Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Scoring Summaries Using Recurrent Neural Networks

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Ruseti, Stefan	University Politehnica of Bucharest	
Dascalu, Mihai	University Politehnica of Bucharest	
Johnson, Amy M.	Arizona State University	
McNamara, Danielle S.	Arizona State University	
Balyan, Renu	Arizona State University	
McCarthy, Kathryn S.	Arizona State University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 17 May 2018

Check type of content being submitted and complete one of the following in the box below:

- ☐ If article: Name of journal, volume, and issue number if available
- ☑ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

14th international conference on Intelligent Tutoring Systems, ITS 2018, June 11-15 2018, Montreal, QC, Cana

DOI or URL to published work (if available) https://doi.org/10.1007/978-3-319-91464-0_19

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** National Center for Educat through **[Grant number]** R305A1301 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** National Center for Educa or the U.S. Department of Education.

Scoring Summaries Using Recurrent Neural Networks

Stefan Ruseti[1], Mihai Dascalu[1,2,3](✉), Amy M. Johnson[4], Danielle S. McNamara[4], Renu Balyan[4], Kathryn S. McCarthy[4], and Stefan Trausan-Matu[1,2,3]

[1] University Politehnica of Bucharest, Splaiul Independenței 313, 60042 Bucharest, Romania
{stefan.ruseti,mihai.dascalu,stefan.trausan}@cs.pub.ro

[2] Academy of Romanian Scientists, Splaiul Independenţei 54, 050094 Bucharest, Romania

[3] Cognos Business Consulting S.R.L., Bd. Regina Maria 32, Bucharest, Romania

[4] Institute for the Science of Teaching and Learning, Arizona State University, PO Box 872111, Tempe, AZ 85287, USA
{amjohn43,dsmcnama,renu.balyan,ksmccar1}@asu.edu

Abstract. Summarization enhances comprehension and is considered an effective strategy to promote and enhance learning and deep understanding of texts. However, summarization is seldom implemented by teachers in classrooms because the manual evaluation requires a lot of effort and time. Although the need for automated support is stringent, there are only a few shallow systems available, most of which rely on basic word/n-gram overlaps. In this paper, we introduce a hybrid model that uses state-of-the-art recurrent neural networks and textual complexity indices to score summaries. Our best model achieves over 55% accuracy for a 3-way classification that measures the degree to which the main ideas from the original text are covered by the summary . Our experiments show that the writing style, represented by the textual complexity indices, together with the semantic content grasped within the summary are the best predictors, when combined. To the best of our knowledge, this is the first work of its kind that uses RNNs for scoring and evaluating summaries.

Keywords: Automated summary evaluation · Recurrent neural network
Semantic models · Word embeddings

1 Introduction

Summarization is an effective strategy to promote and enhance learning and deep understanding of the subject matter among students [1, 2]. Summarizing a text allows readers to differentiate between relevant and irrelevant information within texts, integrate content with pre-existing knowledge, allowing for both better retention of the text content [3], as well as deeper comprehension of the material [4]. Earlier studies have indicated that summary writing helps students retain new information [1]. Summary strategies are also effective for different types of learners including native speakers [5], language learners [6], students with learning disabilities [7] and students with low literacy skills [8]. A meta-analysis indicated that summarization

R. Nkambou et al. (Eds.): ITS 2018, LNCS 10858, pp. 191–201, 2018.
https://doi.org/10.1007/978-3-319-91464-0_19

enhanced comprehension in 18 out of 19 studies [9]. Further, summarization is particularly useful for lower-skilled readers [10].

