CmpE561 Research Project Presentation Text Summarization with Latent Semantic Analysis

Mine Melodi Çalışkan - 2015705009 [1]
Serkan Duman - 2016700039 [2]

10 May 2017

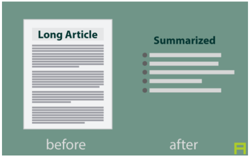
[1] minemelodicaliskan@gmail.com
[2] serkan.duman@gmail.com

Outline

- Automatic Text Summarization
- Latent Semantic Analysis
- Singular Value Decomposition
- Simulation
- Available Data Sets and Corpora
- Evaluation
- State of the art Success Rates
- Approaches for Different Languages
- Discussion

Text Summarization

Automatic Text summarization is a mathematical approach to make a computer program create a representative summary of the given document by finding the most significant sentences.



Input: A text document

Output: A subset of salient sentences from the document

Example Use Cases

- **Summary of online texs and articles:**
  - Example: Reddit's multimedia news summarizer ”auto tl-dr”
- **Collating search engine hits:**
  - Provides more contextual and summary information to retrieve related web pages.
- **Summary of patiences medical data:**
  - Extracts results from multiple medical journal articles returned by a search
  - Filters results that match the patient
  - Merges and orders the remaining facts for the summary.

Types of Text Summarization

- **Extractive Document Summarizers**
  - Extract salient sentences in the original text without modifying them to create a summary.
- **Abstractive Document Summarizers**
  - Build an "internal" semantic representation and then creates summary using natural language generation tecniques.
  - Might contain words that are not explicitly present in the original text.
- **Statistical Summarizers**
  - Use statistical features of the sentences ,e.g title, location, term frequency, assign weights to the keywords.
  - Then calculate scores of the sentences and select the highest scored sentence into the summaries.

Latent Semantic Analysis

- An unsupervised summarization tecnique
- Uses Singular Value Decomposition (SVD)
- Extracts semantically similar words and sentences

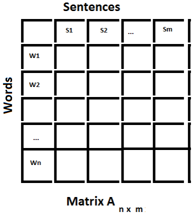
Objective: Discover hidden semantic structures of words and sentences using context of the input document.

Main Steps of LSA

- Input matrix creation
- Singular Value Decomposition (SVD)
- Sentence selection

Input Matrix

- Matrix A : words by sentences matrix
  - Columns represents sentences
  - Rows represents words
- Various methods to represent importance of words:
  - Number of Occurrence
  - Binary Representation
  - Root Type
  - TF-IDF
  - Modified TF-IDF
  - Log-Entropy



Sentences

Words

	S1	S2	...	Sm
W1				
W2				
...				
Wn				

Matrix $A_{n \times m}$

Filling The Input Matrix

- **Number of Occurrence:** Frequency of the word in the sentence.
- **Binary Representation:** If a word occurs in the sentence 1, otherwise 0.
- **Root Type :**If the root type of the word is Noun, cell value is the frequency of the word, otherwise 0.

Filling The Input Matrix

- **Term Frequency-Inverse Document Frequency:**
  TF-IDF is equal to TF×IDF
  - $TF = tf(i,j) = \frac{\text{Frequency of word i in sentence j}}{\text{Sum of frequencies of all words in sentence j}}$
  - $IDF = idf(i,j) = log(\frac{\text{Number of sentences in input text}}{\text{Number of sentences containing word i}})$
- **Modified TF-IDF:**
  - If cell values $<=$ average TF-IDF values in the associated row, then set them zero.
  - Eliminates noise effects.

Filling The Input Matrix

- **Log-Entropy:** Values of the cells are determined by log-entropy of the words i.e the amount of information in each sentence which is calculated as follows:

$$LogEntropy = (1 + \frac{\sum PlogP)}{logn}) * log(1 + f)$$

where

P: The probability of word i appeared in sentence j

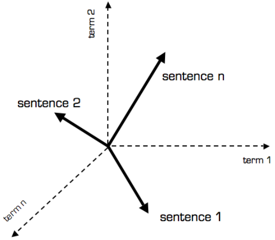
f: The number of times word i appeared in sentence j

n: The number of sentences in the document

Sentences as points in the word space

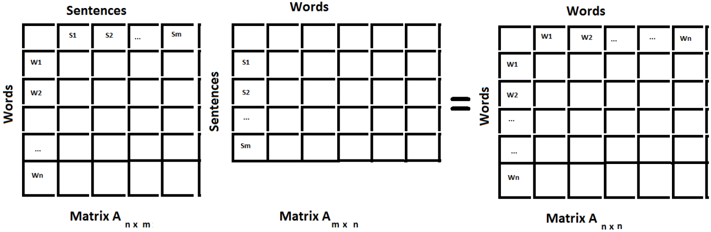
In the words × sentences representation, sentences could be thought of as points in the word space.

