Wordnet Ontology as a Geographical Information Resource

Davide Buscaldi,

Dpto. Sistemas Informáticos y Computación (DSIC)

Universidad Politécnica de Valencia

Valencia, Nov. 15th 2005

Plan of the talk

- The Geographical Information Retrieval task
- WordNet (in brief)
- Exploiting WordNet:
  - Query Expansion
  - Index Terms Expansion
- Results
- Conclusions

The Geographical Information Retrieval Task

- Actually GIR is ambiguous:
  - (Geographic Information) Retrieval**
  - Geographical (Information Retrieval)*
- In this case:
  - "Retrieval of information involving some kind of *spatial awareness*"* (Fred Gey @ GeoCLEF 2005)
  - E.g. "Find news about riots in France."
- Not to be confused with GIR as a particular aspect of Spatial Information Retrieval**
  - E.g. "What is the river flowing through Paris?"

Common GIR issues (1)

- (Almost) The same Geographical Entity can be indicated in several different (and sometimes ambiguous) manners:

- United Kingdom of Great Britain and Northern Ireland
- United Kingdom, UK, U.K. + Ireland, Eire
- Great Britain, GB + Ireland
- Reino Unido, Gran Bretagna
- British Isles

Common GIR Issues (2)

- Missing *explicit* geographical information:
  - E.g., consider the following text:

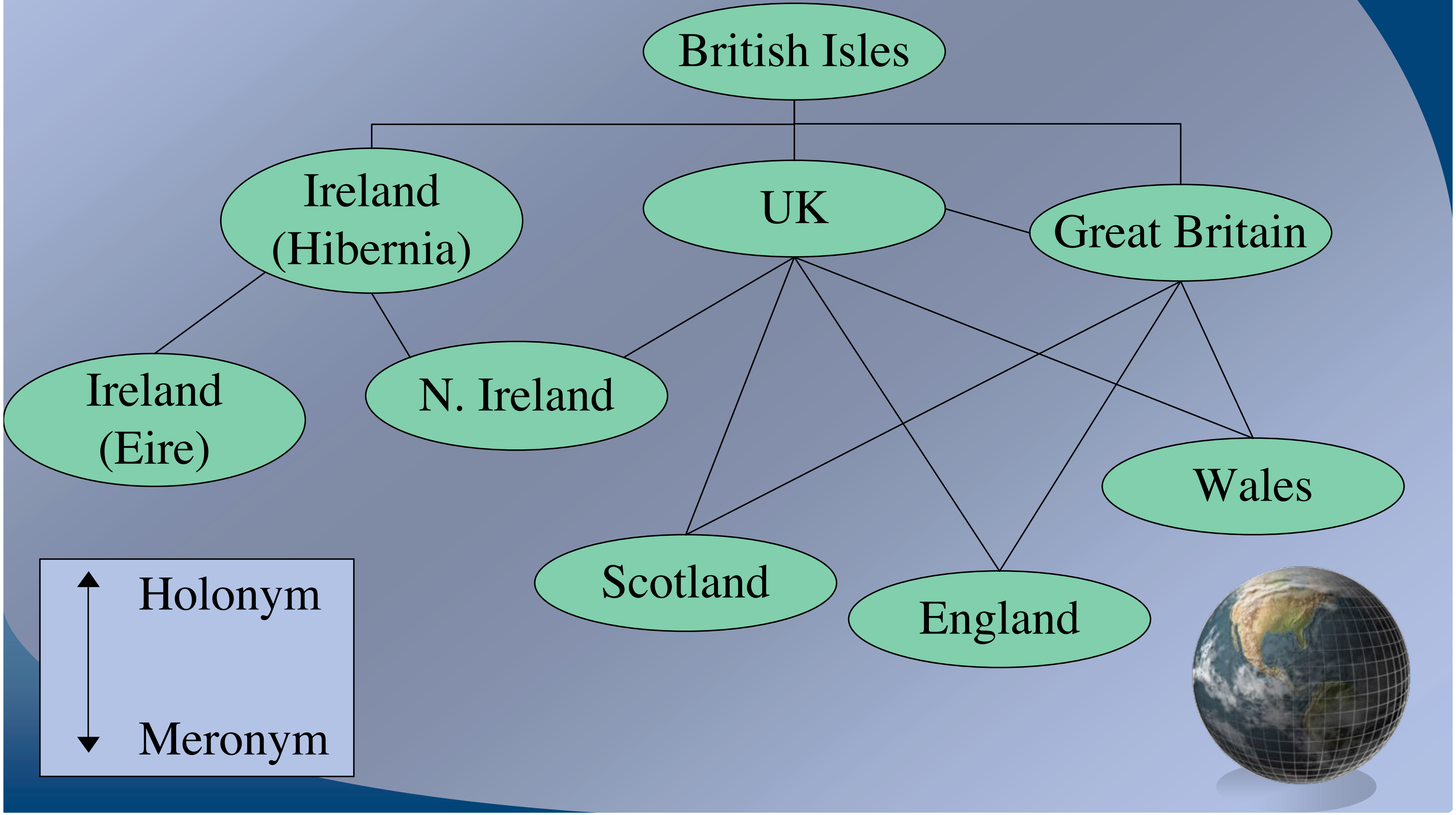
    "On Sunday mornings, the covered market opposite the station in the leafy suburb of Aulnay-sous-Bois - barely half an hour's drive from central Paris - spills opulently on to the streets and boulevards."

