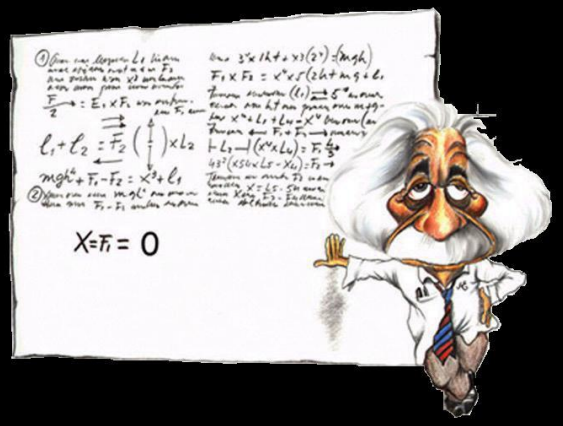
Text Mining Group
@
Department of Computer Science and Engineering
Faculty of Applied Sciences
University of West Bohemia



Update Summarization Based on Novel Topic Distribution

Sutler @ Text Analysis Conference 2008

Josef Steinberger & Karel Ježek
(jstein@kiv.zcu.cz & jezek_ka@kiv.zcu.cz)

November 2008

Outline

- Our previous summarization research
- Summarization approach overview
- The classical latent semantic analysis model and iterative residual rescaling
- Update summarizer
- TAC results
- Conclusion

Our previous summarization research

- Since 2004 we work on a (sentence-extractive) summarization method based on latent semantic analysis
  - Starting point – paper written by Gong and Liu in 2002
  - We improved the method by changing the selection criterion
  - 2006 – method extended to process a cluster of documents
  - 2008 – update summarizer, changes in the core of the summarizer
- Using anaphoric information
  - Since 2005 – with Massimo Poesio and M.A. Kabadjov
  - Tasks: Improving sentence extraction, correcting anaphoric links in the summary, sentence ordering
- Sentence compression
  - Removing unimportant clauses
  - A set of knowledge-poor features
  - A classifier decides if the crucial information was not removed

TAC Update Task

- Update task
- 48 topics, each topic has two sets of 10 documents
- The task is firstly to summarize the set of older documents (multi-document summaries) and then to summarize the set of new documents under the assumption that the reader has already read the set of older documents (update summaries)
- Each participant could submit up to 3 runs, the first two priority runs were annotated
- Main evaluation metric - Pyramids

Summarization approach overview

1. Obtain “older topics” – reader's prior knowledge in the set of older documents
2. Obtain “new topics” – concepts in the set of new documents
3. Specify redundancy of the new topics = how much their information is covered by older topics
4. Specify the significance of the new topics = how important they are in the set of new documents
5. Specify novelty of the new topics = how significant and new they are
6. Create a summary of the sentences that best cover the novel topics

Latent semantic analysis (LSA)

