Egyptian Informatics Journal 20 (2019) 89–96

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

An efficient automated answer scoring system for Punjabi language

Tarandeep Singh Walia [a,*], Gurpreet Singh Josan [b], Amarpal Singh [c]

[a] IKG PTU, Kapurthala, Punjab, India
[b] Punjabi University, Patiala, Punjab, India
[c] BCET, Gurdaspur, Punjab, India

ARTICLE INFO

Article history:
Received 9 July 2018
Revised 28 October 2018
Accepted 19 November 2018
Available online 29 November 2018

Keywords:
Natural Language Processing
Automated scoring system
Ambiguity

ABSTRACT

Automated Scoring is a developing technology. The accuracy and reliability of these systems have been proven to be much higher. Besides being a time-and money-saver, there are a number of studies which are being conducted to provide variety in feedback, not only on grammatical issues, but also on semantic related issues. This will reduce not only the paper load of the teachers, but as well as teachers' assessment related issues. In this paper, we remove those dummy note bins, thereby having 183 dimensions in total. We augmented the input by concatenating it with the first-order difference of the semitone filtered spectrogram. We observed a significant increase of the transcription performance with this addition. Compared to feedforward neural networks, recurrent neural network (RNN) are capable of learning temporal dependency of sequential data, which is the property found in music answer scoring.Also, the Long Short-Term Memory (LSTM) unit has a memory block updated only when an input or forget gate is open, and the gradients can propagate through memory cells without being multiplied each time step. Throughout backward and forward layers together, the networks can access to both history and future of the given time frame. Comparative analyses show that the proposed technique outperforms existing techniques.

1. Introduction

Answer scoring is defined as an act of assigning scores to an answer by a human grader. The Automated scoring system has become a hot issue in the research on Natural Language Processing (NLP). The Automated scoring is performed by extracting the grammatical relations as well as semantic relations from the student answer and reference answer. It is becoming a consistent tool performing for scoring as compared to a human grader in which an answer scored highly by one grader may receive a low score from another [1] Figs. 1–3.

The different scores can be assigned by the same grader for the same answer at different times. Hence, the fairness of answer grading cannot be assured. On the other hand, Automated Scoring will perform fair scoring and can be repeated again and again with consistency [3]. The research in this direction will open new dimensions for researchers as it is an interdisciplinary work. If such a system be developed in any Indian language, it will open the doors for other similar Indian languages. It has been previously mainly implemented with interfaces for problem solving systems for specific domains [5].

1.1. Automated scoring and NLP

This paper reviews and compares NLP based Automated Answer scoring approaches. Semantic similarity plays a vital role in NLP, Informational Retrieval, Text Mining, Q & A systems, text-related research and application area [7]. Today, Automated Scoring is still a difficult and interesting issue for researchers in artificial intelligence and NLP though many English Automated Scoring systems have been proposed and developed but with little success [8].

NLP ensures the most reliable score by mapping student answer into a formal world model. NLP has major tasks such as discourse analysis, morphological segmentation, parsing, word sense disambiguation and information extraction, etc [10].

* Corresponding author.
E-mail address: taran_walia2k@yahoo.com (T.S. Walia).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2018.11.001
1110-8665/

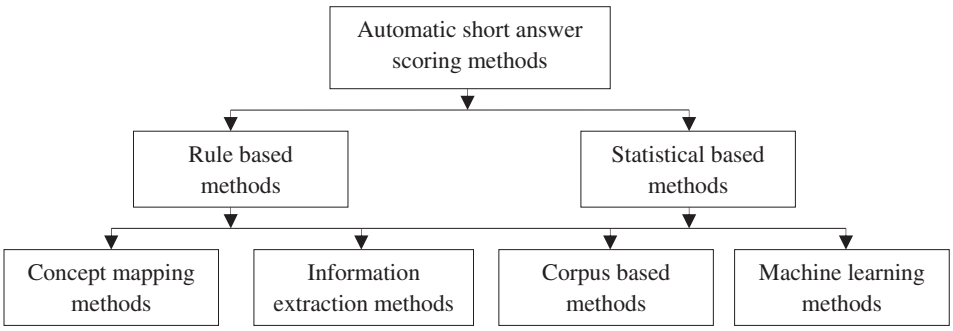

Fig. 1. Methods of Automatic short answer scoring.

