Automatic Annotation Suggestions for Audiovisual Archives: Evaluation Aspects

L.Gazendam,
V.Malaise,A.Jong,C.Wartena,H.Brugman,
G.Schreiber

Evi Kiagia

NLP/Text Mining for historical documents

Framework of the Project

- Initiative forwarded by Netherlands Institute for Sound and Vision
- Archiving and digitizing publicly broadcasted Tv and Radio Programs
- Manual annotation of keywords with the help of cataloguers
- Generating automatic annotation suggestions to assist manual annotation by cataloguers

Overview

- Manual Annotations in Audiovisual Archives
- Usual Techniques of Semantic Annotations
- Pipeline and Core of CHOICE- Project
- Experiments & Evaluation Methods
- Results & Discussion
- Summing Up

Manual Annotation Process

- Cataloguers classify manually TV programs into categories using:

GTAA keywords vocabulary

  - GTAA(Common Thesaurus of Audiovisual Archives)
  - Contains keywords and relations between them
  - Programs are described in terms of these keywords

Manual Annotation Process

IMMiX Metadata Model

- Adaptation of the FRBR data model for library data categorization
- Divides the data into 4 categories
  - Information Content
  - Audiovisual Content
  - Formal Data(intellectual property rights)
  - Document management data(Id number)

Automatic Annotation Tools & Techniques

- Generate automatically GTAA Keywords for quick classification
- Semantic Annotations performed by tools that generate them without human interaction
- Both tools based on GATE * platform.
- * *A generic NLP platform that implements NER modules and a rule language to define specific patterns to expand on simple string recognition.(Cunningham et.al 2002)*

KIM Platform:

Provides a Infrastructure for automatic semantic annotation and customizable IE based on GATE

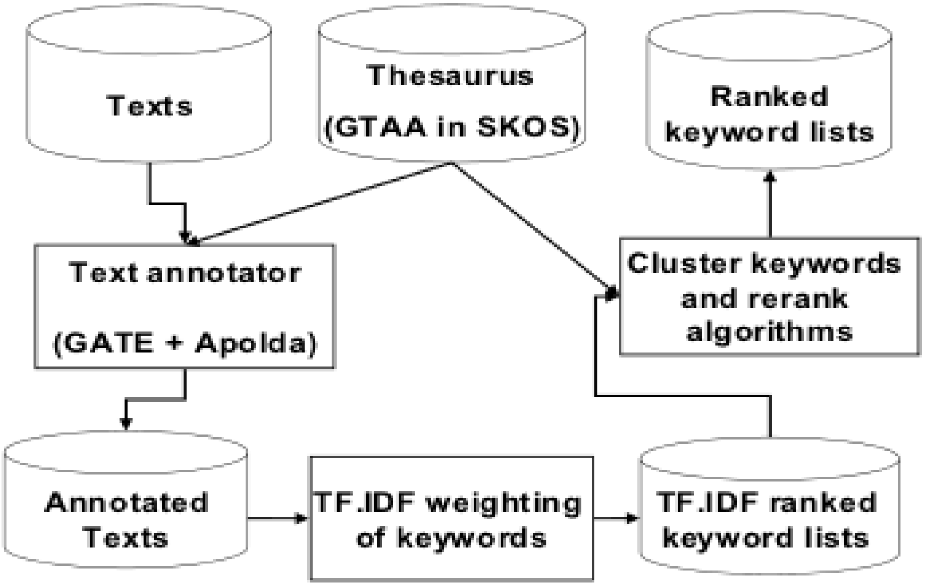
Mnm Tool:

