GikiP: Evaluating Geographical Answers from Wikipedia

- Diana Santos and Nuno Cardoso
- Natural Language Technologies Group
  SINTEF ICT, Oslo, Norway

5th Workshop on Geographic Information Retrieval
CIKM 2008, Napa Valley, CA, USA, 30th October 2008

Goals

- Present the GikiP 2008 pilot evaluation task.
- Motivate you to participate with your GIR system in GikiCLEF 2009.

What is GikiP / GikiCLEF?

- GikiP is a pilot **evaluation task** held in 2008.
  GikiCLEF is the current task planned for 2009.
  http://www.linguateca.pt/GikiCLEF

- Task: **Find Wikipedia articles that answer a particular information need which requires geographical reasoning of some sort.**

GIR + QA + Wikipedia = GikiCLEF

GikiP / GikiCLEF's goals

- **Scientific goal**: Create synergies between geographic information retrieval (GIR) and question answering (QA).
- **Practical goal**: Wouldn't it be good if we had systems that could mediate between us & Wikipedia, and answer our complex questions, no matter the language?

Topic titles in GikiP 2008

Table 1: Topic titles in GikiP 2008

ID	English topic title
GP1	Which waterfalls are used in the film "The Last of the Mohicans"?
GP2	Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany?
GP3	Portuguese rivers that flow through cities with more than 150,000 inhabitants
GP4	Which Swiss cantons border Germany?
GP5	Name all wars that occurred on Greek soil.
GP6	Which Australian mountains are higher than 2000 m?
GP7	African capitals with a population of two million inhabitants or more
GP8	Suspension bridges in Brazil
GP9	Composers of Renaissance music born in Germany
GP10	Polynesian islands with more than 5,000 inhabitants
GP11	Which plays of Shakespeare take place in an Italian setting?
GP12	Places where Goethe lived
GP13	Which navigable rivers in Afghanistan are longer than 1000 km?
GP14	Brazilian architects who designed buildings in Europe
GP15	French bridges which were in construction between 1980 and 1990

Language biases: English German Portuguese Other

Topic titles in GikiP 2008

Table 1: Topic titles in GikiP 2008

ID	English topic title
GP1	Which waterfalls are used in the film "The Last of the Mohicans"?
GP2	Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany?
GP3	Portuguese rivers that flow through cities with more than 150,000 inhabitants
GP4	Which Swiss cantons border Germany?
GP5	Name all wars that occurred on Greek soil.
GP6	Which Australian mountains are higher than 2000 m?
GP7	African capitals with a population of two million inhabitants or more
GP8	Suspension bridges in Brazil
GP9	Composers of Renaissance music born in Germany
GP10	Polynesian islands with more than 5,000 inhabitants
GP11	Which plays of Shakespeare take place in an Italian setting?
GP12	Places where Goethe lived
GP13	Which navigable rivers in Afghanistan are longer than 1000 km?
GP14	Brazilian architects who designed buildings in Europe
GP15	French bridges which were in construction between 1980 and 1990

Kinds of subjects: Places Persons Events Buildings Plays

GikiP 2008 topics

- Geographically challenging
- Aiming for a list of entities (places, persons, cities, etc)
- Close to real users' needs
- Never answered by a single Wikipedia document (open list questions)

GikiP collection: Wikipedia

Wikipedia is a great collection to work on:

