SYNC—Short, Yet Novel Concise Natural Language Description: Generating a Short Story Sequence of Album Images Using Multimodal Network

M. S. Karthika Devi, Shahin Fathima and R. Baskaran

Abstract Image captioning, which aims at generating automated descriptions for an image, is the large focus in current research while most of the previous works have dealt with the association between the single image and single sentences. This paper proposes to take one step further to investigate the summarized version of the narrative description for the image stream and in more generalized form for a normal user. The major challenge in the proposed work is to consider the visual variance in an ordered image collection and in preserving coherence relation among multiple sentences. Our proposed work is aimed to retrieve a coherent flow of multiple sentences that use multimodal neural architecture and ranking-based summarization to generate the summarized description of possibly larger image streams. With qualitative evaluation, the proposed work has attained significant performance improvement over traditional state-of-the-art method for text sequence generation and has captured the relevant context with syntactic meaning with respect to summarized version of the detailed descriptions.

Keywords Image captioning · Bidirectional recurrent neural networks · Convolutional neural networks · Coherence model · Summarization deep learning · Computer vision · Artificial intelligence

1 Introduction

A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene. However, this remarkable ability has proven to be an elusive task of content analysis for visual recognition models.

M. S. Karthika Devi · S. Fathima · R. Baskaran (✉)
Department of Computer Science and Engineering, College of Engineering, Anna University, Chennai, Tamil Nadu, India
e-mail: baaski@annauniv.edu

M. S. Karthika Devi
e-mail: karthikadevi88@gmail.com

S. Fathima
e-mail: shahinfathi1995@gmail.com

S. Fong et al. (eds.), *ICT Analysis and Applications*, Lecture Notes in Networks and Systems 93, https://doi.org/10.1007/978-981-15-0630-7_23

Automatic description of an image is a very challenging task as the system must capture not only what is contained in an image but also how the objects in an image are related to each other, what actions are involved in and how the changes occurred over one image to other are captured. This task is harder than tradition image or object recognition models, as it requires language model to express the semantic knowledge along with the visual understanding.

Visual contents are generally represented by images or videos which is associated with captions or tags as text sentences. This provides the way to learn the representation of multiple data types. The multimodal neural architecture that is used for implementation consists of convolutional neural networks for image descriptions, bidirectional recurrent neural network to represent the content flow of text sentences.

A key requirement for any machine translation system that produces natural language sentences is the coherence of its output. Coherence relations must able to capture the relatedness between the texts with respect to sentence transitions. Local coherence is necessary for global coherence that automatically abstracts a text into a set of entity transition sequences and records distributional, syntactic and referential information about discourse entities, i.e. entity grid representation of discourse, which captures pattern of entity distribution in a text.

General users usually take a large number of images and description of each of the images would possibly create detailed descriptions like paragraphs which could be too long for users. The summarized description must be short, able to preserve the coherent content flow and should be in accordance with the story sequences. However, the shortened story over machine-translated natural language sentences have not been explored in the earlier works.

This proposed work could have a great impact, for instance, helping novice users better understand the context of an image available in a web and can be used for various applications such as remembering the memories of life, recognition of action [1] over a series of events, digital storytelling and many such sequential or historical events through stream of images.

Organization of this paper is as follows. Section 2 presents a literature survey of the work. Section 3 presents the detailed design. Section 4 deals with experimental setup, and Sect. 5 presents the result analysis.

2 Related Works

In recent years, there has been a growing interest in exploring the relation between images and language. Simultaneous progress in the fields of Computer Vision (CV) and Natural Language Processing (NLP) [2–4] has led to impressive results in learning both image-to-text and text-to-image connections. Tasks such as automatic image captioning, image retrieval or image generation from sentences have shown good results.

Image Captioning: Image captioning method is paying drastic attention in the field of computer vision and machine learning community which aims to generate text sentences that describes an input image. Most popular approach of text generation is to retrieve best sentences by learning from training a system through embedding between images and text [3, 5]. This work involves retrieval of text from the learned model and generates story based on compatibility score between the language model and the coherence model.