    Whereas the text is talking about events in France, the GE *France* itself is never mentioned.

The WordNet Ontology

- Lexical resource containing nouns, verbs, adjectives and adverbs organized into synonym sets *(synsets)*
  - each synset represents one underlying lexical concept.
  - various relations link the synonym sets
    - Hypernymy (is-a relation)
    - Meronymy (has-part relation)
    - Holonymy (part-of relation)
- Available at
  - http://wordnet.princeton.edu/perl/webwn

Geographical Conceptual Networks in WordNet



Exploiting WordNet

- WordNet can help in addressing most of GIR issues
- Solve *synonymy*:
  - E.g. synset corresponding to “*U.K.*”:
    - {United Kingdom, UK, U.K., Great Britain, GB, Britain, United Kingdom of Great Britain and Northern Ireland}
- Find missing (geographical) information:
  - Meronymy (“has member/part” relationship)
  - Holonymy (“is member/part of”)
- Two solutions tested:
  - Query Expansion (QE)
  - Index Terms Expansion (ITE)

Query Expansion

- Expand the geographical terms of the query with their synonyms and (some) meronyms
  - Geographical terms are identified through the WordNet ontology (words having the synset {region, location} among their hypernyms
  - Meronyms containing the word "*capital*" in the definition (*gloss*) or in the meronym synset itself

Query Expansion - Example

- "Foreign minorities in Germany"
  - "Germany" appears in the synset: {Germany, Federal Republic of Germany, Deutschland, FRG}
  - The following meronyms contain the word "capital":
    - Berlin, german capital
    - Bonn (was the capital of Germany between 1949 and 1989)
    - Munich, Muenchen (capital of Bavaria)
    - Aachen, Aken, Aix-la-Chapelle (formerly Charlemagne northern capital)

Index Terms Expansion

- Find geographical terms in the text collection
  - – *openNLP* Named Entities detector (http://opennlp.sourceforge.net)
- Put all their holonyms and synonyms into a special *geo* index
  - – Search Engine used: Lucene (http://lucene.jakarta.org)
- Label geographical terms in the query with the *geo* search field:
  - – E.g. “riots in France” -> text:riots geo:France

Index Terms Expansion - Example

"On Sunday mornings, the covered market opposite the station in the leafy suburb of Aulnay-sous-Bois - barely half an hour's drive from central Paris - spills opulently on to the streets and boulevards."

From WordNet:

⇒ Paris, French capital, capital of France, city of light
- ⇒ France, French Republic
  - ⇒ Europe
    - ⇒ Northern hemisphere