Given the effectiveness of summarizing texts, our aim is to develop computer-based summarization strategy training and practice that parallels an existing implementation of self-explanation and comprehension strategy practice within the Interactive Strategy Training for Active Reading and Thinking (iSTART) [11]. iSTART was developed to train comprehension strategies that help students understand complex, informational texts. Previous research demonstrated the effectiveness of iSTART for middle school [11], high school [12, 13], and college students [14, 15]. Currently, iSTART includes lesson videos covering four summarization strategies (deletion, main ideas, replacement, and topic sentences; see [16]). The development of practice modules in which students practice writing and revising summaries in turn necessitates a Natural Language Processing (NLP) algorithm capable of scoring the quality of summaries. ITSs that leverage NLP can provide students immediate, individualized feedback on their constructed (i.e., written) responses. This feedback is indispensable to learners attempting to improve their literacy skills [17].

Although summarization practice has proven effectiveness, teachers can find it challenging to implement practice activities because evaluating student summaries requires a great deal of effort and time [18]. Automated methods for summary evaluation traditionally involve evaluating quality metrics such as readability, content, conciseness, coherence and grammar [19]. In recent years, the research community has been successful in developing various measures for evaluating summaries. Some of the automated summary evaluation tools include Recall-Oriented Understudy for Gisting Evaluation (ROUGE [20]), ParaEval, Summary Input similarity Metrics (SIMetrix [21], QARLA [22], and SEMantic similarity toolkit (SEMILAR [23]).

The purpose of this study is to investigate the use of one of the most recent machine-learning techniques – recurrent neural networks (RNNs) [24] for automated scoring of summaries. To the best of our knowledge, this is the first work of its kind that uses RNNs for scoring and evaluating summaries.

The next section describes existing solutions and approaches used in literature for automated summary evaluation, and general deep-learning methods. In Sect. 3, the corpus, scoring rubric, followed by the proposed solution along with a detailed architecture is discussed. Finally, we report the results and conclude with discussions and future scope of the work.

2 Related Work

Evaluation of summaries is generally classified as intrinsic or extrinsic [25]. Intrinsic evaluation measures the *text quality* of summaries assessed by human annotators for fluency, informativeness and coverage, or evaluates the *content* of the summary using cue-words, term-frequency and inverted document frequency, cohesion methods, and Latent Semantic Analysis (LSA) [26]. By contrast, extrinsic evaluation is mostly task based involving document categorization, question answering and information retrieval [27]. The work described here focuses on intrinsic summary evaluation. Some of the earliest works in intrinsic summary evaluation include evaluation of chemistry documents [28] and electronic news publications [29]. Both of the latter studies used small data sets of 200 to 250 documents for evaluation. However, some early research efforts in large-scale evaluation of text summarization include TIPSTER SUMMAC [30] and the Document Understanding Conference (DUC). Researchers contributing to DUC have claimed that at large scales, even simple manual summary evaluations of content coverage and linguistic traits (e.g., capitalization errors, incorrect word order, unrelated fragments joined into one sentence, unnecessarily repeated information, misplaced sentences) requires a few thousand hours of human efforts [31]. In addition, some studies [32–35] show that human evaluations can be unstable and inconsistent with low inter-annotator agreement.

2.1 Automated Summary Evaluation

Some initial efforts towards developing automated summary evaluation metrics used n-gram overlap [33, 36]. These studies were motivated by the machine translation evaluation metric BiLingual Evaluation Understudy (BLEU) [37]. ROUGE [20] is one of the first and most widely used recall-oriented metrics for summary evaluations. ROUGE compares inputted summaries with one or multiple human written gold-standard summaries. One of the disadvantages of ROUGE is that all n-grams are considered equally important when computing the final score. Hovy et al. [38] proposed another simple metric based on basic elements' overlap, which are represented by one or two words, depending on their syntactic role.

Saggion et al. [39] proposed three content-based similarity measures: *cosine similarity, unit overlap* (unigrams or bigrams), and *longest common subsequence* (LCS). However, they did not discuss how these measures correlated with human evaluation. Another novel semi-automated approach is the *pyramid method* [40] which identifies and compares expert summaries' content units (SCUs) with to-be-evaluated summaries.

