VSMBM: A NEW METRIC FOR AUTOMATICALLY GENERATED TEXT SUMMARIES EVALUATION

Alaidine Ben Ayed[1,3], Ismaïl Biskri[2,3] and Jean-Guy Meunier[3]

[1]Department of Computer Science, Université du Québec à Montréal (UQAM), Canada
[2]Department of Mathematics and Computer Science, Université du Québec à Trois-Rivières (UQTR), Canada
[3]LANCI : Laboratoire d'ANalyse Cognitive de l'Information, Université du Québec à Montréal (UQAM), Canada

ABSTRACT

In this paper, we present VSMbM; a new metric for automatically generated text summaries evaluation. VSMbM is based on vector space modelling. It gives insights on to which extent retention and fidelity are met in the generated summaries. Two variants of the proposed metric, namely PCA-VSMbM and ISOMAP VSMbM, are tested and compared to Recall-Oriented Understudy for Gisting Evaluation (ROUGE): a standard metric used to evaluate automatically generated summaries. Conducted experiments on the Timeline17 dataset show that VSMbM scores are highly correlated to the state-of-the-art Rouge scores.

KEYWORDS

Automatic Text Summarization, Automatic summary evaluation, Vector space modelling.

1. INTRODUCTION

1.1. Automatic Text Summarization

Automatic text summarization (ATS) is the process of creating a short, accurate, and fluent summary from a longer source text [1]. It has been a field of study for decades. [2] provides six reasons why we need ATS. Indeed; 1) Summaries reduce reading time, 2) they make the selection process easier when researching documents, 3) ATS improves the effectiveness of indexing, 4) ATS algorithms are less biased than human summarizers, 5) Personalized summaries are useful in question answering systems as they provide personalized information and 6) Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process. Note that automatically generated summaries should satisfy three criteria:

- Retention: It is a measure of how much the generated summary reports salient topics present in the original text.
- Fidelity: Does the summary accurately reflect the author's point of view?

David C. Wyld et al. (Eds): ITCSE, NLCA, ICAIT, CAIML, ICDIPV, CRYPIS, WiMo - 2020
pp. 121-128, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100510

- Coherence: To which extent, the generated extract is semantically meaningful?

There are mainly two subtasks of ATS [2]: 1) single text summarization: it uses only one source text to build the summary, 2) multi text summarization: it uses a bunch of source texts to create the final output. In both cases, evaluating the generated summaries is still a challenging research area.

In the next two section, we make a short state of the art of most relevant proposed protocols for automatically generated text summarization. Then, we present key features which make the originality of our work.

1.2. Related Work

Evaluating automatically generated summaries is not an effortless task. In the last two decades, significant advances have been made in this research field. Therefore, various evaluation measures have been proposed. SUMMAC [3], DUC (Document Understanding Conference) [4] and TAC (Text Analysis Conference) [5] are the main evaluation campaigns led since 1996. Note that the evaluation process can be led either in reference to some ideal models or without reference [6]. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is the most used metric for automatically generated abstracts evaluation. Summaries are compared to a reference or a set of references (human-produced summaries) [7]. Note that there are five variants of the ROUGE metric: 1) ROUGE-N [8]: it captures the overlap of N-grams between the system and reference summaries, 2) ROUGE-L [9]: it gives statistics about the Longest Common Subsequence (LCS), 3) ROUGE-W: a set of weighted LCS-based statistics that favors consecutive LCSes, 4) ROUGE-S [10]: a set of Skip-bigram (any pair of words in their sentence order) based co-occurrence statistics. COVERAGE is another metric which has been used in DUC evaluations. It gives an idea on to which extent peer summary conveys the same information as a model summary [11]. RESPONSIVENESS has also been used in focused-based summarization tasks of DUC and TAC evaluation campaigns [11]. It ranks summaries in a 5-point scale indicating how well the summary satisfied a set of needed information criteria. The pyramid evaluation approach uses Summarization Content Units (SCUs) to calculate a bunch of weighted scores [12]. A summary containing units with higher weights will be affected a high pyramid score. A SCU has a higher weight if it appears frequently in human-generated summaries. Fresa is another metric [13]. It is the state-of-the-art technique for evaluating automatically generated summaries without using a set of human-produced reference summaries. It computes a variety of divergences among probability distributions. Recently, [14] proposed a new implementation of the ROUGE protocol without human-built model summaries. The new summary evaluation model (ASHuR) extracts most informative sentences of the original text based on a bunch of criteria: the frequency of concepts, the presence of cue-words, sentence length, etc. Then, the extracted set of sentences will be considered as the model summary. [15] gives an overview of challenging issues related to summary evaluation

