Sad horse
WebCLEF — Goodbye!

Maarten de Rijke
Agenda

• History
• WebCLEF this year
• What’s next?
Where it all started ...
Where it all started …

• So you want to do multilingual information access?
Where it all started …

- So you want to do multilingual information access?
- Work with web data, with web data in a European setting!
- For many of us Europeans, dealing with multilingual text is all we know
A bit of history

- Launched as a known-item search task in 2005, repeated in 2006
- Resources created used for a number of purposes
- But there are information needs out there beside navigational ones, even on the web
A bit of history

3. A taxonomy of web searches

In the web context the "need behind the query" is often not informational in nature. We classify web queries according to their intent into 3 classes:

1. **Navigational.** The immediate intent is to reach a particular site.
2. **Informational.** The intent is to acquire some information assumed to be present on one or more web pages.
3. **Transactional.** The intent is to perform some web-mediated activity.

Before we discuss these types in detail, we need to clarify that there is no assumption here that this intent can be inferred with any certitude from the query. The examples below might have alternative explanations.
A bit of history

- Launched as a known-item search task in 2005, repeated in 2006
- Resources created used for a number of purposes
- But there are information needs out there beside navigational ones, even on the web
### A bit of history

Launched as a known-item search task in 2005, repeated in 2006. Resources created used for a number of purposes.

But there are information needs out there beside navigational ones, even on the web.

---

#### Search GOAL

<table>
<thead>
<tr>
<th>DESCRIPTION</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Navigational</td>
<td>aloha airlines, duke university hospital, kelly blue book</td>
</tr>
<tr>
<td>2. Informational</td>
<td>what is a supercharger, 2004 election dates, baseball death and injury, why are metals shiny</td>
</tr>
<tr>
<td>2.1 Directed</td>
<td>color blindness, jfk jr</td>
</tr>
<tr>
<td>2.1.1 Closed</td>
<td>help quitting smoking, walking with weights</td>
</tr>
<tr>
<td>2.1.2 Open</td>
<td>pella windows, phone card</td>
</tr>
<tr>
<td>2.2 Undirected</td>
<td>travel, amsterdam universities, florida newspapers</td>
</tr>
<tr>
<td>2.3 Advice</td>
<td>kaza lite, name roms</td>
</tr>
<tr>
<td>2.4 Locate</td>
<td>xxx porno movies free, live cameras in l.a.</td>
</tr>
<tr>
<td>2.5 List</td>
<td>watchor, measure converter</td>
</tr>
<tr>
<td>3. Resource</td>
<td>free jack o lantern patterns, ellis island lesson plans, house document no. 587</td>
</tr>
</tbody>
</table>
A bit of history

- Launched as a known-item search task in 2005, repeated in 2006
- Resources created used for a number of purposes
- But there are information needs out there beside navigational ones, even on the web
Switching tasks
Switching tasks

- In 2007, launched a multilingual information synthesis task
- “For a given topic, participating systems extract important snippets from web pages found using a web search engine”
Switching tasks

- In 2007, launched a multilingual information synthesis task
  - “For a given topic, participating systems extract important snippets from web pages found using a web search engine”
- Topics and assessments created by participants
  - Assessment environment provided by UAms
In 2007

- Only few participants
- Afraid of the task?
- Afraid of the amount of data to be processed?
- Afraid of the content extraction task?
- Afraid of the implementation effort?
- Poorly advertized?
- Just an odd year?
In 2008

- We gave participants the data
- We gave participants extracted content
- We gave participants a system
- We mailed, we announced, we phoned up, we invited, …
The outcome?
The outcome?

- First things first …
- The user model
  - knowledgable person writing a survey article on specific topic with clear goal and audience
  - user needs to locate items of information to be included in article and wants to use an automatic system for this purpose
  - user only uses online sources found via a Web search engine
The outcome (2)

- Information needs specified as
  - short topic title
  - free text description of goals and intended audience of the article
  - list of languages in which user is willing to accept information found
  - optional list of known sources: resources (URLs of web pages) considered relevant
  - optional list of Google retrieval queries
The outcome (3)