Many earlier research attempts have exploited multimodal networks that combine deep Convolution Neural Networks (CNN) [6, 7] for image description and Recurrent Neural Networks (RNN) [6, 8–10] for language sequence modelling. There are many variants of combinations used as multimodal architecture includes CNN with bidirectional RNN [11], long-term recurrent convolution networks [6], long-short term memory networks [7]. This work takes an advantage of existing models with distinctive extension to multiple dimensions of input and output.

Huang et al. [12] introduced the first dataset of sequential images with corresponding description. They first collect storyable photo albums from Flickr and then outsourced to crowd workers using Amazon’s Mechanical Turk (AMT) to collect the corresponding stories and descriptions.

Retrieval of Image: Most of the existing work involves retrieval of image by keyword for the structured queries. Few earlier works include image ranking and retrieval based on text sentences, multiple attributes and other data structured objects like graphs [13]. In [14], three different data types such as image, text and sketch are combined as a query for image retrieval. Lin et al. [15] proposed a method for video search using a text sentence as a query. Hu et al. [16] retrieve a natural language object that takes an image and associated texts, for each text query that corresponds to the image contouring. Kong et al. [17] took a scene and a sentence as an input to find the relatedness between its regions and text phrases. Similarly in [18], the correspondence between regions to phrases is computed. This work is distinct as it involves images sequences, instead of single image.

Entity-Based Approaches to Local Coherence: Substance based records of neighbourhood intelligence include a long custom inside the phonetic and subjective science writing. Element-based portrayal [19] of talk permits learning the properties of lucid writings from a corpus, without plan of action to manual explanation or a predefined information base. Entity-based theories capture coherence by characterizing the distribution of entities across discourse utterances, distinguishing between salient entities and the other texts.

Text Summarization: Text summarization is an approach that uses Natural Language Processing principles and algorithms to understand the larger text [20] and generate smaller and efficient summaries. Our approach uses certain linguistic elements [21] to identify the most relevant segments of a text and must be able to capture the syntactic and coherent flow of the generated narrative descriptions while reducing it to precise representation of the text.

3 Proposed Architecture

Image captioning technique applied over sequence of images requires learning coherent meaning and the summarization technique aimed at generating concise description of an image. This provides the way to investigate multimodal architecture which has been shown in Fig. 1.

Figure 1 majorly divided into four different components: Convolutional Neural Network (CNN) used for describing an image, Bidirectional Recurrent Neural Network (BRNN) for language modelling, the local coherence learning to capture the smooth flow of sentences, and rank-based summarization technique to produce the crisp story of an image sequences.

3.1 *Text Descriptions*

The text sentences associated with an image are represented in two ways: paragraph vector to represent the text features and parse tree to represent the grammatical roles of the text sentences.

3.2 *Bidirectional Recurrent Neural Network*

The BRNN model is used to represent a content flow of text sequences. This bidirectional model helps to consider the previous and next text sentences while modelling forward and backward processing.

Initialize the weights $W_i^{f_c} \psi W_i^{b_c} \psi W_{f^c} \psi W_{b^c} \psi W_{o\wp\psi}$ and bias $\Leftarrow\leftarrow b_i^{f_c} \psi b_i^{b_c} \psi b_{f^c} \psi b_{b^c} \psi b_o \wp$.

For each paragraph vector p_t, set the activation function f to the Rectified Linear Unit (ReLU)

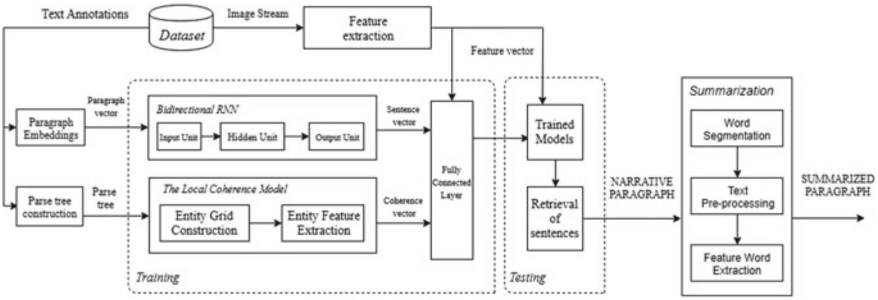

Fig. 1 Overall architecture of the proposed system

$$\mathrm{f(x)} = \max(0, \mathrm{x}) \tag{1}$$

Compute the activation of input units to the forward units $\left(\mathrm{x}_\mathrm{t}^\mathrm{f}\right)$

$$\mathrm{x}_\mathrm{t}^\mathrm{f} = \mathrm{f}\left(\mathrm{W}_\mathrm{i}^\mathrm{f}\mathrm{p}_\mathrm{t} + \mathrm{b}_\mathrm{i}^\mathrm{f}\right) \tag{2}$$

