An Experiment in Automatic Indexing Using the HASSET Thesaurus

Mahmoud El-Haj - Lorna Balkan - Suzanne Barbalet - Lucy Bell - John Shepherdson

Lancaster University

UK Data Archive

SKOS-HASSET

- SKOS-HASSET Project at the UK Data Archive
- Funded by Jisc
- automatically index UK Data Archive/UK Data Service document collection

SKOS-HASSET 

Funded by JISC

Led by UK • DATA ARCHIVE

Purpose and Motivation

- Apply automatic indexing tool, KEA, to some of the UK Data Archive's document collection using HASSET thesaurus with aims to:
- see whether KEA could potentially be used to aid metadata creation.
- develop recommendation for the future use of automatic indexing with an existing thesaurus

Data Collection

Corpus Name	Whole Corpus		Training Corpus	
	# Files	Size MB	# Files	Size MB
Nesstar bank of variables/questions	26,634	5.70	21,307	4.56
Survey Question Bank (SQB)	1,353	88.00	1,082	70.00
ESDS partial data catalogue records	5,610	14.50	4,488	11.60
Case Studies / Support guides	243	4.10	194	3.28

Kea (Keyword Extraction Algorithm)

- an algorithm for extracting keywords from text documents
- calculates feature values for each candidate (TF.IDF, First Occurrence, Length)
- uses a machine-learning algorithm to predict which candidates are good keywords.

