KEYWORD INDEXING WITH A SKOS VERSION OF HASSET THESAURUS

MAHMOUD EL-HAJ

DEVELOPMENTAL SYSTEMS AND DATA MINING DEVELOPER
UK DATA ARCHIVE
UNIVERSITY OF ESSEX

LAC WORKSHOP, UNIVERSITY OF ESSEX
4 OCTOBER 2012
PROBLEM AND MOTIVATION

- Manual indexing of the Archive’s collection (studies)
- Choose keywords that represents the study’s topic
- Select best match from HASSET thesaurus
- Ensure consistency of keyword coverage in similar studies
- Processing time varies
- Gold standards
DATA COLLECTION

- A bank of individual questions
- Full Survey Question Bank (SQB) questionnaires
- ESDS data catalogue records and documentation
- Other full-text documents
  - Case studies
  - User guides
DATA COLLECTION PRE-PROCESSING

- PDF to Text
- Extract metadata (manual keywords)
- Extract the data into two file types:
  .txt: holds the actual text
  .key: holds the gold-standard keywords
EXPERIMENTAL WORK

- TF.IDF Model
- Keyphrase Extraction Algorithm (KEA)
TF.IDF MODEL

- We processed 2,436 SQB documents
- No training data
- No controlled vocabulary
- Keywords with no domain-specific information
KEYPHRASE EXTRACTION ALGORITHM (KEA)

- Keywords Indexing using Controlled Vocabulary
- Use training data (based on keywords coverage)
- Build a classifier training model (WEKA)
- Use features to determine importance
TRAINING

- Create training model using human indexer’s keywords
- 80% for training
- 20% for testing
- Use SKOS version of HASSET controlled vocabulary
- Use stop-word list + method terms ("do you think", “closest to your view”)

JISC

UK DATA ARCHIVE
BUILDING THE MODEL

- Use the training documents with their keywords
- Identify keywords
- Calculate features
- Mark each phrase as keyword, non-keyword using the gold-standard keywords
KEA PRE-PROCESSING

1. Input Cleaning
   • Remove apostrophes
   • Split hyphenated words into two
   • Remove non-token characters (that do not contain letters)

2. Candidate Identification
   • Candidate phrases are limited to a certain maximum length (usually three words)

3. Phrase stemming and case-folding (the distinction between upper and lower-case)
KEYWORD EXTRACTION

- Unseen test documents
- Calculate features:
  1. Tf.idf – calculate frequencies
  2. First occurrence: number of words that precede the phrase’s first appearance / number of words in the document
  3. Length of Keyword: number of its component words
  4. Node Degree: number of keywords semantically related to this keyword
Get PDFs
Extract Manual-Keywords
Convert PDFs to Text
Automatic Indexing
Data Mining/Machine Learning
Extract Automatic Keywords
Automatic Evaluation (Recall/Precision/Overlap)
Manual Evaluation
QUESTIONS?