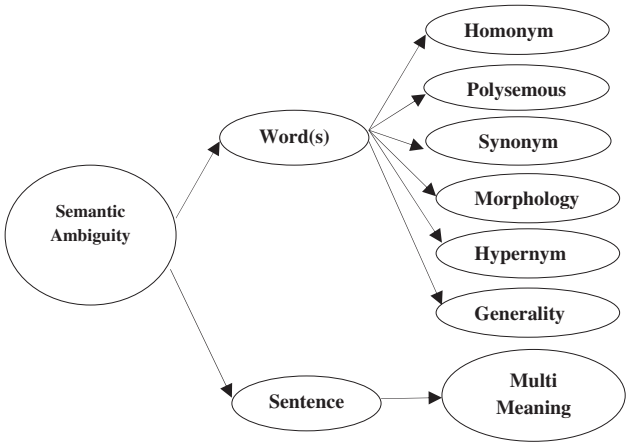

Fig. 2. Ambiguity Types.

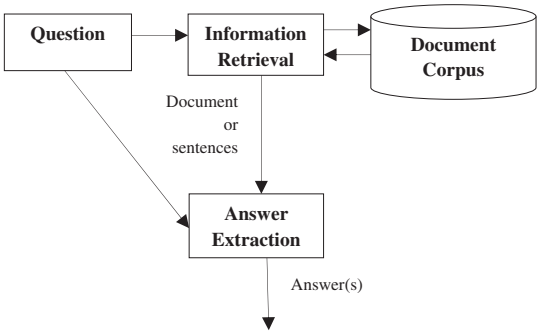

Fig. 3. Architecture of Question Answering System.

Automated Scoring can choose some tasks from NLP to the scoring process. Automated Scoring systems are a combination of various techniques such as - NLP (NLP) along with, Linguistics, Artificial Intelligence (Machine Learning), Statistics and Web Technologies, etc [12]. Current automatic essay grading mechanism relies on two aspects such as machine learning techniques and grammatical measures of quality. However, none of them which identifies statements of meaning (propositions) in the text. Hence, it proves to be an inappropriate for scoring the content of an answer [13].

1.2. Ambiguity types

Ambiguity occurs in SQA because a word or a sentence can have more than one meaning or multiple words in the same language can share the same meaning. Semantic ambiguity is concerned with the meaning of word(s) and sentence(s) [14].

A word, a phrase or a sentence may have more than one meaning in a given situation. Lexical ambiguity can arise because of the following: -

- Homonym: A homonym is when a word has more than one meaning.
- Polysemous: It is another case where a word has more than one meaning. Unlike homonym, the meanings of the word share the same concept.
- Synonym: A synonym occurs when two or more words have the same meaning.
- Morphology: Morphology refers to a word ambiguity.
- Hypernym: A noun that has a subtype or an instance that is also a noun.
- Generality: A word is general with respect to another word.
- Semantic ambiguity of the whole sentence occurs if there are multiple meanings within a sentence.

1.3. Existing system

There are different AES systems covering various aspects to evaluate a number of different features [19]. Currently there are four major developers of automated essay scoring, which are widely used by universities, schools, and testing companies viz: Project Essay Grader (PEG) [23], Intelligent Essay Assessor (IEA), E-rater, and IntelliMetric [24].

1.4. Existing application

Some of examples of this application are: In TOEFL exam and In the Graduate Record Examination (GRE) analytical writing section. At present, the few Automated Scoring system is available, but with little success to evaluate human written (typed) essays automatically.

2. Related work

Different Automated Scoring systems have different approaches. But in all, the most related approach to this research includes:

2.1. Unsupervised approach

Chen et al. (2010) [7] demonstrates the AES, which uses a voting algorithm based on an unsupervised learning approach. To build computational learning model, they use an unsupervised approach which does not use any reference text. Their approach is evaluated on a set of essays, written by different people, on a same topic. The

scoring scheme relies upon further information and the similarities between essays.