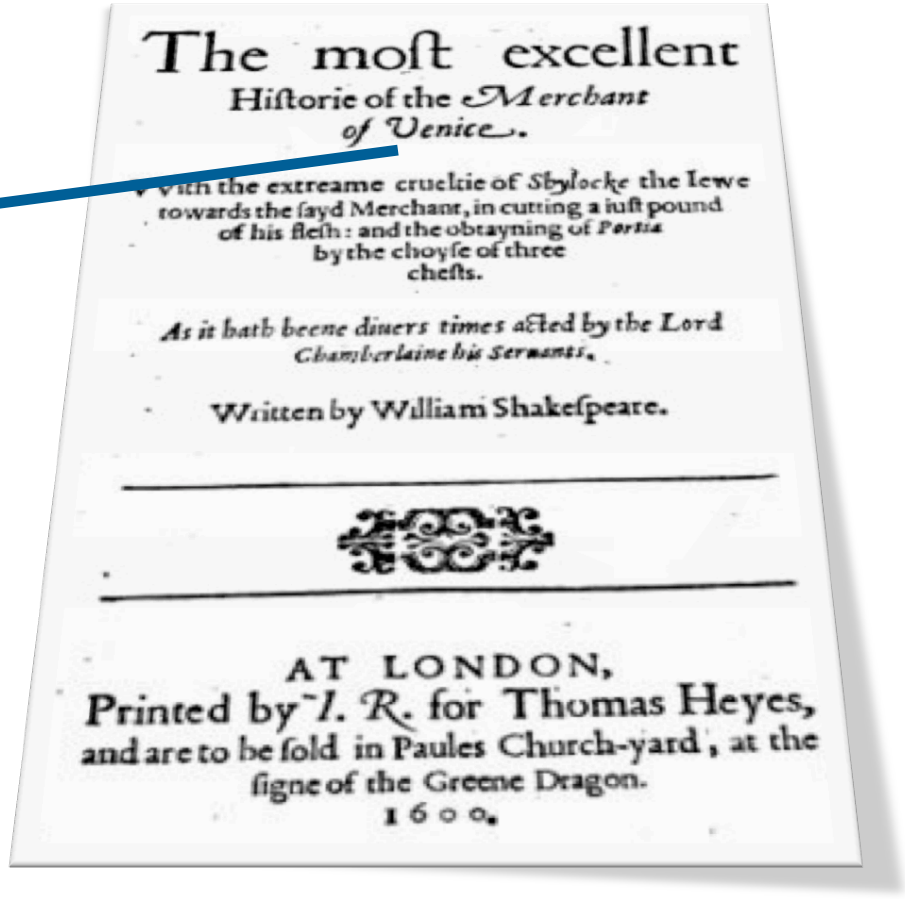
- Truly multilingual (dozens of languages)
- Spans several subjects
- Large
- Documents often well written and reviewed
- Rich content, structure and metadata
- Multimedia (text, audio, images)

Reasoning over the geographic domain

- Topic GP11: "*Which plays of Shakespeare take place in an Italian setting?*"

"Is Venice in Italy?"

Easy question for humans, but not so straight-forward for machines...


The moſt excellent
Hiſtorie of the Merchant of Venice.
With the extreame crueltie of Shylocke the Iewe towards the ſayd Merchant, in cutting a iuſt pound of his fleſh: and the obtayning of Portia by the choyſe of three cheſts.
As it hath beene diuers times acted by the Lord Chamberlaine his Seruants.
Written by William Shakeſpeare.
AT LONDON,
Printed by I. R. for Thomas Heyes, and are to be ſold in Paules Church-yard, at the ſigne of the Greene Dragon.
1600.

Systems in GikiP should ideally...

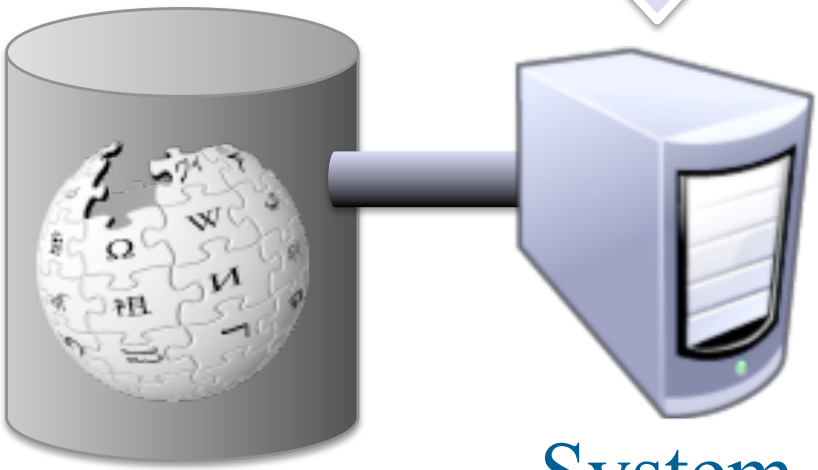
- ...**understand** what the topic really wants (a list of cities, people or events), and its restrictions (a given population/job/time threshold)
- ...**reason** over the Wikipedia collection and over the geographic domain (e.g., "*is Venice in Italy?*")
- ...**return** Wikipedia article names as **answers**: not lists, not overview pages, just the answers.

GikiP in an image…

GP9:

Komponisten von Renaissancemusik, die in Deutschland geboren wurden.
Composers of Renaissance music born in Germany.
Compositores renascentistas nascidos na Alemanha

Topic



System

Answers

(de) Arnolt Schlick
(de) Christoph Demantius
(de) Conrad Paumann
…
*(en) **Arnolt Schlick***
*(en) **Pierre Alamire***
…
(pt) Heinrich Finck
(pt) Hieronymus Praetorius
(pt) Leonhard Kleber
(pt) Martin Agricola
(pt) Michael_Praetorius
…

GikiP 2008 evaluation

- Wikipedia collection: DE, EN & PT ’06 snapshot.
- Runs: lists of answers, **manually** assessed.
- Multilingual systems are rewarded.
- Encourage systems to return short, accurate answers.