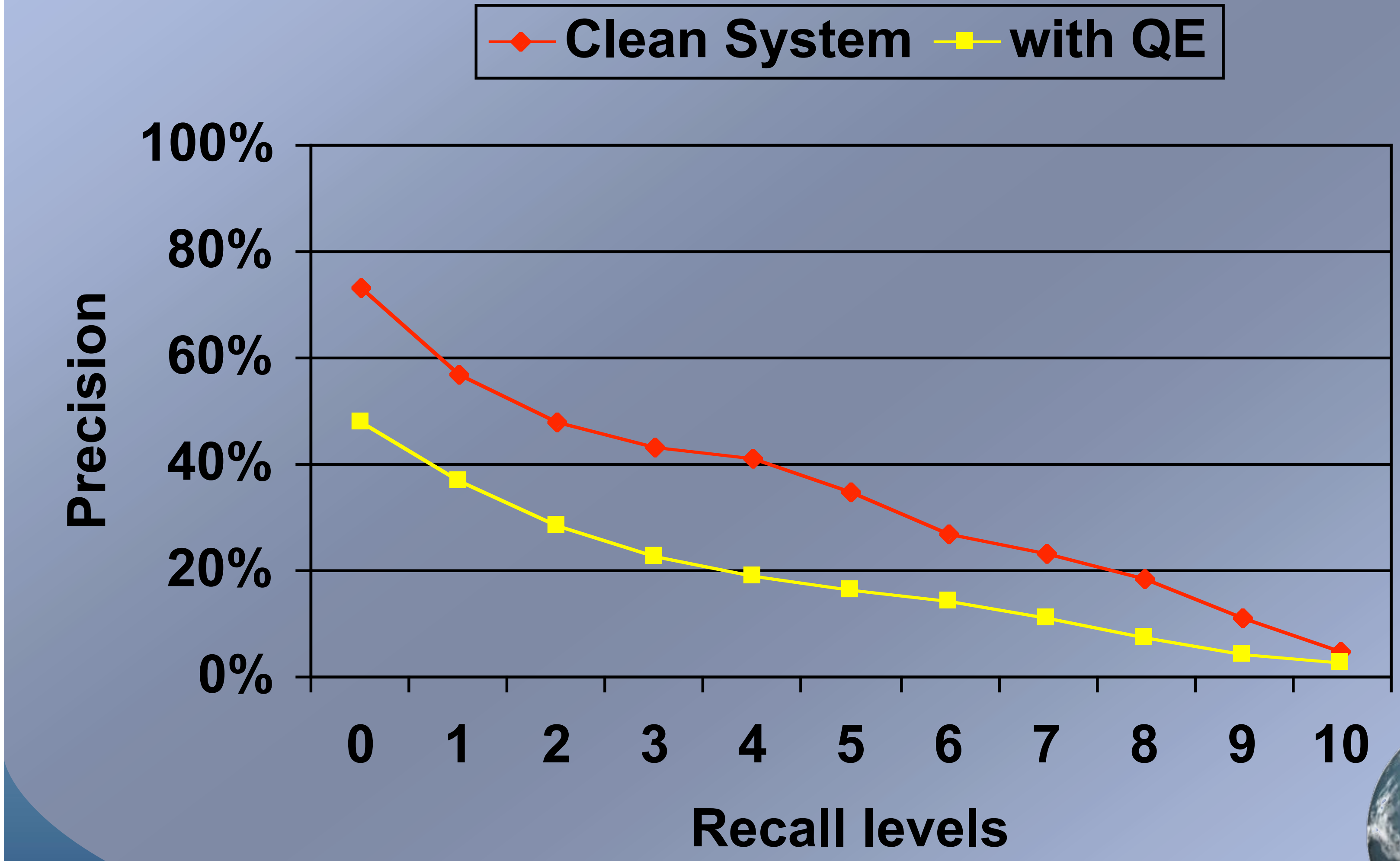
🟨 - To standard index

🟦 - To geographical index

Experiment Setup

- GeoCLEF 2005 collection and queries
  - Los Angeles Times 1994
  - Glasgow Herald 1995
- “Topic Description” runs:
  - Typical TD from queries:
    - “Shark attacks near California and Australia”
    - “Vegetable exporters of Europe”
    - “Holidays in the Scottish Trossachs”
- 1000 results returned for each query

