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, .I.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).*

Bar chart — Feeds Using a Module (%) vs Modules: georss 38%, thr 35%, openSearch 32%, dc 13%
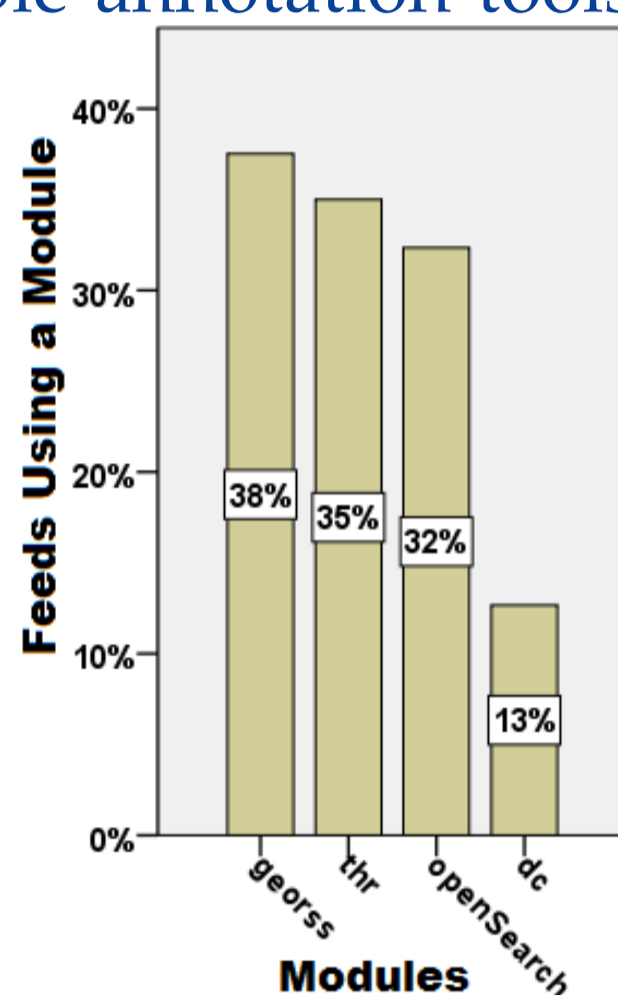
Contribution vs. Previous Work

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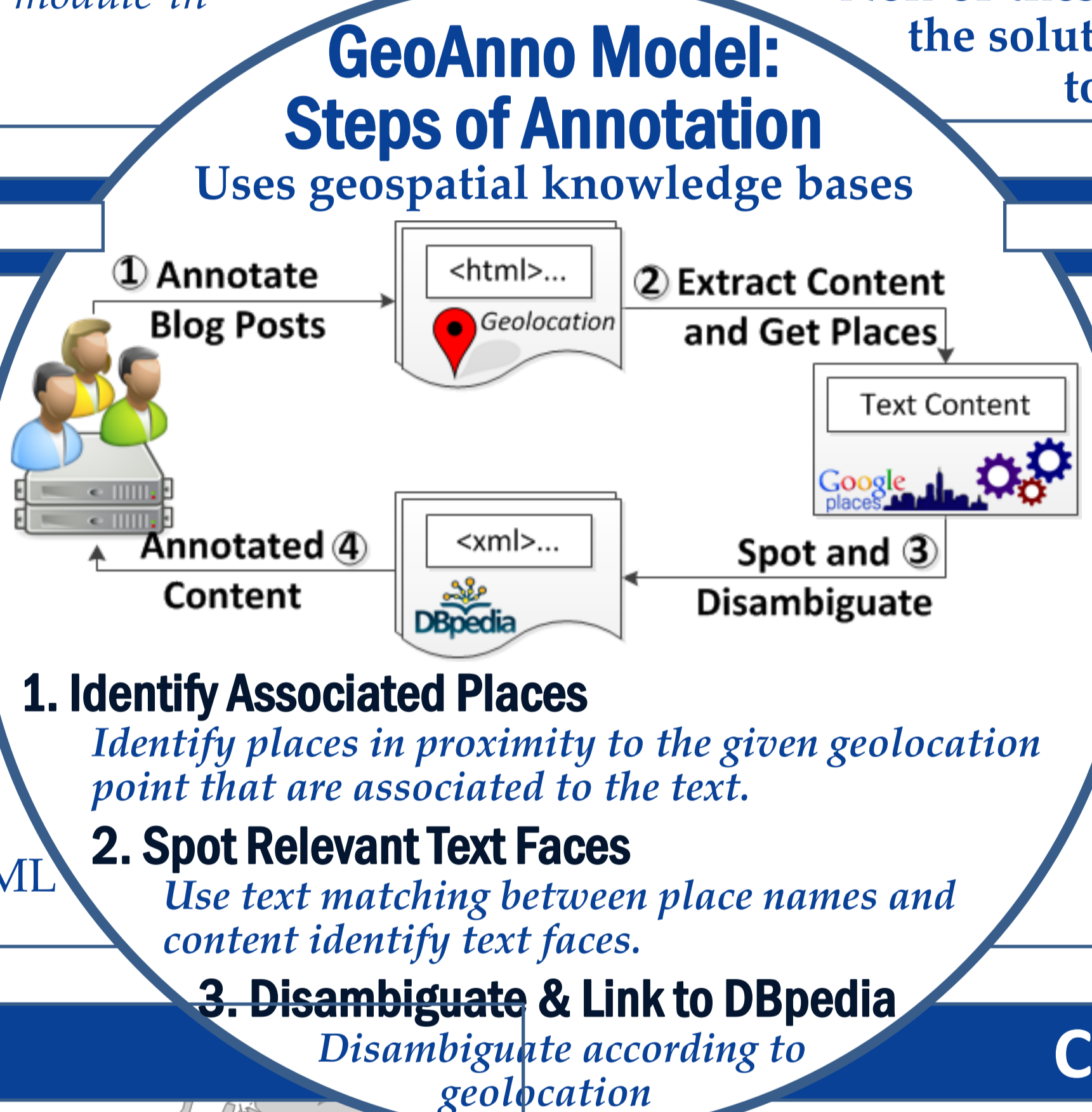- Evaluated GeoAnno, which outperforms existing solutions by exploiting geolocation data.