- Dimensions of the word space correspond to various words
- Coordinates of the sentence are determined by e.g. the number of occurrences of the particular word in the sentence



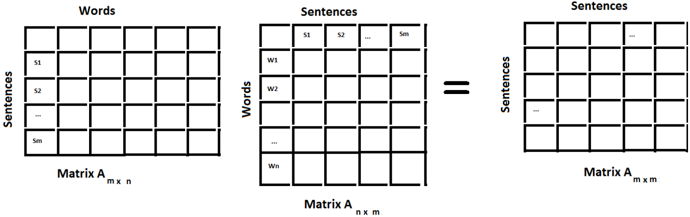
Word Similarity Matrix

Let $\mathbb{W}$ be a $n \times n$ word similarity matrix $\mathbb{W} = AA^T$. In the binary-valued model, element $w_{ij} = \alpha_i \, \alpha_j$ represents number of sentences in which word i and word j co-occur , i.e appear together.



Sentence Similarity Matrix

Let $\mathbb{S}$ be a $m \times m$ sentence similarity matrix $\mathbb{S} = A^T A$. In the binary-valued model, element $s_{ij} = \beta_i \ \beta_j$ represents a number of distinct words sentence i and sentence j have in common.



Singular Value Decomposition

Singular Value Decomposition is a data dimensionality reduction technique.

- Provides an exact representation for the input data matrix as a product of three matrices
- Allows to eliminate less important parts of the data that are linearly independent to produce a much smaller matrix that approximate it with the desired number of dimensions.

$\Rightarrow$ This approximation is called as ”Low-rank approximation”.
$*$ Accuracy is directly proportional to the number of the dimensions we choose.

Computing the SVD of a Matrix

To calculate SVD of a matrix A:

- Find eigenvalues and eigenvectors of AA^T and A^TA
- Resulting eigenvectors of AA^T and A^TA will be columns of U and V, respectively.
- Square roots of eigenvalues (from AA^T or A^TA) will be the singular values of $\sum$.

Singular value decomposition of the given matrix A is:

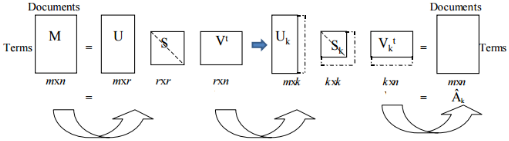
$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5.47 & 0 \\ 0 & 0.37 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.40 & 0.91 \\ -0.91 & 0.40 \end{bmatrix}$$

$= U\Sigma V^T$

Remark. $\sigma_1 > \sigma_2$

Singular Value Decomposition in LSA

LSA performs SVD on A to obtain the singular value matrix and select top k sentences as a summary of a given document.



SVD:
Convert term-by-document matrix into 3matrices U, S and V

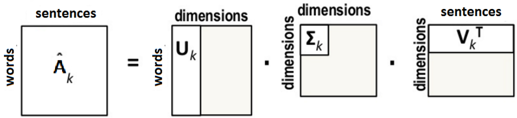
Dimension Reduction:
Ignore zero and low-order rows and columns

Reconstruct Matrix:
Use the new matrix to process queries
OR, map query to reduced space

Singular Value Decomposition in LSA

The given input matrix A decomposed into three matrices U, Σ and V^T where

$$U : \text{Words} \times \text{Concepts matrix}$$
$$\Sigma : \text{Scaling values, diagonal descending matrix}$$
$$V^T : \text{Concepts} \times \text{Sentences matrix}$$



Singular Value Decomposition in LSA

Procedure can be itemized as follows:

- Obtain SVD of A as:
  $A_{n \times m} = U_{n \times r} \Sigma_{r \times r} V_{r \times m}$
- Keep only k eigen values from Σ
- Approximate A with reduced dimensionality :
  $A_{n \times m} \sim U_{n \times k} \Sigma_{k \times k} V_{k \times m}$
- Convert words and sentences to points in k-dimensional space

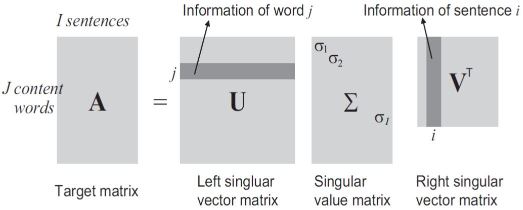
Meaning of columns and rows of U,V

The columns of $U_{n\times r}$ are the r eigenvectors of the $n \times n$ word similarity matrix $\mathbb{W}$,

$$\mathbb{W} = AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T \qquad (1)$$

The columns of $V_{m\times r}$ are the r eigenvectors of the $m \times m$ sentence similarity matrix $\mathbb{S}$,

$$\mathbb{S} = A^T A = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T \qquad (2)$$



Sentence Selection

Using V^T matrix, the matrix of concepts × sentences:

- In V^T matrix row order shows the degree of the importance of the concepts.
- The relation between the sentence and the concept is proportional to the cell values of V^T.
- The summarization process chooses the most informative sentence for each word.

Sentence Selection

Using both V and Σ matrices:

- Length of each sentence vector, represented by the row of V matrix, is used for sentence selection.
- The length of the sentence i is calculated using the words whose indexes are less than or equal to the given dimension.
- Σ matrix is used as a multiplication parameter in order to give more emphasis on the most important words.
- The sentence with the highest length value is chosen to be a part of the resulting summary.

Sentence Selection

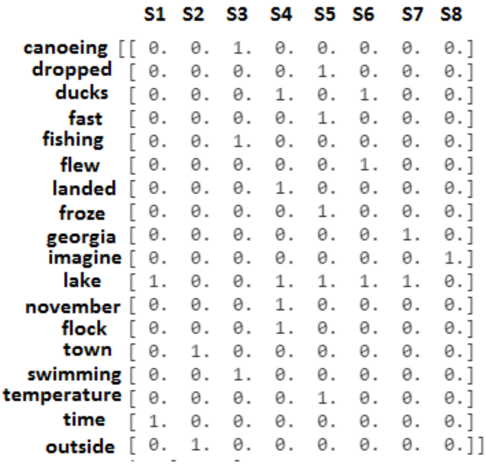
Using V^T and Σ matrices:

- Calculate the percentage of the related singular value over the sum of all singular values, for each concept using Σ matrix.
- Collect multiple sentences from each word using the provided result.

Simulation

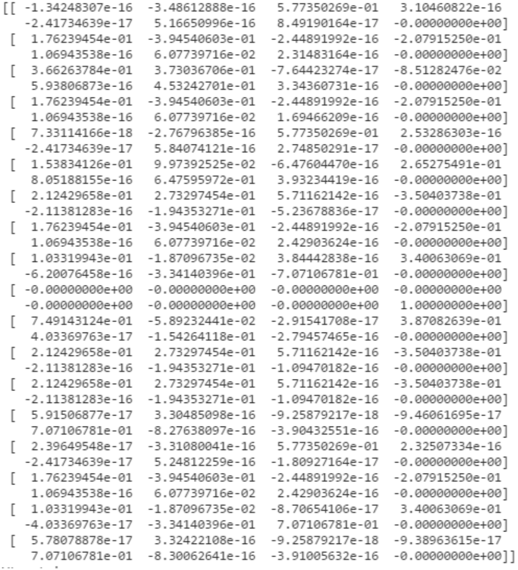
Text to summarize: The lake story from the book "Fried Green Tomatoes- Fannie Flagg"
"One time there was this lake.
And it was right outside of town.
We used to go fishing and swimming and canoeing in it.
One november this big flock of ducks came in and landed on that lake.
And then the temperature dropped so fast that the lake just froze right there.
And they the ducks they flew off you see and they took that lake right with them.
Now they say that lake is somewhere over in georgia.
Can you imagine that? "

