World Applied Sciences Journal 35 (9): 2074-2080, 2017
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2017.2074.2080

Automatic Hausa LanguageText Summarization Based on Feature Extraction using Naïve Bayes Model

Muazzam Bashir, Azilawati Rozaimee and Wan Malini Wan Isa

Faculty of Informatics & Computing, Universiti Sultan Zainal Abidin, Malaysia

Abstract: Automatic text summarization, a branch of natural language processing, is a technique where a computer summarizes a long text into non-redundant form in order to reduce the problem of information overload. Although there are some language independent summarizers, there is a limited research to automatically summarize text in Hausa language. Hausa, a Chadic language generally spoken in West Africa is a low resource language. This study is conducted to develop a model to automatically summarize Hausa Language text based on feature extraction using Naive Bayes model. A dataset of 10 Hausa Language documents is used in this study. The study adopts five features such as keyword, title and cue phrases in the summarization process. Moreover, Naive Bayes model is used to weigh each sentence based on its features. The system produces a set of summary sentences at 30% compression rate. Moreover, experiments are conducted to summarize the dataset using online summarizers such as Text Compactor and Free summarizers. The overall system testing having an average F-score of 78.1% outperforms the online summarizers. The result shows that automatic text summarization tested on the Hausa Language dataset is better if morphological analysis is considered.

Key words: Text summarization · Hausa language · Feature extraction · Naïve Bayes model

INTRODUCTION

The advances in information technology escalate the well-known global problem called information overload. Every field of human endeavors such as daily life activities and decision making depends on information. The main demand is to access the invaluable information at the right time. This information overload leads to the development of automatic analysis systems adapted to automatic processing of personal data. These systems such as machine translation, automatic text summarization and speech recognition fall into the field of natural language processing (NLP) [1]. The systems are necessary in order to reduce the problem of information overload. The present study focuses on automatic text summarization (ATS).

Automatic text summarization is a technique where a computer summarizes a long text into short and non-redundant form [2]. There are two types of ATS systems, namely: Extractive and abstractive systems. The extractive systems are based on verbatim extraction of sentences from the source as a summary. On the other hand, abstractive systems are based on modification and combination of the extracted sentences from the source as a summary. This is achieved by using some complex linguistic tools. Therefore, extractive method is widely used in the field of ATS as abstractive method is expensive [1]. The major challenge of extractive approach is producing incoherent summaries. In [3] also identify two challenges in this field, namely Problems related to NLP and the application of different approaches used in ATS. Problems regarding NLP have to do with the statistical or semantic analysis of the text depending on the implemented approach. The analysis varies from one language to another depending on the semantic complexity.

Several studies in this field are based on feature extraction. Feature extraction produces a set of features by breakdown of the original text. A feature is a mixture of elements that captures vital information of the text [4]. Naive Bayes model is a conditional probability model that assumes feature independence. However, the model works well for certain nearly functional feature dependence. The model is found to outperform other

Corresponding Author: Azilawati Rozaimee, Faculty of Informatics & Computing, Universiti Sultan Zainal Abidin, Malaysia.

models such as Support Vector Machine (SVM) and Hidden Markov Model (HMM) by using such set of extracted features [5].

There exist language independent summarizers such as Text Compactor and Free Summarizer. The language independent summarizers are developed without morphological or semantic analysis of any language [6]. In this regard, several studies are conducted in this field for many languages such as Malay, Sinhala and Oromo [7-9]. Although language dependent summarizers are expensive, literature shows that the summarizers outperform language independent ones [10, 11]. There is a limited research to automatically summarize text in Hausa language. Therefore, there is a need for a study that considers features that require Hausa Language morphological analysis. As a result, this study focuses on developing a model to automatically summarize Hausa Language text based on feature extraction using Naive Bayes model.

Hausa is a Chadic language generally spoken in West African countries such as Nigeria, Niger and Ghana. It has about 52 million native and non-native speakers with Nigeria having the highest number of Hausa Language speakers. It has two forms of writing system namely Ajami and Boko. Ajami uses most letters of Arabic alphabet, while Boko uses the most letters of Latin alphabet. Furthermore, Boko style is widely used in Hausa literature [12]. Hausa language has acquired some computational linguistic tools. The tools include Facebook chats normalizer [13], spell corrector [14] as well as word stemmers [15-16].