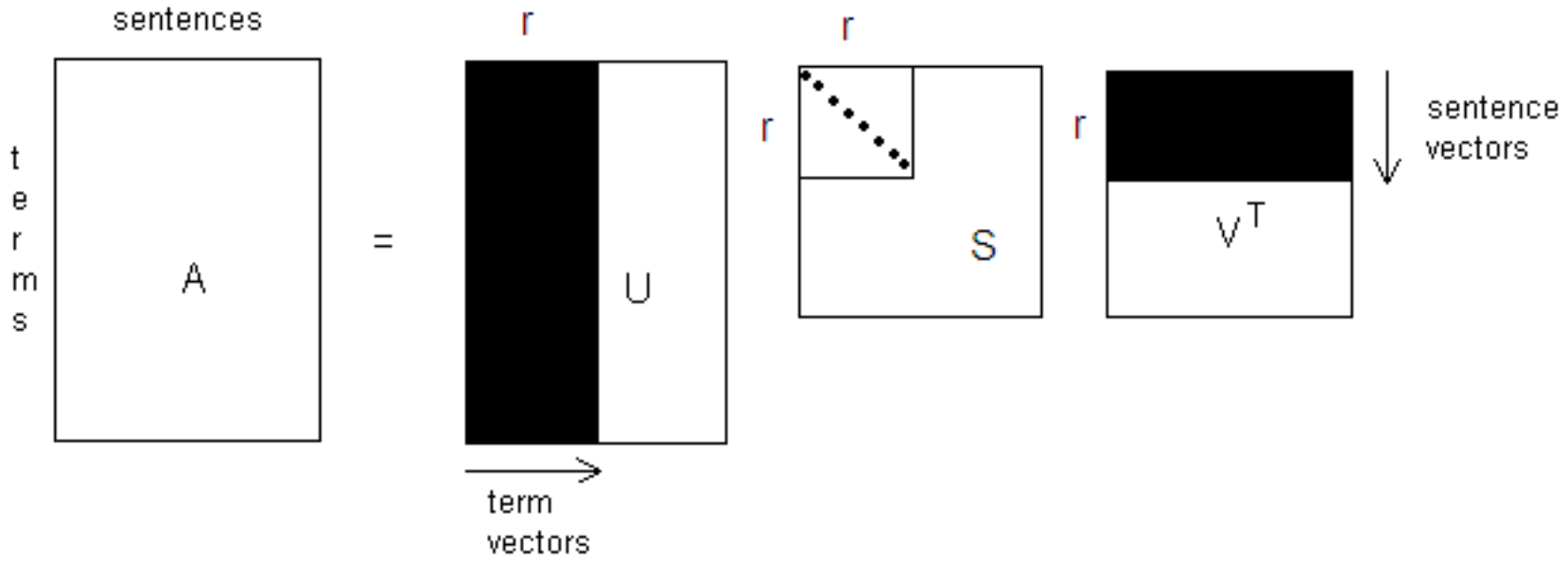
- Technique for extracting hidden dimensions of the semantic representation of terms, sentences, or documents, on the basis of their contextual use (Landauer, 1997)
- Can cluster terms and sentences into topics
- Topics ordered according to their significance
- Dimensionality reduction – insignificant dimensions are removed
- Used in various NLP applications
  - Information retrieval – Berry et al., 1995
  - Text segmentation – Choi et al., 2001
- Gong and Liu (2002) – the first LSA-based summarization approach

The classical LSA model

- Creation of terms by sentences matrix *A*
  - Term weighting:

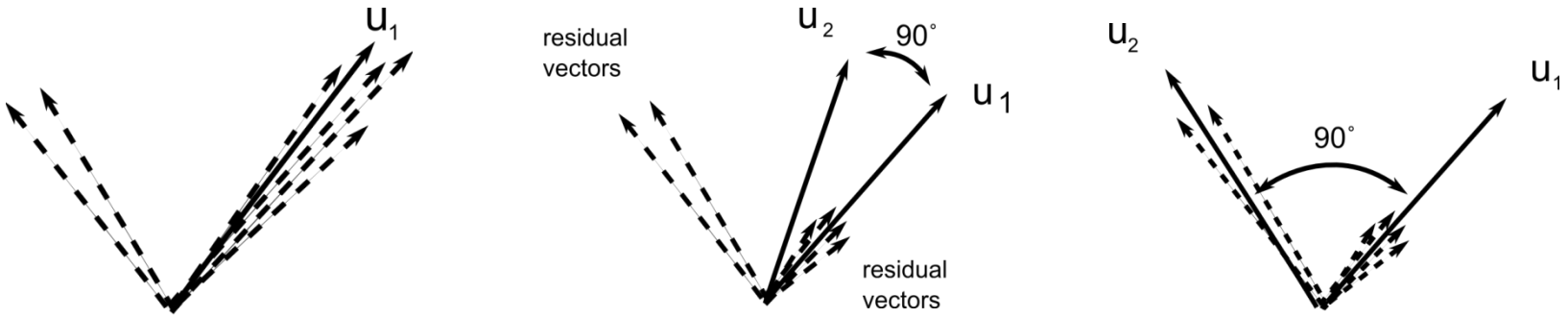
    Local x Global weight – best results with the Boolean local weight and the entropy-based global weight
- Apply SVD (Singular Value Decomposition) on matrix *A*
  - Result – matrix *A* is decomposed into three matrices where information about the most important topics (linear combinations of original terms) can be found

