



# **Co-occurrence and place name disambiguation.**

*GeoCLEF 2006*

Simon Overell  
João Magalhães  
Stefan Ruger

A world map with a light blue and green color scheme, showing continents and oceans. The map is slightly faded and serves as a background for the text.

# Introduction

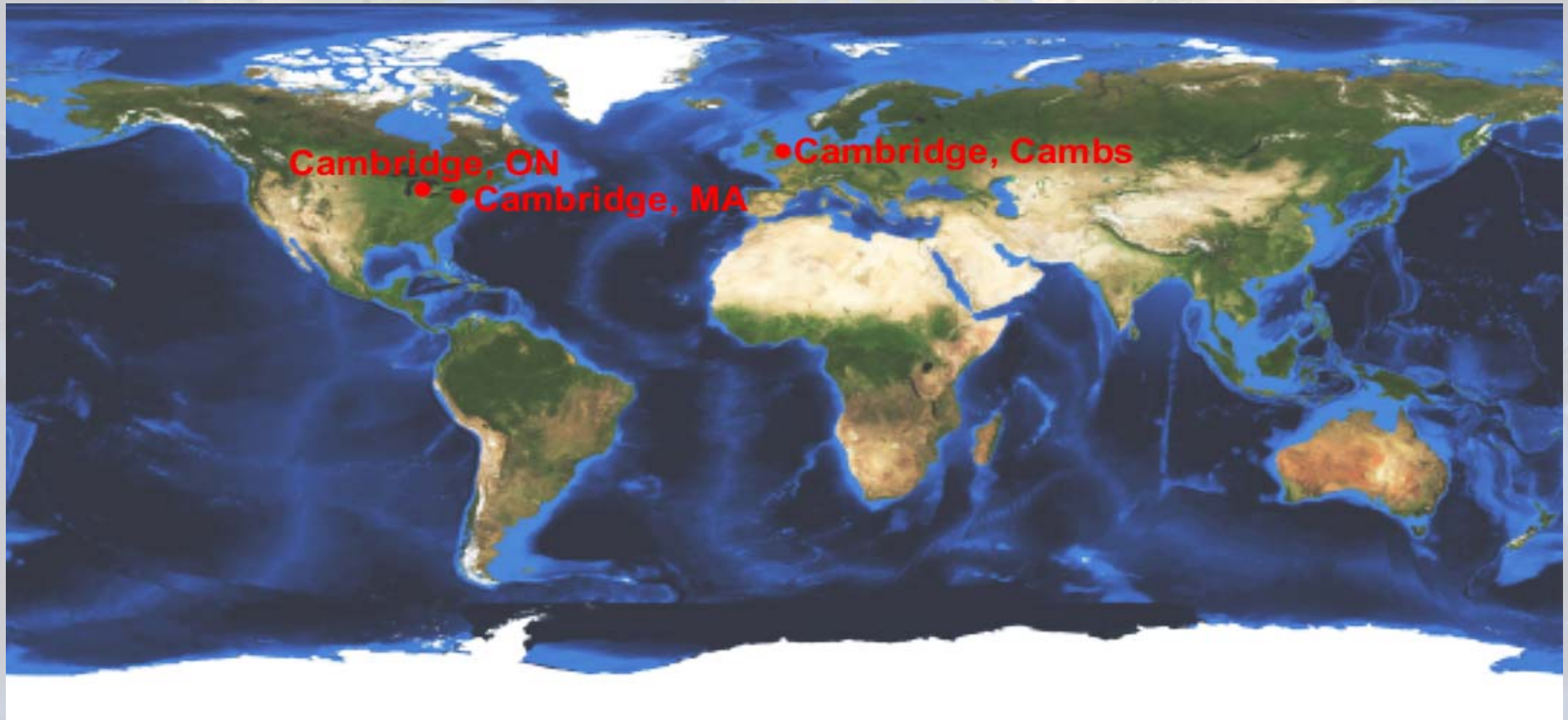
- The Problem
  - *Place name disambiguation*
- The Structure of this Talk
  - Methods of place name disambiguation
  - Previous work
  - Our Geographic Information Retrieval System
  - Results
  - Conclusions and Future work

A world map showing the continents of North America, South America, Europe, Africa, Asia, and Australia. The map is rendered in a light blue and green color scheme, with the oceans in a darker blue. The title 'Disambiguation' is centered at the top of the map.