The first acknowledged study in ATS is that of [17] that depends on term frequency. With the unprecedented increase in electronic textual information in 1990s to date, several studies are conducted to summarize such textual information. The studies are commonly based on text feature extraction using different approaches such as machine learning and discourse structure. The present study reviews some recent studies as follows. In [9] design a summarizer for Afan Oromo news text to use 3 methods (A, B and C). The method 'A' is designed to use position method and term frequency, without stemming and using lexicons. Method 'B' is developed to use the method 'A' with stemming and language-specific dictionary. Finally, method 'C' is designed as a modification of method 'B' with an improved position method. The research evaluates these methods using both subjective and Objective evaluation methods. The results show that method 'C' with 81% and 75% outperforms the other two methods in both objective and subjective evaluations respectively.

Fuzzy logic is a technique used to represent and utilize data that have non-statistical uncertainty. The technique has been used in ATS to provide powerful reasoning capabilities to decision support and expert systems. In this regard, in [10] introduce some new features such as alphanumeric and morphological sentences in addition to the most common ones. The researchers use fuzzy logic to weigh each sentence using 14 different features. The system outperforms some online summary tools such as Copernic summarizer.

Rhetorical Structure Theory (RST) proposed by [18] has also been used in ATS. RST is able to address text organization through relations that exist between parts of the text. In [19] combine RST with neural network in order to develop a powerful summarizer. The researchers train the neural network to learn relevant features of sentences by using back propagation method.

Furthermore, in [20] uses Naive Bayes method to automatically summarize Indonesian text based on latent semantic analysis. The researcher uses a set of 100 documents of a single genre to evaluate the model. It is found that the semantic feature increases precision and F-measure values by 9.8% and 2.4% respectively.

MATERIAL AND METHOD

Proposed Model for Automatic Hausa Language Text Summarization: The language independent summarizers do not consider morphological or semantic analysis of any language [6]. Therefore, the present study adopts five features including some that require Hausa language morphological analysis in sentence extraction method to bridge this gap. In this regard, the following approaches for extracting these features from sentences are adopted.

Sentence Length: This feature is valuable to filter short sentences including datelines and sub-titles found in news articles that are not appropriate in a summary. The present study adopts the approach used by [21] as follows:

$$P(S, Len) = \frac{\text{No of words in the sentence}}{\text{No of words in the longest sentence}} \quad (1)$$

where P(S, Len) is the probability of the sentence S based on its length Len.

Cue Phrase: This feature, pioneered by [22] is effective in indicating invaluable sentences. The present study is informed by the Hausa linguistic expert about its importance in making summary. A manual compilation of the Hausa Language cue phrases is carried out with the help of linguistic expert. Therefore, the probability of the cue phrase occurring in a sentence is determined as follows:

$$P(S, Cue) = \frac{\text{No of occurences of the cue phrases in the sentence}}{\text{Length of the sentence}} \quad (2)$$

where P(S, Cue) is the probability of the sentence S having cue phrases.

Title Words: The probability of the sentence having title word is determined as follows:

$$P(S, Title) = \frac{\text{No of title words in the sentence}}{\text{Length of the sentence}} \quad (3)$$

where P(S, Title) is the probability score given to the sentence S based on its title words.

Keyword Feature: Words in the text are weighed using Term Frequency-Inverse Sentence Frequency (TF-ISF) proposed by [23] for a single document summarization. A set of keywords is formed by choosing ten words with the highest TF-ISF values. The probability of the sentence given its keywords is determined as follows:

$$P(S, keywords) = \frac{\text{No of unique keywords in the sentence}}{\text{Length of the sentence}} \quad (4)$$

where P(S, keywords) is the probability score for the sentence given its keywords.

Sentence Location: The present study is based on the hypothesis that important sentences lie on the beginning and end of a paragraph. Meanwhile, the redundant ones are within the paragraph. Therefore, the probability score given to a sentence is determined as follows:

$$P(S_i, n) = \begin{cases} \frac{n-(2\times i)+1}{n-1}, & i < \frac{n+1}{2} \\ \frac{(2\times i)-n-1}{n-1}, & i > \frac{n+1}{2} \\ 0.1, & i = \frac{n+1}{2} \end{cases} \quad (5)$$

where P(Si, n) is the probability given to a sentence based on its location. Si is the sentence location within the paragraph and n is the total number of sentences in the paragraph.

