Eventful Event Extraction from Wikipedia
Final Report
Chris Schlechty and Kansu Dincer
UW CSE 454 Advanced Internet Systems

Introduction

As part of our Advanced Internet Systems coursework we chose to work on building a system that would employ machine learning techniques on a large body of data to extract temporal information. In this paper we will present our goal, our methods, findings, and conclusions.

Project Goal

There is a large amount of information presented through Wikipedia with many entries containing important dates and events. Our goal was to develop a system to extract temporal information from this large body of data and possibly match this information to a search query subject. We did not aim to develop a front-end to this system but the obvious choice would be to build a web application that allows creating timelines for certain subjects.

Design and Development

Early on into the project we had to identify what we meant by temporal information. We classified time and date expressions into three: absolute, relative, and vague. Those references that are formatted in a calendar system, such as 17 September 1975 or Feb 3rd, 2001 we called absolute expressions. Any expression that refers to an absolute expression or needs an absolute expression in order to be correctly evaluated, such as today or two years later we called a relative expression. The relative expressions that lack the precision to identify a certain date, such as a few months later or several years before we called vague expressions. We chose to begin with absolute expressions and to try to add the relative expressions later on. We chose not to deal with vague expressions at all.

We decided to analyze the data on a sentence by sentence basis to first identify the sentences that contained date and time expressions and then to extract the context of these sentences by performing a parts of speech analysis.

The data that we had was a slightly preprocessed version of the actual Wikipedia entries as they were edited by the authors. Although this marked up version contained some useful information such as hyperlinks and tables, we chose to strip away these markups and other Wikipedia-specific tags to reveal the natural language material only. To this end, we decided to use simple string matching and replacement techniques that relied mostly on regular expressions.

De-chunking the material into sentences was another step that needed to be taken, one that proved to be not trivial. We decided to use the BreakIterator class which is part of the Java standard library in order to avoid excessively complicating the project. Since we had a very large body of data we decided we could afford having only a portion of the data correctly broken into sentences.

For processing the sentences, we chose to use the toolkit LingPipe which contains a public api for a number of different linguistic models after originally experimenting with MALLET. We chose LingPipe due to the fact that it was able to use available tagged data sets to quickly train an HMM to determine a given words part of speech.

Conclusion and Evaluation

We believe it was a poor decision that we chose to work with individual sentences out of the context of the article that they belonged to. We started with hopes of using a parts of speech approach to separate out the subject and verbs that are pertinent to the date but there is a major problem with the amount of pronouns that are in the Wikipedia data. We saw of no clear way to resolve pronouns from the individual sentences once they were separated and with the high number of pronoun groups, it was not possible to form statements that would have data independent enough to have any practical use.

We originally looked at using the CRFs built into MALLET to do the parts of speech analysis but MALLET proved to be difficult to work with, given the lack of documentation and scarcity of tutorials or example code. We failed to find alternative methods or toolkits and stubbornly spent a lot of time trying to massage into our project the little that we found out studying the MALLET tutorials.

As a result of the flaws in our design and due to the poor time managing skills of our duo we were unable to integrate the parts of our system into a concrete pipeline that would produce coherent output and enable us to experiment.

We believe the time we spent trying to understand the various implementations of Conditional Random Fields was not a total loss and that this project contributed to our understanding of the machine learning techniques.

Group Dynamics

Our group, or pair rather, had a relatively poor dynamic overall. We are friends and our personalities work well together, but our communication skills at times were poor and it was difficult to brainstorm new ideas or stay focused with so few people. Having a limited number of people in our group made it so that we would need to focus our attention to one area, yet we had no well defined strategies which made it more difficult.

Outside Resources

We used LingPipe and part of an associated tagged corpus.