# Disambiguation

- Rule-Based Methods
  - Based on heuristics
- Data Driven
  - Require a large annotated corpus
- Hybrid (Bootstrapping) Methods
  - Semi-supervised methods requiring only a small annotated corpus

# Disambiguation

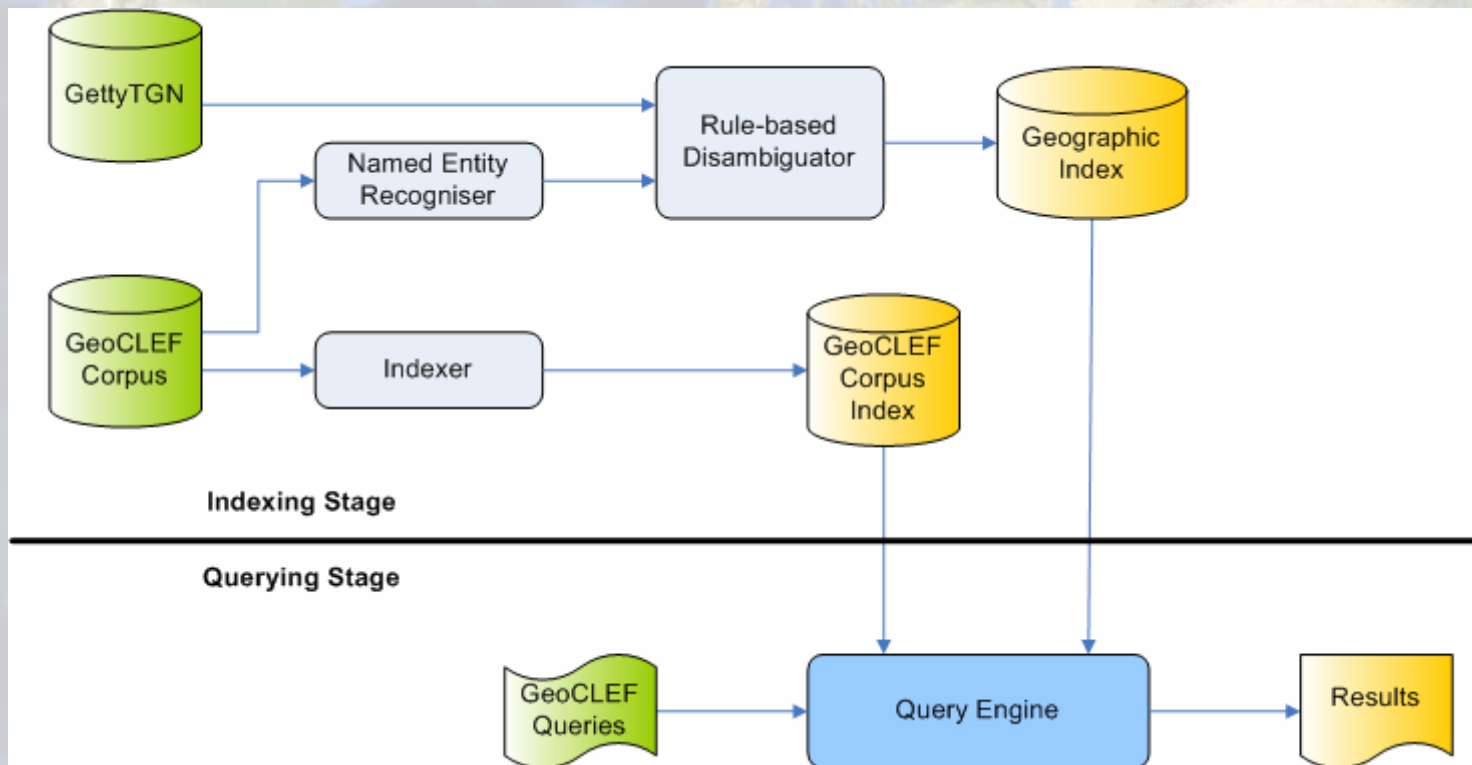


# Previous Work

- Identifying and grounding descriptions of places\*
  - Demonstrated and evaluated a hierarchical rule based method of place-name disambiguation
  - Used Wikipedia as a corpus
  - Proposed using co-occurrence for place-name disambiguation

