GikiCLEF overview

Diana Santos
Luís Miguel Cabral
Nuno Cardoso
Linguateca

GikiCLEF in a nutshell

- Asking open list questions to Wikipedia
- A difficult task for people and for computers
- A realistic situation where crosslingual and multilingual skills may be of real interest

http://www.linguateca.pt/GikiCLEF

- A merger of QA and GIR, a follow-up of the GikiP pilot
- 10 Wikipedia collections, 50 culturally motivated topics
- 8 participants in 2009

Examples of GikiCLEF-like systems in use

- In a mountain trip in the middle of Norway, an Italian family...
  - In which European countries is the bidet usual?
    - Portugal
    - Spain
    - Andorra
    - Italy

- In the middle of a football match on TV in a foreign station
  - What south American countries have yellow in their football team?
    - Brazil
    - Colombia
    - Ecuador
Examples of GikiCLEF-like systems in use

- Which German cities have more than one university?
  - Berlin
  - Bonn
  - Cologne
  - Aachen
  - Bremen
  - Augsburg
  - Hamburg
  - ...

Examples of GikiCLEF-like systems in use

- Where was this American museum which had a Picasso painting on which we saw that program last Winter?

Which American museums have Picasso works?
- Museum of Modern Art (MOMA)
- Museum of Fine Arts (Boston)
- Solomon R. Guggenheim Museum
- Metropolitan Museum of Art
- National Gallery of Art
- Art Institute of Chicago
- Denver Art Museum
- (..)

Examples of GikiCLEF-like systems in use

- Which Romanian poets published volumes with ballads before 1931?  ... which may have influenced Mircea Eliade? Preliminary research for a MA Thesis in Romanian literature...

Results from the topic group work

- We gathered a lot of more or less culturally-laden topics (70 -> 50)
  - Some of them cross-cultural, in fact (like the influence of Italy in Hemingway or the spread of Picasso’s influence in the American continent)
  - Most around themes such as thinkers/writers/famous people, and places
  - Not really too geographical for the most part
  - Still, Alps are subdivided differently in different countries/languages
  - Flemish towns turn out to be a not clear concept internationally
  - We allowed visual clues to decide (for snow, colours and/or left in a map)
  - A lot of little-developed and inaccurate Wikipedia pages were found a posteriori, showing that it may not be so obvious as initially thought that searching Wikipedia crosslingually is a good idea

The task as seen by a participant system

- In order to provide a correct answer, a system had to produce justification in at least one of the languages returned
- Even if correct, an answer would not be rewarded in GikiCLEF if it were not possible to assess by a human reader
- Justification could be in the page itself, or in a set of other pages provided together
  - Example question: Name places where Goethe fell in love.
  - One correct answer is Leipzig. But in the article about Leipzig this is not stated ☺
  - Where to find the justification? In "Johann Wolfgang von Goethe" page we find the following excerpt: "In Leipzig, Goethe fell in love with Käthchen Schönkopf and wrote cheerful verses about her in the Rococo genre."
  - Correct and justified GikiCLEF answer: Leipzig, Johann Wolfgang von Goethe

GikiCLEF evaluation measures computation

- C: Number of correct answers
- N: Total number of answers provided by the system
- GP: score per language: precision vs. C: C*C/N (so the score of Romanian will be CRO*CRO/NRO)
- Total score: Sum of all scores per languages: ΣGP_{lang}
Behind the scenes: SIGA

A system to help manage this evaluation track
SIGA: Sistema de Gestão e Avaliação do GikiCLEF

SIGA: GikiCLEF organization workflow

Administrator

Assessor

Topic manager

Participant

GikiCLEF topic management: edit topic

GikiCLEF topic management: find answers

GikiCLEF assessment system: assess answers

GikiCLEF conflict resolution

To check assessment coherence and understand better the whole assessment process, we assigned some overlap in the answers that the assessors should evaluate.

This caused (at least for some assessors) a total repetition of their assignments, but was deemed very useful.

Thanks to all 30 assessors: Sören Auer, Anabela Barreiro, Gisoe Bruma, Luís Miguel Cabral, Nuno Carvalho, Leda Canzona, Luís Costa, Inutin Dornezue, Ana Eifh, Cristina Fonseca, dPamela Ferton, Fredric Gey, Danilo Giampiccolo, Sven Hartrumpf, Katrin Lamm, Ray Larson, Laska Laskova, Johannes Leveling, Thomas Marsili, Cristina Mota, Constantin Onaxan, Petra Osenova, Anselmo Peñas, Erik Tjong Kim Sang, Diana Santos, Julia Scholz, Yvonne Schulten, Rosário Silva, Kiril Simov, Alvaro Rodrigo Yuste.
GikiCLEF assessment system: conflict resolution

- Only the minimum number of answers were assessed, so we had to propagate the scores to the runs (to the other languages).
- As well as confirm whether there were true multilingual conflict results -> another round of conflict resolution.
- For those, language propagation was inhibited.

