Thesaurus based Keyword Extraction
TIR 2010 Bilbao 29 August 2010
Luit Gazendam (Novay)
Christian Wartena (Novay)
Rogier Brussee (Univ. of Applied Sciences Utrecht, presenter)
Problem description

- Keywords used for organising and retrieval of documents (including non textual ones)
- Problem:
  - Determine keywords automatically

- Operational problem:
  - Define relevance measure of terms
  - Select collection of terms based on relevance
    - Here, just rank
Keywords, world knowledge, informativity

• Relevance of term as keyword depends on:
  – **Importance** of term for the *document*
  – **Discriminative power** of term within *document collection*
  – **A priori criteria**
    • in a thesaurus
    • right wordclass,
    • non stopword,
    • …
Example: TF.IDF

- Consider TF.IDF
  - Input:
    - document collection \( \{d_1 \ldots d_N\} \)
    - Term collection \( \{t_1 \ldots t_n\} \)
    - Count's
      - \( n(d,t) = \# \) occurrences of term \( t \) in document \( d \)
      - \( df(t) = \# \) documents containing \( t \) (doc frequency)
      - \( N = \# \) documents

\[
\text{tf.idf}(d,t) = n(d,t) \log(N/df(t))
\]

\( n(d,t) \): weighs importance of term in the document
\( N/df(t) \): weighs importance of term in the doc collection
World knowledge from thesaurus structure

• Problem: What can we do if we do not have access to large document collection?  
  – or there is no natural document collection

• Importance in the doc collection is really a proxy for importance of terms in “the world”.
  – Importance w.r.t. everything
    • ever written, English web, Information retrieval literature

• Thesauri are alternative sources of world knowledge
  – Also, required by many archives
Mission to Afghanistan uncertain
More and more parties are beginning to doubt the planned mission of 1100 Dutch soldiers to Afghanistan. Tomorrow, representatives of the Pentagon and the State department will come to the Hague for talks with high ranking civil servants. The Dutch cabinet will make its final decision on Friday.
Mission to Afghanistan uncertain
More and more parties are beginning to doubt the planned mission of 1100 Dutch soldiers to Afghanistan. Tomorrow, representatives of the Pentagon and the State department will come to the Hague for talks with high ranking civil servants. The Dutch cabinet will make its final decision on Friday.

……
Count lexical representations of Th-terms.

Prisons (1)
Missions (6)
Prisoners of war (1)
Civil servants (1)
Soldiers (5)

Camps (1)
Governments (5)
Ministers (1)

Voting (1)
Democratisation (1)

Prime ministers(1)
Ranking: frequency of terms (?).

- Prisons (1)
- Camps (1)
- Voting (1)
- Missions (6)
- Governments (5)
- Soldiers (5)
- Prisoners of war (1)
- Civil servants (1)
- Ministers (1)
- Democratisation (1)
- Prime ministers (1)
Prisons

Ministers

Governments

Voting

Democratisation

Missions

Prisoners of war

Civil servants

Soldiers

Professions

Government ministers

Heads of state

Forming of cabinets

Public authorities

Professions

Service professions

Prisoners

Prison camps

Camps
Ranking: frequency of terms (????)

- Prisons (1)
- Ministers (1)
- Governments (5)
- Civil servants (1)
- Soldiers (5)
- Prisoners of war (1)
- Camps (1)
- Voting (1)
- Democratisation (1)

- Missions (6)
- Public authorities
- Forming of cabinets
- Heads of state
- Professions
- Service professions
- Ministries

- Prisons
- Prison camps
- Governments
Ranking: connectedness of terms!

Prisons (1) → prison camps → Camps (1)

Prisoners of war (1) → prisoners

Missions (6)

Civil servants (1)

Soldiers (5)

Professions

Ministers (1)

Governments (5)

Ministries

Service professions

Forming of cabinets

Elections

Voting (1)

Democratisation (1)

Heads of state

Prime ministers (1)
TF.RR: Term frequency, Realised relations

- Select words in text that are concepts in the thesaurus
- Determine weight of (key)words by
  - Frequency
  - Number of thesaurus relations to other words in the text: central words in the text become higher weights

\[ \text{tf}.rr(t, d) = \text{tf}(t, d) \text{rr}(t, d) \]
\[ \text{tf}(t, d) = 1 + \log(n(t, d)) \]
\[ \text{rr}(t, d) = 1 + \mu r_1(t, d) + \mu^2 r_2(t, d) \]

- with \( n(t, d) \) the number of occurrences of \( t \) in \( d \),
- \( \mu = \alpha/\text{avlinks} \) where \( \text{avlinks} \) is the average out degree of the thesaurus
- average number of relations a term has in the thesaurus
- We need \( 0 < \alpha < 1 \), we set \( \alpha = \frac{1}{2} \)
Evaluation

- Generate and rank keyword suggestions for TV-programs from contextual resources
  - 258 TV broadcasts
  - 362 context documents,
    - Length = 25 -- 7000 words, av = 1000
  - Thesaurus, so called GTAA (Common Thesaurus Audiovisual Archives)
    - #keywords = 3860, #relations = 20 591
    - #keywords = 1 -- 15, av = 5.7
  - Manual keywords ground truth for evaluation
    - Inter-annotator consistency 13% -- 77%, av = 44%
Results

![Graph showing precision vs. recall for different models: tf.idf (terms), tf.idf (concepts), tf.rr, and Concept frequency. The graph illustrates differences in performance across these models.]
Conclusion

- Using thesaurus relations improves on just counting concepts/syn-sets
- Results comparable to using a corpus. i.e.

  A *good* thesaurus is a reasonable alternative to having access to representative corpus