Using Latent Semantic Analysis in Text Summarization and Summary Evaluation

Josef Steinberger*
jstein@kiv.zcu.cz

Karel Ježek*
Jezek_ka@kiv.zcu.cz

Abstract: This paper deals with using latent semantic analysis in text summarization. We describe a generic text summarization method which uses the latent semantic analysis technique to identify semantically important sentences. This method has been further improved. Then we propose two new evaluation methods based on LSA, which measure content similarity between an original document and its summary. In the evaluation part we compare seven summarizers by a classical content-based evaluator and by the two new LSA evaluators. We also study an influence of summary length on its quality from the angle of the three mentioned evaluation methods.

Key Words: Generic Text Summarization, Latent Semantic Analysis, Summary Evaluation

1 Introduction

Generic text summarization is a field that has seen increasing attention from the NLP community because effective automatic summarization would be useful in a variety of areas, especially in the explosively growing world-wide web. This paper discusses the use of latent semantic analysis in this field. We mention here classes of summarization methods and a method based on LSA which has been recently published. We have further modified and improved this method. One of the most controversial parts of the summary research is its evaluation process. Next part of the article deals with possibilities of summary evaluation. We propose there two new evaluation methods based on LSA, which measure a content similarity between an original document and its summary. At the end of the paper we present evaluation results and further research directions.

2 Generic Text Summarization Methods

Generic text summarization approaches are divided into four classes. The first class we call heuristic approaches. This extraction methods use for scoring sentences easy techniques as for example the sentence position within the document or an occurrence of a word from the title in a sentence [6]. The next group includes approaches based on a document corpus (corpus-based methods) [7]. An example of such a method is TF.IDF (term frequency · inverse document frequency). The third class consists of methods which take a discourse structure into account. An example is the lexical chains method which searches for chains of context words in the text [8]. The last group is called knowledge-rich approaches. They are the most advanced but can be used only in particular domains (e. g. STREAK – summaries of

* Department of Computer Science and Engineering, Univerzitní 22, CZ-306 14 Plzeň

basketball games [9]). A quite new approach in text summarization uses the latent semantic analysis.

3 LSA Summarization

Yihong Gong and Xin Liu have published the idea of using LSA in text summarization in 2002 [1]. They, inspired by the latent semantic indexing, applied the singular value decomposition (SVD) to generic text summarization. The process starts with creation of a term by sentences matrix $\mathbf{A} = [A_1, A_2, \ldots, A_n]$ with each column vector A_i, representing the weighted term-frequency vector of sentence i in the document under consideration. If there are a total of m terms and n sentences in the document, then we will have an $m \times n$ matrix $\mathbf{A}$ for the document. Since every word does not normally appear in each sentence, the matrix $\mathbf{A}$ is sparse.

Given an $m \times n$ matrix $\mathbf{A}$, where without loss of generality $m \geq n$, the SVD of $\mathbf{A}$ is defined as:

$$A = U\Sigma V^T,$$

where $\mathbf{U} = [u_{ij}]$ is an $m \times n$ column-orthonormal matrix whose columns are called left singular vectors; $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is an $n \times n$ diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order, and $\mathbf{V} = [v_{ij}]$ is an $n \times n$ orthonormal matrix, whose columns are called right singular vectors (see figure 1). If rank($\mathbf{A}$) = r, then (see [5]) $\mathbf{\Sigma}$ satisfies:

$$\sigma_1 \geq \sigma_2 \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0.$$