Some researchers have used random indexing [41, 42], that reduces terms by considering synonyms, hence allowing greater variations in summaries. Others have used distribution-similarity measures such as Kullback–Leibler (KL) divergence and Jensen Shannon (JS) divergence [21, 43], textual entailment [44] and crowdsourcing based LSA [18] for evaluating summaries. However, relatively few studies have used machine-learning techniques for summary evaluation beyond the aforementioned regression-based approaches [45–47].

2.2 Deep Neural Networks and Summary Evaluation

A common architecture used for text representation consists of recurrent neural networks, in particular Long Short-Term Memory networks (LSTM) [48] and Gated Recurrent Unit (GRU) [49]. These networks are capable of "memorizing" information, thus being able to better represent longer segments of text, without the danger of vanishing/exploding gradients encountered in traditional, normal recurrent neural networks [50]. These types of networks have been successfully used in most NLP tasks [51].

Recurrent neural networks have been improved further by considering different networks for the forward and backward directions [52]. This is especially useful when dealing with long text segments, because not all words in the text will have the same weight (e.g., depending on the language, the ones at the end are in most cases more important than the ones at the beginning). When using two different networks, the output for each word is usually represented by the concatenation of the outputs from the two directions. This way, all the words in the text influence the output for a single word.

We could not find any work that uses deep-learning techniques such as RNNs in particular for evaluating and scoring summaries. As a result, in order to explore the performance and success of these latest techniques for summary scoring and evaluation, we performed several experiments using RNNs.

3 Method

3.1 Corpus Description

We collected a corpus of 636 summaries for 30 texts (range: 20–24 summaries per text) using the Amazon Mechanical Turk online research service. The 30 texts used for the summary corpus collection were attained from the California Distance Learning Project (CDLP)[1], with permission from the Sacramento County Office of Education. The CDLP texts are real, simplified news stories that can be used by low-literate adults to improve their comprehension skills. The texts cover life-relevant topics, such as health and safety, housing, family, and money. Each text was between four and eight paragraphs and ranged from 128 to 452 words ($SD = 73.9$ words). Flesch-Kincaid grade level was between 4th and 8th grade ($SD = 1.1$) for all texts. The participants read and summarized three texts, randomly selected from the full set of 30 texts. Most of the participants (210/214) completed the entire summary task, producing three summaries total, for three separate texts. However, summaries submitted by the four participants who did not complete the entire task were also included in the corpus.

3.2 Scoring Rubric

Two trained researchers scored the summaries in the corpus on two major dimensions: (a) main ideas and (b) accuracy of main ideas. Before applying the coding scheme, the researchers individually examined the original texts, identifying the main ideas from each. Through discussions, they finalized a list of main ideas for each text. During coding of the summaries, the trained coders referenced this list of main ideas. For the main ideas dimension, each summary was scored from 0 (none of the main ideas from the text are included in the summary) to 3 (all of the main ideas from the text are included in the summary). For the accuracy of main ideas dimension, each summary was scored from 0 (main ideas present in the summary are completely inaccurate, or no main ideas are present in the summary) to 3 (all the main ideas in the summary are accurate representation of the content from the text).

1 www.cdlponline.org.

Two trained raters scored the three dimensions for all 636 summaries. Inter-rater agreement for the Main Idea dimension was $kappa_{linear\ weighted} = .67$, $r = .78$, 71% exact agreement, and 99% adjacent agreement. Agreement for the Accuracy of Main Ideas dimension was $kappa_{linear\ weighted} = .44$, $r = .52$, 76% exact agreement, and 91% adjacent agreement. Differences between the ratings from the two researchers were resolved through discussions.

The distribution of the scores for main ideas and accuracy of main ideas is presented in Table 1. Due to the highly unbalanced distribution of the accuracy of main ideas dimension, it was not included in our follow-up experiments. Moreover, all 14 examples with a score of 0 for the main ideas dimension were ignored as there were not sufficient test cases in order to train a classifier.

Table 1. Distribution of output classes.