1.3. Originality of our work

Most of the above described metrics only focus on the overlap of N-grams between the original text and the generated summary. In other words, they reflect the coverage ratio meanwhile they don’t give insights on to which extent fidelity is met, i.e. if a long source text contains six concepts and a first summary focuses on the four last most important ones, it will be assigned a higher score than another summary focusing on the most important two concepts present in the original text. In this case retention is met. However, it is not the case for the fidelity criterion

In this paper we present a new vector space modelling-based metric for automatic text summaries evaluation. The proposed metric gives insights on to which extent both retention and fidelity are met. We assume that fidelity is met if we assign higher weights to text units related to most important concepts reported in the original text. The next section describes technical and mathematical details of the proposed metric. The third one describes conducted experiments and obtained results. Conclusion and future work are exposed in the fourth section.

2. VECTOR SPACE MODELLING BASED METRIC (VSMBM) FOR AUTOMATICALLY GENERATED TEXT SUMMARIES EVALUATION

From a computational point of view, the main idea is to project the original text onto a lower dimensional space that captures the essence of concepts present in it. Unitary vectors of the latter space are used to compute the two proposed *VSMbM* metrics. Mathematical and implementation details of *PCA-VSMbM* and *ISOMAP-VSMbM* will be expanded in the coming two subsections.

2.1. The *PCA-VSMbM*

First, source text is segmented onto *m* sentences. Then a dictionary of all nouns is constructed and filtered in order to remove all generic nouns. Text is then represented by an $m \times z$ matrix, where *m* is the number of segments and z is the number of unique tokens. Next the conceptual space is being constructed. It will be used later to compute the *PCA-VSMbM* metric.

2.1.1. Construction of the conceptual space

Each sentence S_i is represented by a column vectorζ_i. $\zeta_i$is a vector of Z components. Each component represents the *tf-idf* of a given word. Afterwards, mean concept vector τ is computed as follows:

$$\tau = \frac{1}{m}\sum_{i=1}^{m} \zeta_i \tag{1}$$

Note that each ζ_i should be normalized to get rid of redundant information. This is performed by subtracting the mean concept:

$$\Theta_i = \zeta_i - \tau \tag{2}$$

In the next step, the covariance matrix is computed as follows:

$$C = \frac{1}{m}\sum_{n=1}^{m} \Theta_n \Theta_n^T = AA^T \tag{3}$$

Where $A = [\Theta_1, \ldots, \Theta_m]$. Note that C in (3) is a $z \times z$ matrix and A is a $z \times m$ matrix. Eigen concepts are the eigenvectors of the covariance matrix$C$. They are obtained by performing a singular value decomposition of A:

$$A = U.S.V^T \tag{4}$$

Where dimensions of matrix U, S and V are respectively $z \times z$, $z \times m$ and $m \times m$. Also, U and V are orthogonal ($UU^T = U^TU = Id_z$ and $VV^T = V^TV = Id_m$). In addition to that;

- Columns of V are eigenvectors of A^TA.
- Columns of U are eigenvectors AA^T.
- Squares of singular values s_k of S are eigenvalues λ_k of AA^T and A^TA.

