CLEF 2000 – 2014: Lessons Learnt from Ad Hoc Retrieval

Nicola Ferro and Gianmaria Silvello

Information Management Systems Research Group
Department of Information Engineering
University of Padua

{ferro, silvello}@dei.unipd.it
Outline

- Motivations
- Standardization
- Analysis of CLEF Ad-Hoc Campaigns
- Wrapping up
Motivations
Why Evaluation?

“To measure is to know.”

“If you cannot measure it, you cannot improve it.”

Lord William Thompson, first Baron Kelvin (1824-1907)
How Does Experimental Evaluation Work?

- Cranfield Paradigm
  - Dates back to mid 1960s

- Makes use of experimental collections
  - Documents, topics, relevance judgments

- Experimental collections and the results of the evaluation are shared and re-usable

- Ensures comparability and reproducibility of the experiments
Experimental Evaluation: Economic Impact

- The **TREC 2010 Economic Impact** study estimated in about **30 M$** the overall **investment** in **TREC** by NIST.

- probably much much more if we had a means to estimate also the investment by participants in TREC.

- They are the **pillars** for all the subsequent **scientific research** and **technology development**.

- TREC estimated the **return on investment** in the range of **3$-5$** for each invested dollar.
The European contest: CLEF

The CLEF Initiative is a self-organized body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure.

- **multilingual** and multimodal system testing, tuning and evaluation;
- investigation of the use of unstructured, semi-structured, **highly-structured**, and semantically enriched data in information access;
- an open forum for **research groups** and **industries** to discuss and share competencies;
- discussion of results, comparison of approaches, exchange of ideas, and **transfer of knowledge**.
Our goal is to carry out a longitudinal study the evolution of CLEF in the last 15 years in order to understand its impact on monolingual, bilingual, and multilingual search.
Motivations

- There have been very few systematic longitudinal studies about the impact of evaluation campaigns on the overall effectiveness of IR systems

- e.g. SMART system tested on TREC 1 to 8

- C. Buckley, 2005.
The SMART project at TREC, TREC - Experiment and Evaluation in Information Retrieval. MIT Press

– D. Harman, 2011,
Motivations

- It is not easy (or possible) to conduct that kind of study for CLEF, because we would need to:
  - Use different versions of one or more systems
  - Test them on many collections for a great number of tasks

- Today’s systems increasingly rely on-line linguistic resources which continuously change over time, thus preventing comparable longitudinal studies even when using the same systems.
Motivations

- RQ1. Do performances of monolingual systems increase over the years? Are more recent systems better than older ones?

- RQ2. Do performances of bilingual systems increase over the years and what is the impact of source languages?

- RQ3. Do monolingual systems have better performances than bilingual and multilingual systems?
Standardization
Standardization

Inter-collection comparison between systems by limiting the effect of collections and by making system scores interpretable in themselves

– W. Webber, A. Moffat, and J. Zobel. 2008
Score standardization for inter-collection comparison of retrieval systems.
How standardization works

For every run \((r)\) in a collection, we have a measure \(m\) for each topic \(t\) with mean \(\mu_t\) and standard deviation \(\sigma_t\).

These are AP values of all the runs for topic 301 of CLEF Ad-Hoc bilingual English 2006.
How standardization works

For each topic in a collection we can calculate the z-scores of a measure \( m \) as

\[
m' = \frac{m - \mu_t}{\sigma_t}
\]
How standardization works

Normalization in the \([0, 1]\) range by using the cumulative density function:

\[
F_X(m') = \int_{-\infty}^{m'} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx
\]

![Graph showing the standardized AP and sAP](image)
Analysis of CLEF Ad-Hoc Campaigns
RQ 1
Do performances of monolingual systems increase over the years? Are more recent systems better than older ones?

There is not a clear trend showing a steady improvement of sMAP over the years.
RQ 1: Do monolingual systems improve over the years?

The more evident improvement over the years is shown by the languages introduced in 2004 and 2005.
RQ 1: Do monolingual systems improve over the years?

The median sMAP of the monolingual tasks shows several examples of languages for which performances decrease over the years.
RQ 1: Do monolingual systems improve over the years?

The median sMAP is often influenced by the number of groups participating and by the number of newcomers with the positive effect of growing new local IR research communities.

![Box plot showing sMAP improvement over years](image)
RQ 1: Do monolingual systems improve over the years?

... but looking at the best sMAP we find out that in several cases it grows through the years.
RQ 1: Do monolingual systems improve over the years?

But… Are general trends enough to explain phenomena?
RQ 1: Do monolingual systems improve over the years?