Provides both automatic and semi automatic annotations

Integrates an ontology editor with IE pipeline

Ranking Pipeline of CHOICE-Project

Text--->GTAA Keywords--->thesaurus relationships



CHOICE-PROJECT Pipeline

1.Text annotator

Tags the occurences of thesaurus words keywords in the texts

2.TF.IDF computation

Ranks the keywords tagged in the previous method

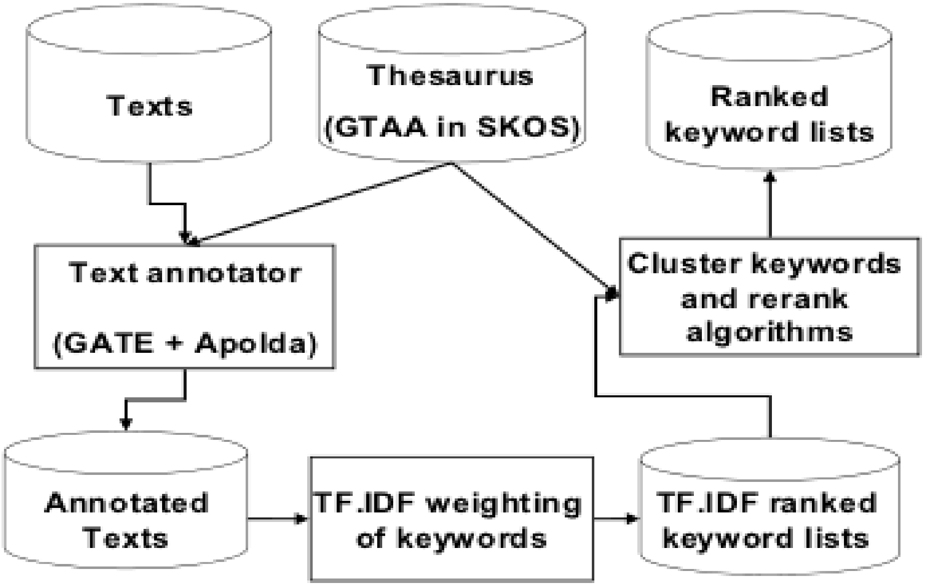
3.Cluster-and-Rank process/Algorithms

Uses thesaurus relations to improve upon the TF.IDF ranked list

- CARROT Algorithm
- Pagerank Algorithm
- Mixed Algorithm using General keyword importance

Ranking Pipeline of CHOICE-Project

Text--->GTAA Keywords--->thesaurus relationships



2. TF.IDF computation

- Information Retrieval measure that reflects the importance of a document in a collection of other documents/corpora.

Term frequency (tf)

- ➢ tf=the number of occurrences of a word in a document

Inverse document frequency(idf)

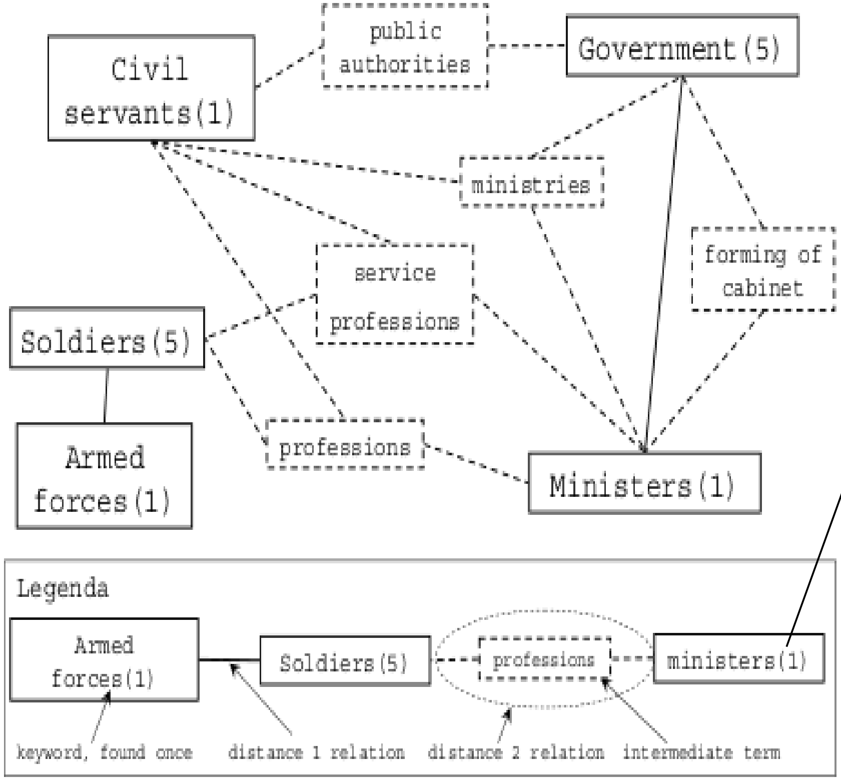
- ➢ idf = a measure of a general importance of word