Arijit De, Sunil Kumar (2011) [8] discussed an unsupervised technique that uses a 'bag of words' approach for selecting a set of good essays from a huge selection of essays written on the same topic. But it does not require deep parsing. The approach is discussed using the help of Kullback-Liebler divergence that is used in text mining application for computing similarities between essays. The accuracy was lower at 30%. The advantage of unsupervised approach is that it can be applied with a little modification to any language as it does not use any specific language feature. The limitation of this approach is that it does not consider style, organization, and grammar features.

2.2. Supervised approach

Hongbo Chen, Ben He, Baobin (2012) [9] introduced a Ranked-based learning approach to AES. They discussed supervised learning problems in detailed steps. It also discusses how to apply learning to rank to automated essay scoring, such as feature extraction and scoring. It regards automated essay scoring as a ranking problem and plan to solve this problem by learning to rank algorithms. Learning to rank that automatically construct a ranking model or function to rank objects.

2.3. Short answer scoring approach

The domain of our research is closed short answer scoring because the short essay answer marking is relying heavily on the contents of the essays only. Marking short answer essay typed examination differs from marking the free text essay, where the score of the latter is the total of the style and contents. Some of related work to short answer are:

MohdJuzaiddin, Fatimah Dato (2009) [19] discussed scoring the short answers, essay typed examination in which similarities of sentences are measured by comparing the answer scripts and the scoring scheme. Short answer essay scoring relies heavily on the contents of the essays only. It consists of the following modules:

i. Shallow syntactic analysis
ii. Pronoun resolute
iii. Morphology
iv. Morphology and negation
v. Matching
vi. Filling in the semantic gaps

The answers will be mapped automatically in the mapping module. The mapping module enables to score specific ideas. It is not suitable for scoring open ended essay.

Luis Tandalla (2012) [16] discussed Scoring Short answer Essay. In their system concepts are spilt into smallest unit which are predefined expertly written model answers to closed end question. Their dependencies are automatically measured by tagging resultant concepts and their dependencies. Dependencies are pattern matched with the scoring scheme. The process is repeated for every student to answer.

Ali Muftah Ben Omran and MohdJuzaiddin Ab Aziz (2013) [3] demonstrate that one of the most complicated domains is marking of short essay answers automatically as it relies heavily on the semantic similarity. Their research proposes two methods: the first proposition is an Alternative Sentence Generator Method, which uses the synonym dictionary to produce the alternative model answer. The second proposition uses a combination of three algorithms together in matching phase, which is Longest Common Subsequence (LCS), Commons Words (COW) and Semantic Distance (SD).

2.4. Question answering (QA) approach

Automated answering scoring is typically based on Question Answering (QA) System. Question answering is a technique of positioning exact answer sentences from wider collections of documents in question answering. Question answering is a sophisticated form of information retrieval characterised by information needs and is one of the most natural forms of human computer interaction. In contrast with classical information retrieval, where complete documents are considered relevant to the information request, in question answering, specific pieces of information are returned as an answer. The ultimate aim of automatic question answering is to detect accurate answers to the natural language questions specified by users. It focuses on the interactions between information extraction, NLP and information retrieval.

The QA system architecture, has a question analysis module which processes a question posed by the user. It constructs a query that is used by an IR module, as well as answer type information that is used by an AE module. The IR module retrieves documents or passages from a corpus and passes them to the AE module. The architecture of our system shown in Fig. Follows this pattern closely, although it doesn't have a separate question analysis module. Question processing is an integral part of the IR and AE modules and only involves tokenization and removal of stop-words.

Oleksandr Kolomiyets et al. (2011) [20] describe a question answering technology from an information retrieval perspective. Different levels of processing, yielding bag-of-words were discussed along with the more complex representations integrating semantic roles, discourse analysis, POS, classification of the expected answer type, translation of the question into a SQL-like language or logical representation.

