Geo-tagging of News Story

Justin Wang

Supervisors: Zhiyong Wang, David Feng
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

INTRODUCTION

Project Aims

- Finding geographic relevance of a news article along with its ranking.
- Exploring statistical approach to geo-tag textual document.
- Enhancing the geographic meaning of a textual document by using information other than toponyms (location words) in the text.

METHODOLOGY

Toponym Recognition

- Extract named entities (location, person, organization) and its frequency from textual document using Stanford Natural Language Processing tool's Named-Entity-Recognition (NER) [1].

Geo-location Booster

- Phrases other than location words, such as person and organization can potentially provide additional geographic meaning especially to text with insufficient location words.
- Person and organization phrases are being passed into the search engine to obtain as many web articles returned.
- The ranking of different web pages is dropped to exclude the performance of the search engine.
- These web articles are then processed with NER again to extract the location terms and its frequency only.

Geo-Relevance

- Score for the location terms found in the text is calculated using maximum likelihood.
- Score for the person and organization terms are found by using the probability of the location terms. Location terms with only one occurrence are being dropped as it is highly likely to be noise and this reduced the number of candidates by more than half.
- These terms and their score are being passed into a gazetteer built using GeoNames [2] to map a location term to its possible location on the map.
- Each individual score would be divided among the different senses equally.
- The output would be a score obtained by combining the geo-location of the location, person and organization phrases with its latitude and longitude coordinates.

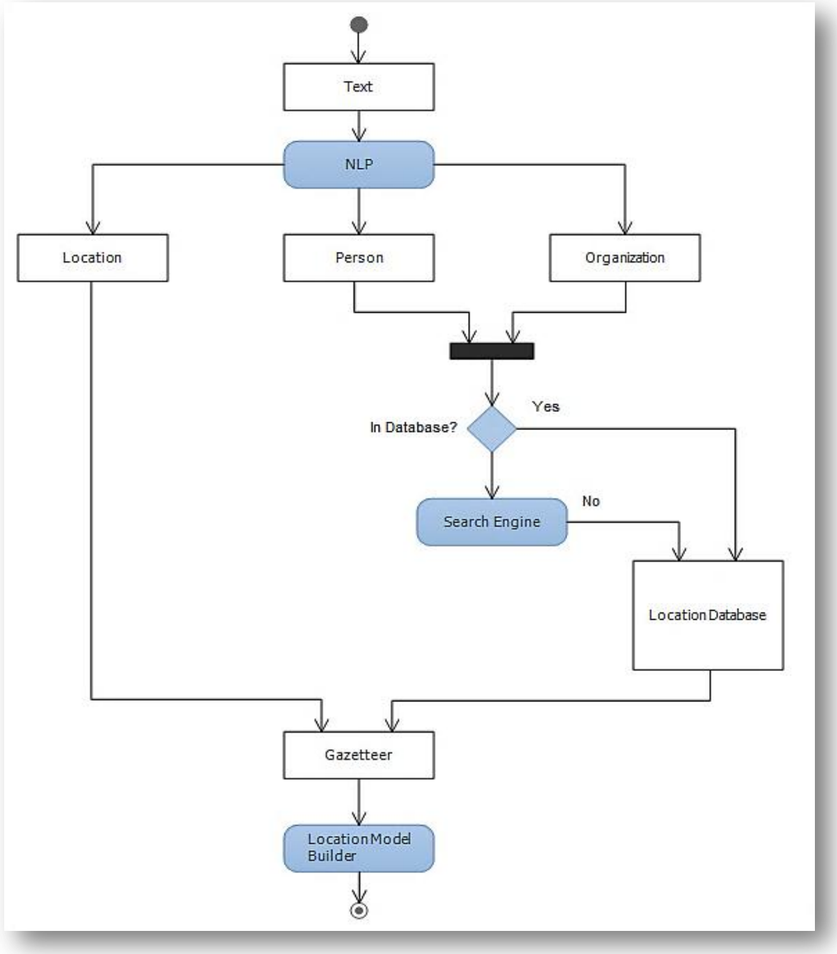
Modeling Locations

- These scores are modeled by adding them to a hierarchy tree with the following structure:

  Earth->Continent->Country->State->City

- The tree is first being built by using detected location terms.
- If the location term is higher up the hierarchy, the score is added to the leave nodes (city level) by dividing the score evenly among its children.
- If the location term is below the city level, the score is simply added to it.

Workflow Chart



RESULTS

Intermediate Results

- The person and organization phrases provide useful information for the geographic location.
- An example is shown in **table 1** for the word "Qantas" and **table 2** for the word "George".
- Table 1 shows that "Qantas" has a strong geographic relevance with Australia.

Location	Percentage (%)
Australia	12.4
Sydney	8.1
Melbourne	5.3
Singapore	3.4
London	2.7
Brisbane	2.2
Perth	2.2
New Zealand	1.8
USA	1.8
Asia	1.5

- Table 2 shows an example of a generic name "George" which has a weaker relation to a geographic location. The higher score for Washington could be due to the relation with the famous person George Washington or there is a large number of George found there.

Location	Percentage (%)
Washington	5.2
U.S.	2.6
London	1.9
United States	1.7
New York	1.7
St. George	1.4
Lake George	1.4
Iraq	1.1
US	1.1
Texas	1.1

DISCUSSION & FUTURE WORK

- Improve on geo-relevance by optimizing weight instead of dividing evenly.
- Future testing and evaluation would be done on Open Directory Project.

REFERENCES

[1] Finkel, J.R., T. Grenager, and C. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics2005, Association for Computational Linguistics: Ann Arbor, Michigan. p. 363-370.

[2] http://geonames.org