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Cluster and Rank Algorithms

Text--->GTAA Keywords--->thesaurus relationships

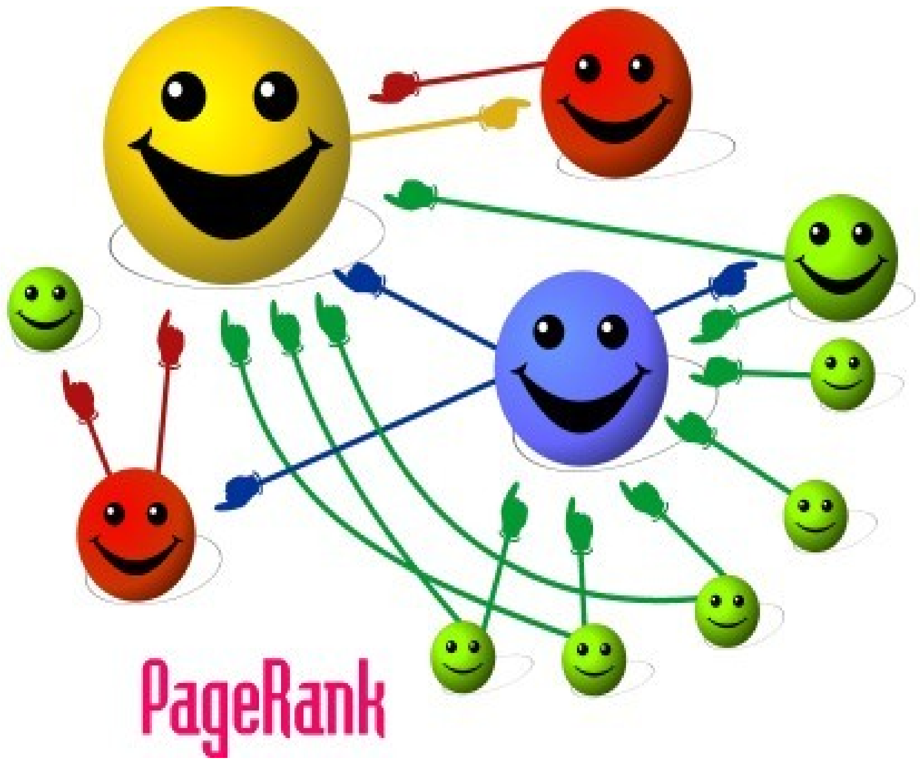
Graph:



Output:

Reranked list of elements
With the help of 3 different algorithms

Cluster &Rank Algorithms
Pagerank Algorithm



Pagerank algorithm(Brin and Page 1998)

- “Assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set “(wikipedia)
- Captures the importance and centrality of a specific keyword in a set by assigning weighting to the edges.
- It can be described as an activation spreading through a network
- The activation on each node is its Pagerank score and shows its importance

Cluster &Rank Algorithms
CARROT **Algorithm**

- Acronym for (Cluster and Rank Related Ontology concepts or Thesaurus terms)
- Constructed for this project
- Combines local connectedness of a keyword and the TF.IDF score
- Each group is sorted on the TF.IDF values

Cluster &Rank Algorithms
Mixed algorithm using general keyword

- Keeps relevancy information through the TF.IDF while performing spreading of activation
- Keywords that are considered important are favoured
- Topics that are considered more important are modelled with many keywords

Keywords with the highest GTAA pagerank:

bussiness, buildings, people, sports,animals

Keywords with the lowest GTAA pagerank:

lynchings,audiotapes,holography,autumn,spring

Experiment 1

- Uses two kinds of evaluations on the algorithms introduced previously
  - Classical precision/recall evaluation
  - Evaluation using semantic overlap:

Automatic Annotations vs. Manual Annotations

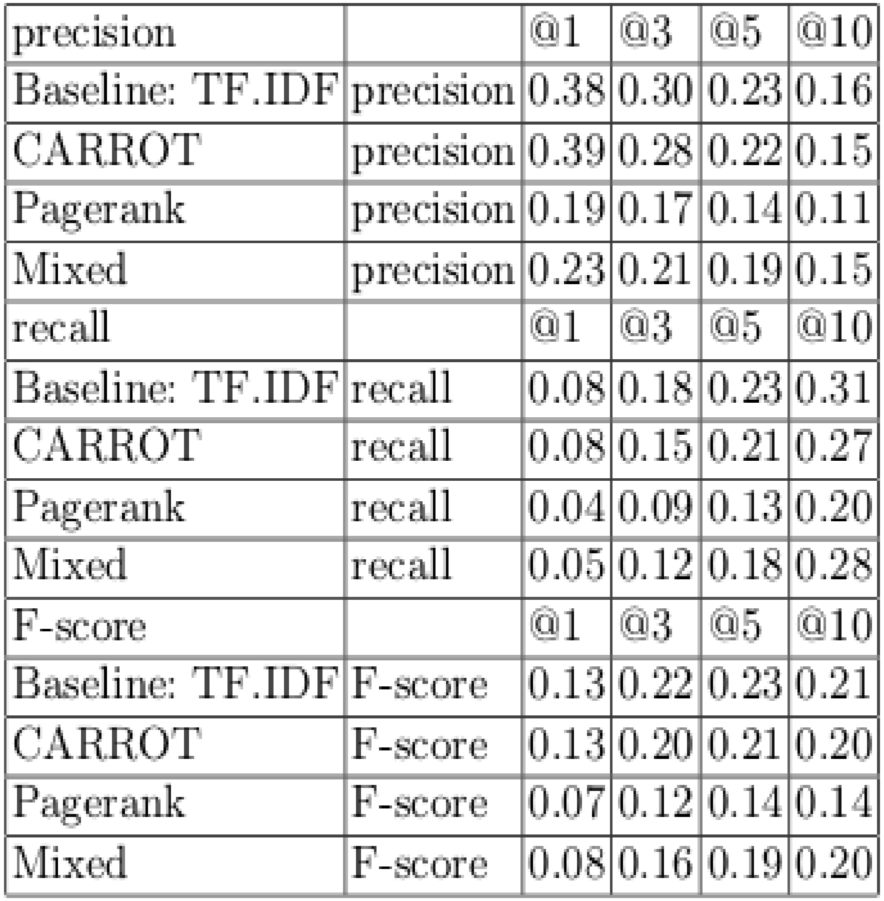
Material:

- 258 tv-documentaries belonging to 3 series of TV-programs
- Each of these documents associated with context documents
- 362 context documents in sum

Evaluation of Experiment 1
Precision/ Recall Evaluation

- Reflects the quality of the automatically derived documents(Manual annotation documents were also used for this reason, serving as the “gold”standard)
- **Precision** in this context: number of relevant keywords suggested by the algorithms,divided by the total number of keywords that are given by our system
- **Recall**:number of relevant keywords suggested by the system for one tv-program , divided by the total number of existing keywords.

Evaluation of Experiment 1
Precision/ Recall Evaluation



precision		@1	@3	@5	@10
Baseline: TF.IDF	precision	0.38	0.30	0.23	0.16
CARROT	precision	0.39	0.28	0.22	0.15
Pagerank	precision	0.19	0.17	0.14	0.11
Mixed	precision	0.23	0.21	0.19	0.15
recall		@1	@3	@5	@10
Baseline: TF.IDF	recall	0.08	0.18	0.23	0.31
CARROT	recall	0.08	0.15	0.21	0.27
Pagerank	recall	0.04	0.09	0.13	0.20
Mixed	recall	0.05	0.12	0.18	0.28
F-score		@1	@3	@5	@10
Baseline: TF.IDF	F-score	0.13	0.22	0.23	0.21
CARROT	F-score	0.13	0.20	0.21	0.20
Pagerank	F-score	0.07	0.12	0.14	0.14
Mixed	F-score	0.08	0.16	0.19	0.20

