Identifying and grounding descriptions of places.

Simon E Overell
Multimedia & Information Systems
Dept of Computing, Imperial College London
London SW7 2AZ, UK
simon.overell@imperial.ac.uk

Stefan Rüger
Multimedia & Information Systems
Dept of Computing, Imperial College London
London SW7 2AZ, UK
s.rueger@imperial.ac.uk

ABSTRACT

In this paper we test the hypothesis *Given a piece of text describing an object or concept our combined disambiguation method can disambiguate whether it is a place and ground it to a Getty Thesaurus of Geographical Names unique identifier with significantly more accuracy than naïve methods.* We demonstrate a carefully engineered rule-based place name disambiguation system and give Wikipedia as a worked example with hand-generated ground truth and bench mark tests. This paper outlines our plans to apply the co-occurrence models generated with Wikipedia to solve the problem of disambiguating place names in text using supervised learning techniques.

Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing

Keywords

Geographic Information Retrieval, Disambiguation, Wikipedia

1. INTRODUCTION

Geographic Information Retrieval is a fast growing area in the broader Information Retrieval discipline. It involves many of the methods generally associated with information retrieval such as searching, browsing, storing and ranking data as well as a series of its own problems.

Generally, Geographic Information Retrieval is split into four stages: Information Extraction, Disambiguation, the User Interface and Information Storage.

In this paper we deal with the problem of disambiguation. Our ultimate aim is to build a place name co-occurrence model; however, we are starting with the more simple problem: given a description of an object or a concept can we disambigate whether it is a place and, if it is a place, ground it to a TGN unique identifier. Wikipedia is used as our test corpus, because the articles are normally carefully written, well-linked with significant geographic names pointing to an article about the place thus disambiguating it.

2. BACKGROUND

Browsing data by time, place and event has been one of the goals of Information Retrieval for decades but it is only in recent years that necessary resources have existed. Larson's seminal paper, *Geographic Information Retrieval and Spatial Browsing,* identifies the advantages of browsing via location over traditional query-then-browse methods [8]. In a geographical query the user is able to specify that they require documents related to places falling within a certain area. In 2004 Sanderson and Kohler analysed Excite's query logs to discover what percentage of queries submitted to a search engine had a geographical term: 18.6% of the queries in their sample had geographical terms, a significant proportion of internet searches [14].

2.1 Mining Wikipedia

Wikipedia is a huge resource that has only recently begun to be mined. The accuracy of Wikipedia has been repeatedly tested with current debates remaining unresolved [5]. Despite controversy regarding its validity, Wikipedia is an excellent example of a huge hyper-linked corpus of textual descriptions in the public domain [16].

Wikipedia's suitability for data mining was evaluated in Kinzler's paper *WikiSense - Mining the Wiki,* where the use of the highly formatted template data, inter-language links and clusters inferred from the hyper-linked structure were highlighted as particularly useful [7].

Data mining Wikipedia is slowly making its way into Geographic Information Retrieval with the XLDB group using it as a source for place names in GeoCLEF 2005 [2].

3. RELATED WORK

The problem of disambiguating place names in text has been approached from several different angles, most methods fit into one of the two categories described below:

3.1 Rule-based methods

The rule-based disambiguation methods apply one or more of the following heuristics either iteratively or in a linear process.

- Unique match – the place is unambiguous!
- Defaults – based on a simple heuristic rule select either the most important place or the place located closest to where the document was published.
- Referents within text – look at the places and descriptions referred to within 2-5 words of the place being disambiguated.
- Minimum bounding polygon – attempt to fit a bounding polygon around the place being disambiguated and the surrounding places referred to, select the smallest polygon to disambiguate.
- Polygonal overlay – map a kernel over each surrounding place mentioned, disambiguate by calculating the minimum distance to the maximum height of overlapping polygons.

These rules can be applied in varying orders with varying parameters. They can either be applied together with each rule voting or returning a probability and the results combined, or applied in order attempting to get an absolute answer with each one [2, 3, 9, 11, 13, 17, 18].

3.2 Data driven methods

The data driven methods of disambiguation generally apply standard machine learning methods to solve the problem of matching place names to locations. The problem with these methods is that they require a large accurate corpus of annotated ground truth; if such a corpus existed naïve methods (e.g. Bayes' theorem) or more complex methods (e.g. Latent Semantic Indexing) could be applied [4, 6].

