(Project Proposal)
Extracting contingent event pairs in blog corpus

Abhinav V Venkataraman
abhinav@soe.ucsc.edu

Keshav Mathur
kemathur@ucsc.edu

February 6, 2015

Abstract

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 fashion. It would be 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. Previous work shows good results in finding the contingent event pairs from film scenes by modeling the likelihood between events[1]. We would like to use the same approach for blogs and see how successful we are.

2 Data

We have annotated a number of stories into two major domains: travel and sports. These stories were taken from The Internet Personal Story Archive[2] 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.

3 Tools

We plan to use Stanford CoreNLP[3] to extract the verbs from the text. To get the bigrams probabilities of verb sequences and building the language model we plan to use the SRILM toolkit[4].

4 Methods

First we like to use the Stanford toolkit to do POS tagging and named entity recognition to extract the verbs and also get an event representation

like the one given in [1].
To calculate a value for CONTINGENCY we would use three measures :

- Pairwise Mutual Information(PMI).
- Causal Potential(CP)[5]
- Bigram probability.

We would like to first extract potential causal event pairs from the document using SRIM toolkit to figure out the bigram probability of event pairs(verbs) that are likely to occur.

Using the probabilities measured in SRIM toolkit for an event(verb) pairs we build a hypothesis that verb sequences with higher probability are likely to be contingent on one another.

After extracting the probable event pairs from these probabilities we would like to find the causal potential and PMI between any two events(in the event representation with arguments) i.e this would answer the question Does event A occur before (or simultaneously) with event B (whether event A causes event B).

5 Evaluation

Having found the contingent event pairs from the different measures, we would like to use it for two different problems to measure its quality. First we would like to use the extracted event pairs as features for a binary classification problem which predicts whether the test document belongs to a particular domain or not. In our case whether the document belongs to hiking or not / sports event or not. We would use precision, recall and f-measure to determine the model's performance.

The other method of evaluating these event pairs is by using a narrative cloze metric as mentioned by Chambers & Jurafsky(2009)[6]. In this measure we remove one event from the chain of events and use the model to predict the missing event.

References

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

[2] Reid Swanson, *The Internet Personal Story Archive.* reid@reidswanson.com.

[3] Manning, Christopher D. and Surdeanu, Mihai and Bauer, John and Finkel, Jenny and Bethard, Steven J. and McClosky, David, *The Stanford CoreNLP Natural Language Processing Toolkit*, *Proceedings of 52nd Annual Meeting*

of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland, June 2014.

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

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

[6] Chambers, Nathanael, and Dan Jurafsky, *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, 2009.