Pagerank : worse than the others (no incorporation of the TF.IDF scores)

Mixed algorithms: f-score(starts very bad at the beginning but catches up with the tf.idf baseline and CARROT)

TF-IDF: Best scoring , but the difference is not statistically big

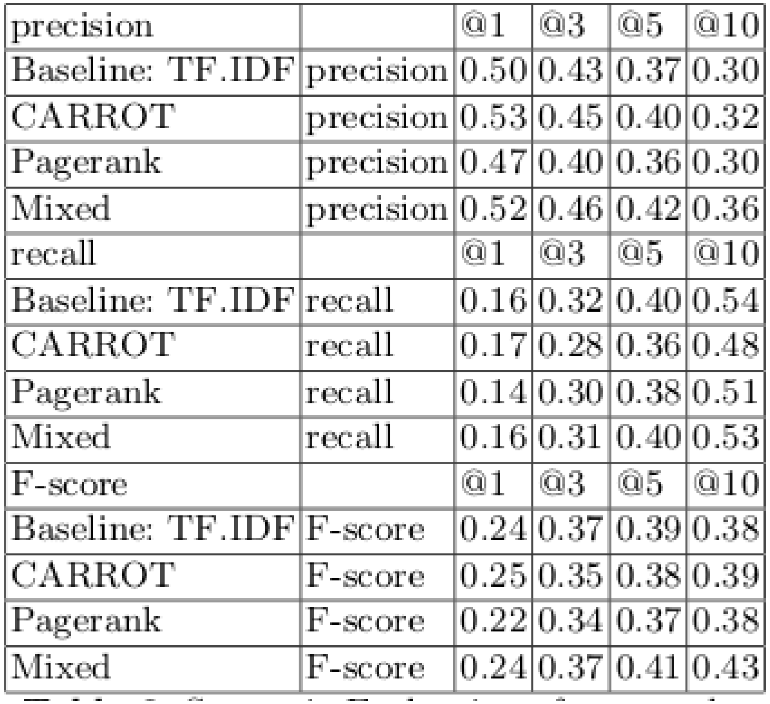
Evaluation of Experiment 1
Semantic Evaluation

Semantic evaluation employed to measure the quality of suggestions better than the precision/recall evaluation

- Automatic suggested keywords similar with the manually annotated ones.
- All terms within one thesaurus relationship are considered

Goal: *Conceptual Consistency* of suggested keywords

Evaluation of Experiment 1
Semantic Evaluation



precision		@1	@3	@5	@10
Baseline: TF.IDF	precision	0.50	0.43	0.37	0.30
CARROT	precision	0.53	0.45	0.40	0.32
Pagerank	precision	0.47	0.40	0.36	0.30
Mixed	precision	0.52	0.46	0.42	0.36
recall		@1	@3	@5	@10
Baseline: TF.IDF	recall	0.16	0.32	0.40	0.54
CARROT	recall	0.17	0.28	0.36	0.48
Pagerank	recall	0.14	0.30	0.38	0.51
Mixed	recall	0.16	0.31	0.40	0.53
F-score		@1	@3	@5	@10
Baseline: TF.IDF	F-score	0.24	0.37	0.39	0.38
CARROT	F-score	0.25	0.35	0.38	0.39
Pagerank	F-score	0.22	0.34	0.37	0.38
Mixed	F-score	0.24	0.37	0.41	0.43

Mixed model: Good in precision but normal in recall

Tends to suggest more general terms

Mixed and Pagerank Model: At the end are Improved much more than the other models

Experiment 2
“Serendipitous Browsing”

Lists of Annotation suggestions contain:

- Exact suggestions
- Semantically related suggestions
- Sub topics
- Wrong Suggestions

Experiment 2
“Serendipitous Browsing”

- Created as a new way to evaluate the perceived value of the automatic annotations
- Overlap of list of keywords/annotation suggestions between two broadcasts.
- Overlapping by chance , makes a good measure of relatedness between two broadcasts
- Tests the overlapping of between documents/keywords of automatic vs manual annotations

Serendipitous Browsing:

“Discovering of unsuspected relationships between documents through browsing them, thus creating a “moment of serendipity”(Gazedam et.al

