How to Compare Methods for Automated Keyword Extraction

Maximilian Kähler, National Library of Germany

Setting Up an Evaluation Plan for Method Selection
Three complementary approaches for improving automated keyword extraction

- Find and write better algorithms for keyword extraction
- Reduce the complexity of the problem
- Get better at diagnosing good keyword extraction
Aspects of Evaluation in ML-Projects
Generalisability of Evaluation Results
Generalisability of Results

In every measurement one has to account for **systematic** error and **random** error

<table>
<thead>
<tr>
<th>Examples of systematic error:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Distribution shift between production and training(^1)</td>
</tr>
<tr>
<td>• Information leakage between training and evaluation data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Examples of random error:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Random splitting of data into training and evaluation data</td>
</tr>
<tr>
<td>• “Unknowns” in a complex data generation process</td>
</tr>
</tbody>
</table>

\(^1\)cf. Toepfer M, Seifert C 2020; Fusion architectures for automatic subject indexing under concept drift; https://doi.org/10.1007/s00799-018-0240-3
Generalisability of Results

- while systematic error can only be assessed on a case-by-case basis, random error can be quantified with confidence intervals

- Random error is influenced by the size of the test-set as well as underlying variability of the data
Example: boot-strap confidence intervals to quantify uncertainty

Repeated calculations of the target metric under resampling of test-set result in a range of different results.

Percentiles of the resampled results form a confidence interval that helps to assess the uncertainty due to random error.
Dimensions of Evaluation
Dimensions of Evaluation

- Overall performance metrics produce no insights into why and how good/bad indexing performance is achieved.
- A useful evaluation workflow must enable „drill-down“ analytics to generate hypothesis to improve indexing algorithms.
- Important Dimensions of Evaluation should be agreed upon between subject specialists and data scientists.
Example: Indexing performance stratified by subjects categories
Example: Indexing performance stratified by keyword frequency for different models

Keywords frequency in training data set

1. Models are powered by [annif.org](http://annif.org)
   - **mlm:** Maui like lexical matching (cf. annif.org)
   - **Omikuji-bonsai:** Khandagale et al. 2016 “Bonsai: diverse and shallow trees for extreme multi-label classification”
Dimensions of Evaluation impact Test-Set construction and size

- Dimensions of Evaluation need to be considered before splitting your data into training and evaluation data.

- The overall size of your test set is determined by the error rate and confidence level you require for your metric-estimates in the smallest stratum of your evaluation.

- Stratified sampling techniques can ensure that all strata of evaluation are present in your test-set.
Summary: Where do we meet?

- Plan your evaluation scheme, before you start training models
- Discuss the dimensions of your data that need to be looked at
- Choose metrics that reflect your goals and priorities
- Discuss uncertainty and generalizability of your results
Thank you!

Please get in touch for further questions and discussion:

Maximilian Kähler

m.kaehler@dnb.de

Our Project@DNB: https://www.dnb.de/ki-projekt