Extracting contingent event pairs from Travel and Sports stories

Abhinav Venkataraman
abhinav@soe.ucsc.edu

Abstract

Script learning is one of the most challenging research topic in the field of NLP and many researchers have been tackling this problem by solving different sub domains of the same. One of such important sub domain is extraction of CONTINGENT relations between event pairs i.e deciding whether an event caused another event. There are different contingency measures in establishing this relation and this paper involves building different variations of a bigram language model for a variety of representations (Manshadi et al., 2008) and calculating different contingency measures as used in previous works (Chao et al., 2013). The quality of these measures and the extracted event pairs are evaluated based on experiments like Discriminate, Narrative Cloze and HIT jobs. Results show that an mixture of likelihood of adjacent events and events spanning over more than two events perform better than just the stand alone adjacent event measures. Also, different smoothening techniques can actually improve the performance of any contingency measure over the standard approach of adding one count to an unseen event pair. Interestingly enough, the result also shows how weak narrative cloze as an evaluation measure is.

1 Introduction

A narrative story can be seen as a chain of ordered events. In order to learn them, its very important to understand the internal structure of the stories prevailing on different topics. More generally put, its important to figure out which action and its type characterizes the story belonging to a particular topic. Extracting contingent events is also a prerequisite in many NLP tasks like text coherence, entailment, question answering and information retrieval(Girju and Brandon, 2009). This paper focuses on extracting them,in an unsupervised data driven fashion. Previous work shows good results in finding the contingent event pairs from film scenes by modeling the likelihood between events(Chao et al., 2013).

We feel trying to extract such causal chains from travel and sport may provide for better common sense reasoning and help in script learning for common scenarios about which people often blog. Also trying out different event representations and determining the likelihood of events based on different smoothening techniques could help yield better quality and understanding of which components contribute to the discovery of the CONTINGENT relation between events.

2 Related Works

Girju et al. focused on giving a statistical measure to the events that are in causal relationship by defining causal potential(Girju and Brandon, 2009). Chambers et al. defined such events with causal relationships as narrative event chains and came out with different ways of learning (Chambers and

Jurafsky, 2008). They also build a new method of identifying event semantics that jointly learns event relations and their participants from unlabeled corpora(Chambers and Jurafsky, 2009). Chiacros solved the same problem by providing a method for identifying a discourse connective between different utterances in text(Chiarcos, 2012). Manshadi et al. tried to solve a similar problem by learning a probabilistic model of event sequences using statistical language modeling techniques(Manshadi et al., 2008). Quang Xuan Do et al. followed Chiacros methodology of solving similar problem by feeding discourse connectives and the particular discourse relation in addition to the distributional similarity to identify causal relations between events(Do et al., 2011)

Our hypothesis and methodology is mostly a hybrid of Chao et al. and Manshadi et al. with certain differences. We create a variety of different event representations and bigram language models with different smoothening techniques. There is also a skipped language model thats built to capture CONTINGENT events spanning over more than two events. A mixture of both standard and skipped model is created and different possible computation of the above mentioned contingency measures are evaluated for sports and travel stories. We believe that the underlying assumption of temporal coherence and contingency do exist in these domains.

3 Data

A sample story from our corpus:

> This morning my dad, uncle and I went for a short hike around the Short Hill. I had long known that part of the mountain was part of Harpers Ferry National Park, but only last week discovered access to the site . I told my dad about it and he was interested to check it out. Despite the cold drizzley weather we set out, being careful to not injure our selfs stepping into deep piles of leaves as I had the last time I visited the site. Almost immediately we encountered the ruins of an old lime kiln, which sadly I neglected to take a picture. A little further up the trail we came to the ruins of the old River Mill. We then made our way around the Short Hill on an old road bed, catching good views across and up river. That last pic was taken from a 60ft bluff above the river, and though you can't make it out in the picture, good views of Harpers Ferry 5 miles distant were had. Though the hike was short and we only stuck to the shore and did not attempt the mountain it was enjoyable and left me impatient for the spring when i can mount a full expedition of the area. Definitely going to need a machete because the undergrowth is going to be bad, good shoes and a hiking stick will also be necessary on the steep ungraded slopes. Check out my facebook for a few pics from the hike.