Kong Joo Lee et al. (2011) [13] introduced an automated English sentence evaluation system for students learning English as a second language. Firstly, the input sentence is analysed in order to detect possible errors such as syntactic errors and spelling errors. Secondly, the input sentence is compared with given answers to identify the differences as errors. Then the system evaluates the performance and the output produced by the system is compared with the result given by human ratters. The system still has the problems which most NLP systems experience. The problems are found in ambiguity resolution, handling parsing failure, lexicon incompleteness, and so on.

Paloma Moreda et al. (2011) [21] describe combining semantic information in question answering systems. Their study involves two proposals which are based on semantic rules and WordNet, which are to extract an answer in the general open - domain question answering (QA) system. The main objective of their research is to compare extracted answers with the existing QA systems with the named entities (NEs).

Matthias H. Heie et al. (2012) [17] explain question answering using statistical language modelling. Their work is to develop robust systems for different languages which overcome the need for highly tuned linguistic modules. They embedded this framework with an implementation in the TREC 2006 QA evaluations. As long as appropriated training data is available, this system can rapidly be adapted to any language.

Min-Chul Yang et al. (2015) [18] explain knowledge-based question answering using the semantic embedding space. In their study, the semantic embedding space is used to fulfil the goal to answer questions in any domains in which the embedding encodes the semantics of words and logical properties. This embedding-based inference approach for question answering allows the mapping of factoid questions posed in a natural language ontological representation of the correct answers guided by the knowledge base.

Sofian Hazrina et al. (2016) [25] describe the advancements of disambiguation in the semantic question answering system. In any semantic question answering (SQA) system ambiguity is an obvious problem. Therefore, an SQA system has to select the correct meaning with disambiguation solutions when the linguistic triples matched with multiple knowledge based (KB) concepts. The system mentions similar words in cases where linguistic triples do not match with any knowledge based (KB) concept.

Boris Galitsky et al. (2017) [5] introduced matching parse thickets for open domain question answering, it supports complex question answering, which include multiple sentences, to handle as many constraints expressed in this question as possible. In this study, they explore how parse thickets of questions and answers can be matched. They demonstrate the necessity for using discourse-level analysis to answer complex questions.

KajaZupanc et al. (2017) [11] incorporates additional semantic coherence and consistency attributes by proposing an extension of existing automated essay evaluation systems. The novel coherence attributes were designed by transforming sequential parts of an essay into the semantic space and evaluating changes among them for the estimation of coherence of the text.

2.5. Rule based approach

The advantages of this approach are that it is easy to incorporate domain knowledge into the linguistic knowledge which provides highly accurate results.

Kevin, Eileen & Robert et al. (2010) [12] "Machine Learning and Rule-Based Automated Coding of Qualitative Data". They explore how to implement fully or semi-automatic coding of textual data by leveraging NLP (NLP). In particular, they compare the performance of human-developed NLP rules to those inferred by machine learning algorithms. Both rule-based and machine-learning-based automatic coding seems to offer promise for coding qualitative data, even in the face of the low quality of the text. Overall, the rule-based approach performed better than the machine-learning approach, especially for those codes with few training instances.

Zhang Qiang et al. (2014) [28] describes the application of network automated essay scoring system in the college English writing course. Basing on the network automated essay scoring system, he first introduces the development of this kind of system and then tries to optimize the design of classroom teaching of college English writing. Finally, the paper proposes some potential problems of the automated essay scoring system and provides some useful suggestions for the teaching of college English writing. This paper effectively combines the traditional process writing method with the automated essay scoring system, and creates a highly effective essay training mechanism for the students.

2.6. Latent semantic analysis approach

LSA is a statistical method for automatic document indexing and retrieval. Latent semantic analysis (LSA) is a technique in NLP<https://en.wikipedia.org/wiki/Natural_language_processing>. Its advantage to other indexing techniques is that it creates a latent semantic space. Latent semantic indexing is the application of a particular mathematical technique, called Singular Value Decomposition or SVD, to a word-by-document matrix. LSA assumes that words that are close in meaning will occur in similar pieces of text. Limitation of LSA where a text is represented as an unordered collection of words.