The present study adopts a simple Naive Bayes model used by [24] to assign a total weight for each sentence given its features. The formula is as follows:

$$P(s\ S|F1,F2\&Fk) = \frac{P(F_1, F_2, \ldots F_k \mid s \in S)\ P(s \in S)}{P(F_1, F_2, \ldots F_k)} \quad (6)$$

By assumption of statistical independent of the features, Equation (6) becomes:

$$P(s\ S|F1,F2\&Fk) = \frac{\prod_{x=1}^{k} P(F_x \mid s \in S)\ P(s \in S)}{P(F_x)} \quad (7)$$

By using logarithmic rule, Equation (7) becomes:

$$P(s|F1,F2\&Fk) = \log(P(s \in S)) + \sum \log\left(P(F_k \mid s)\right) \quad (8)$$

$$P(s \in S) = \frac{1}{q} \quad (9)$$

where P(s | F1, F2 … Fk) is the total weight of the sentence s given the set of features F1, F2 … Fk. P(s ∈ S) is the probability of the sentence being in the summary, while q is the number of summary sentences. And P (Fk |s) is the determined probability for the sentence s given the feature Fk.

It is observed that a good compression rate lies between 5% and 30% [25]. In the present study, the compression rate is set to 30% of the original document length. Hence, a threshold value is determined as follows:

$$\text{Threshold} = \frac{30 \times n}{100} \quad (10)$$

where n is the total number of sentences in the document.

Therefore, the following proposed model shown in Fig. 1 can improve the quality of summary for Hausa Language text. The system is developed using a Java platform. The design of this summarization system is divided into 3 stages include pre-processing, feature weight extraction as well as sentence weighing and simplification.

Fig. 1 shows that the system reads an input text file and enters a preprocessing stage. In that stage, the system uses a tokenization module to break down the text into groups of smaller units, namely paragraph, sentence

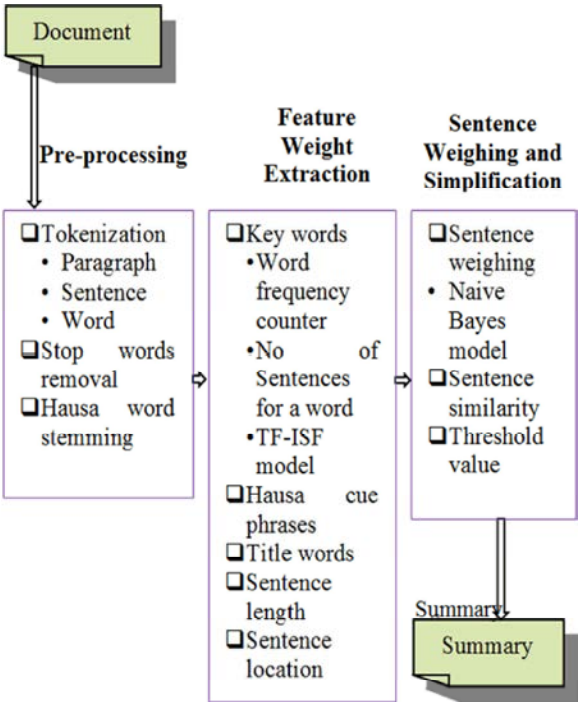

Fig. 1: Hausa Language text summarizer model

and word units. The stop words removal module is applied on the word units to eliminate the stop words. The resulting content words are stemmed using the Hausa word stemming algorithm developed by [15].

In the feature extraction stage, the weight of each word is determined using TF-ISF model. Meanwhile, a set of keywords is formed by choosing the top 10 words with the highest TF-ISF values. The system assesses each sentence using five modules such as Hausa cue phrase, keywords and title words modules.

In the final stage, the system uses Naive Bayes model to determine a total weight for each sentence using the five different values. A sentence similarity module is used to eliminate any duplicate sentence. A threshold value module is used to choose the summary sentences. The system determines the precision, recall and F-score values of the summary based on the 3 corresponding experts' summaries. Finally, the system presents the resulting final summary and its precision, recall as well as F-score values.

Corpus Preparation: Experts' summaries are considered as gold standard based on the assumption that human summaries are perfect or model summaries [26]. In that regard, system generated summaries are usually compared to the experts' ones. Unlike rich resource languages, standard dataset in Hausa is limited for assessing text summarization system. Therefore, the present study assembles 10 Hausa documents from two different newspapers namely Aminiya and Leadership Hausa. The dataset consists of distinctive genres such as sports, social and politics. Each document has at least 600 words. Furthermore, manual summaries of these documents are formed. Each document is given to 3 Hausa linguistic experts for summary that is 30% of the original document. The experts are informed to only select sentences that show the purpose of the document.