Small sets of ground truths have been created for the purposes of evaluation or applying supervised learning methods to small domains [1, 10, 12, 15]; however a large enough corpus does not yet exist in the public domain to apply supervised methods to free text.

4. DISAMBIGUATING DESCRIPTIONS

WikiDisambiguator is the application designed to build our co-occurrence model. The data gathered (collected from a crawl of every Wikipedia article) takes the form of three database tables: links believed to be places and the order in which they occur; links believed to be non-places and the order in which they occur and a mapping of Wikipedia articles to TGN unique identifiers[1].

WikiDisambiguator uses rule-based methods of disambiguation. We have implemented four naïve disambiguation methods to provide an experimental baseline and a more complex method to build the co-occurrence model. All of these disambiguation methods fit into a disambiguation framework which crawls Wikipedia.

The Disambiguation framework is a simple framework to allow different disambiguation methods to be easily tested.

The framework is outlined as follows:

```
The WikiDisambiguator loads the Wikipedia articles
  to be crawled from the database
for each Wikipedia article all the links are extracted
  for each Link
    if it has already been disambiguated as not a
      place - add an entry to the db and continue
    if the page pointed to has already been
      disambiguated as a place - add an entry to
      the db and continue
    otherwise - attempt to disambiguate using the
      Method of Disambiguation specified
  end for
end for
```

The Methods of Disambiguation are passed:

- a list of candidate places
- a list of names of places related to this link
- the text making up the article that this link points to
- the article title
- how the link appeared in the text

The candidate places are taken from the TGN: places with either the same name as the anchor text in the crawled article or the same name as the title of the article linked to. There can either be one or multiple methods of disambiguation, each method can either:

- remove candidate places
- add related places
- mark as definitely a location and return a unique id
- mark as definitely not a location

[1]Our copy of Wikipedia was taken 3/12/2005

4.1 Naïve disambiguation methods

The first baseline method was **Random**, the intention with Random was to maximise recall regardless of precision and to quantify the amount of error caused by ambiguous place names. Each possible place name was mapped to a random matching entry in the TGN.

The second naïve method was **Most Important**; based on the feature type as recorded in the gazetteer, the most important place is returned. We mapped the following ordering across the feature types:
As large as or larger than an average nation $\gg$ Large populated area $\gg$ Large geographical feature $\gg$ Populated place $\gg$ Small geographical feature $\gg$ Small populated Place

Any entity not occurring in one of the above categories was deemed too insignificant to return.

The third naïve method was **Minimum Bounding Box**; the Wikipedia article describing the possible place is looked at and the first four related places (unambiguous if possible) extracted. A minimum bounding box is fitted around these places; if any are ambiguous, multiple boxes are formed with each possible location for the ambiguous place name and the smallest box is selected. The disambiguated place is the candidate place closest to the centre of the box.

The final naïve method was **Disambiguation with Referent**; the Wikipedia article describing the place, the link text and the page title are all searched for place names which *refer* to the place being disambiguated. These candidate referrer names are compared to the containing objects as listed in the gazetteer.

For example if a location appears in text as "London, Ontario", Ontario is only mentioned in reference to the disambiguation of London. The gazetteer is then queried for containing objects of places called London: "Ontario, Canada" and "England, United Kingdom". The candidate London will then be grounded as London, Canada rather than London, United Kingdom.

The intention of this disambiguation method was to maximise precision and the proportion of places correctly grounded regardless of recall.

4.2 Final disambiguation method

Based on the results observed by running our naïve methods on test data, we designed a hierarchical disambiguation system that could exploit the meta-data contained in Wikipedia and strike a balance between precision and recall.

Each disambiguation method is called in turn:

```
Disambiguate with Templates - Extract any Wikipedia
  template data and see if there is enough
  information to disambiguate the place (e.g.
  Latitude or Longitude data) or mark the article
  as not a place (e.g.  Biographic or Taxonomic data)
Disambiguate with Categories - Extract the
  Wikipedia category data and check if the
  information identifies the country / continent
  or identifies the article as not a place
Disambiguate with Referents (as described in the
```

Table 1: Disambiguation method results

	Recall	Precision	Ground	F
Random	87.1	60.5	58.6	71.4
Most Important	84.9	61	66.2	71.0
MBB	79.2	66.6	68.8	72.3
Referents	61.3	87	94.8	71.9
Combination	80.3	80.2	82.8	80.3

```
  naive method)
Disambiguate with Text Heuristics (described below)
```

We have defined our own heuristic method based on a combination of the Minimum Bounding Box method and the Most Important place method (however with slightly lower recall and significantly higher precision). The hypothesis used is *When describing an Important place, only places of equal or greater importance are used as referrers.*

5. GROUND TRUTH

Our ground truth takes the form of a list of all the links extracted from 1,000 Wikipedia articles chosen at random. Each link has been manually annotated as either a place or not a place and is matched to a unique identifier in the Getty TGN; this was all done by hand. The ground truth contains 1,694 locations and 12,272 non locations[2].