Simulation - Count Matrix



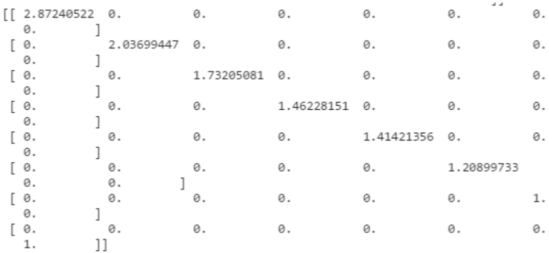
```
              S1  S2  S3  S4  S5  S6   S7  S8

   canoeing [[ 0.  0.  1.  0.  0.  0.  0.  0.]
    dropped  [ 0.  0.  0.  0.  1.  0.  0.  0.]
      ducks  [ 0.  0.  0.  1.  0.  1.  0.  0.]
       fast  [ 0.  0.  0.  0.  1.  0.  0.  0.]
    fishing  [ 0.  0.  1.  0.  0.  0.  0.  0.]
       flew  [ 0.  0.  0.  0.  0.  1.  0.  0.]
     landed  [ 0.  0.  0.  1.  0.  0.  0.  0.]
      froze  [ 0.  0.  0.  0.  1.  0.  0.  0.]
    georgia  [ 0.  0.  0.  0.  0.  0.  1.  0.]
    imagine  [ 0.  0.  0.  0.  0.  0.  0.  1.]
       lake  [ 1.  0.  0.  1.  1.  1.  1.  0.]
   november  [ 0.  0.  0.  1.  0.  0.  0.  0.]
      flock  [ 0.  0.  0.  1.  0.  0.  0.  0.]
       town  [ 0.  1.  0.  0.  0.  0.  0.  0.]
   swimming  [ 0.  0.  1.  0.  0.  0.  0.  0.]
temperature  [ 0.  0.  0.  0.  1.  0.  0.  0.]
       time  [ 1.  0.  0.  0.  0.  0.  0.  0.]
    outside  [ 0.  1.  0.  0.  0.  0.  0.  0.]]
```

Simulation - U Matrix



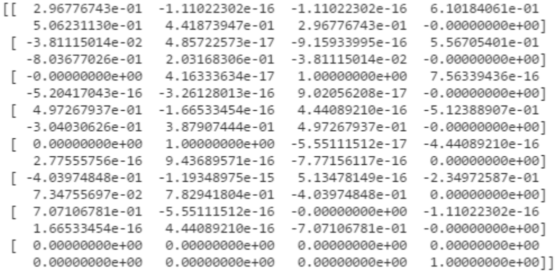
```
[[ -1.34248307e-16  -3.48612888e-16   5.77350269e-01   3.10460822e-16
   -2.41734639e-17   5.16650996e-16   8.49190164e-17  -0.00000000e+00]
 [  1.76239454e-01  -3.94540603e-01  -2.44891992e-16  -2.07915250e-01
    1.06943538e-16   6.07739716e-02   2.31483164e-16  -0.00000000e+00]
 [  3.66263784e-01   3.73036706e-01  -7.64423274e-17  -8.51282476e-02
    5.93806873e-16   4.53242701e-01   3.34360731e-16  -0.00000000e+00]
 [  1.76239454e-01  -3.94540603e-01  -2.44891992e-16  -2.07915250e-01
    1.06943538e-16   6.07739716e-02   1.69466209e-16  -0.00000000e+00]
 [  7.33114166e-18  -2.76796385e-16   5.77350269e-01   2.53286303e-16
   -2.41734639e-17   5.84074121e-16   2.74850291e-17  -0.00000000e+00]
 [  1.53834126e-01   9.97392525e-02  -6.47604470e-16   2.65275491e-01
    8.05188155e-16   6.47595972e-01   3.93234419e-16  -0.00000000e+00]
 [  2.12429658e-01   2.73297454e-01   5.71162142e-16  -3.50403738e-01
   -2.11381283e-16  -1.94353271e-01  -5.23678836e-17  -0.00000000e+00]
 [  1.76239454e-01  -3.94540603e-01  -2.44891992e-16  -2.07915250e-01
    1.06943538e-16   6.07739716e-02   2.42903624e-16  -0.00000000e+00]
 [  1.03319943e-01  -1.87096735e-02   3.84442838e-16   3.40063069e-01
   -6.20076458e-16  -3.34140396e-01  -7.07106781e-01  -0.00000000e+00]
 [ -0.00000000e+00  -0.00000000e+00  -0.00000000e+00  -0.00000000e+00
   -0.00000000e+00  -0.00000000e+00  -0.00000000e+00   1.00000000e+00]
 [  7.49143124e-01  -5.89232441e-02  -2.91541708e-17   3.87082639e-01
    4.03369763e-17  -1.54264118e-01  -2.79457465e-16  -0.00000000e+00]
 [  2.12429658e-01   2.73297454e-01   5.71162142e-16  -3.50403738e-01
   -2.11381283e-16  -1.94353271e-01  -1.09470182e-16  -0.00000000e+00]
 [  2.12429658e-01   2.73297454e-01   5.71162142e-16  -3.50403738e-01
   -2.11381283e-16  -1.94353271e-01  -1.09470182e-16  -0.00000000e+00]
 [  5.91506877e-17   3.30485098e-16  -9.25879217e-18  -9.46061695e-17
    7.07106781e-01  -8.27638097e-16  -3.90432551e-16  -0.00000000e+00]
 [  2.39649548e-17  -3.31080041e-16   5.77350269e-01   2.32507334e-16
   -2.41734639e-17   5.24812259e-16  -1.80927164e-17  -0.00000000e+00]
 [  1.76239454e-01  -3.94540603e-01  -2.44891992e-16  -2.07915250e-01
    1.06943538e-16   6.07739716e-02   2.42903624e-16  -0.00000000e+00]
 [  1.03319943e-01  -1.87096735e-02  -8.70654106e-17   3.40063069e-01
   -4.03369763e-17  -3.34140396e-01   7.07106781e-01  -0.00000000e+00]
 [  5.78078878e-17   3.32422108e-16  -9.25879217e-18  -9.38963615e-17
    7.07106781e-01  -8.30062641e-16  -3.91005632e-16  -0.00000000e+00]]
```

