Evaluating Concept Extraction

Whitepaper by Michael Upshall and Mads Rydahl

One of the first questions we are asked by new clients is how to evaluate the services provided by UNSILO. This is a sensible question, since there is little reliable guidance available on the Web on how to evaluate text analytics. This paper lists some possible approaches and their strengths and weaknesses. There are two main approaches by which the features and concepts extracted as part of the PoC can be evaluated. Quantitative evaluation measures the accuracy of the extracted concepts against a benchmark; whereas qualitative evaluation measure the quality of the extracted concepts by collecting feedback from users on the perceived value of the concepts. Both methods have their benefits, and this document suggests some ideas on how best to use them.

Quantitative evaluation

Measured evaluation tries to quantify the accuracy of extracted concepts by comparing them against a "gold standard", such as pre-existing index terms, manually assigned keywords, or concepts extracted by other software. The UNSILO research team is continuously benchmarking our performance relative to the most popular alternatives, and we can help you design and execute quantitative evaluation tests at a competitive cost. If you prefer to do your own quantitative testing, you should take note of the following common pitfalls and limitations:

Limitations of quantitative evaluation

Limited Knowledge: Human indexers may be skilled at their task, but may not select the terms or concepts that any given user expects. And given that the subject matter of scholarly articles extend the limits of current knowledge, we cannot expect human indexers to have specialist knowledge of all subject matter. And even if they did, human indexing practices may not favour the concepts that most efficiently link or differentiate documents.

Limited Scope: Where human indexers are used, cost considerations mean that human indexers typically create far fewer concepts than an automatic system. UNSILO extracts hundreds of concepts from every document, normalizes them using available dictionaries and ontologies, and ranks them by order of relevance to the document. This collection of concepts provide a much more detailed and flexible "fingerprint" of the article than the typical four or five keywords assigned by a human indexer.

Human Error and Limited Congruence: Two humans indexing the same manuscript are unlikely to agree completely: it is generally thought that more than 85% overlap between two human indexes is unachievable. Which of the human indexes should the machine-based system be compared with?

Human Bias: There is no evidence that human indexers extract the most relevant terms, and no evidence that manually extracted terms are preferred by authors of other relevant research. UNSILO extracts every concept from every indexed document, and calculates recommendations based on the terms that are actually used in similar documents.

F1 Score: The best-known tool for measured evaluation, the F1 score measures the harmonic mean of precision and recall. But the F1 score can only be used to compare different solutions if the scores are calculated using the same query against the same corpus for every solution. Furthermore, depending on the use case, end users may favour a high recall and tolerate low precision, or favour high precision and not bother about low recall. We have written about some of the difficulties to watch out for when using the F1 score on our blog.

Cross Validation: Comparing the output from multiple text analytics tools applied to the same text poses a similar set of use case specific challenges. The UNSILO research team is continuously evaluating the performance of our concept extraction relative to the most widely used alternatives, Watson/Alchemy, Microsoft Text Analytics, Dandelion API, and NLTK/RAKE. In practice, UNSILO identifies all of the concepts that these 4 alternative tools agree on, and about 97% of the concepts that at least two of the alternatives agree on.

Semantic Precision: On average, UNSILO extracts more than four times as many concepts from a document as the average alternative. Most of the concepts that only UNSILO can find, include additional valuable detail that none of the alternative tools pick up, like "*auditory* brainstem response" and "*viral oncogene* homolog" instead of simply "brainstem response" and "homolog". These additional details make our fingerprints much richer and our recommendations more semantically accurate.

Qualitative evaluation

A trial with real users might comprise, for example, qualitative studies of 10 researchers whose area of expertise is within the domain of the articles indexed. One approach would be to ask all ten users the same open-ended questions from an interview guide, to steer them through a structured evaluation of the usefulness of the extracted concepts in specific use cases. Such an interview guide could also be used to perform a comparative evaluation of the UNSILO search and recommender algorithms against another tool or search engine, preferably limited to the same corpus. To reduce bias, you may consider doubling the number of respondents and only exposing each respondent to results from one tool, e.g. either Google Scholar or UNSILO. You may even bypass the UX completely, and run the respondent's own query in advance, print the source PDFs of the top 10 documents returned, and ask them to evaluate based on a deeper analysis of the article relevance.