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.**
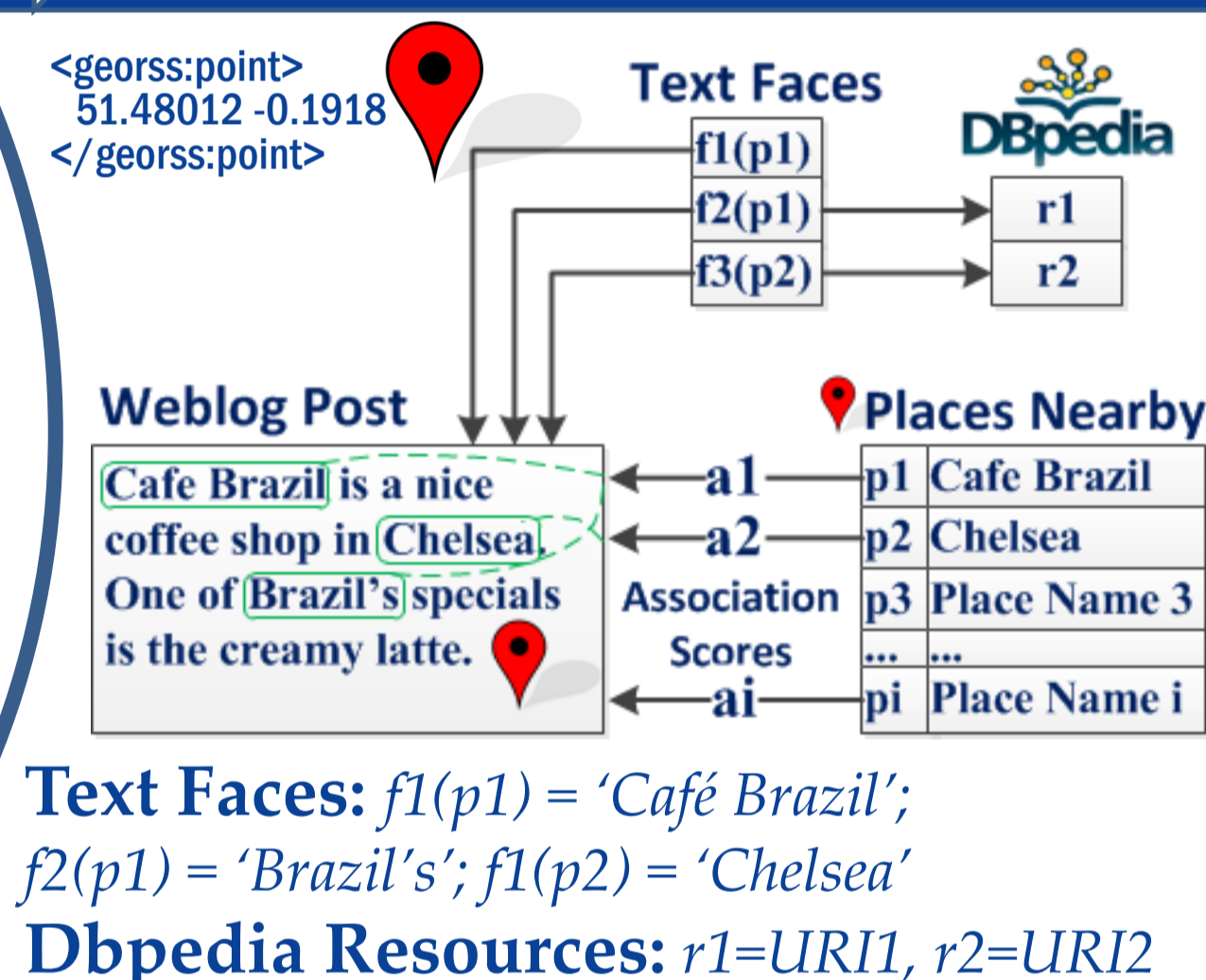
GeoAnno Model: Steps of Annotation
Uses geospatial knowledge bases