The data consists of a number of blog stories divided into two major domains: travel and sports. These stories were taken from The Internet Personal Story Archive(Reid,, 2007) which is an archive of blog posts scraped from the internet. Stories within each domain are also sub categorized into a number of fine grained topics like hiking, skiing, scuba diving etc. for travel and cricket, soccer, swimming for sports. These two domains were chosen because we think these two are characterized by a set of events that occur similarly in the experiences of different people. The corpus used consists of 353 stories for sports and 444 for travel.

4 Event Representations

There are different ways in which an event can be represented. At the core it is an action- a verb. There are other representations which give more information about the action which are : *a*) Just the verb *b*) Verb + Subject+Object *c*) Verb + Subject *d*) Verb + Object.

As an example the following sentence can be seen in any of the 4 forms:

Ronaldo kicked the ball.

a) **Verb + Subj + Obj** : *kick, Ronaldo, ball.*

b) **Verb + Subj** : *kick, Ronaldo*

c) **Verb + Obj** : *kick, ball.*

d) **Verb** : *kick*

Different representations might be helpful in different domains and this why we test them out for both the domains and see which representation is more informative in that context.

So the corpus is converted to a form where each document is represented as a space delimited sequence of event representation. And different documents are delimited by a line break.
So for an example document:

> *John Doe opened the box. He ate the chocolates.*

The above document would be represented in the following way in the different representations:

a) **Verb + Subj + Obj**: open|PERSON|box eat|PERSON|chocolates.

b) **Verb + Subj** : open|PERSON eat|PERSON.

c) **Verb + Obj** : open|box eat|chocolates.

d) **Verb** : open eat.

5 Contingency Measures

In this paper the simplest event representation is considered, which is just the action - verb and different contingency measures were computed which are explained below.

Probabilistic Language Models. We build statistical language model using the events where we represent each document as a sequence of events(Manshadi et al., 2008). We compute the bigram probabilities of verbs that occur in the document and it is defined as :

$$p(w_i|w_{i-n+1}^{i-1}) = \begin{cases} p(w_i|w_{i-n+1}^{i-1}) \\ \quad \text{if } count(w_{i-n+1}^{i}) \geq 1 \\ \\ p(w_i|w_{i-n+2}^{i-1}).bow(w_{i-1}^{i-1n+1}), \\ \quad \text{otherwise} \end{cases} \quad (1)$$

These bigram probabilities are computed using SRI toolkit and it uses a back off model which is $bow(w_{i-1}^{i-1n+1})$ for discounting unseen higher order probabilities with lower order ones as per equation (1).

There are many smoothening techniques available in the toolkit which is used to compute these unseen higher order probabilities and the ones considered for this paper are :

- Additive Smoothing(Add 1 smoothing)
- Good Turing estimate
- Kneser-Ney Smoothing

Additive smoothing is the simplest and traditional type of smoothing and add 1 smoothing is used in (Chao et al., 2013). In this kind of smoothing it is pretended that an event pair occurs once more than it actually does. This type smoothing generally performs poorly which is discussed more in results section.
Good Turing estimates is central smoothing technique do it is basically a way of performing normalization to bigram probabilities(normal likelihood) by finding a *discount ratio*. Good Turing estimates fail because it does not include the combination of higher-order models with lower-order models necessary for good performance.
Kneser-Ney Smoothing is an extension of absolute discounting where the lower order distribution and the higher order distribution are combined in a unique way and this is a standard example of back off model.
The probability computed by 1 is then used to calculate other contingency measures.

Causal Potential. A measure for causality was proposed by (Girju and Brandon, 2009). They define a manipulation test for annotators trying to judge if event A caused event B.

(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?