Singular Values



```
[[ 2.87240522  0.          0.          0.          0.          0.          0.
   0.        ]
 [ 0.          2.03699447  0.          0.          0.          0.          0.
   0.        ]
 [ 0.          0.          1.73205081  0.          0.          0.          0.
   0.        ]
 [ 0.          0.          0.          1.46228151  0.          0.          0.
   0.        ]
 [ 0.          0.          0.          0.          1.41421356  0.          0.
   0.        ]
 [ 0.          0.          0.          0.          0.          1.20899733
   0.          0.        ]
 [ 0.          0.          0.          0.          0.          0.          1.
   0.        ]
 [ 0.          0.          0.          0.          0.          0.          0.
   1.        ]]
```

Simulation - Vt Matrix



```
[[ 2.96776743e-01 -1.11022302e-16 -1.11022302e-16  6.10184061e-01
   5.06231130e-01  4.41873947e-01  2.96776743e-01 -0.00000000e+00]
 [-3.81115014e-02  4.85722573e-17 -9.15933995e-16  5.56705401e-01
  -8.03677026e-01  2.03168306e-01 -3.81115014e-02 -0.00000000e+00]
 [-0.00000000e+00  4.16333634e-17  1.00000000e+00  7.56339436e-16
  -5.20417043e-16 -3.26128013e-16  9.02056208e-17 -0.00000000e+00]
 [ 4.97267937e-01 -1.66533454e-16  4.44089210e-16 -5.12388907e-01
  -3.04030626e-01  3.87907444e-01  4.97267937e-01 -0.00000000e+00]
 [ 0.00000000e+00  1.00000000e+00 -5.55111512e-17 -4.44089210e-16
   2.77555756e-16  9.43689571e-16 -7.77156117e-16  0.00000000e+00]
 [-4.03974848e-01 -1.19348975e-15  5.13478149e-16 -2.34972587e-01
   7.34755697e-02  7.82941804e-01 -4.03974848e-01  0.00000000e+00]
 [ 7.07106781e-01 -5.55111512e-16 -0.00000000e+00 -1.11022302e-16
   1.66533454e-16  4.44089210e-16 -7.07106781e-01 -0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
   0.00000000e+00  0.00000000e+00  0.00000000e+00  1.00000000e+00]]
```

Simulation-Sentence Summary

Using **singular values** $\times V^T$ matrix:
One Sentence Summary
”one november this big flock of ducks came in and landed on that lake”