Score = *mult* * N * N/total

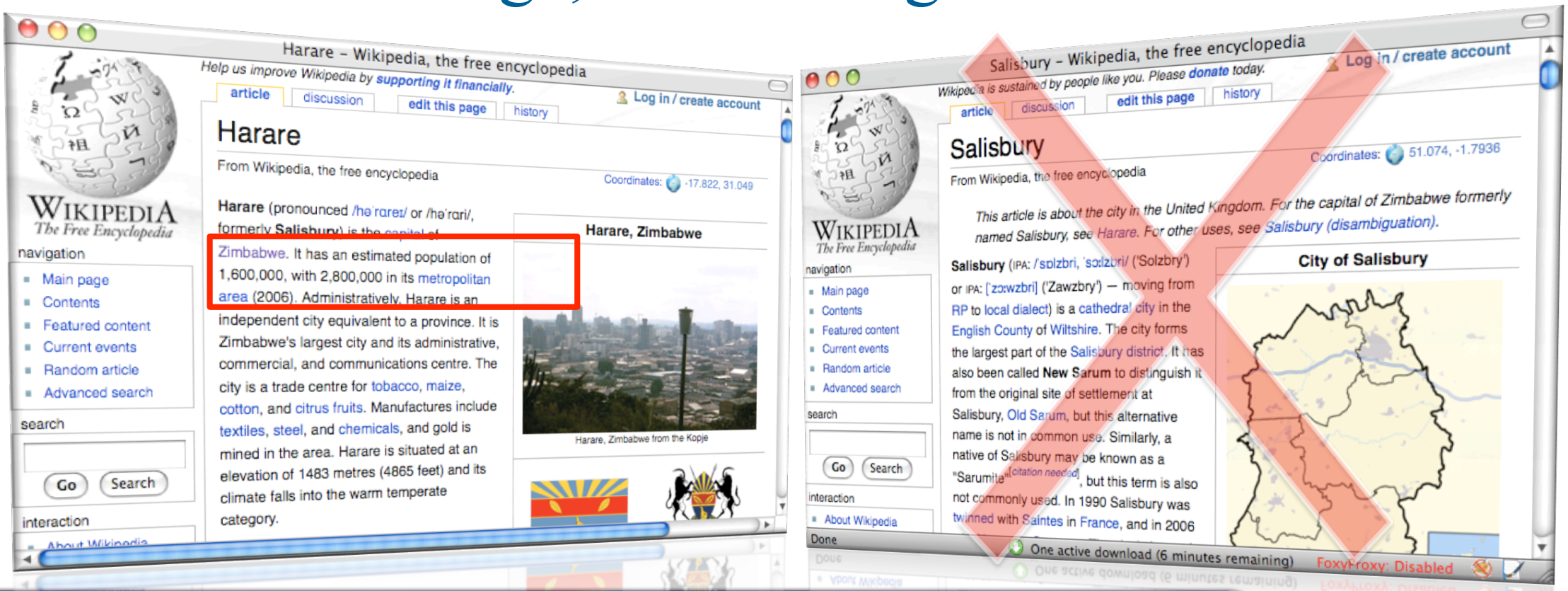
{1-3} 

Number of correct hits

= Precision

Interesting issues (1)

- GP7: "*African capitals with a population of two million inhabitants or more*".
- Names change, roles change!



Interesting issues (2)

- GP6: *"Which Australian mountains (...)"*
- Different languages, different meanings of geo-scope.
- Australia: in PT is only a country, not a continent.

Interesting issues (3)

- Different languages, different data
- Ex: From the Wikipedia pages of Harare…

English

Country	Zimbabwe
Province	Harare
Founded	1890
Incorporated (city)	1935
Government	
- Mayor	Muchadeyi Masunda
Elevation [1]	1,490 m (4,888 ft)
Population (2006)	
- City	1,600,000
- Urban	2,800,111 estimated

German

Wappen	Karte
	

Basisdaten	
Geografische Lage:	17° 51′ 50″ S, 31° 1′ 47″ O
Höhe:	1.490 m ü. NN
Fläche:	872 km²
Einwohner:	1.903.510 (2006)
Bevölkerungsdichte:	2.183 Einwohner/km²

Portuguese

Harare	
Capital	Harare
População	1.903.510 habitantes
Censo	2002
Área	872 km²
Densidade	2.182,92 hab/km²

Mapa

Interesting issues (4)

- GP5: "*Name all wars that occurred on Greek soil*" - Not all questions can be answered easily by a person.
  - There is no straight-forward category in Wikipedia to start with.
  - Even if there were a "Greek War" category, would it really include only wars fought on Greek soil, or all wars involving Greece?
  - Time issues: How was the Greek soil back then? Narrower or longer than today's Greek boundaries?