Results - Query Expansion

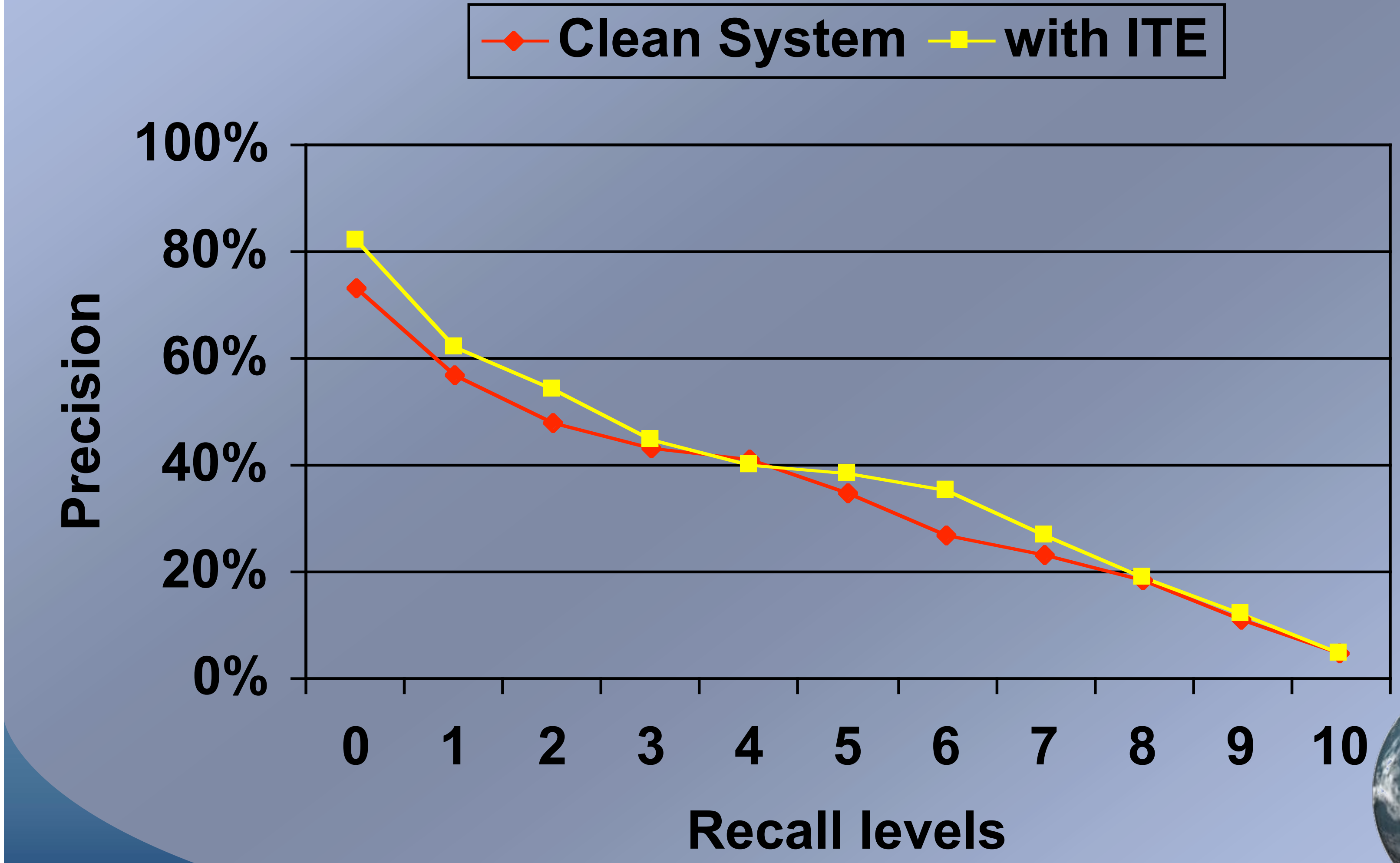


QE - Error Analysis

- Why did it perform so bad?
- Two major errors:
  - Inconsistent expansions
    - E.g. “Sacramento” expanding *California* in the query: “Shark attacks in California”
  - Ambiguity
    - E.g. “Europe” in “Vegetable exporters of Europe”
      - WordNet returns three senses for “Europe”:

      1. Europe as continent
      2. Europe as the European Union
      3. Europe as the set of nations on the European continent

Results - Index Terms Expansion



Conclusions

- ITE better than QE
  - Seems to be less sensitive to ambiguity problems
  - However: it needs NE recognition during the indexing phase (not trivial)
- WordNet *can* be used as a Geographical Information Resource
  - To be evaluated against a specialized resource like the TGN (http://www.getty.edu/research/conducting_research/vocabularies/tgn/)

Thank you!
Grazie!
Gracias!
Dhanyavaad! (Hindi)
Manjuthe! (Telugu)
Shukria! (Urdu)