Systems & Tools Currently Used

- **sumy 0.6.0 :** Automatic summarization of text documents and HTML pages.
  - A Python module
  - Implements LSA summarization method proposed in Steinberger & Jezek, 2004
  - Available at: *https : //pypi.python.org/pypi/sumy*
- **SVDPlag v1.0 :** Automatic plagiarism detection system.
  - uses LSA framework on statistical computations
  - infers associations among the common N-grams contained in the examined documents.
  - Available at download section of *http : //textmining.zcu.cz/*

Avaliable Data Sets and Corpora

- CNN Worldview news programs
  - 243 news stories, each contains $>$10 sentences
  - 3 manual summaries per story, each contains 5 sentences
  - (Possibly) available upon request to Gong & Liu
- Reuters Corpora
  - A large collection of Reuters News stories
  - 3 released corpora: RCV1, RCV2 and TRC2
  - Available upon request to NIST
  - For details: *http://trec.nist.gov/data/reuters/reuters.html*
- ICSI Meeting Corpus
  - Simultaneously recorded audio and transcript
  - 75 meetings of 4 main types, 53 unique speakers
  - Available at: *http://groups.inf.ed.ac.uk/ami/icsi/download/*
- TS Corpus
  - A collection of various corpora in Turkish
  - As of 2017, 9 published corpora, e.g. TweetS, TS Abstract, TS Wikipedia
  - Available at: *http://tscorpus.com/*

The Success Criteria used for Evaluation

- Evaluation of Automatic Summaries: Also a very challenging and active research area
  - What is the ”right” summary?
  - What metrics to measure it (automatically)?
- General Approaches for Summary Evaluation
  - Evaluation by Sentence Co-selection
  - Content-based Methods
  - Relevance Correlation
  - Task-based Evaluations
- ROUGE: Recall-Oriented Understudy for Gisting Evaluation

General Approaches for Summary Evaluation

- Evaluation by Sentence Co-selection
  - Requires a ”right summary” to compute precision, recall and F-measure.
  - Time consuming and highly subjective
- Content-based Methods
  - Content-based similarity measures between a full text & its summary
  - Create a vector space model for full text & its summary
  - e.g. Cosine similarity, given by the formula:

$$\cos(X, Y) = \frac{\Sigma x_i * y_i}{\sqrt{\Sigma(x_i)^2 * \Sigma(y_i)^2}} \tag{3}$$

General Approaches for Summary Evaluation - 2

- Relevance Correlation
  - A measure for accessing the relative decrease in retrieval performance when indexing summaries instead of full documents
- Task-based Evaluations
  - Measure human performance using the summaries for a certain task
    - e.g. Suitability of using summaries instead of full texts for text categorization
  - Requires a classified corpus of texts

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

- Used for evaluating Automatic Summarization and Machine Translation systems
- Based on n-gram co-occurance, longest common subsequence and weighted longest common subsequence between the ideal summary and the extracted summary.
- Various metrics:
    - ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W, etc.

State-of-art Success Rates

- Gong & Liu, 2001:
  - 5 sentence for 5 major topics within a document
  - Used manual summaries as reference: 3 human summarizers
    - Evaluation by Sentence Co-selection

Table: Evaluation Results of Gong & Liu, 2001

Test Data	**R**	**P**	**F**
Assessor 1	0.60	0.62	0.61
Assessor 2	0.49	0.53	0.51
Assessor 3	0.55	0.68	0.61
Majority Vote	0.53	0.61	**0.57**

State-of-art Success Rates - 2

- Steinberger & Jezek, 2004:
  - 20 % summary ratio
  - LSA-based cosine similarity between summary and full text
    - Content-based Automatic Evaluation
  - 2 evaluation methods:
    - Similarity of the main topic
    - Similarity of the term significance

Table: Cos similarity results of the main topic evaluation

	Gong & Liu	**Steinberger & Jezek**
minimum	0,45113	0,45113
maximum	0,90419	0,95839
average	0,75134	**0,78705**