Experiment 2
“Serendipitous Browsing”

- Tests the overlapping of between keywords through comparing automatic vs manual annotations

- Material

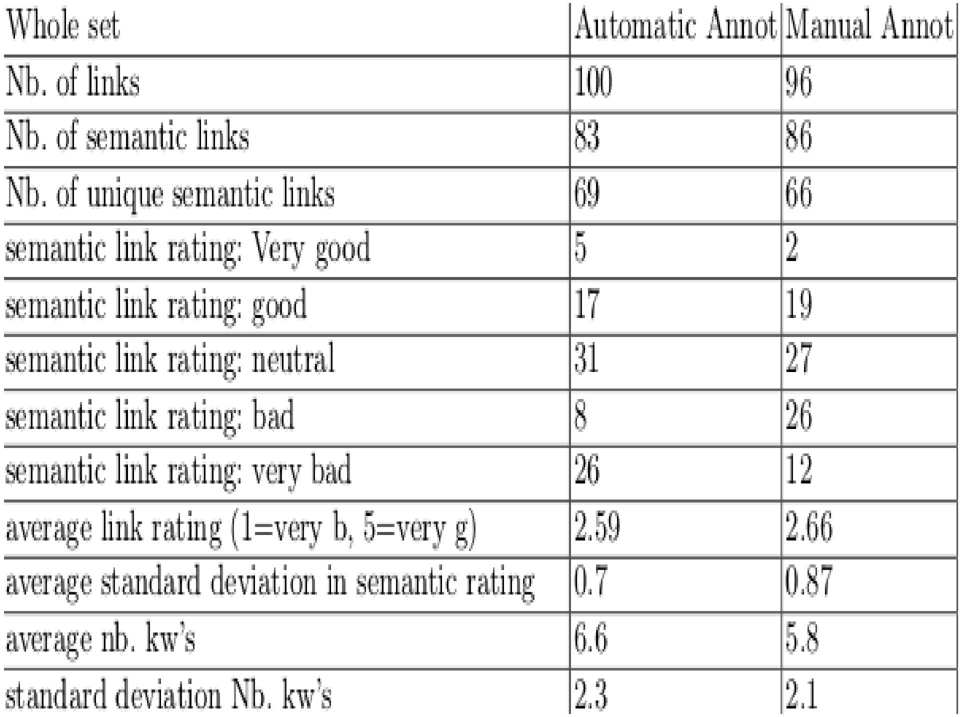
Corpus: 258 programs

Automatic Annotations pairs: 13-5 overlapping keywords

Manual Annotation pairs:9-4 overlapping keywords

- Overlapping keywords for each pair represent the semantics of the link between the two documents

“Serendipitous Browsing” Evaluation



Whole set	Automatic Annot	Manual Annot
Nb. of links	100	96
Nb. of semantic links	83	86
Nb. of unique semantic links	69	66
semantic link rating: Very good	5	2
semantic link rating: good	17	19
semantic link rating: neutral	31	27
semantic link rating: bad	8	26
semantic link rating: very bad	26	12
average link rating (1=very b, 5=very g)	2.59	2.66
average standard deviation in semantic rating	0.7	0.87
average nb. kw’s	6.6	5.8
standard deviation Nb. kw’s	2.3	2.1

- 2 documents appear in the list of 10 best manual annotation pairs

A specific document is the most similar document for twdo differen other programs

- Average quality of semantic links is not very high
- Both automatic and manual annotations have 21 good or very good semantic judgments
- Interesting links between documents can be found between documents in both annotations

Combined Evaluation & Discussion

- Classic evaluation showed TF.IDF best ranking method
- Semantic Evaluation showed Mixed Model perfomed better
- Manual Annotations and automatic Annotations have the same value for finding interesting related documents(Serendipitous Experiment)
- Combined evaluation of these 3 methods make it hard for the manual annotations to serve as a “gold” standard.

Future Work

Apply semantic evaluation

Applying user evaluation of keyword suggestions for cataloguers

Suggestion of keywords based on automatic speech transcripts from broadcasts and compare results with this paper.

Questions?

Thank you !!!!!