Note that $m < z$. So, eigenvalues λ_k of AA^T are equal to zero when $k > m$ and their associated eigenvectors are not necessary. So, matrix U and S can be truncated, and, dimensions of U, S and V in (4) become respectively $z \times m$, $m \times m$ and $m \times m$. Next, conceptual space is being constructed by K eigenvectors associated to the highest K eigenvalues:

$$\Xi_k = [U_1, U_2, \dots, U_k] \tag{5}$$

Each projected sentence onto the conceptual space is represented as a linear combination of K eigenconcepts:

$$\Theta_i^{proj} = \sum_k C_{\Theta_i}(k) U_k \tag{6}$$

Where $C_{\Theta_i}(k)$=$U_k^T \Theta_i$ is a vector providing coordinates of the projected sentence in the conceptual space.

2.1.2. Computation of the PCA-VSMbM score

The goal here is to find out to which extent selected sentences to be part of the generated summary are expressing the main concepts of the original text. Thus, each vector ζ_i representing a given sentence S_i is normalized by subtracting the mean concept $\tau : \Theta_q = \zeta_i - \tau$. Then it is projected onto the newly constructed conceptual space:

$$\Theta_q^{proj} = \sum_k C_{\Theta_q}(k) U_k \tag{7}$$

Next, the Euclidean distance between a given concept q and any projected sentence is defined and computed as follows:

$$d_i\left(\Theta_q^{proj}\right) = \left\|\Theta_q^{proj} - \Theta_i^{proj}\right\| \tag{8}$$

Next, Retention-Fidelity matrix is constructed as follows: First, we fix a window size W. In the bellow example, W is set to 4. The first line gives the index of the four sentences having the smallest distances to the vector encoding the first most important concept. The second line gives the same information related to the second most important concept. Also, the order of a given sentence in each window W depends on its distance to a given concept. For instance, the first sentence is the best one to encode the first most important concept while the 8th sentence is the last one to encode the same concept in a window of four sentences.

$$\begin{array}{c} \\ \text{Concept importance} \downarrow \end{array} \begin{array}{c} \text{Distance (concept)} \longrightarrow \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 1 & 6 & 9 & 8 \\ 1 & 22 & 13 & 11 \\ 10 & 22 & 6 & 1 \\ 22 & 2 & 1 & 11 \\ 9 & 11 & 2 & 8 \end{pmatrix} \end{array}$$

Next, the *Retention* score of each sentence being projected in the conceptual space is defined as follows: it's equal to the number of times it occurs in a window of size W when taking in consideration the most important K concepts. The main intuition behind it, is that a given sentence having a height *Retention* sore should encode as much as possible the K most important concepts expressed in the original text.

$$R_{kw}(s) = \frac{1}{k}\sum_{i=1}^{k} \alpha_i \tag{9}$$

$\alpha_i = 1$ if the sentence S occurs in the i^{th} window. If not, it is equal to zero.

Now, the *PCA-VSMbM* score is defined as shown in the tenth equation as the averaged sum of the retention coefficients of summary sentences. Note that every retention coefficient is weighted according to the sentence's position in a given window of size W . The main intuition behind it is that, single units (sentences) of a given summary whose *PCA-VSMbM* score is high should encode the most important concepts expressed in the original text. So, they should have minimal distances $d_i\left(\Theta_q^{proj}\right) = \left\|\Theta_q^{proj} - \Theta_i^{proj}\right\|$ in equation 8. In other words, the *PCA-VSMbM* score gives insights on to which extent extracted sentences encode concepts present in the original text while taking in consideration the importance degree of each concept

$$PCAVSMbM_{\ kw}(s) = \frac{1}{p}\frac{1}{k}\sum_{j=1}^{p}\sum_{i=1}^{k} \alpha_i \left[1 + \frac{1-\psi_i}{w}\right] \tag{10}$$

p is the number of extracted sentences to construct the summary, $\alpha_i = 1$ if a sentence s occurs in the i^{th}window. If not, it is equal to zero. ψ_i is the rank of s in the i^{th} window.