Evaluation Criteria: In the present study, the intrinsic evaluation technique is adopted (the precision and recall). This is to measure the quality of the system generated summary with respect to the gold standard (experts' summary). Precision determines the fraction of the sentences in system summary that appeared in at least one of the 3 corresponding experts' summaries (Equation 11). On the other hand, recall determines the fraction of the sentences in system summary that appeared in all the 3 corresponding experts' summaries (Equation 12).

$$\text{Precision} = \frac{S_t \cap S_m}{S_m} \times 100 \quad (11)$$

$$\text{Recall} = \frac{S_m \cap S_c}{S_c} \times 100 \quad (12)$$

where S_m is the set of summary sentences generated by the system, S_t is the union set of the 3 sets of summaries manually produced by the experts and S_c is the intersection set of the 3 sets of summaries manually produced by the experts.

Equations 11 and 12 are combined for easy comparison between one summary and another. This combination or harmonic mean of the precision and recall is referred to as the F-score that determines the peak value of these evaluators as shown in Equation 13.

$$\text{F-score} = 2 \times \frac{P \times R}{P + R} \times 100 \quad (13)$$

where P and R are the precision and recall values respectively.

Moreover, coefficient of variation (CV) is adopted in the present study to compare between two or more models. CV is defined as the ratio of the standard deviation to the mean (Equation 14). It can measure the

rate of variation of the F-score values in relation to the mean of the values. The lower the CV value the higher the fitness of a model. Meanwhile, the higher the CV value the lower the fitness of the model [27].

$$C_v = \frac{\sigma}{\mu} \times 100 \quad (14)$$

where C_v is the coefficient of variation (CV) and ó is the standard deviation and µ is the mean..

RESULTS AND DISCUSSION

In Table I, the Hausa Language summarizer produces a summary with a lowest similarity of 57.14% to the 3 corresponding experts' summaries in "doc7.txt". Moreover, the system produces summaries that have 100.00% similarity with the experts' summaries in "doc8.txt" and "doc2.txt". This shows that the system is able to retrieve all sentences agreed by at least one of the 3 corresponding experts' summaries. In addition, the system is also able to retrieve all sentences agreed by all the 3 corresponding experts' summaries.

Experiments are conducted to compare the performance of the Hausa Language summarizer with that of online summary tools (Text Compactor, SMMRY and Free Summarizer). These tools summarize any text written in Roman (Latin) alphabets. The experiments are done by summarizing the dataset with each tool. Moreover, the resulting summaries are compared automatically with the with the 3 corresponding experts' summaries. This is to determine the precision, recall and F-score values of the summaries by these tools (refer to Table II-IV).

In Table 2 shows that text compactor produces its best summary having 66.67% similarity to the corresponding experts' summaries in doc9.txt. On the other hand, it worst performance in doc10.txt.

From Table 3, the Free Summarizer produces its best summary having 80.00% similarity to the corresponding experts' summaries in doc9.txt. It has 29.41% similarity as its worst performance in doc10.txt.

In Table 4, the SMMRY produces its best summary having 92.31% similarity to the corresponding experts' summaries in doc8.txt. It has 27.59% similarity as its worst performance in doc10.txt.

The F-score values of Table 1 are compared graphically to the F-score values in Table 2, 3 and 4 (Fig. 2). This is to quickly determine the set of documents that Hausa Language summarizer outperforms the online tools in producing its summary.

Table I: Precision, recall and F-score values of Hausa summarizer

Document	Precision	Recall	F-Score
doc1.txt	83.33	66.67	74.07
doc2.txt	100.00	100.00	100.00
doc3.txtz	57.14	60.00	58.54
doc4.txt	63.64	66.67	65.12
doc5.txt	87.50	75.00	80.77
doc6.txt	90.00	66.67	76.60
doc7.txt	50.00	66.67	57.14
doc8.txt	100.00	100.00	100.00
doc9.txt	81.82	100.00	90.00
doc10.txt	77.78	80.00	78.87

Table 2: Precision, recall and F-score values for text compactor

Document	Precision	Recall	F-Score
doc1.txt	57.14	33.33	42.11
doc2.txt	33.33	50.00	40.00
doc3.txt	60.00	40.00	48.00
doc4.txt	66.67	33.33	44.44
doc5.txt	66.67	25.00	36.36
doc6.txt	62.50	50.00	55.56
doc7.txt	50.00	33.33	40.00
doc8.txt	100.00	25.00	40.00
doc9.txt	50.00	100.00	66.67
doc10.txt	28.58	20.00	23.53