Xinming Hu et al. (2010) [26] discussed Automated Scoring System for Subjective Questions based on LSI (Latent Semantic Indexing). Specifically, LSI is a statistical method that analyses terms and identifies important associative patterns among them.

Pantulkar Sravanthi et al. (2017) [22] explains "semantic similarity between sentences". In this, they evaluated and tested three different semantic similarities approaches like cosine similarity, path-based approach, and feature based approach. They propose an unsupervised approach to automatically calculate sentence level similarities based on word level similarities, without using any external knowledge.

Govinnage R. et al. (2015) [8] "A Dynamic Semantic Space Modelling Approach for Short Essay Grading". In this paper, the authors presented a novel approach for automating short essay grading using Latent Semantic Analysis and NLP. On most of other LSA based systems, it is required to feed the system with the premarked training essays. This dynamic semantic space was computed using Vector Space Models (i.e. LSA) and it was dynamically built upon each essay question set utilizing student answers. The future direction of this research is to extend this research for handling different languages other than English and to evaluate the system with a large data set.

2.7. Statistical or vector space model (VSM) approach

The Statistical model, i.e. VSM to find semantic similarity among student answer and reference answers and calculate probability of scores to be assigned to the student answer. Following are some work related to VSM:

Li Bin et al. (2008) [15] used the K-Nearest Neighbour (KNN) algorithm for AES. Each vector is expressed by the term frequency and inverse document frequency (TF-IDF) weight. The information gain (IG) and TF methods are used to select features by predefined features. The KNN algorithm for automatic essay scoring based on the well-developed text categorization modelling. With the different methods of feature selection, they are able to achieve 76% accuracy.

Yonghong Yan et al. (2012) [27] demonstrates AES system using vector regression. Each essay is represented by the Vector Space Model (VSM). To implement the system, it gives score on several features, including the surface features such as the number of words, number of sentences, average word length, average sentence length etc. and complex features such as grammar checking, sentences. Both the words and part-of-speech tag are considered. They use the simple model and CVA (content vector analysis) model. They get 86% precision given the two-score deviation compared to human ratters.

Jovita, Linda et al. (2015) [10] introduced using the Vector Space Model in Question Answering System. The aim of their research is to represent knowledge, and retrieve the answer for a given question by using the Vector Space Model. The query is compared to the knowledge base by measuring their similarity. Thus, it suggests for future improvement to use topic modelling or some other approaches like Latent Semantic Analysis to construct the model of document representation.

Alzahrani et al. (2015) [2] developed and compared number of NLP techniques that accomplish the task of automating scoring. Their study is based on a vector space model (VSM) in which after normalization, the baseline-system represents each answer by a vector, and subsequently calculates its score using the cosine similarity between it and the vector of the model answer.

Badr HSSINA et al. (2016) [4] in their study "Evaluation of semantic similarity using the vector space model based on textual corpus" create a semantic similarity calculation system between text documents to contribute to their semantic clustering. The system uses the WordNet lexical database and groups words together based on their meanings.

The advantages of Vector Space Model are that it allows computing a degree of similarity, ranking documents and partial matching, etc. VSM is best suited for short question answering.

But the disadvantages are that long documents are poorly represented, different vocabulary won't be supported and the order of the document is lost.

2.8. Summary

No doubt, there are many existing automated scoring systems and their approaches as described above in this study. But this makes scoring a complex task, and at the same time proper integration and correlation of semantic features of these systems are not fully possible and thereby giving inaccurate results. So, there will be one such system that will combine all above table listed approaches which gives the results that will be equal and nearly equal to the human raters Table 1.

Ensuring the Quality of Automated Scoring Implementation

The following are the measurement of quality of Automated Scoring:

1. Automated scores are reliable and fair.
2. The influence of automated scoring on reported scores is understandable.
3. Automated scores are produced is meaningful and understandable.
4. Automated scores are consistent with the scores from expert human graders.
5. Automated scores have valid scores, when checked against external measures in the same way done with human grading.