\*published at Workshop on Geographic Information Retrieval SIGIR 2006

# Rule-based Disambiguation



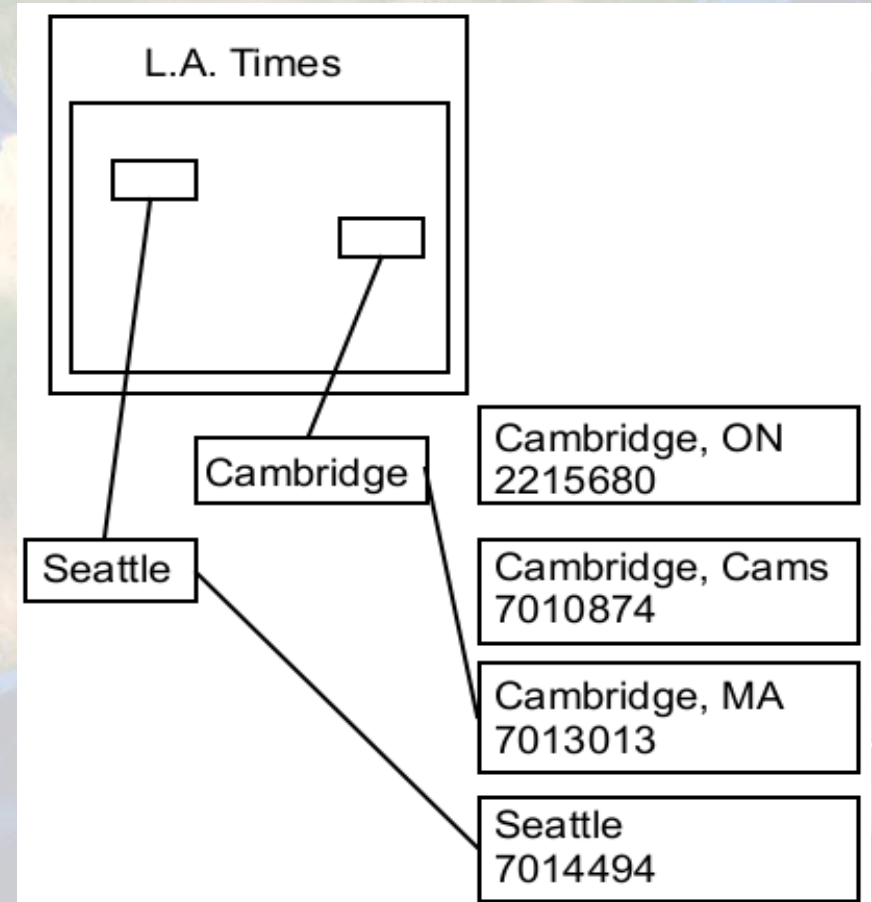
# Co-occurrence model

Mapping Table

Text	TGN Ids	Wiki Pages
Cambridge	2215680	Cambridge, ON
Cambridge	7010874	Cambridge, Cams
Cambridge	7013013	Cambridge, MA
Seattle	7014494	Seattle, WA