Score	No. of summaries	
	Main ideas	Accuracy of main ideas
0	14	28
1	165	22
2	255	61
3	202	525

3.3 Network Architecture

The network receives as input the summary and the original text, represented with pretrained Glove [53] word embeddings of size 100, ignoring words that were not part of the vocabulary. A BiGRU Siamese architecture (Fig. 1) was used to share network weights for the summary and the whole text. Max-pooling is performed on the forward-backward concatenated outputs from each cell. This results in two $2 * d$ vectors (where d is the size of the GRU cell), representing the summary and the text. These two vectors are concatenated ("concat" operator from Fig. 1) and passed through two fully-connected layers (FCN module from Fig. 1).

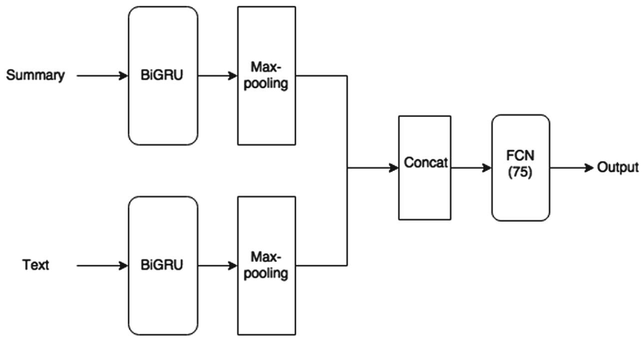

Fig. 1. Siamese recurrent network architecture.

The network produces a real number between 0 and 1, whereas our dataset has 3 output classes. Hence, we represented this task as a linear regression. To avoid force-fitting the network to the boundaries of this interval for the two extreme classes, the output score was multiplied by 4, resulting in a (0, 4) interval. When using the sigmoid activation function as output, it a good practice to avoid values close to 0 and 1 because the gradient for these flat regions is close to 0, making the training process difficult. Therefore, we split this domain and reassigned the predicted classes (i.e., 1, 2, and 3) as shown in Fig. 2, where all the three classes have almost equal range in the global interval.

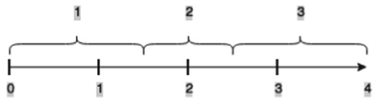

Fig. 2. Regression output.

As a baseline, we tested various complexity indices computed with the *ReaderBench* framework [54], which provides indices related to express the writing style of the text, instead of its content. From the available index categories, we extracted surface, syntax, word complexity, co-reference, connectives, cohesion, semantic dependencies, and word lists indices. Indices with low linguistic coverage (more than 20% of the values were missing) were removed and remaining indices were checked for multi-collinearity (Pearson $r \geq .9$). This cleaning process resulted in 191 features. These features were used to train two different models: a) individually within a 2-layered fully-connected network, and b) together within the recurrent network, as shown in Fig. 3. In both the cases, we tested two ways of using the complexity indices as input to the network: the difference between the two feature vectors (summaries and text) and the concatenation. Difference (marked as “diff” in Fig. 3) refers to the mathematic operator and is useful to highlight discrepancies between each feature or embedding dimension.

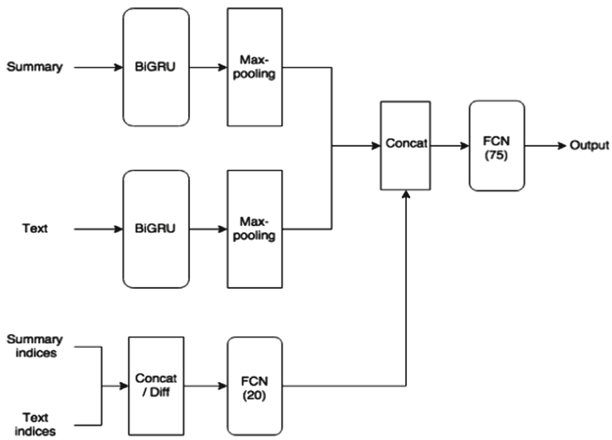

Fig. 3. Hybrid architecture of BiGRU, combined with *ReaderBench* textual complexity indices.

