CLEF Monolingual Grid of Points

Nicola Ferro  @frrncl  Gianmaria Silvello  @giansilv

Information Management Systems Research Group
Department of Information Engineering
University of Padua, Italy
Outline

- The Problem
- The CLEF Monolingual GoP
- Validation of the GoP
- Example of Analyses
- Conclusions and Future Work
The Problem
The Problem

If we want to decide between alternative indexing strategies, we must use these strategies as part of a complete information retrieval system, and examine its overall performance (with each of the alternatives) directly.

[Robertson, 1981]
Typical Situation in Evaluation Campaigns

- Merging Strategies
- Pivot Languages
- Parallel Corpora
- Machine Translation

- Learning To Rank
- Divergence From Randomness
- Query Expansion
- Relevance Feedback

- Language Models
- Probabilistic Model
- Boolean Model
- Vector Space Model

- Components
- Stemmer
- N-grams
- Stop List
- Word de-compounder

- Bulgarian
- Spanish
- Russian
- Portuguese
- Italian
- Hungarian
- German
- French
- Finnish
- English
- Dutch

- Bulgarian
- Learning To Rank
- N-grams
- Stop List

- Swedish
- English
- Probabilistic Model
- Boolean Model

- German
- French
- Hungarian
- Russian
- Czech
- Portuguese
- Spanish
- Finnish
- Dutch
- English
- Russian

N. Ferro and G. Silvello  The CLEF Monolingual Grid of Points
Grid@CLEF: Back in 2009

To conduct a series of systematic and comparable grid experiments across languages and components by performing a community effort to evaluate not only each others components but also their interaction.

[Ferro and Harman, 2010]
To conduct a series of systematic and comparable grid experiments across languages and components by performing a community effort to evaluate not only each others components but also their interaction.

A proper methodology to analyze these data was lacking.

[Ferro and Harman, 2010]
The purpose of the exercise was to invite the developers of open-source search engines to provide reproducible baselines of their systems in a common environment.
- The purpose of the exercise was to invite the developers of open-source search engines to provide reproducible baselines of their systems in a common environment.

<table>
<thead>
<tr>
<th>System</th>
<th>Model</th>
<th>Stop</th>
<th>Stem</th>
<th>bg</th>
<th>de</th>
<th>es</th>
<th>fa</th>
<th>fi</th>
<th>fr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrier</td>
<td>BM25</td>
<td>✓</td>
<td></td>
<td>0.2092</td>
<td>0.2733</td>
<td>0.3627</td>
<td>0.4033</td>
<td>0.3464</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>BM25</td>
<td>✓</td>
<td></td>
<td>0.2081</td>
<td>0.2742</td>
<td>0.3656</td>
<td>0.4022</td>
<td>0.3392</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>BM25</td>
<td>✓</td>
<td></td>
<td>0.3194</td>
<td>0.4347</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>BM25</td>
<td>✓</td>
<td></td>
<td>0.3215</td>
<td>0.4516</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>Hiemstra LM</td>
<td>✓</td>
<td></td>
<td>0.1647</td>
<td>0.2520</td>
<td>0.3016</td>
<td>0.3140</td>
<td>0.3125</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>Hiemstra LM</td>
<td>✓</td>
<td></td>
<td>0.1640</td>
<td>0.2561</td>
<td>0.3081</td>
<td>0.3193</td>
<td>0.3156</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>Hiemstra LM</td>
<td>✓</td>
<td></td>
<td>0.2753</td>
<td>0.3673</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>Hiemstra LM</td>
<td>✓</td>
<td></td>
<td>0.2801</td>
<td>0.3783</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>PL2</td>
<td>✓</td>
<td></td>
<td>0.2043</td>
<td>0.2925</td>
<td>0.3266</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>PL2</td>
<td>✓</td>
<td></td>
<td>0.2001</td>
<td>0.2813</td>
<td>0.3182</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>PL2</td>
<td>✓</td>
<td></td>
<td>0.3138</td>
<td>0.4198</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>PL2</td>
<td>✓</td>
<td></td>
<td>0.3102</td>
<td>0.4211</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>TFIDF</td>
<td>✓</td>
<td></td>
<td>0.2071</td>
<td>0.2709</td>
<td>0.3277</td>
<td>0.3007</td>
<td>0.3046</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>TFIDF</td>
<td>✓</td>
<td></td>
<td>0.2083</td>
<td>0.2729</td>
<td>0.3274</td>
<td>0.3033</td>
<td>0.3393</td>
<td></td>
</tr>
<tr>
<td>Terrier</td>
<td>TFIDF</td>
<td>✓</td>
<td></td>
<td>0.3142</td>
<td>0.4373</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Lucene</td>
<td>BM25</td>
<td>✓</td>
<td></td>
<td>0.3126</td>
<td>0.4256</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Lucene</td>
<td>LM Dirichlet</td>
<td>✓</td>
<td></td>
<td>0.2051</td>
<td>0.1365</td>
<td>0.2586</td>
<td>0.2306</td>
<td>0.3865</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: MAP@1000 scores on the benchmarked CLEF collections. Languages are expressed as ISO 639:1 two letters code. "Stop" indicates if a stop-list was used and "Stem" if stemmer was used.
The CLEF Monolingual GoP
The Monolingual Grid of Points