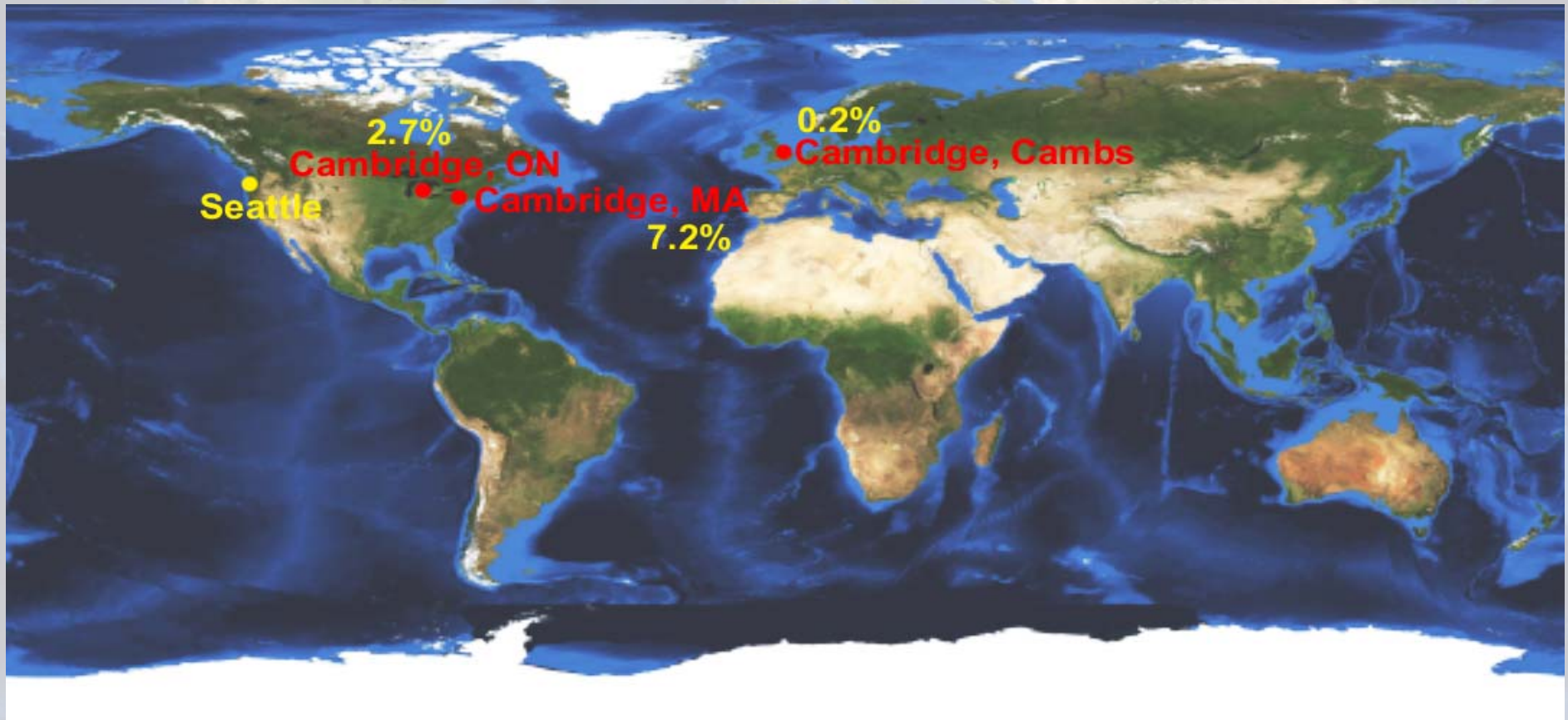
Occurrences Table

From Pages	Location Pages
British Universities	Cambridge, Cams
British Universities	Oxford, Oxon
British Universities	London, UK
University of Washington	Seattle, WA
University of Washington	Washington, US