4 Results

As there were multiple summaries available for each text, the data were split into training and test sets (80–20). There were no common texts in the two partitions in order to avoid overfitting. The reported accuracies in Table 2 for each corresponding model were computed by averaging accuracy over three runs. The results in Table 2 indicate that the concatenation of feature vectors, despite needing more weights for the training process, works better and achieves better accuracy than the difference operator. This shows that important information is lost when the difference between the feature vectors is computed.

Table 2. Cell size and accuracy of models.

Model	Cell size	Accuracy (%)
Indices (difference)	-	41.90
Indices (concatenate)	-	37.14
Siamese	50	50.47
Siamese	100	50.15
Siamese + indices (difference)	50	47.30
Siamese + indices (difference)	100	53.34
Siamese + indices (concatenate)	**100**	**55.24**

In addition, we can observe from results in Table 2 that the complexity indices by themselves have the lowest accuracy, followed by the Siamese BiGRU network when used separately. The highest accuracy was obtained when combining the Siamese BiGRU network with the textual complexity indices from *ReaderBench*. This shows that both semantic features and writing style are important for summary evaluation.

5 Conclusions

This paper introduces a state-of-the-art model based on recurrent neural networks and textual complexity indices to evaluate and score summaries. To the best of our knowledge, this is the first work of its kind and the obtained accuracies of more than 55% is encouraging, given the size of the dataset. Moreover, our experiments show that the semantic content of the summary is more important than the writing style represented by the *Readerbench* textual complexity indices. However, replications with larger corpora should be conducted to support this conclusion.

Follow-up studies will also include an *attention* mechanism proven to be successful when comparing two or more text fragments by weighting the words with values computed based on the remainder of the text [55]. This mechanism is primarily used in question answering, but it can also be applied to summarization tasks by comparing the summary with the original text. However, the added weights may render the network too complex for this dataset, therefore reducing accuracy. In addition, the results might be improved by adjusting the hyper-parameters of the network using a grid-search method that performs cross-validations on the training set.

In sum, there are multiple ways in which this work can be validated and improved upon. However, this study demonstrates important promise in the use of recurrent neural networks to assess the quality of natural language.

Acknowledgment. This research was partially supported by the README project "Interactive and Innovative application for evaluating the readability of texts in Romanian Language and for improving users' writing styles", contract no. 114/15.09.2017, MySMIS 2014 code 119286, the 644187 EC H2020 RAGE project, the FP7 2008-212578 LTfLL project, the Department of Education, Institute of Education Sciences - Grant R305A130124, as well as the Department of Defense, Office of Naval Research - Grants N00014140343 and N000141712300.

