KEA++: Semantically Enhanced Keyphrase Extraction Algorithm

Olena Medelyan, University of Waikato, New Zealand

I. INTRODUCTION

Keyphrases
- single words or multi-word lexemes describing document's subjects
- provide content-based organization of documents, thematic access
- help to represent search results and assist with navigation
- can be freely chosen or from a controlled vocabulary
- usually assigned manually by professional indexers (1h per document)
- time-consuming, expensive and practically impossible

Automatic Keyphrase Indexing
- extraction: select significant word n-grams from the document not much training data required, but error-prone, inconsistent
- assignment: cluster documents in classes labeled with keyphrases higher accuracy, but require large training data, impractical

II. How KEA++ Works

Thesaurus
- controlled vocabulary
- descriptors (allowed keyphrases)
- non-descriptors (not allowed)
- defines semantic relations

Extracting Candidates
- Collects all n-grams that match thesaurus terms
- Replaces non-descriptors by descriptors

Features
- TF-IDF
- First Occurrence (beginning/end)
- Length (two-word phrases)
- Node Degree: semantically related to most other document's phrases

Extracting Keyphrases
- Computes feature values for candidate
- Applies the model to compute probability of each term being a keyphrase
- Selects top requested phrases

Extracting probabilities

III. EVALUATION

Humans vs. KEA vs. KEA++
- 10-fold X-validation 780 docs: F-Measure 26.4 (vs. KEA: 13.0)
- indexing consistency 10 docs with Rolling's Measure 6 professional indexers: 38%, KEA++: 27% (vs. KEA 7%)

IV. LEXICAL CHAINS

Lexical chains are sequences of semantically related words that reflect the cohesive structure of a document. They can be computed with any electronic thesaurus and used in keyphrase indexing for detecting individual topics in the document.

Impact of Natural Disasters in the Highlands of Central Vietnam.
The most damaging hazard experienced in the highlands is flashflood, as it occurs with little warning. Floods with strong currents cause permanent damage to fields, washing away the topsoil. Floodwaters also deposit rock and gravel onto fields. Heavy rains trigger landslides that cut off roads and communication networks.

Result:
These are the top six lexical chains scored according to their length and frequencies of chain members. They reflect topics from keyphrase sets assigned by one or more professional indexers (numbers in circles).