Let’s take a look to the Italian case


<table>
<thead>
<tr>
<th>Year</th>
<th>Groups</th>
<th>Newcomers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>2001</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>2002</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>2003</td>
<td>13</td>
<td>4</td>
</tr>
</tbody>
</table>
RQ 1: Do monolingual systems improve over the years?

Let’s take a look to the Italian case

![Box plot of MAP scores for monolingual Italian systems in CLEF 2000-2003](image)

- **Best system for monolingual Italian**

<table>
<thead>
<tr>
<th>Year</th>
<th>Groups</th>
<th>Newcomers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>2001</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>2002</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>2003</td>
<td>13</td>
<td>4</td>
</tr>
</tbody>
</table>
RQ 1: Do monolingual systems improve over the years?

Let's take a look to the Italian case

RQ 1: Do monolingual systems improve over the years?


![Box plot showing improvements in SMAP over years](image-url)
RQ 1: Do monolingual systems improve over the years?

in 2002 FUB used a full enhanced system with an advanced weighting schema, bigrams and coordination level matching, furthermore they focused only on the title of the queries and used a simple form of stemmer.
RQ 1: Do monolingual systems improve over the years?

In 2003 they focused on testing diverse techniques rather than further tuning a well-working fully-enhanced system.
RQ 1: Outcomes

Performances of best groups tend to steadily increase over time

- Usually research groups that participated in previous years might have been more interested in testing new techniques and retrieval settings rather than tuning already well performing systems

- This hypothesis is corroborated by the best performances analysis, where in the first years of a task research groups dedicated much effort to tuning and enhancing good systems already tested in previous campaigns
RQ 2

Do performances of bilingual systems increase over the years?

- For bilingual tasks we have to consider:
  - The target language: the language of the corpus
  - The source languages: the languages of the topics

- It is not always possible to identify a steady improvement of performances for a given target language over the years
RQ 2

Do performances of bilingual systems increase over the years?
RQ 2: Do bilingual systems improve over the years?

<table>
<thead>
<tr>
<th>Task</th>
<th>Year</th>
<th>Groups(new)</th>
<th>Runs</th>
<th>Best sMAP</th>
<th>Median sMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH Bili DE</td>
<td>2002</td>
<td>6(-)</td>
<td>13</td>
<td>.6674 (-)</td>
<td>.5340 (-)</td>
</tr>
<tr>
<td>TEL Bili DE</td>
<td>2008</td>
<td>6(4)</td>
<td>17</td>
<td>.6268 (-6.08%)</td>
<td>.4599 (-13.88%)</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>6(3)</td>
<td>26</td>
<td>.7179 (14.53%)</td>
<td>.4731 (+2.87%)</td>
</tr>
<tr>
<td>AH Bili EN</td>
<td>2000</td>
<td>10(-)</td>
<td>26</td>
<td>.7463 (-)</td>
<td>.5196 (-)</td>
</tr>
<tr>
<td></td>
<td>2001</td>
<td>19(15)</td>
<td>55</td>
<td>.7725 (+3.51%)</td>
<td>.5618 (+8.12%)</td>
</tr>
<tr>
<td></td>
<td>2002</td>
<td>5(3)</td>
<td>16</td>
<td>.6983 (-9.60%)</td>
<td>.4524 (-19.47%)</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>3(3)</td>
<td>15</td>
<td>.6980 (-0.04%)</td>
<td>.4074 (-9.95%)</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>4(4)</td>
<td>11</td>
<td>.5895 (-15.54%)</td>
<td>.5251 (+28.89%)</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>8(8)</td>
<td>31</td>
<td>.7845 (+33.08%)</td>
<td>.5667 (+7.92%)</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>5(4)</td>
<td>32</td>
<td>.7559 (-3.64%)</td>
<td>.4808 (-15.16%)</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>10(9)</td>
<td>67</td>
<td>.7746 (+2.47%)</td>
<td>.4835 (0.56%)</td>
</tr>
<tr>
<td>TEL Bili EN</td>
<td>2008</td>
<td>8(7)</td>
<td>24</td>
<td>.7611 (-1.74%)</td>
<td>.5382 (+11.31%)</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>10(7)</td>
<td>43</td>
<td>.7808 (2.59%)</td>
<td>.4719 (-12.32%)</td>
</tr>
<tr>
<td>AH Bili ES</td>
<td>2002</td>
<td>7(-)</td>
<td>16</td>
<td>.6805 (-)</td>
<td>.4969 (-)</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>9(7)</td>
<td>15</td>
<td>.6737 (-1.01%)</td>
<td>.5394 (+8.55%)</td>
</tr>
<tr>
<td>AH Bili FR</td>
<td>2002</td>
<td>7(-)</td>
<td>14</td>
<td>.6708 (-)</td>
<td>.5647 (-)</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>7(5)</td>
<td>24</td>
<td>.6015 (-10.33%)</td>
<td>.5211 (-7.72%)</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>9(8)</td>
<td>31</td>
<td>.7250 (+20.53%)</td>
<td>.5703 (+9.44%)</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>4(3)</td>
<td>12</td>
<td>.6273 (-13.47%)</td>
<td>.4886 (-14.33%)</td>
</tr>
<tr>
<td>TEL Bili FR</td>
<td>2008</td>
<td>5(5)</td>
<td>15</td>
<td>.6358 (+1.35%)</td>
<td>.4422 (-9.50%)</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>6(4)</td>
<td>23</td>
<td>.7151 (+12.47%)</td>
<td>.4355 (-1.52%)</td>
</tr>
<tr>
<td>AH Bili IT</td>
<td>2002</td>
<td>6(-)</td>
<td>13</td>
<td>.5916 (-)</td>
<td>.5306 (-)</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>8(5)</td>
<td>21</td>
<td>.7119 (+20.34%)</td>
<td>.5309 (+0.05%)</td>
</tr>
<tr>
<td>AH Bili PT</td>
<td>2004</td>
<td>4(-)</td>
<td>15</td>
<td>.6721 (-)</td>
<td>.4278 (-)</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>8(5)</td>
<td>24</td>
<td>.7239 (+7.71%)</td>
<td>.5020 (+17.34%)</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>6(4)</td>
<td>22</td>
<td>.6539 (-9.67%)</td>
<td>.4804 (-4.30%)</td>
</tr>
<tr>
<td>AH Bili RU</td>
<td>2003</td>
<td>2(-)</td>
<td>9</td>
<td>.6894 (-)</td>
<td>.4810 (-)</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>8(7)</td>
<td>26</td>
<td>.6336 (-8.09%)</td>
<td>.5203 (+8.17%)</td>
</tr>
</tbody>
</table>
RQ 2: Do bilingual systems improve over the years?