3. Proposed work

3.1. Problem definition

Automated Scoring is a developing technology. The accuracy and reliability of these systems have been proven to be much higher. Besides being a time-and money-saver, there are a number of studies which are being conducted to provide variety in feedback, not only on grammatical issues, but also on semantic related issues. This will reduce not only the paper load of the teachers, but as well as teachers' assessment related issues.

At present, the many Automated Scoring systems are available, but with little success to evaluate human written (typed) essays automatically and none of them is in Punjabi language. The proposed framework is illustrated in Fig. 4. The left-hand presents the two independent AMT systems that return either 88 note or chroma output and chroma onset output, respectively. The outputs are concatenated and aligned with the score MIDI through Dynamic Time Warping (DTW). Since our main idea is not improving the performance of AMT system but rather utilizing a neural-network based system that produces features for automatic answer scoring, we borrowed the state-of-art AMT system proposed by Schedl [6]. However, we slighly modified the training setting for our purpose Fig. 5.

3.2. Pre-processing

As aforementioned, our AMT system is based on the existing model. Therefore, we used the same multi-resolution STFT with semitone spaced logarithmic compression in the model. It first receives answer scoring as input and computes two types of short time Fourier transform (STFT), one with a short window (2048 samples, 46.4 ms) and the other with a long window (8192 samples, 185.8 ms), with the same overlap (441 samples, 10 ms). The STFT with a short time window gives temporally sensitive output while the one with a longer window offers better frequency resolution. A Hamming window was applied on the signal before the STFT. We only take magnitude of the STFT, thereby obtaining spectrogram with 100 frames/s. To apply logarithmic characteristics of sound intensity, a log-like compression with a multiplication factor 1000 is applied on the magnitude of spectrograms. We then reduce the dimensionality of inputs by filtering with semi-tone filterbanks. The center frequencies are distributed according to the frequencies of the 88 MIDI notes and the widths are formed with overlapping triangular shape. This process is not only effective for reducing size of inputs but also for suppressing variance in piano tuning by merging neighboring frequency bins. In the low frequency, some note bins become completely zero or linear summation of neighboring notes due to the low frequency resolution of the spectrogram.

We remove those dummy note bins, thereby having 183 dimensions in total. We augmented the input by concatenating it with the first-order difference of the semitone filtered spectrogram. We observed a significant increase of the transcription performance with this addition. The Böck and Schedl model uses a recurrent neural network (RNN) using the Long Short Term Memory (LSTM) architecture. Compared to feedforward neural networks, RNNs are capable of learning temporal dependency of sequential data, which is the property found in music answer scoring.

Also, the LSTM unit has a memory block updated only when an input or forget gate is open, and the gradients can propagate through memory cells without being multiplied each time step. This property enables LSTM to learn longterm dependency. In our task, the LSTM is expected to learn the continuity of onset, sustain and offset within a note as well as the relation among notes. The LSTM units are also set to be bidirectional, indicating that the input sequence is not only presented in order but also in the opposite direction. Throughout backward and forward layers together, the networks can access to both history and future of the given time frame.

While the Böck and Schedl model used a single network that predicts 88 notes, we use two types of networks; one predicts 88 notes or 12 chroma and the other predicts 12 chroma onsets. In

Table 1
Comparative analysis between various categories of automated scoring system.

Existing approaches	Advantages	Limitations/future work
Unsupervised	Can be Applied it to any language Free from language specific features	Organization, style, and grammar features
Supervised	Use of supervised algorithm	Complexity
Short answer scoring	Content scoring	Accuracy needed
Question Answering (QA)	Information extraction, NLP and information retrieval	Need a scoring mechanism to grade answer
Rule based Approach	Linguistic knowledge	Collaboration and integration with a statistical approach
Latent Semantic Analysis	Statistical/Mathematical approach Meaning extraction	Need more focus on semantic features
Statistical or Vector Space Model (VSM)	Degree of similarity, Ranking documents and Partial matching	Not suitable for long answers Collaboration and integration of different syntactic as well as semantic features None of them on Indian language