2.2. The *ISOMAP-VSMbM*

In the *ISOMAP-VSMbM*, we rather use the geodesic distance. The *ISOMAP-VSMbM* approach consists in constructing a k-nearest neighbor graph on n data points each one representing a sentence in the original space. Then, we compute the shortest path between all points as an estimation of geodesic distance D^G. Finally, we compute the decomposition K in order to construct Ξ_k previously defined in equation 5 where:

$$K = \frac{1}{2} H D^G H \tag{11}$$

H is acentering matrix; $H = Id\ \frac{1}{n} ee^T$ and $e = [1,1,\ldots,1]^T$. T is an $n \times 1$ matrix. Note that the decomposition of K is not always possible in the sense that there is no guarantee that K is a

positive semidefinite matrix. We deal with this case by finding out the closest positive semidefinite matrix to K. Then we decompose it. Next we proceed the same way we proceeded previously with *PCA-VSMbM*. *ISOMAP-VSMbM* is defined as *PCA-VSMbM* in equation 10.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

The *Timeline17* dataset is used for experiments [16]. It consists of 17 manually created timelines and their associated news articles. They mainly belong to 9 broad topics: BP Oil Spill, Michael Jackson Death (Dr. Murray Trial), Haiti Earthquake, H1N1 (Influenza), Financial Crisis, Syrian Crisis, Libyan War, Iraq War, Egyptian Protest. Original articles belong to news agencies, such as BBC, Guardian, CNN, Fox news, NBC News, etc. The contents of these news are in plain text file format and noise filtered.

3.2. Results and discussion

In order to evaluate the proposed metric, we compute the Pearson's correlation between *VSMbM* and *ROUGE* (Recall-Oriented Understudy for Gisting Evaluation) sores. Note that Pearson's correlation coefficient measures the statistical correlation, between two signals. Thus, we assume that all the computed scores with a given evaluation approach constitute a signal. Then, we compare obtained averaged *Rouge-1* and *PCA/ISOMAP-VSMbM* scores when using both human-made and automatically generated summaries [17] [18]. Results of the described above experiments are reported in Table 1 and Table 2.

Table 1: Pearson's correlation between *VSMbM* and *ROUGE* scores.

	ROUGE-1	**ROUGE-2**	**ROUGE-S**
PCA-VSMbM	0.79	0.88	0.89
ISOMAP-VSMbM	0.81	0.89	0.91

Table 2: Average*ROUGE-1*, *ISOMAP-VSMbM* and *PCA-VSMbM* scores when using handmade summaries and automatically made ones by MEAD and ETS summarizers.

	MEAD	**ETS**	**Human ms**
ROUGE-1	0.207	0.206	0.211
ISOMAP-VSMbM	0.204	0.205	0.205
PCA-VSMbM	0.189	0.201	0.203

Obtained results in Table 1 show that the *VSMbM* scores are highly positively correlated to the *ROUGE* scores. Indeed, the proposed metric can give a high score when the ROUGE protocol for summary evaluation does. It gives a low score in the inverse case. Also, the *ISOMAP-VSMbM* outperforms *PCA-VSMbM*. Indeed, when using the *PCA-VSMbM*, we assume that we are dealing with a linear dimensional reduction problem (which is not totally true regarding the high dimensionality) and we use Euclidian distance. Meanwhile, with the *ISOMAP-VSMbM*, we use the geodesic distance since we assume that we are dealing with a nonlinear dimensionality reduction problem. Results of Table 2 lead to the same conclusions when using both human-made and automatically generated summaries. Note that, the *VSMbM* protocol do not only check

whether the generated summary reports salient topics present in the original text or not. It also gives insights on to which extent fidelity is met by focusing on the most important ones.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a new metric for automatically generated text summaries evaluation. The proposed metric is based on vector space modelling. It gives insights on to which extent retention and fidelity are met. Conducted experiments on the Timeline17 dataset show that scores of the proposed metric are highly positively correlated to those produced by *ROUGE*: the standard metric for *ATS* evaluation. To deal with the decomposition problem of *K* in equation 11, we are currently implementing a Locally Linear Embedding version of our *VSMbM* metric *(LLE-VSMbM*). Next, we will test our metric with bigger size and multilingual corpora, and we will compare its performance to more *ATS* evaluation metrics.