- Unlike for the monolingual tasks, that the higher median sMAP as well as the best sMAP are achieved in the last years of each task.

- This is an indicator of the improvement of language resources – e.g. dictionaries, external resources like Wikipedia and the use of semantic rather than syntactic resources.

- CLEF Continued Impact!
RQ 2: Do bilingual systems improve over the years?

Effect of the improvement of language resources

- The best bilingual system for the “X2FR” task exploited seven different machine translation systems, three bilingual dictionaries and ten freely available translation tools.

![Box plot of SMAP scores for CLEF 2000 - 2009 Ad-Hoc Bilingual task targeting French.](image-url)
RQ 2: Do bilingual systems improve over the years?

Effect of the improvement of language resources

- The best bilingual system in the TEL “X2DE” task exploited three out-of-the-box retrieval systems (i.e. Lucene, Lemur and Terrier) and the high quality of the Google translation service contributed substantially to achieving the final result.
RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?

In most cases the median sMAP of the monolingual tasks overcome the median sMAP of the corresponding bilingual task.
RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?

Things are different when we consider the best performance ratios than the median ones…
RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?

CLEF 2002 – 2009, Mono/Bili Best MAP comparison
RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?

The gap between top monolingual and top bilingual systems is progressively reduced across the years and in several cases the trend is inverted.
RQ3

<table>
<thead>
<tr>
<th>sMAP</th>
<th>Monolingual</th>
<th>Bilingual</th>
<th>Multilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>.8309</td>
<td>.7845</td>
<td>.8513</td>
</tr>
<tr>
<td>Median</td>
<td><strong>.5344</strong></td>
<td>.5165</td>
<td>.5173</td>
</tr>
<tr>
<td>Mean</td>
<td><strong>.5054</strong></td>
<td>.4898</td>
<td>.4914</td>
</tr>
</tbody>
</table>

Bilingual and multilingual systems have a similar median and mean sMAP even though they are slightly higher for the multilingual and both are exceeded by the monolingual systems.

It is interesting to note that the best system is the multilingual one that has a sMAP 8.52% higher than the top bilingual and 2.46% higher than the top monolingual system.
Wrapping-Up

- RQ1: Steady improvement of best systems and stable median performances taking into account the continuous growth of the community

- RQ2: CLEF had a significant impact in driving the improvement of bilingual systems by continuously stimulating the creation of new and better linguistic resources

- RQ3: Crossing the language barriers requires a big effort but pays off when you compare best systems
MATTERS
MATlab Toolkit for Evaluation of information Retrieval Systems

http://matters.dei.unipd.it/