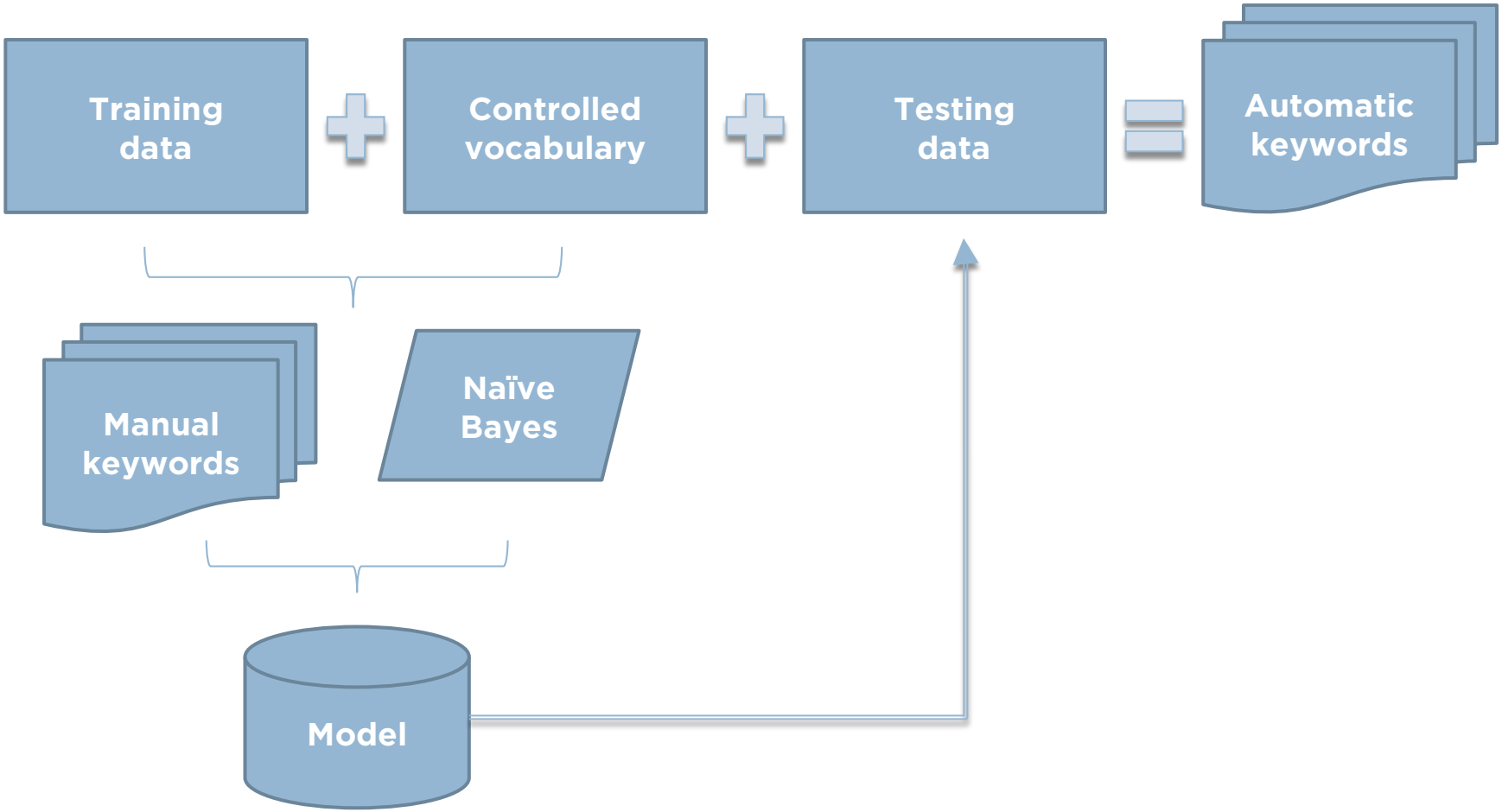
Indexing Process

1
- Get PDFs
- Extract Metadata (Manual-Keywords)
- Convert PDFs to Text

2
- Prepare In/Out (.txt/.key)
- Apply KEA
- Extract Auto Keywords

3
- Automatic Evaluation
- Manual Evaluation (Experts)

Keywords Extraction



Evaluation Methods

- Automatic Evaluation:

  In the automatic evaluation, KEA–generated keywords were compared with the set of manually assigned keywords ('gold standard').

- Human Evaluation:
  - Manually compare auto-keywords with manually assigned keywords on a subset (50 documents) of the test set.
  - How suitable is the KEA term for Information Retrieval?

    5. Extremely suitable = should definitely be keyword

    2. Partially suitable = too narrow or too broad

    0. Unsuitable = far too broad, or completely wrong

Evaluation Metrics

- The main evaluation metrics we used were precision, recall and F1–score, defined as follows:

$$\text{Precision} = \frac{\text{Relevant\_Keywords\_Retrieved\_by\_Auto-indexer}}{\text{All\_Keywords\_by\_Auto-indexer}}$$

$$\text{Recall} = \frac{\text{Relevant\_Keywords\_Retrieved\_by\_Auto-indexer}}{\text{All\_Relevant\_Keywords}}$$

$$\text{F1-Score} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

Stages and Protocol

- Automatic evaluation: the KEA keyword is considered relevant only if it is an exact match of a manual keyword
- Manual evaluation:
  - strictly relevant : 'exact match' of a manual keyword, or 'extremely suitable'.
  - broadly relevant : 'exact match' of a manual keyword, or 'extremely suitable' or 'partially suitable' by the evaluator.

Second Evaluation Stage

- Independent of the first stage
- To what extent is the KEA term semantically related to the
- Gold standard?
  - 5. Totally related (exact match)
  - 4. Closely related: Narrower Terms, Broader Terms or Related Terms to manual keyword
  - 3. Somewhat related: in the same hierarchy as manual keyword
  - 2. Remotely related: related, but not in the same hierarchy as manual keyword
- 1. Unrelated

Results

Corpus Name	Auto	Strict	Broad
Nesstar	0.12	0.14	0.34
SQB	0.14	0.33	0.43
Cat. records (ESDS)	0.11	0.19	0.21
Case Studies / Support Guides	0.06	0.27	0.36

EVALUATION RESULTS (F1–SCORE)

Discussion of the Results

- Best performance overall was seen in the SQB corpus, with a broad F1–score of 0:43.
- Close behind were the Nesstar and case studies/support
- guides corpora, with F1–score c:0:35 each.
- Catalogue records had a low F1–score of 0:21. This was to be expected, given that KEA had relatively little text to index from, compared to the manual indexers.
- This, together with the fact that KEA was applied in non-stemming mode, led to a poor recall score.

Conclusion

- 1) KEA is a useful tool for indexers of full text social science materials; however, KEA would work best as a suggester of new terms, with moderation from a human indexer;
- 2) KEA could also be used as a quality assurance tool, to ensure that terms are not overlooked – some terms it suggested that were highly relevant had not been included in the gold standard, manual indexing;
- 3) more work is needed to investigate KEA further and to see how it could be incorporated technically, and in terms of process, into ingest systems.