Advantages of working with end users

Researchers may misrepresent their own practice: Trials with real users avoid the errors introduced in any survey of perceived benefits only recording an assumption of what the benefit might be.

Fewer preconceptions: End users are likely to be more tolerant of innovative techniques, as long as they produce the results they want, which is to find relevant articles.

Limitations of UX and usability testing

Facilitation is a profession: It is easy to make user studies invalid by suggesting answers, or by leading the respondents to the conclusions the interviewer wants to make. In addition, some respondents try to please the interviewer by giving the answers they think the interviewer wants to hear, rather than their actual thoughts. Hence it is always better to observe user behaviour than to uncritically believe what users say they do or want.

The Primary Findings are usually true: The results of user studies may be as broad as "75% of users interviewed found the tool reduced their searching time and increased its effectiveness". However simplistic such a measure may seem, it eliminates many common sources of error and is typically not less accurate than the quantitative evaluation approach described above.

Questions should be open enough to make respondents evaluate and explain in some detail, they should reference their personal opinion and preference, and each question should be specific enough both to ensure which aspects the user should consider, and to ensure that responses can be normalized and summarized.

Example questions for an Interview Guide

Assessing accuracy

- After searching for a topic of interest and briefly skimming the result set,
  - How many of the top 10 results are relevant?
  - How many of the top 10 results are irrelevant?
- After viewing the most relevant article in the result set
  - How well do the concepts (listed below the article author list) represent the topic of the article?
  - How many of the related articles (listed below the article) are relevant to the topic of the current article?
  - Did you find any relevant articles that you have not seen before?

Assessing fitness to individual needs

- After exploring the interactive features of the related articles section,
  - Were you able to refine the list of related articles?
  - Did the list of related articles appear more relevant after you refined it?
  - Would you recommend this feature to a colleague?

Assessing time savings and discovery improvements

- After completing the same topic search using both UNSILO and at least one other tool such as Google Scholar,
  - Which tool would you say is most efficient?
  - Which tool would save you the most time?
  - Which tool provides the most relevant results?
  - Which tool provides fewer irrelevant results?
  - Which tool provides the most comprehensive insight?
- After trying the interactive Related Content section:
  - Would you have found the same relevant articles on Google Scholar?

Other criteria to consider

Here are some additional high-level questions to ask your team and primary stakeholders when considering the role and use of text analytics software in your business:

Time to Market: What is the estimated total time to implement a solution – not simply configuring the software, but to reach a point at which it can go live? Many analytics tools require a lengthy period of manual ontology creation and curation before the system produces acceptable results.

Maintenance Requirements: Some solutions require a constant manual updating and revisions of the subject ontology in use. How long does it take for a new taxonomy term to be identified, added to the ontology, and for all content to be re-indexed so users can discover other documents using that term? With UNSILO, this process is completely automated, and the solution is always kept in sync with the reference ontologies.

Scalability: Your solution should have sufficient throughput and capability to support constantly growing data sets and corpus sizes. UNSILO is currently being used for live sites with tens of millions of documents and millions of daily user sessions.

Cost-efficiency: What is the total cost of ownership, including recurring editorial resource requirements, software licenses, IT, and data warehousing? The key comparison may be between high quality manual indexing and the high throughput of an automatic tool.

Complexity and Flexibility: Is it easy to integrate with existing content tools? The use of APIs and support for existing structured knowledge, such as hand-built dictionaries and taxonomies, is usually a good sign here.

Support for different document formats: Are your original source documents stored as DOCX, XML, PDF, or other more exotic formats? Ideally, tools should be format neutral to the widest possible extent, so that it can be applied to a wide range of content, and eventually also be applied to internal business processes.