Extracting and Analyzing Semantic Relatedness between Cities Using News Articles

Yingjie Hu¹, Xinyue Ye², and Shih-Lung Shaw¹

¹Department of Geography, University of Tennessee, Knoxville
²Department of Geography, Kent State University, Kent

Apr. 8, 2017
Introduction

• News articles are rich sources of information

• Diverse topics
  • Economy, politics, science, sports, ...

• Various entities
  • Persons, organizations, places, ...

• Timely information
  • Prompt report of latest events
Introduction

- **Cities**, as hubs of human activities, are frequently mentioned in news articles.
- Two or more cities may **co-occur** in the same news article.
- E.g., comparing the **lifestyles** of two cities.
Introduction

- E.g., sports may draw teams from two cities together

- E.g., cities may address environmental issues collaboratively
Introduction

• Cities can be related under a variety of topics (semantic relatedness)

• Such semantic relatedness is partially captured in news articles

• Objective: to develop a computational framework that can automatically process a large number of news articles and extract semantic relatedness
Problem Formalization

**Input**

- **News Articles**
  - based on:
    - A set of cities \( \{c_i\} \)
    - A set of time periods \( \{y_k\} \)

- **Semantic Topics**
  - defined by:
    - A set of topic terms \( \{t_m\} \)

**Output**

- **Semantic Relatedness**: A set of relatedness values \( \{r_{ijkm}\} \)

- City 1
  - Sports
  - Politics
  - 2013
  - 2014
  - 2015

- City 2
  - Politics
  - Sports
  - 2013
  - 2014
  - 2015
Problem Formalization

• Core idea:
  1) Identify the topics of news articles
  2) Assign the topics to the cities
  3) Quantify the semantic relatedness

• Key question: given a news article, which topics is it talking about?

Multi-label classification problem
Framework

Model training:
1. Label Topics
2. Train Model for text data (Labeled LDA)
3. Finish training

Topic extracting:
4. Input
5. Output
6. Identify threshold
7. Extract topics

All news articles with extracted topics

Testing topic accuracies at possible thresholds

Topic scores:
- Culture: 0.313
- Politics: 0.001
- Business: 0.245
- Sports: 0.002

Human annotators

Topics:
- Culture
- Politics
- Business
- Sports
Framework
Experiments

- **Cities**: top 100 cities in the contiguous U.S.
- **Time**: 1/1/2006 and 12/31/2015

- News articles from The Guardian
  - 543,824 news articles
Experiments

- **17 semantic topics from IPTC**
  - E.g., Culture, Politics, Sports, Disaster, Crime, …

- **Obtaining training data**
  - Existing news tags
  - Mapping some tags to topics
  - **141,765 training data records**

<table>
<thead>
<tr>
<th>IPTC Topic</th>
<th>News Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts, Culture and Entertainment</td>
<td>culture, music, film, media, books, artanddesign, television, art, fashion, festivals, history, comedy, museums, opera, drama, poetry, documentary, painting, theatre, sculpture</td>
</tr>
</tbody>
</table>
Experiments

- **Training** the LLDA model

- **Topic extracting**
  - Applying the trained LLDA model to all news articles

**A training data record**

<table>
<thead>
<tr>
<th>IPTC labels</th>
<th>Processed text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts culture and entertainment</td>
<td><em>la hard city accept reality suffer personal setback illness sick gorgeous setting move santa monica beach...</em></td>
</tr>
<tr>
<td>Lifestyle and leisure</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

- Identifying suitable **threshold**

\[
\text{Precision} = \frac{|\text{Extracted Relevant Topics}|}{|\text{All Extracted Topics}|}
\]

\[
\text{Recall} = \frac{|\text{Extracted Relevant Topics}|}{|\text{All Relevant Topics}|}
\]

\[
\text{F score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]
Experiments

• **Visualize city relatedness**
• Based on **semantic topics**
Experiments

• **Visualize city relatedness**
• **Based on publication time**

Conflicts, War and Protests

NYC and Washington DC

Los Angeles and New Orleans
Experiments

- Both Los Angeles and New Orleans were enrolled in the program in 2014
Distance Decay Analysis

- A weak distance decay effect was found in a previous research based on place co-occurrence in news articles (Liu et al. 2014, Transactions in GIS)

\[ c_{ij} \propto \frac{c_i c_j}{d_{ij}^\beta} \]

- \( \beta \) is the friction coefficient; \( \beta = 0.2 \) in Liu et al. 2014

- City relatedness under different topics might have different distance decay effects
## Distance Decay Analysis

All news: $\beta = 0.23$

<table>
<thead>
<tr>
<th>Topic</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts, Culture and Entertainment</td>
<td>0.21</td>
</tr>
<tr>
<td>Sport</td>
<td>0.08</td>
</tr>
<tr>
<td>Crime, Law and Justice</td>
<td>0.37</td>
</tr>
<tr>
<td>Science and Technology</td>
<td>0.19</td>
</tr>
<tr>
<td>Politics</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Conclusions

- **News articles** partially capture the **semantic relatedness** between **cities**

- A **computational framework** is developed to “read” a large number of news articles and extract semantic relatedness

- An experiment based on more than 500,000 news articles shows **different network structures** and **temporal variations**

- **Varied distance decay effects** were observed for the different semantic relatedness
Thank You!

Questions?

Yingjie Hu
yhu21@utk.edu