Answering yes to both these questions would imply causality.
CP is thus defined as :

$$\phi(e_1, e_2) = pmi(e_1, e_2) + \log \frac{p(e_1 \rightarrow e_2)}{p(e_2 \rightarrow e_1)} \quad (2)$$

Here $pmi(e_1, e_2)$ is defined as *Principal Mutual Information*PMI which is a symmetric measure of two events occurring adjacent to each other in a document and it is defined as:

$$pmi(e_1, e_2) = \log \frac{p(e_1, e_2)}{p(e_1)p(e_2)} \quad (3)$$

in which e_1, e_2 are two events. $p(e_1)$ is the probability that event e_1 occur in corpus which is determined by equation (1).

In addition to this,one skip languages models are also built. The basic intuition behind this approach is to capture the CONTINGENT event pairs that span over more than one events. For eg, if a document is represented as a series of events e_1, e_2, e_3, e_4 and e_5 then we create a model considering e_1, e_3, e_5 as a sequence of events and e_2, e_4 as a sequence of another set of events. A language model following the this structure of event representation is created. We interpolate this model with the language model that was defined above for normal bi gram probabilities.
Given a pair of events e_1 and e_2 present in both original language model and skipped model. The interpolated model is represented as :

$$\lambda p_{normal}(e_1, e_2) + (1 - \lambda) p_{skipped}(e_1, e_2) \quad (4)$$

We tune the value for λ for each of the different types of event representation over a development set.

All words in the test data that were not present in the vocabulary of the language models were mapped to the symbol *unk* which is already present in the language model and the probability is calculated based on that.

6 Experiment set up and Evaluation

In this paper, only one event representation is being focused and the experiments are run for only verbs alone event representation.

6.1 Evaluation

The major tasks used for evaluation are Discriminate, Narrative Cloze and Human Intelligence Tasks. We also use perplexity as an initial measure to find an estimate of how good the language models are.

6.1.1 Perplexity

Perplexity is one of the ways to evaluate language models over test data. It shows how surprised is the model in seeing a new event. Perplexity of a sentence is denoted by :

$$PP_p(S) = 2^{H_p(S)} \quad (5)$$

where $H_p(S)$ is the cross entropy and its defined as :

$$H_p(S) = \frac{-1}{W_s} * \log_2 p(S) \quad (6)$$

where W_s is the number of words in the sentence in other words number of bits required to encode the sentence.

6.1.2 Discriminate

In this task, random permutations of the events of a document are generated and it is seen if the language models assign a higher score to the random permutations or the original ordering of events. If the score of the original order of events are higher than the random permutations of events then its determined as a win otherwise its treated as a loss. In case the scores are the same, then its a tie :
For each contingency measure, the desired property is :

- $\Sigma_{i=1}^{n-1} Org\_Order\ \phi(e_i, e_{i+1}) > \Sigma_{i=1}^{n-1} Permuted\_Order\ \phi(e_i, e_{i+1})$
- $\Sigma_{i=1}^{n-1} Org\_Order\ \log p(e_i, e_{i+1}) > \Sigma_{i=1}^{n-1} Permuted\_Order\ \log p(e_i, e_{i+1})$

The goal of this evaluation is to compare the different contingency measures and different language models to see combination performs the best.

6.1.3 Narrative Cloze

The cloze task(Taylor,, 1953) is used to evaluate a system (or human) for language proficiency by removing a random word from a sentence and having the system attempt to fill in the blank . Depending on the type of word removed, the test can evaluate syntactic knowledge as well as semantic(Chambers and Jurafsky, 2008). Similarly, there is an event which is plucked out at each position in the document and it is rated by each of the contingency measure. The position rated best is taken to be the predicted position and we calculate the value delta (for each measure) as,

$$\delta = |actual\_position - predicted\_position| \quad (7)$$

This delta value is averaged over the number of events in the corpus to get the Mean positional value per event and Mean positional value per document is calculate as :

$$mean\_pos\_score = \frac{\sum_{i \in docs} avg_i}{n} \quad (8)$$

where n is the number of documents and avg_i is defined as :

$$avg_i = \frac{\sum_{j \in events(i)} \delta_j}{length(i)} \quad (9)$$

6.2 Experiment set up

We use Stanford CoreNLP(Manning et al., 2014) toolkit to get the annotations. We feed into the parser the entire corpus of stories and yield corresponding XML files which contain necessary annotations. We parse the XML output to get the verbs in their lemmatized form, which are actions and also extract the subject and object of the verb from the dependency parse. We are trying to use the SRILM toolkit for getting the necessary counts and probabilities for each of these measures(Stolcke,).