- The language of the Vatican: Italian and/or Latin?
- Height of a Norwegian waterfall: EN: Rjukanfossen is a waterfall of 104 meter... NO: Fossen har en total fallhøyde på 238 meter og høyeste loddrette fall er 104 meter.

Overview of GikiCLEF 2009

Participants (8):
- Richard Flemmings et al. (BBK_UFRGS)
- Gosse Bouma & Sérgio Duarte (JoostER)
- Nuno Cardoso et al. (GREASE/XLDB)
- Iustin Dornescu (EQUAL)
- Ray R. Larson (CHESHIRE)
- Sven Hartrumpf & Johannes Leveling (GIRSA-WP)
- Daniel Ferrés & Horácio Rodríguez (GIKITALP)
- Adrian Iftene et al. (UAICGIKI09)

17 runs submitted.

Participants (8):
- Richard Flemmings et al. (BBK_UFRGS)
- Gosse Bouma & Sérgio Duarte (JoostER)
- Nuno Cardoso et al. (GREASE/XLDB)
- Iustin Dornescu (EQUAL)
- Ray R. Larson (CHESHIRE)
- Sven Hartrumpf & Johannes Leveling (GIRSA-WP)
- Daniel Ferrés & Horácio Rodríguez (GIKITALP)
- Adrian Iftene et al. (UAICGIKI09)

17 runs submitted.

Scores per language

<table>
<thead>
<tr>
<th>System</th>
<th>language</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG</td>
<td>9.757</td>
<td>18.980</td>
</tr>
<tr>
<td>NL</td>
<td>6.722</td>
<td>8.258</td>
</tr>
<tr>
<td>EN</td>
<td>13.657</td>
<td>12.007</td>
</tr>
<tr>
<td>DE</td>
<td>1.333</td>
<td>2.250</td>
</tr>
<tr>
<td>IT</td>
<td>3.030</td>
<td>1.798</td>
</tr>
<tr>
<td>NO</td>
<td>2.065</td>
<td>1.299</td>
</tr>
<tr>
<td>PT</td>
<td>3.030</td>
<td>1.798</td>
</tr>
<tr>
<td>RO</td>
<td>1.356</td>
<td>1.356</td>
</tr>
<tr>
<td>ES</td>
<td>1.635</td>
<td>0.267</td>
</tr>
<tr>
<td>Total</td>
<td>23.998</td>
<td>41.717</td>
</tr>
</tbody>
</table>

Final score

<table>
<thead>
<tr>
<th>System</th>
<th>#Answers</th>
<th>#Corrects</th>
<th>Precision</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQUAL</td>
<td>813</td>
<td>385</td>
<td>0.4736</td>
<td>181.9329</td>
</tr>
<tr>
<td>GREASE/XLDB</td>
<td>1161</td>
<td>332</td>
<td>0.2860</td>
<td>96.0070</td>
</tr>
<tr>
<td>Cheshire</td>
<td>564</td>
<td>214</td>
<td>0.3794</td>
<td>80.9247</td>
</tr>
<tr>
<td>GIRSA-WP</td>
<td>38</td>
<td>31</td>
<td>0.8158</td>
<td>24.7583</td>
</tr>
<tr>
<td>GIRSA-WP</td>
<td>985</td>
<td>142</td>
<td>0.1442</td>
<td>23.3919</td>
</tr>
<tr>
<td>GIRSA-WP</td>
<td>994</td>
<td>107</td>
<td>0.1076</td>
<td>14.5190</td>
</tr>
<tr>
<td>JoostER</td>
<td>638</td>
<td>36</td>
<td>0.564</td>
<td>2.4053</td>
</tr>
<tr>
<td>GIKITALP</td>
<td>356</td>
<td>26</td>
<td>0.0730</td>
<td>1.9018</td>
</tr>
<tr>
<td>GIKITALP</td>
<td>295</td>
<td>20</td>
<td>0.0678</td>
<td>1.3559</td>
</tr>
<tr>
<td>GIKITALP</td>
<td>526</td>
<td>18</td>
<td>0.0342</td>
<td>0.6964</td>
</tr>
<tr>
<td>bbk_ufrgs</td>
<td>726</td>
<td>8</td>
<td>0.0110</td>
<td>0.0882</td>
</tr>
<tr>
<td>UAICGIKI09</td>
<td>6420</td>
<td>8</td>
<td>0.0012</td>
<td>0.0156</td>
</tr>
<tr>
<td>bbk_ufrgs</td>
<td>734</td>
<td>3</td>
<td>0.0041</td>
<td>0.0123</td>
</tr>
<tr>
<td>UAICGIKI09</td>
<td>1133</td>
<td>2</td>
<td>0.0018</td>
<td>0.0062</td>
</tr>
<tr>
<td>JoostER</td>
<td>272</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>bbk_ufrgs</td>
<td>686</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>UAICGIKI09</td>
<td>4910</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Scores per language