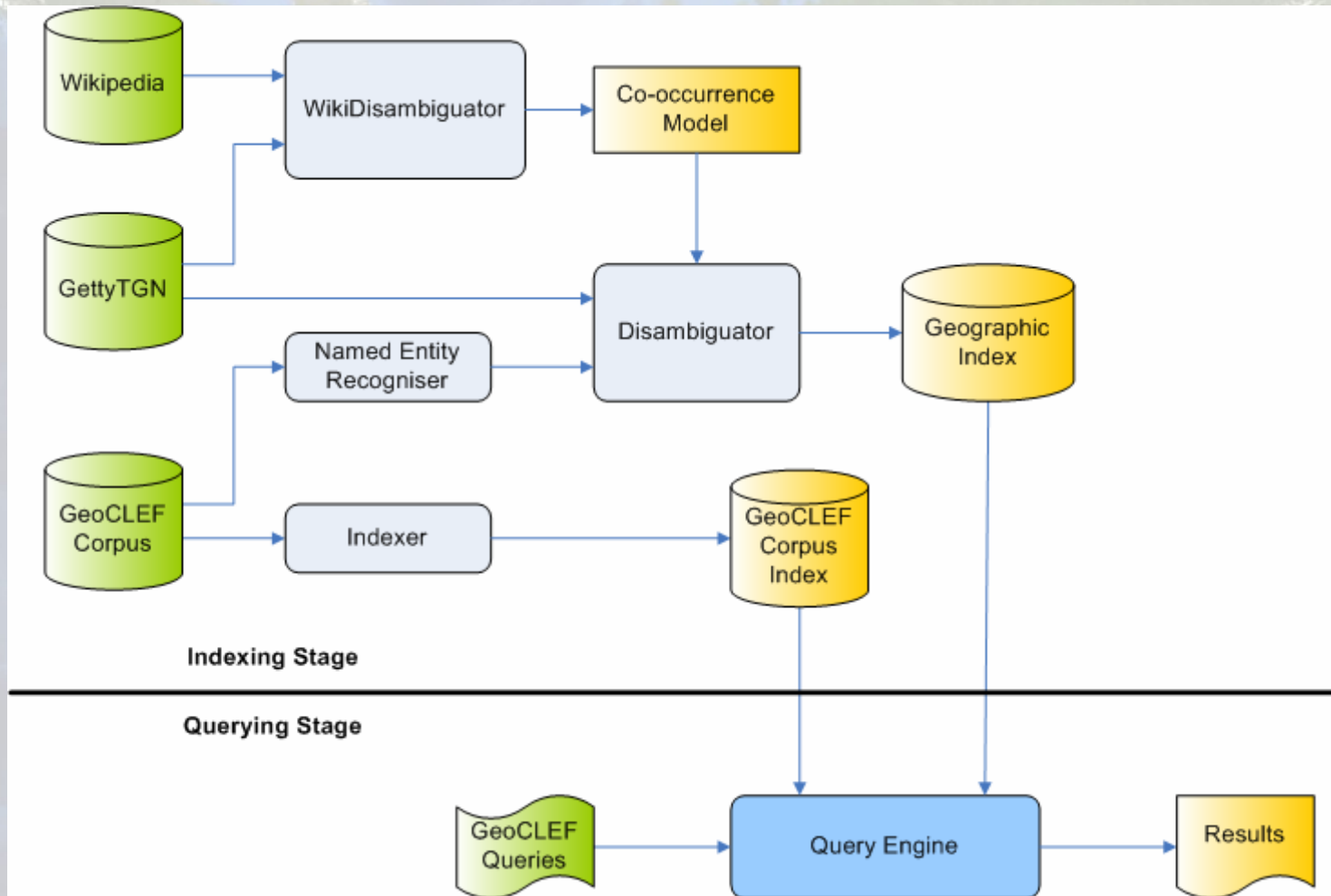
- Disambiguate Place tuples based on the likelihood of a co-occurrence as reflected in the generated model

# Co-occurrence model





# Co-occurrence based Disambiguation



# Results

- Runs
  - Title, Description and Narrative
  - Title and Description

Mean Average Precision						
TDN	TD	Worst	Q1	Median	Q3	Best
19.53%	16.49%	4%	15.64%	21.62%	24.59%	32.23%

A world map showing the continents of North America, South America, Europe, Africa, Asia, and Australia. The map is rendered in a light blue and green color scheme, with the oceans in a pale blue and the landmasses in a light green and yellowish-brown. The map is centered on the Atlantic Ocean.

# Summary

- **Conclusions**
  - The system model is valid
  - Co-occurrence models are a viable method of place name disambiguation
  - The accuracy of the model agrees with previous experiments
  - Further tuning is needed!
- **Future Work**
  - Larger scale co-occurrence model evaluation
    - Larger co-occurrence model
    - Compare multiple methods of applying the model



# Co-occurrence and place name disambiguation.

[www.doc.ic.ac.uk/~seo01](http://www.doc.ic.ac.uk/~seo01)

[www.doc.ic.ac.uk/~seo01/groundtruth](http://www.doc.ic.ac.uk/~seo01/groundtruth)

[www.doc.ic.ac.uk/~seo01/geowiki](http://www.doc.ic.ac.uk/~seo01/geowiki)

Simon Overell  
João Magalhães  
Stefan Ruger