- We release as an open resource the first fine-grained grids of points for many of the CLEF monolingual Adhoc tasks over a range of several years.

- The goal of these grids is to facilitate research in the MLIA field, to provide a set of standard baseline on standard collections, and to offer the possibility of conducting deeper analyses on the interaction among components in multiple languages.
Experimental Setup

- A Grid of Points (GoP) consisting of 160 systems for each monolingual adhoc task at hand has been created with all the possible combinations of the following components by using the Terrier v4.1 system

  - **Stop list**: nostop, stop;

  - **Lexical Unit Generator (LUG)**:
    - nostemmer, weak stemmer, aggressive stemmer;
    - nograms, 4grams, 5grams;

  - **Model**: BB2, BM25, DFRBM25, DFRee, DLH, DLH13, DPH, HiemstraLM, IFB2, InL2, InexpB2, InexpC2, LGD, LemurTFIDF, PL2, TFIDF.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian (bg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French (fr)</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>German (de)</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian (hu)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
</tr>
<tr>
<td>Italian (it)</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish (sv)</td>
<td>AdHoc</td>
<td>AdHoc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Stoplist and Stemmers

<table>
<thead>
<tr>
<th>Language</th>
<th>Stoplist</th>
<th>Weak Stemmer</th>
<th>Aggressive Stemmer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian (bg)</td>
<td>UNINE 258 words</td>
<td>UNINE light stemmer</td>
<td>UNINE stemmer</td>
</tr>
<tr>
<td>German (de)</td>
<td>UNINE 603 words</td>
<td>UNINE light stemmer</td>
<td>Snowball stemmer</td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>UNINE 307 words</td>
<td>UNINE light stemmer</td>
<td>Snowball stemmer</td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td>UNINE 747 words</td>
<td>UNINE light stemmer</td>
<td>Snowball stemmer</td>
</tr>
<tr>
<td>French (fr)</td>
<td>UNINE 463 words</td>
<td>UNINE light stemmer</td>
<td>Snowball stemmer</td>
</tr>
<tr>
<td>Hungarian (hu)</td>
<td>UNINE 737 words</td>
<td>Snowball stemmer</td>
<td>UNINE light stemmer</td>
</tr>
<tr>
<td>Italian (it)</td>
<td>UNINE 399 words</td>
<td>UNINE light stemmer</td>
<td>Snowball stemmer</td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td>UNINE 356 words</td>
<td>UNINE light stemmer</td>
<td>Snowball stemmer</td>
</tr>
<tr>
<td>Swedish (sv)</td>
<td>UNINE 386 words</td>
<td>UNINE light stemmer</td>
<td>Snowball stemmer</td>
</tr>
</tbody>
</table>

- UNINE stands for University of Neuchâtel

- [http://members.unine.ch/jacques.savoy/clef/index.html](http://members.unine.ch/jacques.savoy/clef/index.html)
Grid of Points (GoP)

http://gridofpoints.dei.unipd.it/
Grid of Points (GoP)

http://gridofpoints.dei.unipd.it/

- **Data**: The Average Precision values of the GoP runs for all the considered tasks
- **Scripts**: shell scripts to calculate the GoP
- **Java**: stemmer and n-grams extensions to Terrier v4.1
- **Matlab**: scripts and functions to import the runs and calculate the measures
Validation of the GoP
MAP Comparison

Original Data
MAP Distribution

GoP Data
MAP Distribution

N. Ferro and G. Silvello
The CLEF Monolingual Grid of Points
slide 16
Comparison across years is not possible: different collections and different systems

Comparison across years begins to be viable: different collections but same systems
Comparison across years is not possible: different collections and different systems.