State-of-art Success Rates - 3

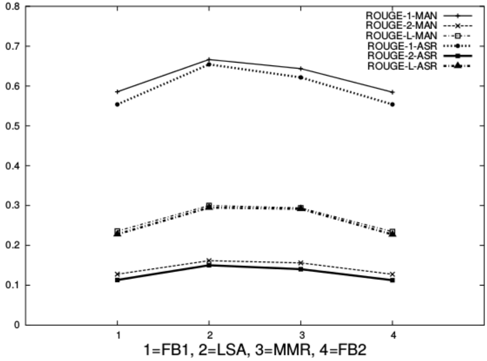
- Steinberger & Jezek, 2004: (cnt'd)

Table: Cos similarity results of the term significance evaluation

	Gong & Liu	**Steinberger & Jezek**
minimum	0,73751	0,73751
maximum	0,94336	0,94336
average	0,82392	**0,85123**

State-of-art Success Rates - 4

- Murray, Renals & Carletta, 2005:
  - Level of dimensionality reduction is learned from data
  - Used human transcriptions and output of an Automatic Speech Recognizer
  - Used various ROUGE scores for evaluation
    - Content-based Automatic Evaluation



ROUGE Scores for the Summarization Approaches

State-of-art Success Rates - 5

- Ozsoy, Cicekli & Alpaslan, 2010:
  - Cross & Topic : LSA-based Summarization Methods
  - Used 2 datasets separately, DS1 & DS2
    - DS2 contains longer articles
  - Used ROUGE-L scores for evaluation
    - Content-based Automatic Evaluation

Table: ROUGE-L Scores on DS1, Ozsoy et al. 2010

	G&L	S&J	MRC	**Cross**	Topic
frequency	0,236	0,250	0,244	**0,302**	0,244
binary	0,272	0,275	0,274	**0,313**	0,274
tf-idf	0,200	0,218	0,213	**0,304**	0,213
log-entropy	0,230	0,250	0,235	**0,302**	0,235
root type	*0,283*	*0,282*	*0,289*	***0,320***	*0,289*
mod. tf-idf	0,195	0,221	0,223	**0,290**	0,223

State-of-art Success Rates - 6

- Ozsoy, Cicekli & Alpaslan, 2010: (cnt'd)

Table: ROUGE-L Scores on DS2, Ozsoy et al. 2010

	G&L	S&J	MRC	**Cross**	Topic
frequency	0,256	*0,251*	0,259	0,264	0,259
binary	0,191	0,220	0,189	***0,274***	0,189
tf-idf	0,230	0,235	0,227	0,266	0,227
log-entropy	*0,267*	0,245	***0,268***	0,267	***0,268***
root type	0,194	0,222	0,197	0,263	0,197
mod. tf-idf	0,234	0,239	0,232	0,268	0,232

Approaches for different languages

- Inspected researches targeting English & Turkish documents
  - Sentence selection methods are not language specific
  - Scheme used to construct word-sentence matrix is a concern
    - In agglutinative languages such as Turkish, root type performs better

Discussion

- Generic text summarization (GTS) and its evaluation are very challenging research areas
  - Neither query nor topic are provided
  - Performance judgments tend to lack consensus
- LSA: An uninformed clustering algorithm
  - Clusters latent topics of a document
  - Reveals k most strong latent concepts (i.e. topics) in a document
- There is still room to improve the performance of GTS methods in general
  - Graph based approaches, or future-based approaches may be used together with LSA

References

- Gong, Y. and Liu, X. 2001. *Generic Text Summarization using Relevance Measure and Latent Semantic Analysis*
- Steinberger, J. and Jezek, K. 2004. *Using Latent Semantic Analysis in Text Summarization and Summary Evaluation*
- Murray, G., Renals, S. and Carletta, J. 2005. *Extractive Summarization of Meeting Recordings*
- Ozsoy, Makbule G., Cicekli, I. and Alpaslan, Ferda N. 2010. *Text Summarization of Turkish Texts using Latent Semantic Analysis*

References - 2

- Makoto Hirohata, Yousuke Shinnaka, Koji Iwano and Sadaoki Furui *Sentence Extraction-based Presentation Summarization Techniques And Evaluation Metrics* Department of Computer Science, Tokyo Institute of Technology
- Bob Glushko *Dimensionality Reduction and Latent Semantic Analysis*
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze *Introduction to Information Retrieval* Cambridge University Press. 2008.
- Finny G Kuruvilla, Peter J Park, Stuart L Schreiber *Vector atalgebra in the analysis of genome-wide expression data* Genome Biology