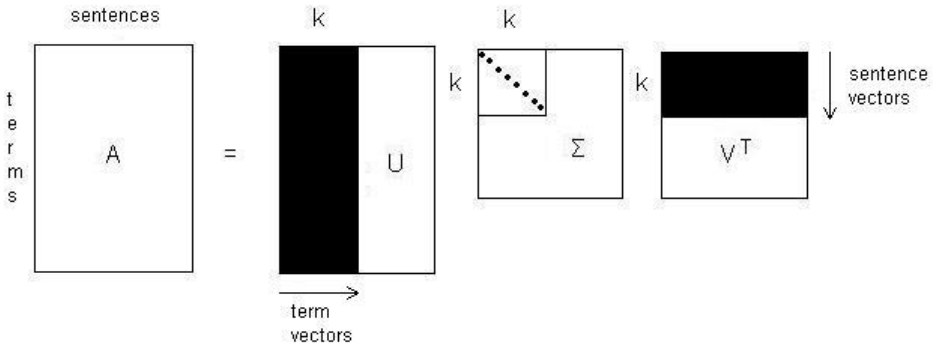

Figure 1: Singular Value Decomposition

The interpretation of applying the SVD to the terms by sentences matrix $\mathbf{A}$ can be made from two different viewpoints. From transformation point of view, the SVD derives a mapping between the m-dimensional space spawned by the weighted term-frequency vectors and the r-dimensional singular vector space. From semantic point of view, the SVD derives the latent semantic structure from the document represented by matrix $\mathbf{A}$. This operation reflects a breakdown of the original document into r linearly-independent base vectors or concepts. Each term and sentence from the document is jointly indexed by these base vectors/concepts. A unique SVD feature is that it is capable of capturing and modelling interrelationships among terms so that it can semantically cluster terms and sentences. Further-more, as

demonstrated in [5], if a word combination pattern is salient and recurring in document, this pattern will be captured and represented by one of the singular vectors. The magnitude of the corresponding singular value indicates the importance degree of this pattern within the document. Any sentences containing this word combination pattern will be projected along this singular vector, and the sentence that best represents this pattern will have the largest index value with this vector. As each particular word combination pattern describes a certain topic/concept in the document, the facts described above naturally lead to the hypothesis that each singular vector represents a salient topic/concept of the document, and the magnitude of its corresponding singular value represents the degree of importance of the salient topic/concept.

Based on the above discussion, authors [1] proposed a summarization method which uses the matrix $\mathbf{V}^T$. This matrix describes an importance degree of each topic in each sentence. The summarization process chooses the most salience sentence for each topic. It means that the *k*'th sentence we choose has the largest index value in *k*'th right singular vector in matrix $\mathbf{V}^T$.

4 Enhanced LSA Summarization

The above described summarization method has two significant disadvantages. At first it is necessary to use the same number of dimensions as is the number of sentences we want to choose for a summary. However, the higher is the number of dimensions of reduced space, the less significant topic we take into a summary. This disadvantage turns into an advantage only in the case when we know how many different topics has the original document and we choose the same number of sentences into a summary. The second disadvantage is that a sentence with large index values, but not the largest (it doesn't win in any dimension), will not be chosen although its content is for the summary very suitable.

In order to clear out the discussed disadvantages we propose following modifications in the SVD-based summarization method. Again we need to compute SVD of a term by sentences matrix. We get the three matrices as shows the figure 1. For each sentence vector in matrix **V** (its components are multiplied by corresponding singular values) we compute its length. The reason of the multiplication is to favor the index values in the matrix **V** that correspond to the highest singular values (the most significant topics). Formally:

$$s_k = \sqrt{\sum_{i=1}^{n} v_{k,i}^2 \cdot \sigma_i^2} \ ,$$

where s_k is the length of the vector of *k*'th sentence in the modified latent vector space. It is its salience score for summarization too. *n* is a number of dimensions of the new space. This value is independent on the number of summary sentences (it is a parameter of the method). In our experiments we chose the dimensions whose singular values didn't fall under the half of the highest singular value (but it is possible to set a different strategy). Finally, we put into a summary the sentences with the highest values in vector *s*.

5 Summary Evaluation

Evaluation of automatic summarization in a standard and inexpensive way is a difficult task. It is the equally important area as the own summarization process and that's why many evaluation approaches were developed [2].

