VECTOR SPACE MODELING BASED EVALUATION OF AUTOMATICALLY GENERATED TEXT SUMMARIES

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

Evaluating automatically generated summaries is not an effortless task. Despite the fact that significant advances have been made in this context during the last two decades, it still remains a challenging resaerch problem. 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. Three variants of the proposed metric, namely PCA-VSMbM, ISOMAP-VSMbM and tSNE-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 ones.

KEYWORDS

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

1. INTRODUCTION

1.1. Automatic Text Summarization

Abstracts are common, and their use has been adopted to the daily running of affairs. According to [1], paper abstracts, book reviews, headlines on TV news, movie trailers and shopping guidelines on online stores are some of the examples of summaries that we have to interact with on a daily basis. A summary has commonly been defined as ‘a text produced from one or more texts with an intention of passing on key information from the original script and is usually less than the original version’ [2]. Notwithstanding the use of the word ‘text’, summaries too apply to other forms of media including audio, hypertext and video. The special case of *Automatic text summarization* (ATS) refers the process of creating a short, accurate, and fluent summary from a longer source text [3].

Following developments in technology, huge amounts of text resources are available at any one’s discretion. This calls for automatic text summarization so that users can access only relevant information they are looking for. [4] argues that automatic summarization has issues worth of address despite having been around for more than five decades. Also, it identifies six main justifications why we need automatic text summarization. The first reason is that summaries reduce the amount of time that one would have spent reading a longer document. They make it possible to consume content faster and more effectively. Second, automatic summaries make the

DOI: 10.5121/ijnlc.2020.9303

selection process easier when researching documents. Automatic summarization can also make the process of indexing text more effective. Next, these approaches also make it possible to prepare summaries that are fairer compared to those prepared by humans. Summaries generated automatically contain a lot of personalized information, which can be a useful addition to question-answering systems. Lastly, we need these processes to increase the number of texts that can be processed by commercial abstract services.

There isn't any indicated scientific classification or arrangement of summary types. Indeed, the arrangement of the types of summary changes is dependent on the angle of perception. [1], introduced nine parameters to find out the various classifications of summaries. One is a parameter based on relationship, and in this case, summary can be considered as either an extract or an abstract. Extractive summation here implies that the most significant parts are combined together from the original text minus any modification to the text selected. On the other hand, abstractive summarizations imply that the significant issues in the original format are paraphrased and presented in a grammatical way to produce a summary that is more coherent. Additionally, considering the readership parameters, the process of summarization can lead to production of generic summaries that is if it depends on the original documents which might have been produced from query driven summaries and this has an interest on getting information that is related to the query. Then there is a span parameter that categorizes the summarization process into one document from a number of documents. Language is one parameter that is considered very important. The language parameter is divided into the monolingual parameter which summarizes documents presented in one language and multi-lingual of cross lingual which presents a summary of texts presented in more than one language.

[5] have pointed out key challenges associated with automatically generated summaries evaluation which is an open subject in text summarization.In the next two section, we make a short state of the art of most relevant proposed evaluation protocols for automatically generated text summarization and 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 [6], DUC (Document Understanding Conference) [7] and TAC (Text Analysis Conference) [8] 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 [9]. 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) [10]. Note that there are five variants of the ROUGE metric: 1) ROUGE-N [11]: it captures the overlap of N-grams between the system and reference summaries, 2) ROUGE-L [12]: 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 [14]. RESPONSIVENESS has also been used in focused-based summarization tasks of DUC and TAC evaluation campaigns [14]. 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 [15]. 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 [16]. 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, [17] 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. [18] gives an overview of challenging issues related to summary evaluation

1.3. Originality of our work

Aautomatically generated summaries should satisfy three criteria: 1) Retention: It is a measure of how much the generated summary reports salient topics present in the original text, 2) Fidelity: Does the summary accurately reflect the author's point of view? and, 3) Coherence: To which extent, the generated extract is semantically meaningful?

Most of the described metrics in the above sub-section 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 given 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 protocol 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 three variants of the proposed *VSMbM* metric namely *PCA-VSMbM*,*ISOMAP-VSMbM* and*tSNE-VSMbM.*Mathematical and implementation details of the proposed metric will be expanded in the coming three 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 × 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:

$$\tau = \frac{1}{m}\sum_{i=1}^{m} \zeta_i \qquad (1)$$

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

$$\Theta_i = \zeta_i - \tau \qquad (2)$$

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

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

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:

$$A = U.S.V^T \qquad (4)$$

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 eigenvectorsAA^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:

$$\Xi_k = [U_1, U_2, \ldots, U_k] \qquad (5)$$

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

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

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 conceptτ : $\Theta_q = \zeta_i - \tau$ .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:

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

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 best one to encode the same concept in a window of four sentences.

$$\begin{array}{cc} & \xrightarrow{\text{Distance (concept)}} \\ \text{Concept importance} \left|\begin{array}{c}1\\2\\3\\4\\5\end{array}\right.\!\!\downarrow & \begin{pmatrix} 1 & 6 & 9 & 8 \\ 1 & 22 & 13 & 11 \\ 10 & 22 & 6 & 1 \\ 22 & 2 & 1 & 11 \\ 9 & 11 & 2 & 8 \end{pmatrix} \end{array}$$

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 $d_i(\Theta_q) = \|\Theta_q - \Theta_i^{proj}\|$ 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 $e = [1,1,\dots,1]^T$. 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.

