The Initial Study of Term Vector Generation Methods for News Summarization

Michal Rott

Speechlab
Institute of Information Technology and Electronics
Technical University of Liberec

RASLAN 2015

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Summaries

Type: Abstract and. Extract

Length: Indicative and Informative

Task: Single/Multi-document, Actualization, Comparative...

Why do we want summaries?

Summec - Count Methods

- Heuristic methods
  - remove Stop list
  - sentence position
  - sentence length
  - count of words
- TFxIDF

$$score(sent) = \sum_{t \in sent} tf(t) \times idf(t, D) \qquad (1)$$

  - Extract the best scoring sentences
  - What do they contains? Repetitive information?

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics and Interdisciplinary Studies

Summec - From Counts to Space

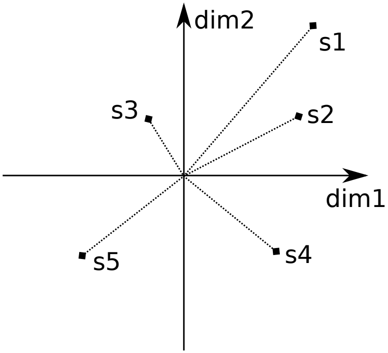
Every dimension represents one term.

Sentence is combination of terms.

$$\vec{s_i} = \sum_{t \in s_i} \vec{v_t}$$

The longest sentence is the most informative.

What is the second best sentence?



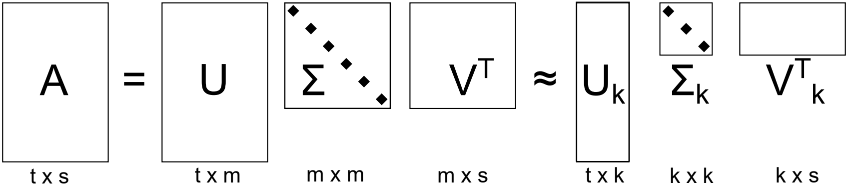
Vector model

The Curse of Dimensionality - spares high dimensional space

- Heuristic selection
  - Stop list
  - Lemmatization
  - Synonyms
- Data transformation
  - Matrix reduction - Latent Semantic Analysis
  - Random projection - Random Manhattan Indexing
- Neural network - Skip-gram model (word2vec)

Latent Semantic Analysis

Term-Document matrix **A** decomposed by SVD.
Dimensions with the lowest variance are throws away.



$\mathbf{U}_k$ - reduced matrix of term vectors
$\mathbf{V}_k^T$ - reduced matrix of sentence vectors

Summarization - Evaluation data

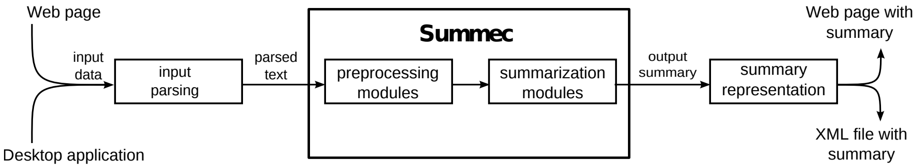
Experiment: Generate Informative Extract of Article

Test data:

- 50 Czech newspaper articles
- 15 annotators
- informative extract (25 % of original text)

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Summec - Scheme & Results



method	ROUGE-1		
	Recall [%]	Prec. [%]	F-score [%]
Heuristic	57.2	54.3	55.3
TFxIDF	62.6	53.3	57.3
LSA	55.4	55.2	55.1

New vector generation methods

Skip-gram model (word2vec)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. (2013)

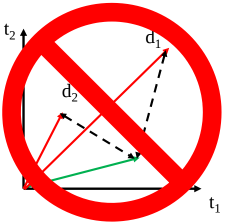
Random Manhattan Indexing

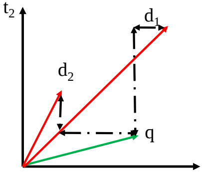
Zadeh, B.Q., Handschuh, S.: Random manhattan indexing. In: Proceedings - International Workshop on Database and Expert Systems Applications, DEXA. (2014) 203–208

Random Manhattan Indexing

Advantage of Random Projection - Euclidian distance

Not suitable for text vectors → Manhattan distance





Random Manhattan Indexing - algorithm

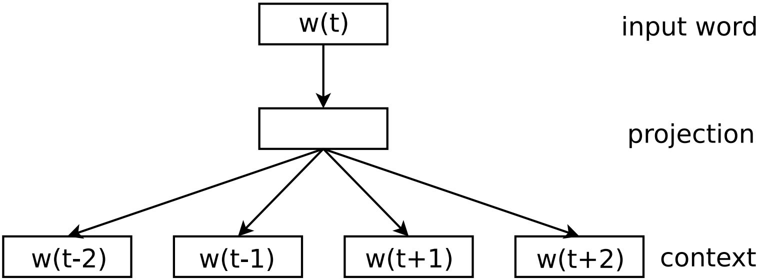
1. Extract index terms from text
2. Generate vector to each index term (2)

$$v_i = \begin{cases} \frac{-1}{U_1} & \text{with prob.} \quad \frac{s}{2} \\ 0 & \text{with prob.} \quad 1-s \\ \frac{1}{U_2} & \text{with prob.} \quad \frac{s}{2} \end{cases} \qquad s = \frac{1}{\sqrt{\beta|\{t\}|}} \tag{2}$$

3. Compute sentence vector (3)

$$\vec{s_j} = \sum_{t \in s_j} \vec{v_t} \tag{3}$$

Skip-gram model



Objective function: maximize value of (4)

$$\frac{1}{T}\sum_{t\in T}\sum_{-c\leq j\leq c, j\neq 0}\log(p(w_{t+j}|w_t)) \tag{4}$$

Results

Training data:
RMI - 706 033 (1 025 815) lemmas
SGM - 8.6 GB lemmatized ASR training data

Table: Comparison of ROUGE-1 score of summarization methods

method	Recall [%]	Precision [%]	F-score [%]
LSA	55.4	55.1	55.2
RMI	50.7	56.7	53.3
SGM	50.7	56.7	53.3
TFxIDF	62.6	53.3	57.3

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Conclusion

- Proposed schemes do not perform better than TFxIDF.
- TFxIDF is still the best performing method.

- RMI - poor results are understandable (random vectors)
- SGM - high expectations - maybe different approach

Current Work, Future Paper?

- Sentences are not sufficient - long and without context.
  Example:
  Title: Věci Veřejné krituzijí akreditační komisi za zveřejnění usnesení.
  Extract: Podle ní má komise povinnost svá rozhodnutí zveřejňovat a u tohoto sledovaného případu chtěla rychlým vyjádřením předejít spekulacím.
- Extracting clauses is more interesting way.
  Podle Dvořákové má komise povinnost zveřejňovat svá rozhodnutí.
  Dvořáková chtěla předejít spekulacím u tohoto sledovaného případu rychlým vyjádřením.
  (Tento případ = Mají se rušit plzeňská práva)

Summec and Aara

Summec + Aara = Sumara

- Evaluate Aara's accuracy.
- How to evaluate abstract? Read and mark vs. automatic

The End

Thank you for attention