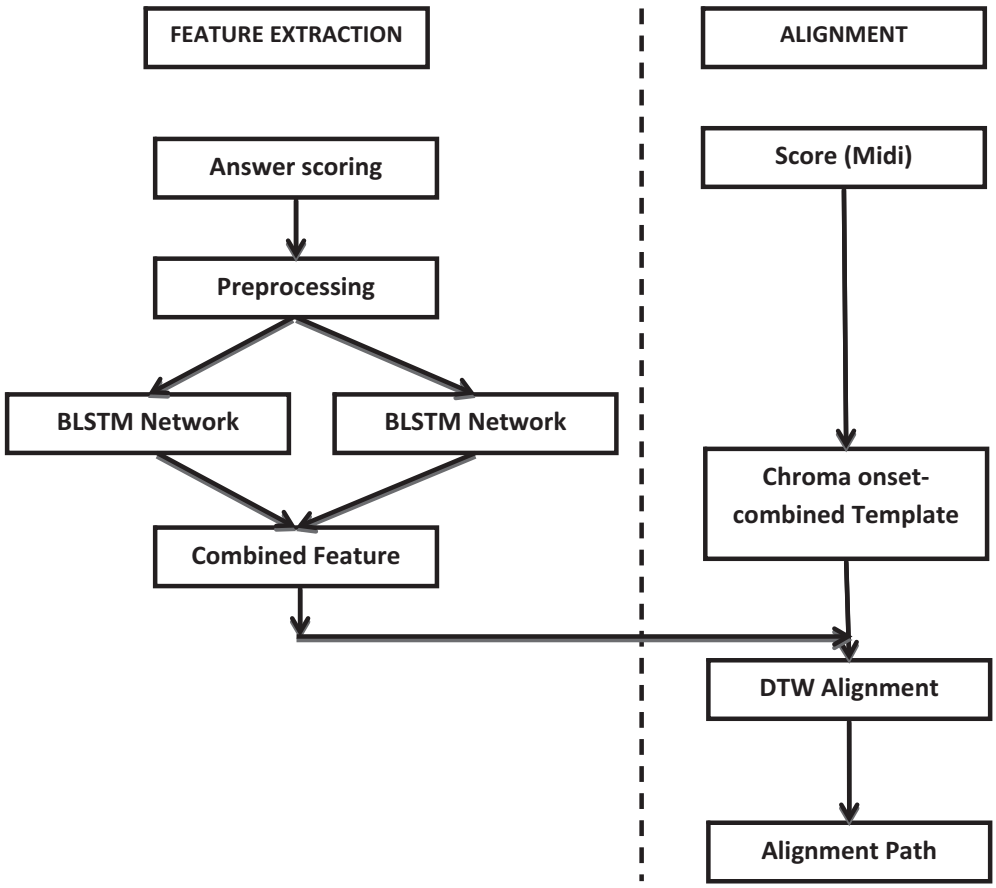

Fig. 4. Flow diagram of proposed automatic answer scoring system.

the 88-note network, we reduced the size to two layers of 200 LSTM units as it performed better in our experiments. In the 12-chroma, we downsized it further having 100 LSTM units on the first layer and 50 LSTM units on the second layer. On top of the LSTM networks, a fully connected layer with sigmoid activation units are added as the output layer. Each output unit corresponds to one MIDI note or chroma (i.e. pitch class of the MIDI note).

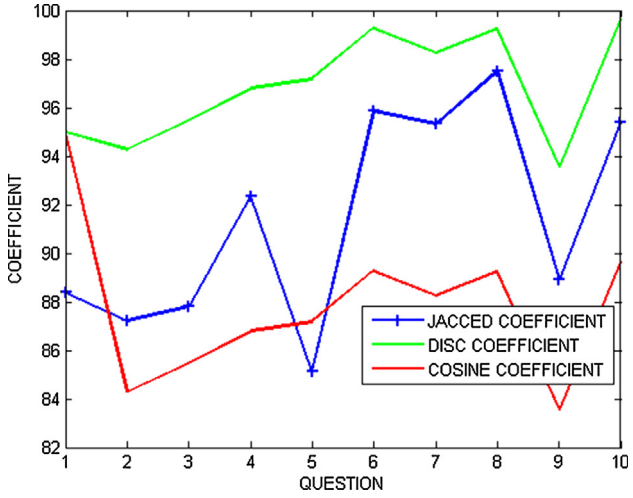

Fig. 5. Correlation between different scoring methods.

