Extracting contingent event pairs in blog corpus

Abhinav Venkataraman
abhinav@soe.ucsc.edu

Keshav Mathur
kemathur@ucsc.edu

Abstract

We implement and evaluate different unsupervised methods for learning event pairs that are likely to be CONTINGENT on one another(Chao et al., 2013). We build language model based on SRILM toolkit rather than pure Manshadi's approach. We compare and contrast different techniques for those data whose temporal coherence are yet to be proved. We would evaluate these models using MTurks, Narrative Cloze and a novel and unique Discriminative method.

1 Introduction

Any narrative story can be seen as a chain of ordered set of events. In this project, we would like to focus on extracting the event pairs which are contingent on each other, in an unsupervised data driven fashion. It would be nice to try to understand more about the internal structure of the different types of stories, basically to find out what kinds of action sequences characterize them. Extracting contingent events is a prerequisite in many NLP tasks like text coherence, entailment, question answering and information retrieval(giriju).Previous work shows good results in finding the contingent event pairs from film scenes by modeling the likelihood between events[chao].

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

2 Related Works

Giru et al. focussed 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 discource connective between different utterences 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 distirbutional similarity to identify causal relations between events(Do et al., 2011)

Our hypothesis and methodology is completely based from Chao et al.(Chao et al., 2013) by creating the same event representation and evaluate some of the methodologies done by others which are mentioned above and see how they perform for sports and travel stories. We believe that the underlying assumption of temporal coherence do exist in the sports and travel stories.

3 Methods

In this project we want to test different measures of contingency and event representations to extract contingent event pairs from the corpus and try to compare and contrast them. An event can be represented in many forms. The ones we want to test out are: *a*) Just the verb *b*) Verb + Subject+Object *c*) Verb + Subject *d*) Verb + Object *e*) Multi-argument event representation as stated by Mooney (Pichotta and Mooney, 2014)

For each of these representations we will calculate three different measures of contingency used in previous works:

Point wise Mutual Information. It is a symmetric measure of two events occurring adjacent to each other in a document. So this finds the probabilities of two events appearing close to one another but not actually imply causality(Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009).
The PMI between two events is defined as:

$$pmi(e_1, e_2) = \log \frac{P(e_1, e_2}{P(e_1)P(e_2)} \qquad (1)$$

in which e_1, e_2 are two events. $P(e_1)$ is the probability that event e_1 occur in corpus :

$$P(e_1) = \frac{count(e_1)}{\Sigma_x count(e_x)} \qquad (2)$$

The joint probability of both events occuring together($P(e_1, e_2)$) is given by :

$$P(e_1, e_2) = \frac{count(e_1)}{\Sigma_x \Sigma_y count(e_x, e_y)} \qquad (3)$$

Causal Potential. A more refined 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 :

$$\phi(e_1, e_2) = pmi(e_1, e_2) + \log \frac{P(e_1 \to e_2)}{P(e_2 \to e_1)} \qquad (4)$$

where $pmi(e_1, e_2)$ is given by (1).

PMI and Causal Potential are found for events adjacent in a document. We plan to define adjacency as a 2-skip model. If a document is represented as a series of events. e_1, e_2, e_3, e_4 and e_5 then the ordered pairs for which the counts will be generated are: $(e_1, e_2), (e_1, e_3), (e_1, e_4), (e_2, e_3), (e_2, e_4), (e_2, e_5), (e_3, e_4), (e_3, e_5), (e_4, e_5)$.
The event pairs rated highly on these measures will be used for the different evaluation experiments that we define in the coming section.

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

$$P(w_1, w_2) = \frac{count(w_1, w_2)}{count(w_1)} \qquad (5)$$

So for an example document:

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

This sentence is converted to `open,ate`. Using this sequence of verbs we calculate bigram probabilities which might imply contingency between the events.

3.1 Difficulties to be faced

In future we anticipate few difficulties that are to be faced in our methodologies. They are :

- We are trying to use the SRILM toolkit for getting the necessary counts and probabilities for each of these measures. It is going to be really straight forward for building the bigram language model but non-trivial for getting the statistics for PMI and CP as there events are considered over a 2-skip model and we dont yet know how to incorporate that into SRILM. If not possible we will be writing code for for finding the counts and the necessary probabilities.

- The data we have may not not be sufficient for finding the necessary statistics and we may have to annotate more.

- The representation used in (Pichotta and Mooney, 2014) may prove to be complicated and time consuming to implement.

4 Data

A sample story from our corpus looks like this :

> 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 we are using consists of a number of stories divided into two major domains: travel and sports. These stories were taken from The Internet Personal Story Archive(Reid,) which is a collection of blog posts taken 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. At present we have 440 stories for travel blogs and 280 for sports. We have considered 67 stories of travel as our held out test set and built the bigram model on 357 stories. Similarly for sports, 238 for building the language model and 42 for testing. After running the initial experiments on this corpus we will assess if there is a need for more data and if need be, annotate more data into each category.