- **topic title:** *Paul Verhoeven*
- **description:** I'm looking for information on similarities, differences, connections, influences between Paul Verhoeven's movies of his Dutch period and his American period
- **language:** English, Dutch
- **known source(s):**
  - http://en.wikipedia.org/wiki/Paul_Verhoeven, ...
- **retrieval queries:**
  - paul verhoeven (dutch AND american)''
  - paul verhoeven (nederlandse AND amerikaanse OR hollywood OR VS)
The outcome (4)

- 61 multilingual topics
  - 21 UK-ES
  - 21 EK-NL
  - 10 UK-RO-ES
  - 6 RU-EN
  - 2 EN-GE-NL
  - 1 RU-EN-NL
Test collection:

- Web docs found using Google with queries provided by topic providers
- For each topic:
  - all “known” sources
  - top 100 hits (or less, depending on actual availability)—down from 1000 in 2007
- for doc
  - URL, original content, plain text conversion (UTF-8), the query/queries that retrieved it, the rank it was returned
The outcome (6)

- Similar to, but simpler than in 2007
  - For a given topic all responses of all system were pooled into an anonymized randomized sequence of text segments
  - To limit amount of assessments required, for each topic only first 7,000 characters of each response were included (according to the ranking of the snippets in response)
  - For the pool created in this way for each topic, the assessors were asked to mark text spans that either (1) repeat the information already present in the known sources, or (2) contain new important information.
  - Unlike in 2007, assessors were not asked to group such text snippets into subtopics (by using nuggets), as the 2007 assessment results proved inconsistent with respect to nuggets.
  - The assessors used a GUI to mark character spans in the responses
So?
Runs and results

- 9 runs submitted by 3 research groups
- Twente, UNED, Salamanca
- 1 run by best performing 2007 system
- As of yesterday morning, 31 topics have been assessed, remainder expected by end of next week
Runs and results

- 9 runs submitted by 3 research groups: Twente, UNED, Salamanca
- 1 run by best performing 2007 system
- As of yesterday morning, 31 topics have been assessed, remainder expected by end of next week

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.0978</td>
<td>0.0808</td>
</tr>
<tr>
<td>ip2008</td>
<td>0.2537</td>
<td>0.2097</td>
</tr>
<tr>
<td>ipt2008</td>
<td>0.2651</td>
<td>0.2191</td>
</tr>
<tr>
<td>ipu2008</td>
<td>0.2218</td>
<td>0.1833</td>
</tr>
<tr>
<td>Uned_RUN1</td>
<td>0.2720</td>
<td>0.2248</td>
</tr>
<tr>
<td>Uned_RUN2</td>
<td>0.2243</td>
<td>0.1854</td>
</tr>
<tr>
<td>Uned_RUN3</td>
<td>0.2445</td>
<td>0.2020</td>
</tr>
<tr>
<td>usal_0</td>
<td>0.2334</td>
<td>0.1927</td>
</tr>
<tr>
<td>usal_1</td>
<td>0.1299</td>
<td>0.1039</td>
</tr>
<tr>
<td>usal_2</td>
<td>0.1032</td>
<td>0.0825</td>
</tr>
</tbody>
</table>
Runs and results (2)

- Preliminary findings
  - runs consistently outperform baseline
  - there are some peaks, but there are no consistently “easy” topics
  - no run consistently outperforms other runs (unlike last year)
- This was WebCLEF’s final year
- Parallel session tomorrow
  - 9.00–9.45
  - a look back & a look forward re: web ir
Let’s step back
Why do you participate?
Why do you participate?

- Because your funding agency tells you to do task-based evaluation?
Why do you participate?

- Because your funding agency tells you to do task-based evaluation?
- To create resources?
Why do you participate?

- Because your funding agency tells you to do task-based evaluation?
- To create resources?
- To publish papers?
  - CLEF SIGIR, TOIS
Why do you participate?

- Because your funding agency tells you to do task-based evaluation?
- To create resources?
- To publish papers?
- CLEF SIGIR, TOIS
- To learn something?
Why do you participate?

- Because your funding agency tells you to do task-based evaluation?
- To create resources?
- To publish papers?
- CLEF SIGIR, TOIS
- To learn something?
- Whatever it is, you need mass and innovation
Oh yes, ...
Oh yes, ... 

- Mass and innovation will help you, your CV and your field
Oh yes, …

- Mass and innovation will help you, your CV and your field
- But what if you could make a difference elsewhere?
Thanks!

Thanks!