6. EVALUATION

We tested each of the disambiguation methods on the evaluation set and compared the returned results to the ground truth. In the results table we record three numbers from each run:

• **Recall** – The proportion of places correctly identified as locations.

• **Precision** – The proportion of the results returned that are locations.

• **Grounding** – The proportion of the correctly identified locations that are matched to the correct TGN unique identifiers.

• **F-measure** – Two times the product of precision and recall divided by the sum of precision and recall.

We provided the system with the following *world knowledge*: 50 places regarded as too large or too important to ever be referred to with disambiguating data (e.g. United States, Pacific Ocean etc.); 20 non-places that cause very common disambiguation errors (e.g. English Language, Law etc.); and in the Combination method of disambiguation, 50 categories that would aid the disambiguation.

The results table shows, as expected, that to maximise recall any article which shares its name with a place must be marked as a place (as in Random). To maximise either precision or correct grounding, only to return candidate places where a referent place is explicitly mentioned. The Combination method gives a suitable middle ground for all three values with a significantly higher F-measure; this should be accurate enough to form the basis for a co-occurrence model.

[2]The ground truth and the sample set are available for academic purposes at `http://www.doc.ic.ac.uk/~seo01/groundtruth` or by contacting the author.

7. FUTURE WORK AND CONCLUSIONS

We have shown that our place name disambiguation heuristic allows us to disambiguate and ground place name descriptions to a usable degree of accuracy. We have also produced a publicly available ground truth for others to test similar systems against.

Our next step is to run the WikiDisambiguator across the entirety of Wikipedia to build a large co-occurrence model. This model will be used in supervised learning methods to disambiguate place names in free text.

8. REFERENCES

[1] B. Bucher, P. Clough, D. Finch, H. Joho, R. Purves, and A. Syed. Evaluation of SPIRIT prototype following integration and testing. Technical report, 2005.

[2] N. Cardoso, B. Martins, M. Chaves, L. Andrade, and M. Silva. The XLDB group at GeoCLEF 2005. In *GeoCLEF 2005 Workshop*, 2005.

[3] P. Clough, M. Sanderson, and H. Joho. Extraction of semantic annotations from textual web pages. Technical report, 2004.

[4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. In *Journal of the Society for Information Science*, 1990.

[5] J. Giles. Internet encyclopaedias go head to head. *Nature*, 2005.

[6] D. Grossman and O. Frieder. *Information Retrieval.* Second edition, 2004.

[7] D. Kinzler. Wikisense - mining the wiki. In *Proceedings of Wikimania 05*, 2005.

[8] R. Larson. Geographic information retrieval and spatial browsing. In *In GIS and Libraries*, 1996.

[9] J. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *HLT-NAACL*, 2003.

[10] J. Leveling, S. Hartrumpf, and D. Veiel. University of Hagen at GeoCLEF 2005: Using semantic networks for interpreting geographical queries. In *GeoCLEF 2005 Workshop*, 2005.

[11] H. Li, R. Srihari, C. Niu, and W. Li. InfoXtract location normalization: A hybrid approach to geographic references in information extraction. In *HLT-NAACL*, 2003.

[12] M. Nissim, C. Matheson, and J. Reid. Recognising geographical entities in Scottish historical documents. In *SIGIR Workshop on GIR*, 2004.

[13] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *HLT-NAACL*, 2003.

[14] M. Sanderson and J. Kohler. Analyzing geographic queries. In *SIGIR Workshop on GIR*, 2004.

[15] D. Smith and G. Mann. Bootstrapping toponym classifiers. In *HLT-NAACL*, 2003.

[16] Wikipedia. http://www.wikipedia.org, 2006.

[17] A. Woodruff. Gipsy: Georeferenced information processing system. Technical report, 1994.

[18] W. Zong, D. Wu, A. Sun, E. Lim, and D. Goh. On assigning place names to geography related web pages. In *Proceedings of JCDL*, 2005.