5.1 Evaluation by Sentence Co-selection

Co-selection measures include precision and recall of co-selected sentences. These methods require having at disposal a "right extract" (to which we could compute precision and recall). We can obtain this extract in several ways. The most common way is to obtain some human (manual) extracts and to declare the average of these extracts as "ideal (right) extract". However, obtaining of human extracts is usually problematic. Another problem is that two manual summaries of the same input do not in general share many identical sentences.

5.2 Content-based methods

We can clear out the above discussed weakness of co-selection measures by content-based similarity measures. These methods compute the similarity between two documents at a more fine-grained level than just sentences. The basic method is to compute the similarity between the full text document and its summary with the cosine similarity measure, computed by the following formula:

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

where X and Y are representations based on the vector space model.

5.3 Relevance Correlation

Relevance correlation is a measure for accessing the relative decrease in retrieval performance when indexing summaries instead of full documents [2].

5.4 Task-based evaluations

Task-based evaluations measure human performance using the summaries for a certain task (*after* the summaries are created). We can for example measure a suitability of using summaries instead of full texts for text categorization [3]. This evaluation requires a classified corpus of texts.

6 Evaluation based on Latent Semantic Analysis

We classify this new method to a content-based category because, like the classical cosine content-based approach (see 5.2), it evaluates a summary quality via content similarity between a full text and its summary. Our method uses Singular Value Decomposition of a terms by sentences matrix (see 3.), exactly matrix **U**. This matrix represents the degree of importance of terms in salient topics/concepts. In evaluation we measure the similarity between the matrix **U** derived from the SVD performed on the original document and the matrix **U** derived from the SVD performed on the summary. For appraising this similarity we have proposed two measures.

6.1 Similarity of the Main Topic

This method compares first left singular vectors (see figure 2) of the full text SVD (i. e. SVD performed on the original document) and the summary SVD (i. e. SVD performed on the summary). These vectors correspond to the most salience word pattern in the full text and its summary (we can call it the main topic).

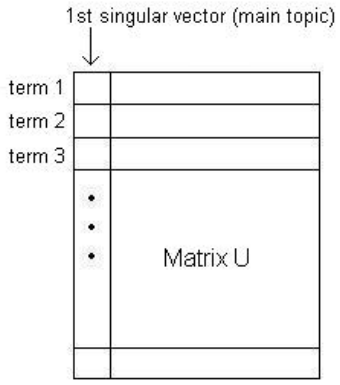

Figure 2: 1st singular vector (main topic)

Then we measure the angle between the first left singular vectors. They are normalized, so we can use the following formula:

$$\cos\varphi = \sum_{i=1}^{n} ue_i \cdot uf_i \ ,$$

where *uf* is the first left singular vector of the full text SVD, *ue* is the first left singular vector of the summary SVD (values, which correspond to particular terms, are sorted up the full text terms and instead of missing terms are zeroes), *n* is a number of unique terms in the full text.

6.2 Similarity of the Term Significance

This evaluation method compares a summary with the original document from an angle of *n* most salient topics. We propose the following process:

- Perform the SVD on a document matrix (see 3.).
- For each term vector in matrix **U** (its components are multiplied by corresponding singular values) compute its length. The reason of the multiplication is to favor the index values in the matrix **U** that correspond to the highest singular values (the most significant topics). Formally:

$$s_k = \sqrt{\sum_{i=1}^{n} u_{k,i}^2 \cdot \sigma_i^2} \ ,$$

  where s_k is the length of the *k*'st term vector in the modified latent vector space, *n* is a number of dimensions of the new space. In our experiments we chose the dimensions whose singular values didn't fall under the half of the highest singular value (but it is possible to set a different strategy).
- From the lengths of the term vectors (s_k) make a resulting term vector, whose index values hold an information about the term significance in the modified latent space (see figure 3).
- Normalize the resulting vector.