Compute the activation of input units to the backward units $\left(\mathrm{x}_\mathrm{t}^\mathrm{b}\right)$

$$\mathrm{x}_\mathrm{t}^\mathrm{b} = \mathrm{f}\left(\mathrm{W}_\mathrm{i}^\mathrm{b}\mathrm{p}_\mathrm{t} + \mathrm{b}_\mathrm{i}^\mathrm{b}\right) \tag{3}$$

Compute the activation of forward hidden units $\left(\mathrm{h}_\mathrm{t}^\mathrm{f}\right)$

$$\mathrm{h}_\mathrm{t}^\mathrm{f} = \mathrm{f}\left(\mathrm{x}_\mathrm{t}^\mathrm{f} + \mathrm{W}_\mathrm{f}\mathrm{h}_{\mathrm{t}-1}^\mathrm{f} + \mathrm{b}_\mathrm{f}\right) \tag{4}$$

Compute the activation of backward hidden units (htb)

$$\mathrm{htb} = \mathrm{f}\ (\mathrm{xtb} + \mathrm{Wb\ ht} - 1\mathrm{b} + \mathrm{bb}) \tag{5}$$

Compute the final activation of BRNN—output unit (ot)

$$\mathrm{o}_\mathrm{t} = \mathrm{W}_\mathrm{o}\left(\mathrm{h}_\mathrm{t}^\mathrm{f} + \mathrm{h}_\mathrm{t}^\mathrm{b}\right) + \mathrm{b}_\mathrm{o} \tag{6}$$

3.3 The Local Coherence Model

To learn the coherence among the texts, this work includes a local coherence model. The sequenced parse trees are concatenated, from which an entity grid for the whole sequence is represented. Each text is represented by an entity grid, a two-dimensional array where each row of the grid corresponds to sentences, while the column corresponds to discourse entities. This representation helps to capture the distribution of discourse entities across text sentences.

Each grid column thus corresponds to a string from a set of categories reflecting the entity's presence or absence in a sequence of sentences. Our set consists of four symbols: S (subject), O (object), X (neither subject nor object) and—(gap which signals the entity's absence from a given sentence). It first identifies the entity classes, fills out the grid entries with relevant syntactic information and then determines the constituent structure for each sentence, from which the syntactic roles are identified.

An entity transition is a sequence $\{\mathrm{S}, \mathrm{O}, \mathrm{X}, -\}^\mathrm{n}$ that represents entity occurrences and their syntactic roles in n adjacent sentences. Local transitions can be easily obtained from a grid as continuous subsequences of each column. After entity grid construction, entity transition is enumerated and the ratio of the occurrence frequency of each transition is calculated.

Zero padded coherence representation is forwarded as input to Rectifier Linear Unit (ReLU) which would output the vector of the coherence model (q).

3.4 Multimodal Network

The outputs of BRNN $\{o_t\}_{t=1 \text{ to } N}$ and the coherence model (q) are given together as a input to two fully connected layers to decide proper language and coherence match. Dropout rates and the dimensions of the variables are set accordingly.

$$W_{f2}W_{f1}\left[O|q\right] = \left[S|g\right] \tag{7}$$

where $O = [o_1|\, o_2| \cdots\, o_N];\ S = [s_1|\, s_2| \cdots\, s_N]$.

3.5 Training and Retrieval of Sentences

To train the model, define the compatibility score between an image comprising an album and the corresponding text sequence. The algorithm considers corresponding score between sentence and image of all possible combinations to find out the best matching.