References

1. Spirgel, A.S., Delaney, P.F.: Does writing summaries improve memory for text? Educ. Psychol. Rev. **28**, 171–196 (2016)
2. van Dijk, T.A., Kintsch, W.: Strategies of Discourse Comprehension. Academic Press, New York (1983)
3. Rinehart, S.D., Stahl, S.A., Erickson, L.G.: Some effects of summarization training on reading and studying. Read. Res. Q. **21**, 422–438 (1986)
4. Wade-Stein, D., Kintsch, E.: Summary Street: Interactive Computer Support for Writing (2004). http://www.tandfonline.com/doi/abs/10.1207/s1532690xci2203_3
5. Leopold, C., Sumfleth, E., Leutner, D.: Learning with summaries: effects of representation mode and type of learning activity on comprehension and transfer. Learn. Instr. **27**, 40–49 (2013)
6. Chiu, C.-H.: Enhancing reading comprehension and summarization abilities of EFL learners through online summarization practice. J. Lang. Teach. Learn. **5**(1), 79–95 (2015)
7. Rogevich, M.E., Perin, D.: Effects on science summarization of a reading comprehension intervention for adolescents with behavior and attention disorders. Except. Child. **74**, 135–154 (2008)
8. Perin, D., Lauterbach, M., Raufman, J., Kalamkarian, H.S.: Text-based writing of low-skilled postsecondary students: relation to comprehension, self-efficacy and teacher judgments. Read. Writ. **30**, 887–915 (2017)
9. Graham, S., Hebert, M.: Writing to read: a meta-analysis of the impact of writing and writing instruction on reading. Harv. Educ. Rev. **81**, 710–744 (2011)
10. Gil, L., Bråten, I., Vidal-Abarca, E., Strømsø, H.I.: Summary versus argument tasks when working with multiple documents: Which is better for whom? Contemp. Educ. Psychol. **35**, 157–173 (2010)
11. McNamara, D.S., O'Reilly, T., Rowe, M., Boonthum, C., Levinstein, I.: iSTART: a web-based tutor that teaches self-explanation and metacognitive reading strategies. In: Reading Comprehension Strategies: Theories, Interventions, and Technologies, pp. 397–420 (2007)
12. Jackson, G.T., McNamara, D.S.: Motivation and performance in a game-based intelligent tutoring system. J. Educ. Psychol. **105**, 1036–1049 (2013)
13. Snow, E.L., Jackson, G.T., McNamara, D.S.: Emergent behaviors in computer-based learning environments: computational signals of catching up. Comput. Hum. Behav. **41**, 62–70 (2014)
14. Magliano, J.P., Todaro, S., Millis, K., Wiemer-Hastings, K., Kim, H.J., McNamara, D.S.: Changes in reading strategies as a function of reading training: a comparison of live and computerized training. J. Educ. Comput. Res. **32**, 185–208 (2005)

15. O'Reilly, T., Sinclair, G.P., McNamara, D.S.: iSTART: A web-based reading strategy intervention that improves students' science comprehension. In: IADIS International Conference Cognition and Exploratory Learning in Digital Age, pp. 173–180 (2004)
16. Johnson, A.M., Guerrero, T.A., Tighe, E.L., McNamara, D.S.: iSTART-ALL: confronting adult low literacy with intelligent tutoring for reading comprehension. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 125–136. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_11
17. McNamara, D.S., Crossley, S.A., Roscoe, R.: Natural language processing in an intelligent writing strategy tutoring system. Behav. Res. Methods **45**, 499–515 (2013)
18. Li, H., Cai, Z., Graesser, A.C.: Computerized Summary Scoring: Crowdsourcing-Based Latent Semantic Analysis (2017). http://link.springer.com/10.3758/s13428-017-0982-7
19. Mani, I.: Automatic Summarization. John Benjamins Publishing, Amsterdam (2001)
20. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of Workshop Text Summarization Branches Out (WAS 2004), pp. 25–26 (2004)
21. Louis, A., Nenkova, A.: Automatically assessing machine summary content without a gold standard. Comput. Linguist. **39**, 267–300 (2013)
22. Amigó, E., Gonzalo, J., Penas, A., Verdejo, F.: QARLA: a framework for the evaluation of text summarization systems. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 280–289 (2005)
23. Rus, V., Lintean, M., Banjade, R., Niraula, N., Stefanescu, D.: SEMILAR: the semantic similarity toolkit. Assoc. Comput. Linguist. **2013**, 163–168 (2013)
24. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**, 179–211 (1990)
25. Spärck Jones, K., Galliers, J.R.: Evaluating Natural Language Processing Systems, An Analysis and Review. Springer Science & Business Media, Heidelberg (1996). https://doi.org/10.1007/BFb0027470
26. Steinberger, J., Jezek, K.: Evaluation measures for text summarization. Comput. Inf. **28**, 1001–1025 (2009)
27. Jing, H., Barzilay, R., McKeown, K.R., Elhadad, M.: Summarization evaluation methods: experiments and analysis. In: AAAI Symposium on Intelligent Summarization, pp. 51–59 (1998)
28. Edmundson, H.P.: New methods in automatic extracting. J. ACM **16**, 264–285 (1969)
29. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. Inf. Process. Manag. **31**, 675–685 (1995)
30. Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B.: The TIPSTER SUMMAC text summarization evaluation. In: 9th Conference on EACL, p. 77. Association for Computational Linguistics, Morristown (1999)
31. Over, P., Yen, J.: An Introduction to DUC-2003 Intrinsic Evaluation of Generic News Text Summarization Systems (2003). http://www-nlpir.nist.gov/projects/duc/pubs/2003slides/duc2003intro.pdf
32. Donaway, R.L., Drummey, K.W., Mather, L.A.: A comparison of rankings produced by summarization evaluation measures. In: NAACL-ANLP 2000 Workshop on Automatic summarization, pp. 69–78. Association for Computational Linguistics (2000)
33. Lin, C.-Y., Hovy, E.: Manual and automatic evaluation of summaries. In: Proceedings of ACL02 Workshop on Automatic Summarization, vol. 4, pp. 45–51 (2002)
34. Rath, G.J., Resnick, A., Savage, T.: The formation of abstracts by the selection of sentences. J. Am. Soc. Inf. Sci. Technol. **12**, 139–141 (1961)
35. van Halteren, H., Teufel, S.: Examining the consensus between human summaries. In: Proceedings of the HLT-NAACL 2003 on Text Summarization Workshop, pp. 57–64. Association for Computational Linguistics, Morristown (2003)

