GikiCLEF topics and Wikipedia articles: Did it blend?

Motivation
Using Wikipedia to answer GikiCLEF topics proved to be quite difficult. Where is the problem, in the task or in the systems? Did the systems mined Wikipedia conveniently, or did they manage to scrape the surface? Are we aware of the difficulties in finding the right answers in Wikipedia? Let’s have a look on the Wikipedia collection and in the GikiCLEF task.

GikiCLEF topics

<table>
<thead>
<tr>
<th>Language bias of topics</th>
<th>none</th>
<th>de</th>
<th>el</th>
<th>es</th>
<th>fr</th>
<th>it</th>
<th>pt</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Geoscope distribution of topics

Expected Answer types

Wikipedia solutions

Solutions: entities found in the pool of all GikiCLEF participant’s answers, merged to a single identifier
Example: Italy is one solution for a topic, regardless of the different answers given by Wikipedia articles from different languages.
Italy, Itália, Italia, Italien, Италия → Italy (country)

Language coverage over the solutions

Number of solutions per topic

Solution location in Wikipedia

Best languages per topic bias

Salient blending disturbers:
- Wikipedia is still too pivoted on English language, which performed better on biased topics.
- The solutions of GikiCLEF 2009 topics were mostly found on the text body. Structured elements, like infoboxes, tables or ordered lists, are helpful but are rare and still machine-unfriendly.
- There was a significative amount of translated pages (EN / X or X / EN) as well as solutions that are normally on two languages (EN and X), thus no information gain of processing multiple languages.
- Wikipedia categories were difficult to work on – they differ on usage patterns per language, they are noisy and too structured (ex: Cities of Germany).

Acknowledgments: Linguistic project is funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/3391-35MAC, by UMIC and FCCN. Nuno Cardoso is supported by FCT scholarship grant SFRH/BD/45480/2008.