Retrieval of best sentence sequence for a given query image stream is as follows:

1. Select the k-nearest images for each query image from training database using Euclidean distance on the image features.
2. The sentences associated with k-nearest images at location are concatenated as a paragraph sentences. This represents the candidate sentences.
3. Compatibility score between an image stream and a paragraph sequence is computed based on the following method:

    a. The ordered and paired compatibility score between a sentence sequence and an image sequence are defined as:

$$S_t^k * V_t^l. \tag{8}$$

    b. The coherence relevance relation between an image sequence and a textsequence are defined as:

$$G^{k*}V_t^l. \tag{9}$$

c. The score $\mathrm{S_{kl}}$ for a sentence sequence k and an image stream l are defined as:

$$\mathrm{S_{kl}} = \sum_{t=1,..N} \left(\mathrm{S_t^k} * \mathrm{V_t^l}\right) + \left(\mathrm{G^k} * \mathrm{V_t^l}\right) \tag{10}$$

where $\mathrm{V_t^l}$ denotes the 4096-dimensional CNN feature vector for tth image of stream l, and $\mathrm{G^k}$ and $\mathrm{S_t^k}$ are the output of Eq. (2.7) for a sentence sequence k.

d. The cost function to train the model are defined as follows:

$$\begin{aligned} \mathrm{C}(\theta) = \sum_{k}\Big[\sum_{l} \max(0, 1 + \mathrm{S_{kl}} - \mathrm{S_{kk}}) + \\ \sum_{l} \max(0, 1 + \mathrm{S_{lk}} - \mathrm{S_{kk}})\Big] \end{aligned} \tag{11}$$

3.6 Text Summarization

PageRank algorithm which is used for text summarization is known as Text Rank. It is an unsupervised method for computing the extractive summary of a text. PageRank algorithm is applied over sentence graph, where the graph is symmetrical. The algorithm then built the PageRank transition by building the sentence similarity.

1. Preprocess the text: It includes removing stop words and stemming the remaining words.
2. Create a graph where vertices are sentences.
3. Each sentence is connected by an edge. The weight of the edge is defined by the similarity of the two sentences.
4. Run the PageRank algorithm on the graph.
5. Pick the vertices which represent the sentences with the highest PageRank score.

4 Experiment

Dataset. The Visual Storytelling (VIST) is the first-ever dataset created particularly for sequential image-to-language. The dataset includes 81,743 unique photos in 20,211 sequences, aligned to descriptive and story language. The image streams are extracted from Flickr and the text stories are crowdsourced for written to Amazon Mechanical Turk (AMT).

4.1 Retrieval Task

For experimental evaluation, the dataset is split into 8-1-1 ratio as a training set validation set, and test set, respectively. Each input query image is represented as a query I_q and the corresponding text annotated sentences as groundtruth T_G. The algorithm retrieves the text sequences from training set for each input query album images that should match well with groundtruth sentences.

Given an input album image and the text sequences, an algorithm computes the compatibility score as in Eq. (11). The low-cost text sequence has given more priority and that is the best-matched sequence retrieved. The generation tasks for our approach are evaluated using quantitative measures. The proposed work performs in test set to produce narrative paragraph and the corresponding shortened story. This work exploits two metrics of language similarity (i.e. BLEU [22] and METEOR [23]) which are popularly used in text generation. A better performance is indicated by higher BLEU and METEOR values.

Figure 2 shows various examples of sentence sequences on VIST dataset. Three different stories are generated for each query image stream: *Image description* represents the single image context. *Narrative story* is generated based on other images comprising a query image sequence, and *Summarized story* is generated for the corresponding narrative story. The difference between single image description and the corresponding narrative story is the coherence among the sentences which are indicated by highlighted words.

5 Result and Discussion

The quantitative results of story generation are shown in Table 1. The methods involved in proposed work is partitioned into three groups: (i) image captioning corresponds to the implementation of Recurrent Convolutional Network (RCN), (ii) generation of narrative paragraphs corresponds to RCN and entity-based coherence model, and (iii) generation of summarized results.

Figure 3 clearly demonstrates that executing a coherent model over language modelling has a significant exhibition as for bleu score and has improved execution concerning meteor score, while having summarization capability has improved execution as for both bleu and meteor score.