36. Lin, C.-Y., Hovy, E.: Automatic evaluation of summaries using N-gram co-occurrence statistics. In: Proceedings of 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2003, pp. 71–78 (2003)
37. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: ACL 2002, p. 311. Association for Computational Linguistics, Morristown (2001)
38. Hovy, E., Lin, C.-Y., Zhou, L., Fukumoto, J.: Automated summarization evaluation with basic elements. In: Proceedings of 5th International Conference on Language Resources and Evaluation, pp. 899–902 (2006)
39. Saggion, H., Radev, D., Teufel, S., Lam, W.: Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In: Proceedings of International Conference on Computational Linguistics, pp. 849–855 (2002)
40. Nenkova, A., Passonneau, R.: Evaluating content selection in summarization: the pyramid method. In: Proceedings of HLT-NAACL 2004, pp. 145–152 (2004)
41. Kanerva, P., Kristofersson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In: Proceedings of 22nd Annual Conference of the Cognitive Science Society, vol. 1036, pp. 16429–16429 (2000)
42. Sahlgren, M.: Vector-based semantic analysis: representing word meaning based on random labels. In: ESSLI Workshop on Semantic Knowledge Acquistion and Categorization (2002)
43. Lin, C.-Y., Cao, G., Gao, J., Nie, J.-Y.: An information-theoretic approach to automatic evaluation of summaries. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of ACL, pp. 463–470. Association for Computational Linguistics, Morristown (2006)
44. Bhaskar, P., Pakray, P.: Automatic evaluation of summary using textual entailment. In: RANLP 2013, pp. 30–37 (2013)
45. De, A., Kopparapu, S.K.: An unsupervised approach to automated selection of good essays. In: 2011 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2011, pp. 662–666. IEEE (2011)
46. Ellouze, S., Jaoua, M., Belguith, L.H.: Machine learning approach to evaluate multilingual summaries. In: Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, pp. 47–54 (2017)
47. Perez-breva, L., Yoshimi, O.: Model Selection in Summary Evaluation, pp. 0–12 (2002)
48. Hochreiter, S., Urgen Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
49. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014, pp. 1724–1734 (2014)
50. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**, 157–166 (1994)
51. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: EMNLP 2017 (2017)
52. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**, 602–610 (2005)
53. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)

54. Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I.C., Dessus, P., McNamara, D.S., Crossley, S.A., Trausan-Matu, S.: ReaderBench: a multi-lingual framework for analyzing text complexity. In: EC-TEL 2017, pp. 495–499 (2017)
55. Santos, C. dos, Tan, M., Xiang, B., Zhou, B.: Attentive Pooling Networks. CoRR, abs/1602.03609, no. 2, p. 4 (2016)