2.2. The *tSNE-VSMbM*

At the begenning, we proceed the same way as *PCA-VSMBM* to constructthe set of $\zeta_1, \zeta_2, \dots, \zeta_m$ feature vectors decribing the$m$ sentences of the text to summarize. Then, we construct a feature matrix whose lines are made up by the ζ_i feature vectors ($1 \leqslant i \leqslant m$). columns of the feature matrix are$x_1, x_2, \dots, x_Z$ where x_i is a word feature vector and Z is the number of unique words used in the the text.*tSNE-VSMBM* first computes probabilities P_{ij} that are proportional to the similarity of wordsX_i and $X_j$for $i \neq j$ as follows:

$$P_{j|i} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2{\sigma_i}^2)}{\sum_{i \neq k} \exp(-\|x_i - x_k\|^2 / 2{\sigma_i}^2)} \tag{12}$$

Note that the similarity of word$x_j$to wordx_i is the conditional probability $P_{j|i}$, that, word $x_j$would be among word x_i 's neighbours if neighbors were chosen based on their probability density under a Gaussian distribution centered at x_i[19].

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2Z} \tag{13}$$

Moreover, the probabilities with $i = j$ are set to zero ($P_{ii} = 0$). The bisection approach is used to set the bandwidth of the Gaussian kernels $\sigma_i$thus and thus the perplexity of the conditional distribution equals a predefined perplexity. Therefore, the bandwidth is adapted to the density of the word feature vectors: In other words, smaller values of Gaussian kernels$\sigma_i$are used in denser parts of the word feature vectors space.

Note that the Gaussian kernel is highly sensitive to dimensionality since it uses the Euclidian distance. It means that the $P_{ij}s$would asymptotically converge to a constant when we deal with long texts. In other words, they become similar.Thus, a power transform, based on the intrinsic dimension of each word feature vector is used to adjust the distances[19].

*tSNE-VSMBM*isbased on the t-distributed stochastic neighbor embedding technique to construct the conceptual space of equation 5. The latter approachconstructs a $d$dimensional map $y_1, y_2, \dots, y_N$ (with$y_i \in R^d$) that reflects perfectly the similarities p_{ij} by measuring similarities q_{ij} between two word feature vectors in the map y_i and $y_j$for $i \neq j$, as follows:

$$q_{ij} = \frac{(1 + |y_i - y_j|^2)^{-1}}{\sum_{k \neq i}(1 + |y_i - y_k|^2)^{-1}} \tag{14}$$

For $i = j, q_{ij} = 0$. In order to allow dissimilar word feature vectors to be modeled far apart in the map, a Cauchy distribution (a kind of Student t-distribution with one-degree of freedom) is used to measure similarities between low-dimensional word feature vectors. Thus, locations of word feature vectorsy_i in the map are obtained by minimizing the Kullback–Leibler divergence of the distribution$Q$from the distribution$P$ as follows:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log(\frac{p_{ij}}{q_{ij}}) \tag{15}$$

The gradient descent approach is used tominimizee the aboveKullback–Leibler divergence. The result of this optimization is a map that reflects the similarities between the high-dimensional word feature vectors. Now, constructed y_i vectors will be set as unitary vectors of the Ξ_k conceptual space of equation 5.

Once, the conceptuel space is constructed,we proceed the same way we proceeded previously with *PCA-VSMbM*. *tSNE-VSMbM* is defined as *PCA-VSMbM* in equation 10.

3. Experiments and Results

3.1. Dataset

The *Timeline17* dataset is used for experiments [20]. 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 inplain 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/t-SNE-VSMbM* scores when using both human-made and automatically generated summaries [21] [22]. Results of the described above experiments are reported in Table 1 and Table 2.

	ROUGE-1	ROUGE-2	ROUGE-S
PCA-VSMbM	0.79	0.88	0.89
ISOMAP-VSMbM	0.81	0.89	0.91
tSNE-VSMBM	0.82	0.89	0.93

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

	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
tSNE-VSMBM	0.190	0.202	0.202

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

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* and tSNE-VSMBMoutperform *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). Also, note that the Euclidian distance used by the *PCA-VSMBM* approach is very sensitive to dimensionality. In other words, the performance decreases when we deal with long texts in this case.Meanwhile, the*ISOMAP-VSMbM* uses the geodesic distance assuming that we are dealing with a nonlinear dimensionality reduction problem which is the case. Also tSNE performs well when dimension of the target space is between 2 and 4 (when we deeal with two to four enngencocepts). 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.

Previously proposed approaches for automatically generated text evaluation only focus on the overlap of N-grams between the original text and the generated output. In other words, they reflect the coverage ratio meanwhile they don't give insights on to which extent fidelity is met. Our proposed approach gives insights on to whichextent boath 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. Currently we are implementing a Locally Linear Embedding [23] version of our *VSMbM* metric *(LLE-VSMbM*) to deal with the decomposition problem of *K* in equation 11. Next, we will test our metric with bigger size and multilingual corpora, and we will compare its performance to more automatically generated summaries evaluation frameworks.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Mani I, Automatic summarization, vol 3. John Benjamins Publishing, Amsterdam,2001

[2] Radev DR, Hovy E, McKeown K (2002) Introduction to the special issue on summarization. Comput Linguist28(4):399–408

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

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

[5] Saggion H, Poibeau T (2013) Automatic text summarization: Past, present and future. In: Poibeau T, SaggionH, Piskorski J, Yangarber R (eds) Multi-source, multilingual information extraction and summarization,theory and applications of natural language processing. Springer, Berlin, pp 3–21

[6] 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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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

[18] 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

[19] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008

[20] 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

[21] 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

[22] 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.

[23] Saul, Lawrence & Roweis, Sam. An introduction to locally linear embedding. Journal of Machine Learning Research. 7, 2001

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).