The sequence of text annotated sentences for each test image sequence is represented as groundtruth $\mathrm{T_G}$, and the generated summarized stories are evaluated with reference to $\mathrm{T_G}$. Since the retrieval method for summarized story is based on the generated narrative story and the evaluations are performed with $\mathrm{T_G}$, the same has captured the coherent meaning and can only generate the similar sentences at best. This can be inferred from Fig. 3 that summarized story is most similar with $\mathrm{T_G}$ and from Fig. 2, the coherent meaning which are preserved from the image descriptions and generated narrative stories are shown by highlighted words.

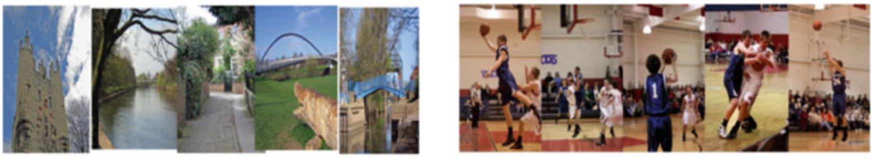

Image description

A castle is underneath a clear blue sky. The calm river runs underneath a suspension bridge. The cobblestone drive way up to the house was fitting. A wooden log shaped like an alligator with a bridge in the background. A river with blue bridge above it and someone is riding a bike across the bridge.

Narrative story

We went to visit a town that is centered around the castle. The river **leads straight to the castle.** Quaint streets lead to gracious neighbourhood and the streets looked like cobblestone. A local wit **carved this log into friendly crocodile into an open field;** it leads the way to **suspension bridge**. There are smaller walking bridges for pedestrians and we enjoyed the views of **blue skyand old town.**

Summarized story

A trip to an old town has river leads to the castle while cobblestone streets drive to the houses. There is a field finds a way to suspension bridge and there are small bridges for pedestrians to enjoy the view.

Image description

A basketball player is saving the play. A basket ball player is shooting the ball. A group of players in blue defending the player in white. A basket ball player wearing the blue shirt is passing the ball. A player steals the ball from the other player.

Narrative story

During the basketball game, the player **attempt to save the ball for the team** from being out of bounds. Then **the blue team player took a shot to score**. After that, he was guarded by the players in opposite team. Then the **blue team player passed the ball to his team**. The blue team tried to get the basketball back.

Summarized story

A blue team and white team **playing a basketball game** with players in **blue team took a shot to score** for the team.

Fig. 2 Examples of generated story on VIST dataset

Table 1 Evaluation of story sentence generation with language similarity metrics (BLEU and METEOR)

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR
Image captioning	0.140	0.027	0.000	0.000	0.026
Narrative story	0.106	0.027	0.000	0.000	0.046
Summarized story	0.162	0.161	0.160	0.159	0.153

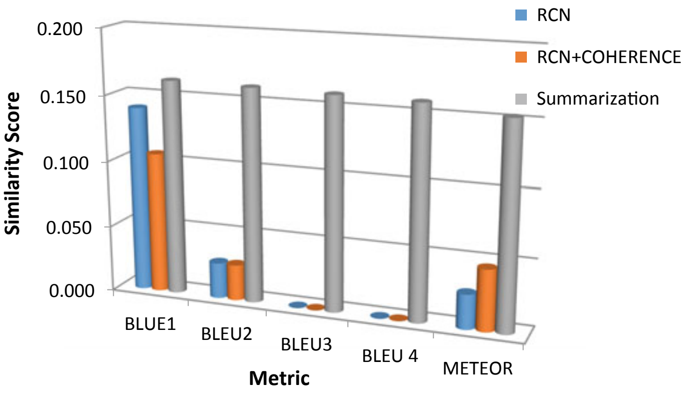

Fig. 3 Comparison of scores for narrative and summarized paragraph

6 Conclusion

Capturing the coherent meaning of a set of images is an important task for generating narrative paragraphs, instead of retrieving a text sentence associated with each image of an image set. Thus, the proposed work implemented a method for generating precise, yet concise story that best describes a sequence of images. With quantitative evaluation, this work demonstrates that generating summarized story from the narrative description has improved performance, and however, it preserves the context of an image set with syntactic and referential information.