The data in both the domains was divided in three parts: train (85%), development (5%) and test (10%). So for travel we have 378 stories for train, 44 for testing and 22 for development. And sports is split into 300 train, 35 test and 17 for development.

7 Results and Discussion

We tuned the mixture model which is a linear combination of skip and the original model with the help of 4. We ran them with λ values ranging from 0 to 1 with a *step size* of 0.01 to figure the best possible value for λ. The λ values change for each representation and domain. For travel, the best possible λ value was in the range of $0.77 - 0.82$ and for sports it was in the range of $0.75 - 0.85$. After tuning the mixture mode with the development set, we run the discriminate task on the original model as well as the mixture model for all the event representations on the with held test test. The tables below show the discriminate and cloze results of all the experiments.

7.1 Perplexity Results

Representation	Perplexity
V + S + O	56.0863
V+ S	132.118
V+ O	72.486
V	87.374

Table 1: Initial perplexity results for Travel(67 stories)

Representation	Perplexity
V + S + O	54.9574
V+ S	141.279
V+ O	96.4696
V	95.1299

Table 2: Initial perplexity results for Sports(53 stories)

7.2 Discriminate Results

Model	Representation	CP				logp			
		Wins	Losses	Ties	Win %	Wins	Losses	Ties	Win %
Original	V+S+O	164	56	5	75.11	178	42	5	**81.3**
	V+S	184	41	0	81.7	199	26	0	**88.4**
	V+O	194	24	6	88.8	202	17	6	**92.4**
	V	186	39	0	**82.7**	176	49	0	78.2
Combined	V+S+O	164	56	5	76	171	49	5	**78.2**
	V+S	183	49	0	**81.3**	176	42	0	78.2
	V+O	184	36	5	84.9	192	28	5	**87.6**
	V	169	56	0	**75.1**	155	70	0	68.9

Table 3: Discriminate results for travel stories(225 samples)

Model	Representation	CP				logp			
		Wins	Losses	Ties	Win %	Wins	Losses	Ties	Win %
Original	V+S+O	140	40	0	77.8	161	19	0	**89.4**
	V+S	158	22	0	87.8	168	12	0	**93.3**
	V+O	158	22	0	87.8	164	16	0	**91.1**
	V	153	27	0	85	165	15	0	**91.7**
Combined	V+S+O	134	46	0	74.4	157	23	0	**87.2**
	V+S	143	37	0	79.4	156	24	0	**86.7**
	V+O	151	29	0	83.9	157	23	0	**87.2**
	V	152	28	0	84.4	157	23	0	**87.2**

Table 4: Discriminate results for sports stories(180 samples)

It is very evident from the table that the discriminate results of $\log p$ does better than CP. For travel, its interesting to see that CP for verbs does better than $\log p$ where as it relatively does bad when compared to $\log p$ for other representations. This could be because since there are a lot of *unks*'s mapped in the representation which shows how sparse the representations are and that is impacts the difference in performance of discriminate.
For sports, the $\log p$ simply outperforms CP in every representation which clearly shows CP is weaker for sports.

7.3 Cloze Results

We ran the cloze task from with the help of probabilities calculated for both mixture and original model.

