Event Ordering of News Articles

James Friel

Supervisor: Shay Cohen

Objectives

The aim of this thesis is to construct a system that predicts the most probable path through an edge-weighted complete digraph of events. we aim to construct this graph by extracting data from Wikipedia for each event and build a date estimate from this. Using this data we will conduct several experiments to optimise accuracy of a graph traversal.

Introduction

Nominal data is descriptive in nature, making it difficult to assign a Canonical ordering to. The problem tackled in this dissertation is the ordering of news article headlines to generate a most-probable traversal of a weighted directed graph of these events. This problem is based off of the paper [1] and some techniques discussed and used henceforth are based off of this paper.

Our data comes from the Wikipedia "Today in History" data set and the nominal data is retrieved from Wikipedia articles.

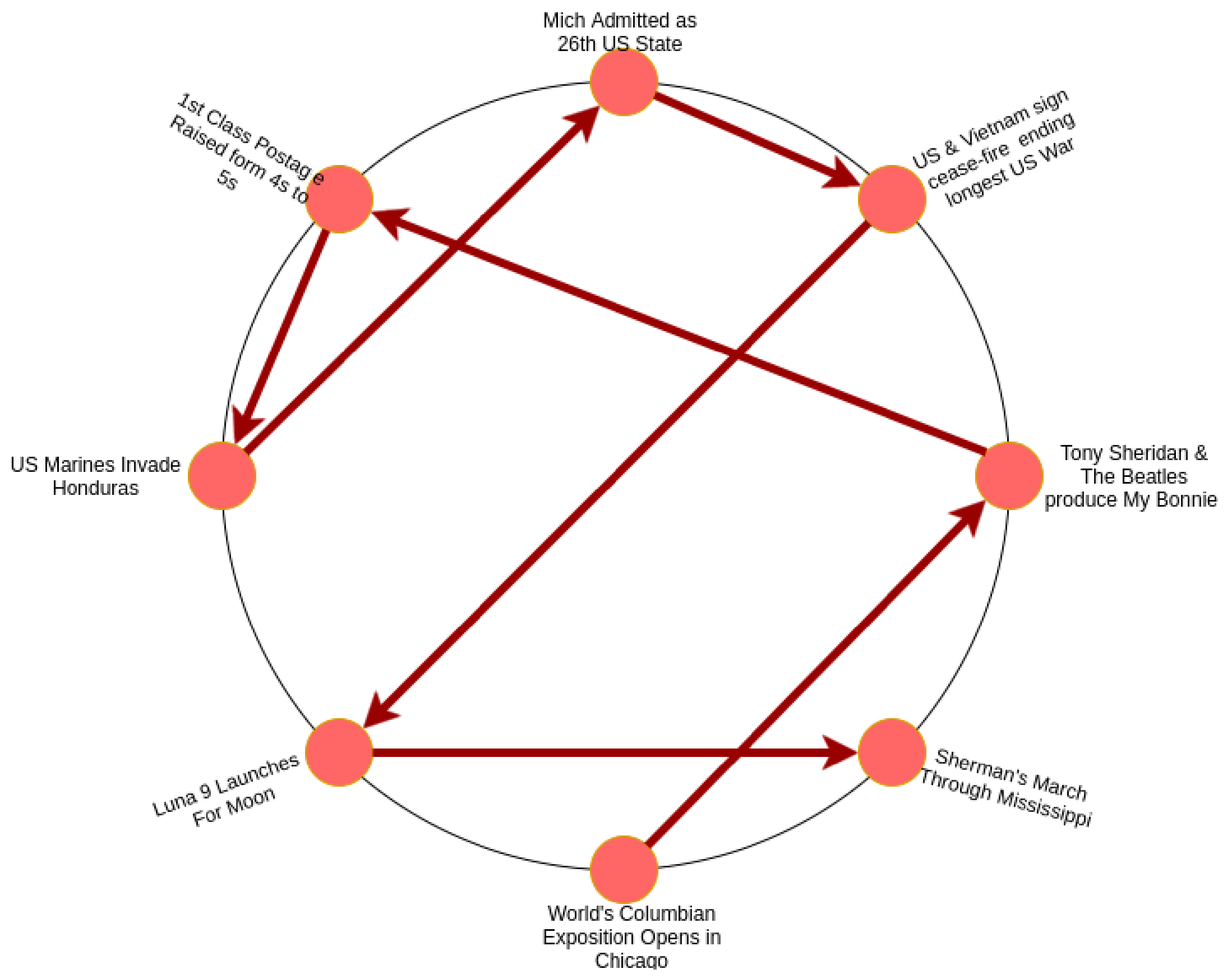
Mich Admitted as 26th US State
US & Vietnam sign cease-fire ending longest US War
Tony Sheridan & The Beatles produce My Bonnie
Sherman's March Through Mississippi
World's Columbian Exposition Opens in Chicago
Luna 9 Launches For Moon
US Marines Invade Honduras
1st Class Postage Raised form 4s to 5s

Figure 1: A Graph of the predicted ordering

Wikipedia As a Data Source

Wikipedia was chosen as the data source for the project as it has over 40 million articles [2] and maintains an unbiased presence through community moderating [3].