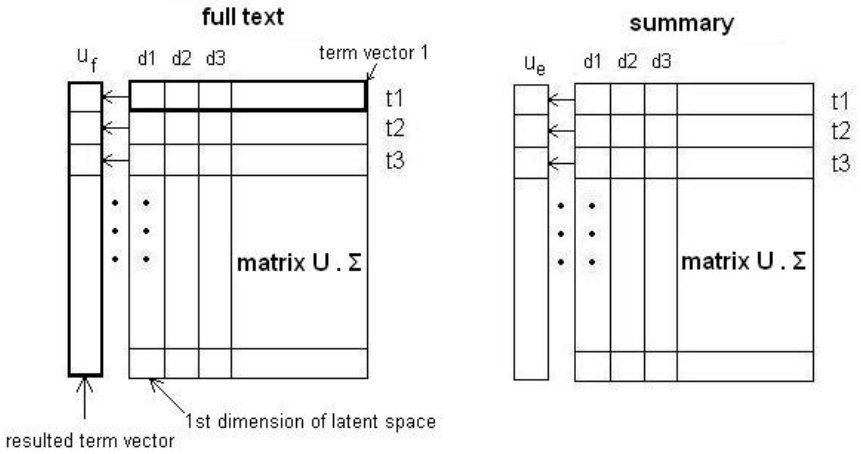

Figure 3: Creation of a resulting term vectors of a full text and a summary

This process is performed on the original document and on its summary (for the same number of dimensions according to the summary) (see figure 3). In the result, we get one vector corresponding to the term vector lengths of the full text and one of its summary. As a similarity measure we use again the angle between resulting vectors (see 6.1).

This evaluation method has the following advantage above the previous one. Suppose, an original document contains two topics with the relatively same significance (corresponding singular values are almost the same). When the second significant topic outweighs the first one in a summary, the main topic of the summary will not be consistent with the main topic of the original. Taking more singular vectors (than just one) into account removes this weakness.

7 Results

We evaluated the following summarizers:

- Gong + Liu LSA summarizer
- LSA summarizer based on our approach
- Random summarizer – evaluation based on the average of 10 random extracts
- 1-itemsets – summarizer based on itemsets method [5]
- 1-itemsets + positional heuristic [5]
- 1-itemsets + mutual reinforcement heuristic [5]
- tf.idf – summarizer based on frequency method [5]

These summarizers were evaluated by the following three evaluation methods:

- Cosine similarity – classical content-based method, in result tables marked as (1)
- LSA similarity
    - Similarity of the main topic, (2)
    - Similarity of the term significance, (3)

We tested documents from Reuters collection. Their required minimum length was 20 sentences. The summarization ratio was set to 20 %. Results are presented in tables 1-3. Values are cosines of angles between a full text and its summary.

	Summary method						
	LSA - L+G	LSA - our	Random	Positional	Mut. Reinf.	1-itemsets	TF.IDF
minimum	0,64446	0,64446	0,52528	0,63692	0,6359	0,6262	0,61351
maximum	0,8505	0,88552	0,80109	0,86818	0,89266	0,89266	0,89266
average	0,76101	**0,77153**	0,64686	0,74494	0,7589	0,76248	0,75171

Table 1: Cosine similarity evaluation – classical content-based evaluation (1)

	Summary method						
	LSA - L+G	LSA - our	Random	Positional	Mut. Reinf.	1-itemsets	TF.IDF
minimum	0,45113	0,45113	0,33566	0,42926	0,53881	0,49791	0,43326
maximum	0,90419	0,95839	0,75969	0,95511	0,95839	0,95927	0,89823
average	0,751344	**0,78705**	0,48795	0,73014	0,77059	0,7635	0,75801

Table 2: Similarity of the main topic evaluation (2)

	Summary method						
	LSA - L+G	LSA - our	Random	Positional	Mut. Reinf.	1-itemsets	TF.IDF
minimum	0,73751	0,73751	0,42923	0,64803	0,71072	0,66116	0,66442
maximum	0,94336	0,94336	0,71599	0,90527	0,93304	0,91204	0,93304
average	0,82392	**0,85123**	0,54244	0,77124	0,81749	0,8112	0,80357