Model	Representation	CP		PMI		logp	
		Mean by line	Mean by event	Mean by line	Mean by event	Mean by line	Mean by event
Original	V+S+O	9.775	12.668	9.589	12.680	9.727	12.838
	V+S	9.535	11.909	**8.619**	10.645	9.224	11.539
	V+O	10.332	12.180	9.258	10.789	9.592	11.169
	V	9.708	10.791	9.582	10.846	**8.749**	9.599
Combined	V+S+O	9.250	11.840	9.199	11.961	8.996	11.678
	V+S	9.653	12.069	**8.314**	10.190	8.936	11.005
	V+O	10.275	11.961	9.610	11.145	9.315	10.708
	V	8.851	9.877	8.665	9.794	**8.586**	**9.415**

Table 5: Cloze results for travel stories

Model	Representation	CP		PMI		logp	
		Mean by line	Mean by event	Mean by line	Mean by event	Mean by line	Mean by event
Original	V+S+O	12.789	13.373	11.796	12.413	13.983	15.153
	V+S	13.351	14.555	12.810	13.905	13.320	14.207
	V+O	12.869	13.462	12.925	13.573	12.057	12.951
	V	11.050	11.870	9.259	9.755	9.477	10.299
Combined	V+S+O	13.515	14.22	12.664	13.75	14.274	15.196
	V+S	11.962	12.732	10.696	11.333	11.588	12.284
	V+O	12.488	13.088	11.853	12.333	10.758	11.373
	V	10.225	10.944	10.204	10.842	**9.246**	**9.818**

Table 6: Cloze results for sports stories

Verbs along with $\log p$ does best at performing the cloze task irrespective of the domain and CP performs the worst.

One common inference from both the tasks is that CP does not perform better than $\log p$ and Verbs in general has the highest expressive power, this could be because others get sparse since we tend to capture more information.

8 Future Work

There are many variants of the task to ascertain some of the inferences made in the paper. One is to set up a Human Intelligence Task on Mechanical Turk. This would actually validate our results. As seen in the (Chao et al., 2013) paper, the verbs alone representation did the best on MTurk and this matches to our inferences but we would need to validate it.

The other important task to do is to run the same procedure with more data, which is to get much cleaner data by using the patterns generated from AutoSlog(Riloff, 1999). Doing such high precision extraction of stories would help us in getting to know more insight about the representations.

References

Beamer, Brandon, and Roxana Girju,. 2009. *Using a bigram event model to predict causal potential*, Computational Linguistics and Intelligent Text Processing Springer Berlin Heidelberg, 2009. 430-441

Chambers, Nathanael, and Daniel Jurafsky, 2008. *Unsupervised Learning of Narrative Event Chains.* ACL. Vol. 94305.

Chambers, Nathanael, and Dan Jurafsky, 2009 *Unsupervised learning of narrative schemas and their participants* In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP Volume 2-Volume 2, pp. 602-610,Association for Computational Linguistics

Chiarcos, Christian., 2012, *Towards the unsupervised acquisition of discourse relations*, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics.

Manshadi, Mehdi, Reid Swanson, and Andrew S. Gordon, 2008, *Learning a Probabilistic Model of Event Sequences from Internet Weblog Stories*, FLAIRS Conference. 2008.

Do, Quang Xuan, Yee Seng Chan, and Dan Roth. 2011, *Minimally supervised event causality identification*, Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Hu, Zhichao and Rahimtoroghi, Elahe and Munishkina, Larissa and Swanson, Reid and Walker, Marilyn A.,, October, 2013. *Unsupervised Induction of Contingent Event Pairs from Film Scenes.* In Conference on Empirical Methods in Natural Language Processing, Seattle, WA

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

Reid Swanson, 2007, *First Person Narrative Story Extraction and Retrieval. Masters*, University of Southern California.

Manning, Christopher D. and Surdeanu, Mihai and Bauer, John and Finkel, Jenny and Bethard, Steven J. and McClosky, David, June 2014. *The Stanford CoreNLP Natural Language Processing Toolkit*, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland

A. Stolcke, *SRILM – An Extensible Language Modeling Toolkit*, Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver.

Taylor, Wilson L, *"Cloze procedure": a new tool for measuring readability*, Journalism quarterly.

Riloff, Ellen., 1999, *"Information extraction as a stepping stone toward story understanding."*, Understanding language understanding: Computational models of reading,435-460.