Acknowledgements This Publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya Ph.D. Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

1. Gowsikhaa D, Abirami S, Ramachandran B (2014) Automated human behavior analysis from surveillance videos: a survey. Artif Intell Rev 42(4):747–765 https://doi.org/10.1007/s10462-012-9341-3
2. Park CC, Gunhee K (2015) Expressing an image stream with a sequence of natural sentences. In: Proceedings of the international conference on neural information processing systems, vol 1(NIPS 15), pp 73–81
3. Gowsikhaa D, Abirami S, Baskaran R (2014) Construction of image ontology using low-level features for image retrieval. In: International conference on computer communication and informatics, pp 1–7
4. Richard S, Karpathy A, Le QV, Manning CD, Ng AY (2014) Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comput Linguist 2 1:207–218 https://doi.org/10.1162/tacl_a_00177

5. Farhadi A, Hejrati M, Amin Sadeghi M, Young P, Rashtchian C, Hockenmaier J, Forsyth D (2010) Every picture tells a story: generating sentences from images. In: European conference on computer vision, pp. 15–29. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15561-1_2
6. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634. https://doi.org/10.1109/cvpr.2015.7298878
7. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164 https://doi.org/10.1109/cvpr.2015.7298935
8. Ramisa A, Yan F, Moreno-Noguer F, Mikolajczyk K (2018) Breakingnews: article annotation by image and text processing. IEEE Trans Pattern Anal Mach Intell 40(5):1072–1085 https://doi.org/10.1109/tpami.2017.2721945
9. Srivastava Nitish, Salakhutdinov Ruslan R (2012) Multimodal learning with deep boltzmann machines. Adv Neural Inf Process Syst 25:2222–2230
10. Feng Y, Lapata M (2013) Automatic caption generation for news images. IEEE Trans Pattern Anal Mach Intell 35, 4:797–812 https://doi.org/10.1109/tpami.2012.118
11. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137 https://doi.org/10.1109/cvpr.2015.7298932
12. Huang TK, Ferraro F, Mostafazadeh N., Misra I, Agrawal A, Devlin J, Girshick R, He X, Kohli P, Batra D, Zitnick CL (2016) 'Visual storytelling', North American Chapter of the association for computational linguistics: human language technology, pp 1233–1239 https://doi.org/10.18653/v1/n16-1147
13. Deborah LJ, Baskaran R, Kannan A (2010) A survey on internal validity measure for cluster validation. Int J Comput Sci Eng Surv 1(2):85–102. https://doi.org/10.5121/ijcses.2010.1207
14. Siddiquie B, White B, Sharma A, Davis LS (2014) Multi-modal image retrieval for complex queries using small codes. In: Proceedings of international conference on multimedia retrieval, pp 321. ACM https://doi.org/10.1145/2578726.2578767
15. Lin D, Fidler S, Kong C, Urtasun R (2014) Visual semantic search: retrieving videos via complex textual queries. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2657–2664 https://doi.org/10.1109/cvpr.2014.340
16. Hu R, Xu H, Rohrbach M, Feng J, Saenko K, Darrell T (2016) Natural language object retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4555–4564 https://doi.org/10.1109/cvpr.2016.493
17. Kong C, Lin D, Bansal M, Urtasun R, Fidler S (2014) What are you talking about? text-to-image coreference. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3558–3565 https://doi.org/10.1109/cvpr.2014.455
18. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S (2015) Flickr30 k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, pp 2641–2649 https://doi.org/10.1109/iccv.2015.303
19. Barzilay R, Lapata M (2008) Modeling local coherence: an entity-based approach. Comput Linguist 34 1:1–34 https://doi.org/10.1162/coli.2008.34.1.1
20. Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268
21. Mihalcea R, Tarau P (2004) TextRank: bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 404–411, Association for Computational Linguistics
22. Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp 311–318. Association for Computational Linguistics https://doi.org/10.3115/1073083.1073135
23. Lavie A, Denkowski MJ (2009) The METEOR metric for automatic evaluation of machine translation. Mach Trans 23, 2–3:105–115 https://doi.org/10.1007/s10590-009-9059-4