Using a modified version of Stanford's natural language processing suite, we are able to build features from our retieved data.

Using this toolkit, the Open Information Extraction from The University of Washington is used to extract subjects, objects and their relations from our event titles.

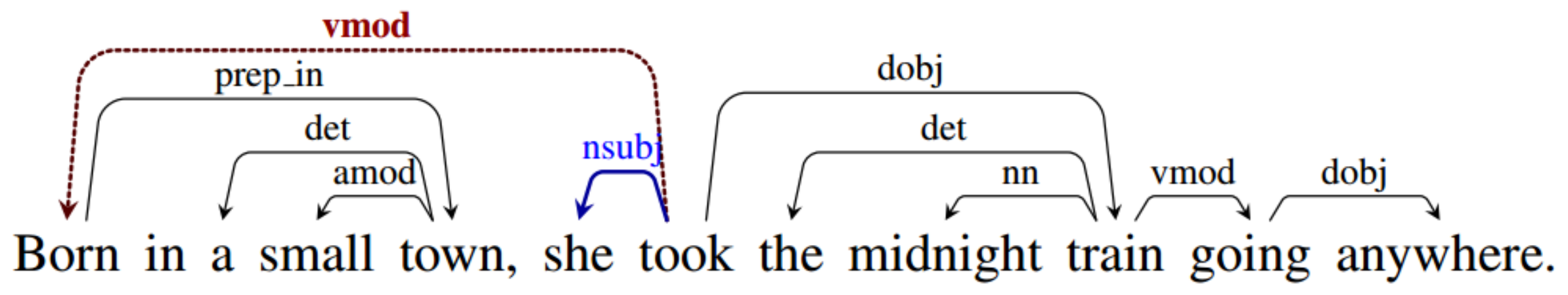
vmod
prep_in
det
amod
nsubj
dobj
det
nn
vmod
dobj
Born in a small town, she took the midnight train going anywhere.

Figure 2: Example OpenIE Extraction [4]

Classification

Given that each event in our data set is of the form

$$(t_i, d_i)\ for\ i\epsilon[M], \tag{1}$$

where t is the title, d is the associated date and M is the original data set, we constructed a new data set

$$\{(t_i, s_i, d_j)\}\ for\ i, j[M] \tag{2}$$

This will form the basis of our data to generate features.

From this we built a new data set

$$\{(t_i, s_i, t_j, s_j, b_{ij})\}\ for\ i, j\epsilon[M] \tag{3}$$

where $b_{ij} = [y_i > y_j]$ indicating which event came first. A similar data set was constructed using only the article headlines.

With this new data set we began to experiment with techniques to classify a test set of the data.

Important Result

The use of external data sources, such as Wikipedia, greatly improves the ability to order saline events in chronological order.

Graphing

From our classifier we produce a complete digraph. In order to construct the most likely path through this graph we must solve a variation of the travelling salesman problem.

> Starting from any node, find the node that will generate the most probable path through every other node without visiting any node twice.

We solve this problem by using the minimum spanning path algorithm whilst inverting the edge weights of the graph. The results of which can be seen in Figure 1.

Results

We measure our results using Kendall's Tau Correlation Coefficient.

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}[5]$$

Accuracy	DT* (tuple)	DT* (triple)	SVM	Logistic Regression
With Articles	53%	48%	66%	76%
With Titles	43%	36	51%	52%

Table 1: Local Learning Results
DT = Decision Tree

Accuracy	Perceptron	MLP
With Articles	66%	83%
With Titles	46%	54%

Table 2: Global Learning Results

Conclusion

As we can see from the results, the addition of external data sources greatly aids in automatic ordering of saline events.

Future Work

If I was to take this project forward, there would be a several things I would look into

- The use of news websites as data sources
- The use of DNNs in classification
- Improved Graphing Techniques

References

[1] Omri Abend, Shay B Cohen, and Mark Steedman. Lexical event ordering with an edge-factored model. In *Proceedings of NAACL*, 2015.

[2] Wikipedia: Size comparison. `https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons`. Accessed: 23-01-2017.

[3] Based on the research of Shane Greenstein and Feng Zhu. Is wikipedia biased?, Dec 2012.

[4] Openie extraction example. `http://nlp.stanford.edu/software/openie.html`. Accessed: 27-07-2017.

[5] Kendal tau metric. https://www.encyclopediaofmath.org/index.php/Kendall_tau_m

Accessed: 25-01-2017.

Acknowledgements

I would like to thank Shay Cohen for all of the advice and guidance given during the project.

I would also like to thank Greggs Bakery for helping me survive through coffee and morning rolls.

Contact Information

- Web: jamesfriel.uk
- Email: s1332298@ed.ac.uk