 2007), pp. 75–80. Association for Computational Linguistics (2007)

Woodward, J.: Causation and manipulability. In: The Online Stanford Encyclopedia of Philosophy (2008)