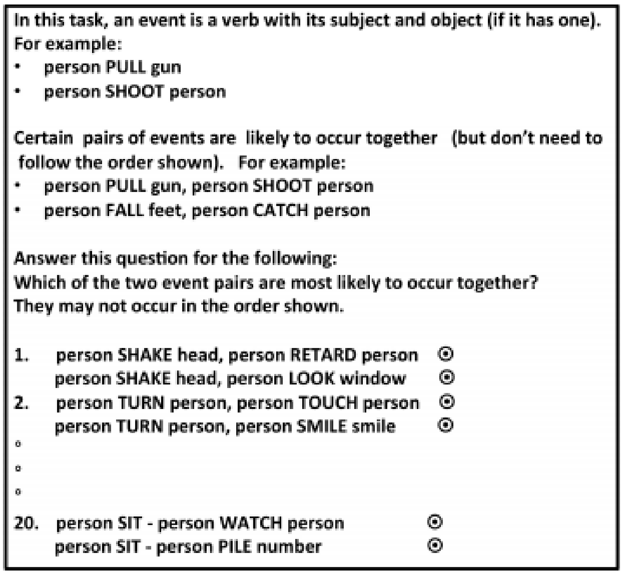
We use Stanford CoreNLP(Manning et al., 2014) toolkit to get the annotations. We feed into the parser the entire corpus of stories and yield corresponding XML files which contain necessary annotations. We parse the XML output to get the verbs in their lemmatized form, which are actions and also extract 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,).

5 Evaluation

We have three evaluations in mind :

5.1 MTurks

We have planned to set up three HITs where each task would have involving the presence and absence of arguments for events and also ordering of events whether it matters or not. We plan to more or less replicate the set up that was mentioned in (Chao et al., 2013). Each HIT would have the task for which the annotator needs to say whether the events given would occur together or not. The same objective is done with order maintained and also done with no preference on the order of events. The other HIT task would to present event with arguments and ask the annotators how likely would given set of events occur together. Sample piece of a format of turk job is given in the picture below.[1]



In this task, an event is a verb with its subject and object (if it has one).
For example:
- person PULL gun
- person SHOOT person

Certain pairs of events are likely to occur together (but don't need to follow the order shown). For example:
- person PULL gun, person SHOOT person
- person FALL feet, person CATCH person

Answer this question for the following:
Which of the two event pairs are most likely to occur together?
They may not occur in the order shown.

1. person SHAKE head, person RETARD person ⊙
   person SHAKE head, person LOOK window ⊙
2. person TURN person, person TOUCH person ⊙
   person TURN person, person SMILE smile ⊙

∘
∘
∘

20. person SIT - person WATCH person ⊙
    person SIT - person PILE number ⊙

5.2 Cloze Task

One event is removed from a narrative sequence and the task for a model is to predict the missing event , its and typed dependency. We plan to implement this to evaluate our models and see how well it performs (Chambers and Jurafsky, 2008).

[1] Obtained from (Chao et al., 2013) paper

5.3 Discriminative Task

The third task that we plan to use for evaluation is the discriminative task. In this task we generate a random permutation of the events of a document and see if the models assign a higher probability to the random permutations or the original ordering of events. To our knowledge this task hasnt been used much for the kind of problem we are tackling. So it will be a good experiment to see how these models perform.

6 Schedules and Responsibilities

6.1 Milestones Achieved

We have successfully built the language models for both categories sports and travel. We represented each document as a sentence of verbs and fed each category into the SRILM toolkit which gave out the bigram model. We then computed the preplexity score for each of them on kndiscount smoothening after discounting the words and here goes the results:

Results of Bigram Model

Test	**Preplexity**
Travel LM on Travel	87.497
Travel LM on Sports	152.325
Sports LM on Sports	117.409
Sports LM on Travel	105.323

Results of Trigram Model

Test	**Preplexity**
Travel LM on Travel	92.6553
Travel LM on Sports	154.198
Sports LM on Sports	121.555
Sports LM on Travel	110.959

Travel LM on Travel means language model built on travel data was tested on travel's test data. Similarly, Travel LM on Sports means the language model built on travel data was tested on sports. Perplexity can be defined as the as the number of guesses you need to make about the next word/verb/event given the context. Therefore, lower the perplexity score, the higher the coherence.It is to be noted that Sports LM on Travel yields lower perplexity in both cases which means Sports LM is not that coherent when compared to travel as Travel LM on sports yields greater perplexity than Travel LM on travel.
In order to accomplish this task the work was split accordingly which is given below:

Keshav:

- Processed the stories using the Stanford CoreNLP pipeline.
- Wrote parsers for the XML output given by StanfordCoreNLP in python.
- Created the text input for the SRILM toolkit in which each line represents all the verbs in a document.

Abhinav :

- Set up the SRILM toolkit
- Compute the verb pair counts for the corpus.
- Built various language models with and without smoothening for bigrams and trigrams.
- Find the perplexity scores.

6.2 Timeline for Future

Although our next steps are evidently explained before, the timeline for the future are :
Keshav:

- Get the statistics for the Verb + Object representation and the Verb + Object + Subject representation.
- Perform the experiments for the discriminative task.

Abhinav:

- Get the statistics for the Subject + Verb representation
- Perform the experiments for the cloze task and set up the HITs on Mechanical turk.

Both of us would work together on mooney's event representation and build the models as discussed in methods section for that representation

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, *The Internet Personal Story Archive.* reid@reidswanson.com.

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.