<table>
<thead>
<tr>
<th>System</th>
<th>#Participations</th>
<th>Total language score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG</td>
<td>8</td>
<td>23.998</td>
</tr>
<tr>
<td>NL</td>
<td>9</td>
<td>41.717</td>
</tr>
<tr>
<td>EN</td>
<td>12</td>
<td>78.340</td>
</tr>
<tr>
<td>DE</td>
<td>8</td>
<td>52.690</td>
</tr>
<tr>
<td>IT</td>
<td>8</td>
<td>31.602</td>
</tr>
<tr>
<td>NO</td>
<td>8</td>
<td>43.106</td>
</tr>
<tr>
<td>PT</td>
<td>8</td>
<td>29.138</td>
</tr>
<tr>
<td>RO</td>
<td>11</td>
<td>34.883</td>
</tr>
<tr>
<td>ES</td>
<td>9</td>
<td>34.647</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>427.193</td>
</tr>
</tbody>
</table>
Information and Communication Technologies

Results overview per language

Topics with most answers (by the participant set)
- GC-2009-17 (788): List the 5 Italian regions with a special statute
- GC-2009-20 (674): List the name of the sections of the North-Western Alps
- GC-2009-02 (646): Which countries have the white, green and red colors in their national flag?
- GC-2009-07 (640): What capitals of Dutch provinces received their town

Topics with least answers (by the participant set)
- GC-2009-04 (78): Chefs born in Austria who have received a Michelin Star.
- GC-2009-03 (130): In which countries outside Bulgaria are there opinions on Petar Dunov’s (Beinsa Duno) ideas?
- GC-2009-01 (240): List the Italian places which Ernest Hemingway visited during his life.

Number of answers per topic

Topics result overview: by precision
- Highest precision
  - GC-2009-34 (43.9%): What eight thousands are at least partially in Nepal?
  - GC-2009-07 (25.3%): Which capitals of Dutch provinces received their town privileges before the fourteenth century?
  - GC-2009-11 (23.8%): What Belgians won the Ronde van Vlaanderen exactly twice?
  - GC-2009-25 (17.13%): Name Spanish drivers who have driven in Minardi.

Precision by topic
Authoritative languages: where is the justification

For each language, was the justification found in the corresponding Wikipedia?

Public resources

http://www.linguatexta.pt/GikiCLEF/
- Topic lists per language, in XML and text format
- Topic descriptions (in English, together with assessment decisions)
- Lists of correct answers
  - Correct and justified in some language: GikiCLEF_answers_correct_justified.txt
  - Correct (because we know): GikiCLEF_answers_correct.txt
- Results (global, per language) and general statistics
- SIGA system, open source
- GikiCLEF collections

Final discussion: Was GikiCLEF worthwhile?

- Recurrent comment: too difficult!
- Often not clear which (Wikipedia) pages were sought
  - Flag pages or country pages?
  - Team pages or country pages?
- Most (organizers) were probably not expecting so much work

- It is possible to create a system capable of answering many languages
- It pays to process English! The amount provided by each language, even in culturally-aware topics, is negligible
- The quality of other languages’ Wikipedias is in strong need of improvement

Further details

Size of collections (pages * 1k)

All collections made available were dated from June 2008, except for the English collection SQL dump, from May and July instead (since the SQL dump of June was not available)

Number of authoritative answers per wiki size
"Overlap" ratio of repeated vs. total, but some answers were assessed by three or even four assessors, since repetition was automatically computed by SIGA.

Correct and justified answers VS correct not justified answers

Only 2 systems made use of the justification field (4 runs) with a total of 229 answers with another justification document.