Singular value decomposition



IRR – generalization of SVD

- Iterative Residual Rescaling
- Ando & Lee (2001)
- When the topic-sentence distribution is non-uniform in the analyzed text, the dominant topics take more than one dimension in the latent space, although the dimensions are orthogonal
- Minor topics are ignored and topic distribution is negatively biased by residual vectors from the dominant topic
- Iterative residual rescaling fights against this problem



Update summarization based on LSA (1)

- Reader's prior knowledge is assumed (represented by a set of older documents)
- Set of new documents is intended for own summarization
- We create a set of "old" topics and a set of "new" topics = separate LSA of both sets
- In matrices U (U_{old} and U_{new}) we can see term/topic distributions
- For each new topic t we measure its redundancy (red_t) = cosine similarity with the most similar old topic

$$red_t = \max_{i=1}^{k_1} \frac{\sum_{j=1}^{m} U_{new}[j,t] \cdot U_{old}[j,i]}{\sqrt{\sum_{j=1}^{m} U_{new}[j,t]^2} \cdot \sqrt{\sum_{j=1}^{m} U_{old}[j,i]^2}}$$

Update summarization based on LSA (2)

- Significance of the topic is determined by its singular value (*sing*)
- Topic novelty (*nov*) is done by:

  $nov = (1 - red) * sing$
- Topics with a high novelty value are considered interesting and thus get greater weights
- From topic novelties we create diagonal matrix *NOV*
- Final matrix *F* can be then computed as $NOV*V_{new}^{T}$
- In *F* both topic novelty and importance are taken into account

Update summarization based on LSA (3)

- Sentence selection starts with the sentence that has the longest vector in F (f_{best})
- The information contained in the sentence is then subtracted from F:
  $F = F - (f_{best} \cdot f_{best} / |f_{best}|^2) * F$
- The values that correspond to similar sentences are decreased, thus preventing inner summary redundancy
- After the subtraction the process of selecting the sentence that has the longest vector in matrix F and subtracting its information from F is iteratively repeated until the required summary length is reached.

TAC results

Overall TAC results of our summarizer.

Evaluation metric	**Rank of run 25 (Total No. of runs)**	**Rank of run 51 (Total No. of runs)**
Average modified (pyramid) score	10 (58)	16 (58)
Average num. of SCUs	12 (58)	17 (58)
Average num. of repetitions	55 (58)	22 (58)
Macroavg. modified score with 3 models	10 (58)	16 (58)
Average linguistic quality	10 (58)	8 (58)
Average overall responsiveness	9 (58)	14 (58)
ROUGE-2	17 (71)	22 (71)
ROUGE-SU4	17 (71)	18 (71)
BE	13 (71)	15 (71)

TAC results – update summaries

Separate TAC results of our update summaries

Evaluation metric	**Rank of run 25 (Total No. of runs)**	**Rank of run 51 (Total No. of runs)**
Average modified (pyramid) score	7 (58)	12 (58)
Average num. of SCUs	9 (58)	12 (58)
Average linguistic quality	5 (58)	12 (58)
Average overall responsiveness	15 (58)	16 (58)
ROUGE-2	18 (71)	25 (71)
ROUGE-SU4	17 (71)	21 (71)
BE	13 (71)	25 (71)

Conclusion

- We created an update summarization method which is independent on the language of the source.
- Next directions:
  - Use anaphora resolution >> co-reference resolution in the case of multi-documents
    - Improving sentence selection
    - Reference correction
    - Sentence ordering
  - Working on sentence compression
- Poster miniboaster
  - Guided example how our summarization approach works
  - You can see some numbers how topics look like in our sense