Table 3: Precision, recall and F-score values for free summarizer

Document	Precision	Recall	F-score
doc1.txt	66.67	33.33	44.44
doc2.txt	60.00	50.00	54.55
doc3.txt	71.43	40.00	51.28
doc4.txt	72.73	66.67	69.57
doc5.txt	50.00	25.00	33.33
doc6.txt	80.00	66.67	72.73
doc7.txt	75.00	66.67	70.59
doc8.txt	57.14	25.00	34.78
doc9.txt	66.67	100.00	80.00
doc10.txt	55.56	20.00	29.41

Table 4: Precision, recall and F-score values for SMMRY

Document	Precision	Recall	F-Score
doc1.txt	50.00	33.33	40.00
doc2.txt	100.00	50.00	66.66
doc3.txt	71.43	60.00	65.23
doc4.txt	72.73	50.00	59.26
doc5.txt	62.50	50.00	55.56
doc6.txt	45.00	60.00	51.97
doc7.txt	50.00	33.33	40.00
doc8.txt	85.71	100.00	92.31
doc9.txt	57.14	100.00	72.73
doc10.txt	44.44	20.00	27.59

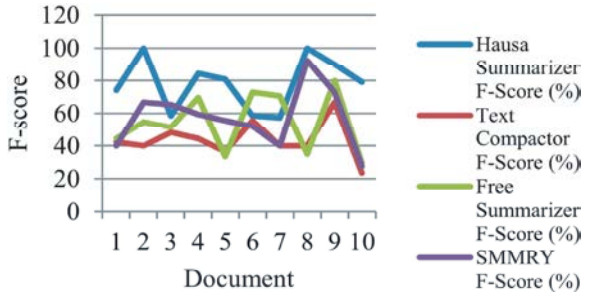

Fig. 2: Graphical comparison of the F-score values in Table-1 and (Table 2-4)

Table 5: CV values of the summarizers

Summarizer	Standard Deviation (%)	Mean (%)	CV
Text Compactor	11.51	43.67	26.35
Free Summarizer	18.39	54.07	34.02
SMMRY	18.65	57.13	32.64
Hausa Summarizer	16.24	78.19	20.77

From Fig. 2, the y-axis represents the F-score values, while x-axis represents the documents. The line with blue color represents the F-score values of the Hausa Language summarizer for the dataset. On the other hand, the dark red, olive green and purple lines represent the F-score values of text compactor, free summarizer and SMMRY for the dataset. Each curve on the lines represents an F-score value for a particular document. It can be observed that, the blue line is above the other 3 lines in many curves. This implies the dominance of the Hausa Language summarizer over the other 3 summarizers in dataset. The present study determines the CV value of each summarizer given its F-score values. This is to quickly determine the best model among the four models as shown in Table 3.

In Table 5, the Hausa Language summarizer has the lowest value of CV of 20.77. This indicates that the Hausa Language summarizer outperforms the other summarizers as far as the dataset is concerned.

CONCLUSION

The Hausa Language summarizer is capable of achieving an average F-score of 78% using 30% compression rate for the defined dataset. On the other hand, the text compactor, SMMRY and Free Summarizer attain average F-score values of 44%, 57% and 54% respectively in summarizing the dataset. This shows that the Hausa Language summarizer is well-designed compared to the results obtained by the online tools. This is achieved by developing the system to use Hausa Language morphological analysis in feature extraction.

Although the result is good compared to the results of online tools, the evaluation is on a relatively small dataset. Therefore, the system needs further enhancement and testing. The feature work will focus on expanding the dataset and dividing it into training and testing datasets. More semantic features such as part of speech (POS) tagging will be introduced in future.

ACKNOWLEDGEMENT

The authors express heartfelt thanks to Kano State Government for sponsorship. The authors also express gratitude to Dr. Tijjani Shehu and Muhammad Ammani of Linguistics Department of Bayero University Kano Nigeria. This is for their tireless effort and support in building the dataset for the present study. The authors' appreciation also goes to Isyaku Hassan of Faculty of Languages and Communication (FBK), UniSZA for competently proofreading this paper.