Table 3: Similarity of the term significance evaluation (3)

Tables 4-6 show the dependencies of a summary quality on the summarization ratio and the evaluation methods.

summary type	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
LSA - our	0,614	0,757	0,833	0,888	0,929	0,952	0,969	0,982	0,992	1
LSA - L+G	0,635	0,757	0,834	0,878	0,92	0,944	0,963	0,979	0,989	1
Random	0,475	0,643	0,743	0,798	0,857	0,899	0,932	0,955	0,978	1

Table 4: Dependency of evaluator (1) on summary length

summary type	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
LSA - our	0,614	0,775	0,875	0,917	0,92	0,968	0,98	0,99	0,995	1
LSA - L+G	0,59	0,75	0,839	0,885	0,888	0,938	0,955	0,971	0,982	1
Random	0,376	0,487	0,565	0,669	0,748	0,789	0,88	0,918	0,948	1

Table 5: Dependency of evaluator (2) on summary length

summary type	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
LSA - our	0,619	0,829	0,887	0,931	0,963	0,977	0,987	0,992	0,997	1
LSA - L+G	0,599	0,826	0,876	0,911	0,947	0,959	0,971	0,98	0,986	1
Random	0,376	0,553	0,645	0,708	0,785	0,838	0,885	0,922	0,961	1

Table 6: Dependency of evaluator (3) on summary length

8 Conclusion

The practical tests proved that our summarizing method outperforms the other examined methods. The classical cosine evaluator shows only small differences between summarizers (the best summarizer – 0,77 and the worst (random) - 0,65). It's caused by a shallow level of this evaluation method which takes into account only term counts in compared documents. The evaluation based on LSA is a more fine-grained approach. It is possible to say that it evaluates a summary via term co-occurrences in sentences. In the evaluation by the main topic we noticed the disadvantage discussed in 6.2 (proved in 10% of documents). The evaluation by the term significance removes this weakness. There is also a big difference between random and other summarizers. Next information we observed from the evaluation was that the LSA summarizer has been shown as the expressively best with the evaluator (3). This property was expected. Our experiments showed that LSA is very sensitive on a stoplist and a lemmatization process. In future research we plan to try other weighing schemes and a normalization of a sentence vector on the SVD input. Of course, other evaluations are needed, especially on longer texts than the Reuters documents are.

References

1. Y. Gong, X. Liu: *Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis*. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States 2001, pp. 19-25
2. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Celebi, D. Liu, E. Drabek: *Evaluation Challenges in Large-scale Document Summarization.* Proceeding of the 41th annual meeting of the Association for Computational Linguistics, Sapporo, Japan 2003, pp. 375-382
3. K. Ahmad, B. Vrusias, P. C. F. Oliveira: *Summary Evaluation and Text Categorization*. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada 2003, pp. 443-444
4. J. Hynek, K. Ježek: *Practical Approach to Automatic Text Summarization*. Proceedings of the ELPUB '03 conference, Guimaraes, Portugal 2003, pp. 378-388
5. M. W. Berry, S. T. Dumais, G. W O'Brien: *Using Linear Algebra for Intelligent Information Retrieval*. SIAM Review 1995
6. H. P. Edmundson: *New Methods in Automatic Extracting*. Journal of the Association for Computing Machinery 16(2):264-228
7. J. Kupiec, J. Pedersen, F. Chen: *A trainable Document Summarizer*: Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, United States 1995, pp. 68-73
8. R. Barzilay, M. Elhadad: *Using Lexical Chains for Text Summarization*. Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL Madrid, Spain 1997
9. K. McKeown, J. Robin, K. Kukich: *Generating Concise Natural Language Summaries.* Information Processing and Management: an International Journal, Volume 31, Issue 5, 1995