GikiP ➔ GikiCLEF 2009

- Goal: more participants,
  more languages, more challenges.
- Organizing committee (so far):
  Sören Auer, Gosse Bouma, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Corina Forascu, Sven Hartrumpf, Johannes Leveling, Constantin Orasan, Diana Santos, Yvonne Skalban

http://www.linguateca.pt/GikiCLEF

Collections for GikiCLEF 2009

- Wikipedia snapshots from 2008.
- Languages so far: Dutch, English, German, Norwegian, Portuguese, Romanian.
- Format on discussion. XML (using UvA's WikiXML tool) is a good candidate:
  - Converts MediaWiki format
  - Parses metadata to XML (infobox, crosslingual links, categories, sections. etc).

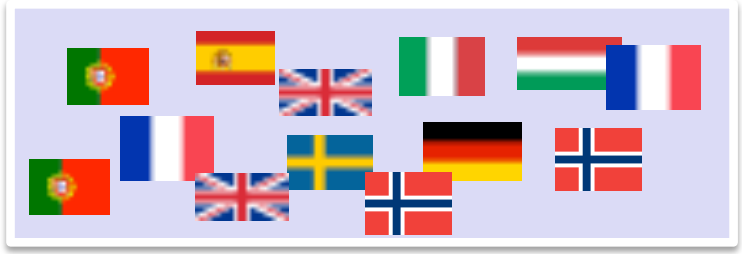
Culture-driven topics for GikiCLEF 2009

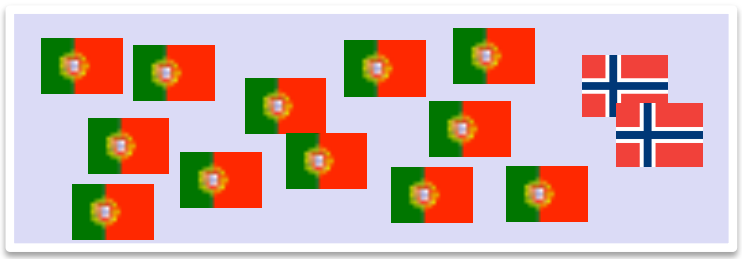
- *"Which Spanish writers lived in America in the XIX century?"*
- *"Which Mexican dishes contain tomato?"*
  - Promote multilinguality and crosslinguality: **culturally-aware topic choice**
  - (Possibly distinct) answers in a lot of Wikipedia languages

Evaluation for GikiCLEF 2009

New evaluation metrics on the forge…

- Rewarding multilingual & accurate answers.
- Encourage geographic diversity (?)
  - Example: “Name European cities with a Gothic cathedral”.





Evaluation for GikiCLEF 2009

Presentation of results:

- Lists with a given order?
  - Places visited by Magellan in a temporal order
  - Mountains ordered by height
  - Places ordered by specificity (Museum of Modern Art, New York, USA) – requires clustering based on inclusion relations

Task definition

- Questions/topics: all topics are provided in all GikiCLEF languages
- Exact answers: mainly NE (names), have to be a title of a Wikipedia page
- Assessment relatively easy and comparable (translation checking of named entities)
- Emphasis on complex geographic reasoning
- Continuation of temporal reasoning issues

Why should the GIR community participate in GikiCLEF?

- This is the only evaluation forum for GIR that we know of (follow-up of GeoCLEF):
  - A concrete task where geographical knowledge is called for;
  - Wikipedia: A realistic, heavily location-oriented collection.

GikiCLEF Schedule

- **Until the end of 2008** - Wikipedia collections made available to all participants
- **November 2008 - February 2009** - Final definition of the GikiCLEF task. Publication of the details of the task
- **March 2009 -** Topic release
- **(2 weeks after)** - Deadline for run submission
- **June 2009** - Assessment and results made available
- **September 2009** - CLEF workshop at Corfu, Greece

Join the GikiCLEF mailing-list in
http://www.linguateca.pt/GikiCLEF

- Questions?

GikiCLEF: Evaluating Geographical Answers from Wikipedia

Participating systems in GikiP 2008:

Countries:

- **GIRSA-WP**- Sven Hartrumpf and Johannes Leveling, Intelligent Information and Communication Systems (IICS), FernUniversität in Hagen (Germany)

- **RENOIR** - Nuno Cardoso, University of Lisbon, Faculty of Sciences, LaSIGE (Portugal)

- **WikipediaListQA@wlv** - Iustin Dornescu, Research Group in Computational Linguistics (CLG), University of Wolverhampton (UK)

Acknowledgements

- The organization work was done in the scope of Linguateca, contract no. 339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union, and administratively led by FCCN.
- This presentation was also partially funded by SINTEF ICT in the scope of GikiP follow-up that was submitted to CLEF by Nuno Cardoso (Univ. of Lisbon, Linguateca, and SINTEF ICT)