1. Annotate Blog Posts
2. Extract Content and Get Places
3. Spot and Disambiguate
4. Annotated Content

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

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

3. Disambiguate & Link to DBpedia
Disambiguate according to geolocation

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

Example

Text Faces: *f1(p1)* = 'Café Brazil'; *f2(p1)* = 'Brazil's'; *f1(p2)* = 'Chelsea'

Dbpedia Resources: *r1=URI1, r2=URI2*

Evaluation

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.

- 📍 **3,165 geolocation points**
- 📍 **1,775 web feeds**
- 📍 **World-wide distribution**

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

Added value:
- 27.8% gain in capturing more annotations

Judgment Result	Judgments (%)	Avg. Confidence
Correct	87.7%	0.97
Incorrect	11.4%	0.79
Ambiguous	0.9%	0.45

Unique Annotations	Number of Annotations for 311 posts	
	DBpedia Spotlight	GeoAnno
Distinct:	751	209
Shared:	86	86
Total:	837	295
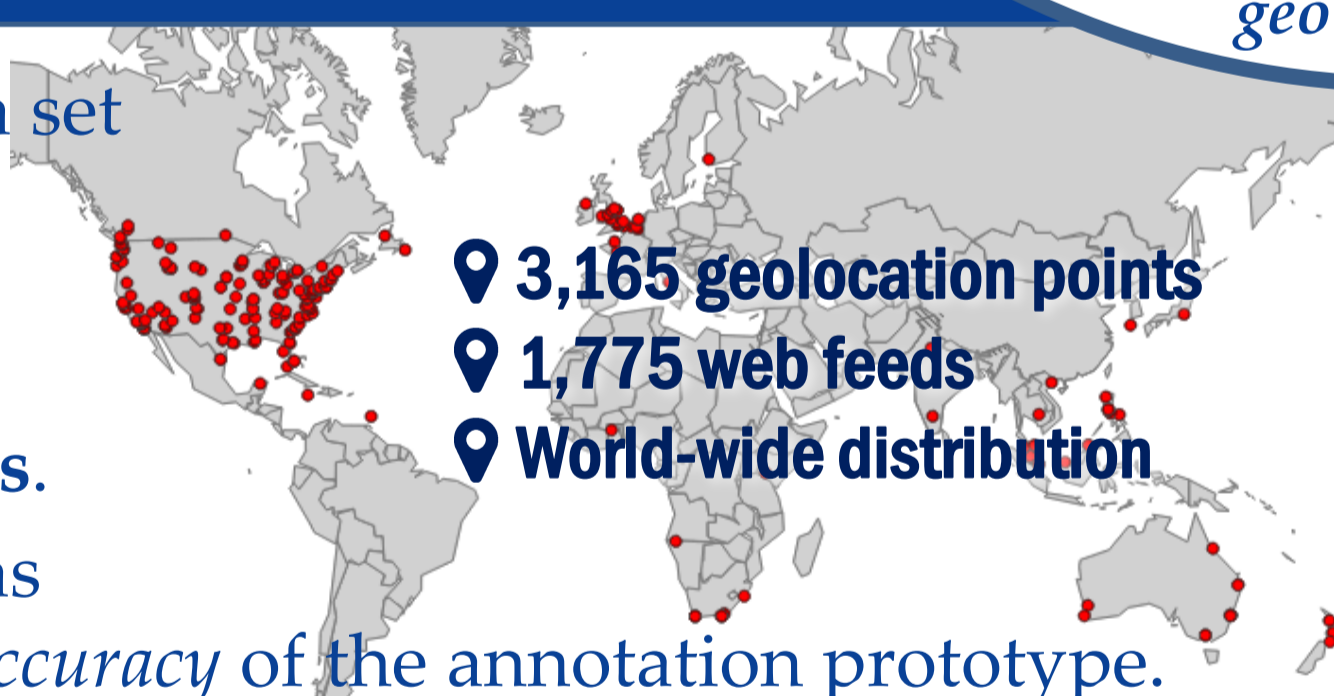
Conclusions and References

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.

Get in Touch

Grab a Leaflet!

🌐 www2.warwick.ac.uk/fac/sci/dcs/people/karen_stepanyan/
✉ Karen.Stepanyan@gmail.com
G+ plus.google.com/u/0/105514866316046533487
🐦 @kstepanyan

Snatch the Paper!

Drop a Line 💬!