ACKNOWLEDGEMENTS

The authors would like to thank **N**atural **S**ciences and **E**ngineering **R**esearch **C**ouncil (**NSERC**) of Canada for financing this work.

REFERENCES

[1] Gambhir, Mahak; Gupta, Vishal. Recent automatic text summarization techniques: a survey, TheArtificial Intelligence Review; Dordrecht Vol. 47, N 1: 166. DOI:10.1007/s10462-016-9475-9, 2017

[2] Torres-Moreno, Juan-Manuel, Automatic Text Summarization, London, Wiley 2014

[3] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, "Summac: a text summarization evaluation," Natural Language Engineering, vol. 8, no. 1, pp. 43–68, 20028

[4] P. Over, H. Dang, and D. Harman, "DUC in context," IPM, vol. 43, no. 6, pp. 1506–1520, 2007.

[5] Proceedings of the Text Analysis Conference. Gaithesburg, Maryland, USA: NIST, November 17-19, 2008.

[6] K. Sparck Jones and J. Galliers, Evaluating Natural Language Processing Systems, An Analysis and Review, ser. Lecture Notes in Computer Science. Springer, 1996, vol. 1083.

[7] Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

[8] Lin, Chin-Yew and E.H. Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In Proceedings of Language Technology Conference (HLT-NAACL), Edmonton, Canada, 2003.

[9] Lin, Chin-Yew and Franz Josef Ochs. 2004a. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip Bigram Statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, July 21 - 26, 2004.

[10] Lin, Chin-Yew and Franz Josef Ochs. 2004a. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, July 21 - 26, 2004.

[11] P. Over, H. Dang, and D. Harman, "DUC in context," IPM, vol. 43, no. 6, pp. 1506–1520, 2007.

[12] A. Nenkova and R. J. Passonneau, "Evaluating Content Selection in Summarization: The Pyramid Method," in HLT-NAACL, 2004, pp. 145–152.

[13] Juan Manuel Torres Moreno, Horacio Saggion, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-Morales, Summary Evaluation with and without References, Polibits (42), 2010

[14] Alan Ramırez-Noriega, Reyes Juarez-Ramırez Samantha Jimenez, Sergio Inzunza, Ashur: Evaluation of the relation summary-content without human reference using rouge, Computing and Informatics, Vol. 37, 509–532, doi: 10.4149/cai 2018 2 509, 2018

[15] Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. Lang. Resour. Eval. 52, 1, 101-148. DOI: https://doi.org/10.1007/s10579-017-9399-2, March 2018

[16] Tran G. B., Tran T.A., Tran N.K.,Alrifai M. and Kanhabua N.: Leverage Learning to rank in an optimization framework for timeline summarization. In TAIA workshop, SIGIR 13, 2013

[17] D. R. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. ÃGelebi, S. Dimitrov, E. DrÃabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. Mead - a platform for multidocument multilingual text summarization. In Proceedings of LREC'04, 2004

[18] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In Proceedings of SIGIR'11, pages 745–754, 2011.

AUTHORS

Alaidine Ben Ayed is a PhD. candidate in cognitive computer science atUniversité du Québec à Montréal (UQAM), Canada. His research focuses onartificial intelligence, natural language processing (Text summarization andconceptual analysis) and information retrieval.

Ismaïl Biskri is a professor in the Department of Mathematics and ComputerScience at the Université du Québec à Trois-Rivières (UQTR), Canada. Hisresearch focuses mainly on artificial intelligence, computational linguistics,combinatory logic, natural language processing and information retrieval

Jean Guy Meunier PhD. is a research professor at UQAM,co-director of the Cognitive Information Analysis Laboratory (LANCI), member of the Institute of Cognitive Sciences at UQAM and member ofthe Centre for Research in Digital Humanities (CRHN), full member ofthe International Academy of Philosophy of Science (Brussels).