Comparison across years begins to be viable: different collections but same systems.

By now, a viable way to do an inter-collection analysis is standardization.

William Webber, Alistair Moffat, Justin Zobel: 
*Score standardization for inter-collection comparison of retrieval systems.*
SIGIR 2008: 51-58

Nicola Ferro, Gianmaria Silvello: 
*3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc-ish CLEF?*
Comparison Between Original and GoP runs

- GoP runs are a good approximation of the original runs

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian (bg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French (fr)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German (de)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian (hu)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian (it)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish (es)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish (sv)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

KL-Divergence

- KLD varies between 0 and +inf: the lower the value, the closer the probability distribution between original and GoP runs
Comparison Between Original and GoP runs

- GoP runs are a good approximation of the original runs

- KLD varies between 0 and $+\infty$: the lower the value, the closer the probability distribution between original and GoP runs
Analyses
Effects of the Stemmers/n-Grams/Models

Multivari plot for CLEF 2003, Monolingual French (Average Precision)
Breakdown of Components Contributions

\[ Y_{ijkl} = \mu + \tau_i + \alpha_j + \beta_k + \gamma_l + \]

Main Effects

\[ \alpha\beta_jk + \alpha\gamma j_l + \beta\gamma kl + \alpha\beta\gamma jkl \]

Interaction Effects

Error

**Factor A (Stop Lists)**

<table>
<thead>
<tr>
<th>( A_1 )</th>
<th>( A_2 )</th>
<th>( \cdots )</th>
<th>( A_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_1 )</td>
<td>( B_2 )</td>
<td>( \cdots )</td>
<td>( B_q )</td>
</tr>
<tr>
<td>( B_1 )</td>
<td>( B_2 )</td>
<td>( \cdots )</td>
<td>( B_q )</td>
</tr>
<tr>
<td>( B_1 )</td>
<td>( B_2 )</td>
<td>( \cdots )</td>
<td>( B_q )</td>
</tr>
</tbody>
</table>

**Factor B (Lexical Unit Generator)**

<table>
<thead>
<tr>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( \cdots )</th>
<th>( C_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_1 )</td>
<td>( T_2 )</td>
<td>( \cdots )</td>
<td>( T_q )</td>
</tr>
<tr>
<td>( T_1 )</td>
<td>( T_2 )</td>
<td>( \cdots )</td>
<td>( T_q )</td>
</tr>
<tr>
<td>( T_1 )</td>
<td>( T_2 )</td>
<td>( \cdots )</td>
<td>( T_q )</td>
</tr>
</tbody>
</table>

**Factor C (Models Subjects Topics)**

<table>
<thead>
<tr>
<th>( T_1 )</th>
<th>( T_2 )</th>
<th>( \cdots )</th>
<th>( T_q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_1 )</td>
<td>( T_2 )</td>
<td>( \cdots )</td>
<td>( T_q )</td>
</tr>
<tr>
<td>( T_1 )</td>
<td>( T_2 )</td>
<td>( \cdots )</td>
<td>( T_q )</td>
</tr>
</tbody>
</table>

---

[Ferro and Silvello, SIGIR 2016]

N. Ferro and G. Silvello
Main Effects Analysis

Italian Monolingual 2003 - Stemmer group

Italian Monolingual 2003 - n-grams group
Interaction Effects Analysis

Italian Monolingual 2003 - Stemmer group

Stop List*IR Model Interaction

Stemmer*IR Model Interaction

Stop List*Stemmer Interaction

Italian Monolingual 2003 - n-grams group

Stop List*IR Model Interaction

N-grams*IR Model Interaction

Stop List*N-grams Interaction
Conclusions
Conclusions

- The CLEF monolingual GoP presents many MLIA systems as a combination of different IR components
- The tested MLIA systems are open-source and allow for easy reproducibility of results
- GoP is the most complete grid for analyzing MLIA system components
- GoP is openly available for further analyses and studies
Future Work

- Extend the monolingual CLEF GoP to new components; e.g., more stop lists, stemmers and n-grams

- Apply the SIGIR 2016 methodology (GLMM) to decompose and determine single component main and interaction effects by removing topic variance*

- Define a GoP for bilingual and multilingual systems