A Hybrid Approach for Spotting, Disambiguating and Annotating Places in User-Generated Text

Karen Stepanyan, George Gkotsis, Vangelis Banos*, Alexandra I. Cristea, Mike Joy
Computer Science, University of Warwick, UK and
*Department of Informatics, Aristotle University of Thessaloniki, Greece
K.Stepanyan@warwick.ac.uk, G.Gkotsis@warwick.ac.uk, *Vbanos@gmail.com, A.Cristea@warwick.ac.uk & M.S.Joy@warwick.ac.uk

Motivation

Problem: The advancement of the Semantic Web requires automation in semantic annotation.
Challenge: The performance of the available annotation tools fails with user-generated content.

Potential: Current web trends are:
1. Growth of User-Generated Content
   - Twitter corpus is twice the size of the printed collection of the Library of Congress
   - Facebook processes ~500 Tb of new data daily
2. Prevalence of Geolocation Data in User-Generated Content
   - GeoRSS is the most widely used RSS module in weblog feeds (used in 38.2% of feeds).

Contribution vs. Previous Work

Contribution:
- Developed a geolocation-aware semantic GeoAnno annotation model for spotting and disambiguation of places in user-generated texts.
- Implemented prototype to annotate places and toponyms.
- Evaluated GeoAnno, which outperforms existing solutions by exploiting geolocation data.

Prior Solutions Use:
- Mainly Wikipedia content, intra/external-wiki links, redirect pages and category hierarchies for disambiguation.
- Windowing strategy that leads to a loss of discriminative features.
- Topic models and proximity to improve performance.

Non of these use geolocation, to further improve the solutions for annotating places and toponyms.

Detailed Description

The model:
(a) takes HTML documents with embedded geolocation as input
(b) identifies associations between the published post and nearby places
(c) exploits each found association to spot relevant text faces and disambiguate entities
(d) returns the annotated content as XML

GeoAnno Model:
Steps of Annotation

1. Annotate
   - Blog Posts
   - GeoRSS

2. Extract Content
   - and Get Places

3. Identify Associated Places
   - Identify places in proximity to the given geolocation point that are associated to the text.

4. Spot Relevant Text Faces
   - Use text matching between place names and content identify text faces.

5. Disambiguate & Link to DBpedia
   - GeoAnno determines the geolocation of the place and attempts to disambiguate and link to DBpedia.

Example

Weblog Post

<table>
<thead>
<tr>
<th>Text Faces</th>
<th>fp1 = ‘Chelsea’</th>
<th>fp2 = ‘Brazil’</th>
</tr>
</thead>
</table>

Places Nearby

<table>
<thead>
<tr>
<th>Place Name 1</th>
<th>Place Name 2</th>
</tr>
</thead>
</table>

Conclusions and References

The evaluation used a set of randomly selected weblog posts.
Found and annotated 334 places in 311 posts.
CrowdFlower.com was used to evaluate the accuracy of the annotation prototype.

Accuracy:
- GeoAnno: 87.7%
- DBpedia Spotlight: 81.3%

Confusion Matrix

<table>
<thead>
<tr>
<th>Judgment Result</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judgment (%)</td>
<td>87.7%</td>
<td>11.4%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Avg Conf.</td>
<td>0.97</td>
<td>0.79</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Added value:
- 27.8% gain in capturing more annotations

Improvement of semantic annotation is possible by exploiting embedded geolocation data.

Benefits: GeoAnno model:
- Annotates social media content that contains geolocation
- Is demonstrated to improve the performance by annotating additional 27% places
- It performed with a high accuracy of 87.7%
- It offers a finer granularity of annotation

We suggest integration of the model into the existing tools for improved performance.

Potential Application:
- Automate annotation of social media content
- Prompt to annotate content at writing time.

Evaluation

Get in Touch

Grab a Leaflet! Dropdown

Snatch the Paper!