REFERENCES

1. Mani, I., 2001. Automatic Summarization, Amsterdam, Netherlands: John Benjamins Publishing.
2. Radev, D.R., E. Hovy and K. McKeown, 2002. Introduction to the special issue on summarization, Computational Linguistics, 28: 399-408.
3. Alami, N., M. Meknassi and N. Rais, 2015. Automatic texts summarization: Current state of the art, Journal of Asian Scientific Research, 5: 1-15.
4. Oracle. 2011. Oracle data mining concepts, 11g release 2 (11.2). [Online]. Available: http://www.cs.utexas.edu/ ~cannata/ dataSci/ Class%20Notes/Data%20Mining%20Concepts_e16808.pdf.
5. Rish, I., 2001. An empirical study of the naive Bayes classifier, IJCAI Workshop on Empirical Methods in Artificial Intelligence, 3: 41-46.
6. Hassel, M., 2007. Resource lean and portable automatic text summarization," Phd thesis, KTH School of Computer Science and Communication, Stockholm, Sweden, 2007.
7. Zamin, N. and A. Ghani, 2010. A hybrid approach for Malay text summarizer, in Proc. IMCETI'10, pp: 1-6.
8. Welgama, W.V., 2012. Automatic text summarization for Sinhala, Master thesis, University of Colombo, Sri Lanka, 2012.
9. Dinegde, G.D. and M.Y. Tachbelie, 2014. Afan Oromo news text summarizer, International Journal of Computer Applications, 103: 1-6.

10. Abhiman, B.D. and P.P. Rokade, 2015. A text summarization using modern features of sentences, International Journal of Innovative Research in Computer and Communication Engineering, 3: 4757-4764.
11. Dixit, R.S. and S.S. Apte, 2012. Improvement of text summarization using fuzzy logic based method, IOSR Journal of Computer Engineering, 5: 5-10.
12. Newman, P., 2000. The Hausa Language: An Encyclopedic Reference Grammar, Connecticut, USA: Yale University Press.
13. Maitama, J.Z., U. Haruna, A.Y. Gambo, B.A. Thomas, N. Idris, A.Y. Gital and A.I. Abubakar, 2014. Text normalization algorithm for Facebook chat in Hausa language, in Proc. IEEE ICICTMW'14, 2014, pp: 1-4.
14. Salifou, L. and H. Naroua, 2014. Design of a spell corrector for Hausa language, International Journal of Computational Linguistics, 5: 14-26.
15. Bashir, M., A.B. Rozaimee and W.M.B.W. Isa, 2015. A word stemming algorithm for Hausa language, IOSR Journal of Computer Engineering, 17: 25-31.
16. Bimba, A., N. Idris, N. Khamis and N.F.M. Noor, 2015. Stemming Hausa text: Using affix-stripping rules and reference look-up, Language Resources and Evaluation, 50: 687-703.
17. Luhn, H.P., 1958. The automatic creation of literature abstracts, IBM Journal of Research Development, 2: 159-165.
18. Mann, W.C. and S.A. Thompson, 1988. Rhetorical structure theory: Towards a functional theory of text organization, Text-Interdisciplinary Journal for the Study of Discourse, 8: 243-281.
19. Sarda, A.T. and A.R. Kulkarni, 2015. Text summarization using neural networks and rhetorical structure theory, International Journal of Advanced Research in Computer and Communication Engineering, 4: 49-52.
20. Najibullah, A., 2015. Indonesian text summarization based on naive Bayes method, in Proc. ISC'15, 2015, pp: 67-78.
21. Suanmali, L., N. Salim and M.S. Binwahlan, 2009. Fuzzy logic based method for improving text summarization, International Journal of Computer Science and Information Security, 2: 1-6.
22. Edmundson, H.P., 1969. New methods in automatic extracting, Journal of the ACM, 16: 264-285.
23. Neto, J.L., A.D. Santos, C.A. Kaestner and A.A. Freitas, 2000. Document clustering and text summarization, in Proc. ICPAKDDM'00, pp: 41-55.
24. Kupiec, J., J. Pedersen and F. Chen, 1995. A trainable document summarizer, in Proc. ACM SIGIR CRDI'95, pp: 68-73.
25. Mani, I., 2001. Recent developments in text summarization, in Proc. ACM ICIKM'01, pp: 529-531.
26. Louis, A. and A. Nenkova, 2013. Automatically assessing machine summary content without a gold standard, Computational Linguistics, 39: 267-300.
27. Abdi, H., 2010. Coefficient of Variation, ser. Encyclopedia of Research Design. California, USA: SAGE Publications Inc., 2010.