3.3. Alignment

The AMT systems return two types of MIDI-level features. The resulting features are combined by concatenation with either 88-note or 12-chroma AMT output features. The corresponding score MIDI was also converted into 88 note (or chroma) and the chroma onsets are elongated in the same manner before combined. We used euclidean distance to measure similarity between the two combined representations. We then applied the FastDTW algorithm [8] which is an approximate method to dynamic time warping (DTW). FastDTW uses iterative multi-level approach with window constraints to reduce the complexity. Because of the high frame rate of the features, it is necessary to employ low-cost algorithm. While the original DTW algorithm has O(N2) time and space complexity, FastDTW operates in O(N) complexity with almost the same accuracy. M̈uller et al. [9] also examined a similar multi-level DTW for the automatic answer scoring task and reported similar results compared to the original DTW. The radius parameter in the fastDTW algorithm, which defines the size of window to find an optimal path for each resolution refinement, was set to 10 in our experiment.

4. Result

The performance of the proposed method is measured by comparing it with the result of Jaccard Coefficient (JC), Dice Coefficient (DC), Cosine Coefficient. The result of the comparison is shown in Table. The table shows the correlation value of the score produced by the system and the score given by the evaluator. The evaluator's

Table 2
Comparative analyses between Jaccard Coefficient, Dice Coefficient and Cosine Coefficient.

No. of questions	Jaccard Coefficient	Dice Coefficient	Cosine Coefficient
1	0.74	0.70	0.79
2	0.48	0.44	0.45
3	0.19	0.19	0.21
4	0.26	0.20	0.26
5	0.30	0.20	0.33
6	0.67	0.65	0.60
7	0.48	0.45	0.44
8	0.60	0.60	0.62
9	0.40	0.43	0.41
10	0.30	0.28	0.30

score is obtained from the average of the score given by two evaluators Table 2.

The performance of the system is evaluated by comparing the output generated by the system with the result given by human raters. The correlation coefficient is calculated and it proves that system score and human raters score are highly correlated with each other i.e. near to one (positive correlation). The value between human raters is quite close to the score agreement value achieved between a human rater and the system. The Table shows the final score of 50 students awarded by the system Table 3.

The following Fig. 6 clearly illustrates the final score of 50 student answer in comparison with the score awarded by 5 human raters. The graph demonstrates that the result produced by the system is the same as that of human raters. A small deviation can be conveniently ignored that demonstrates the accuracy of this system. The system produced score is well within the range of scores awarded by human raters that shows the accuracy and reliability of this system.

Table 3
Final_Score_Table for 50 students.

Student	System's score	Human Rater1	Human Rater2	Human Rater3	Human Rater4	Human Rater5
1	8.7	9	8	8.5	9	7
2	7.2	7	6	7	7.5	6.5
3	10	9	10	8	7	8
4	5.7	8	6	6	6.5	7
5	0	2	0	1	2	1
6	7.5	6.5	8	6	7	8
7	4	3	5	2	4	5
8	9	10	10	8	8	8
9	8.5	7	9	7.5	8.5	8.5
10	2	1	0	2	0	2
11	8	9	6	8.5	9	7
12	3	4	4.5	5	4	3.5
13	8	8	6	8	6.5	8
14	4	6	3	6	6	5.5
15	5.5	5	6	4	4.5	6
16	7.5	6	8	6	8	8
⋮						
49	0	2	0	1	2	1
50	8.5	8	6	8	6.5	8
Correlation coefficient (r)		0.90	0.89	0.87	0.86	0.94

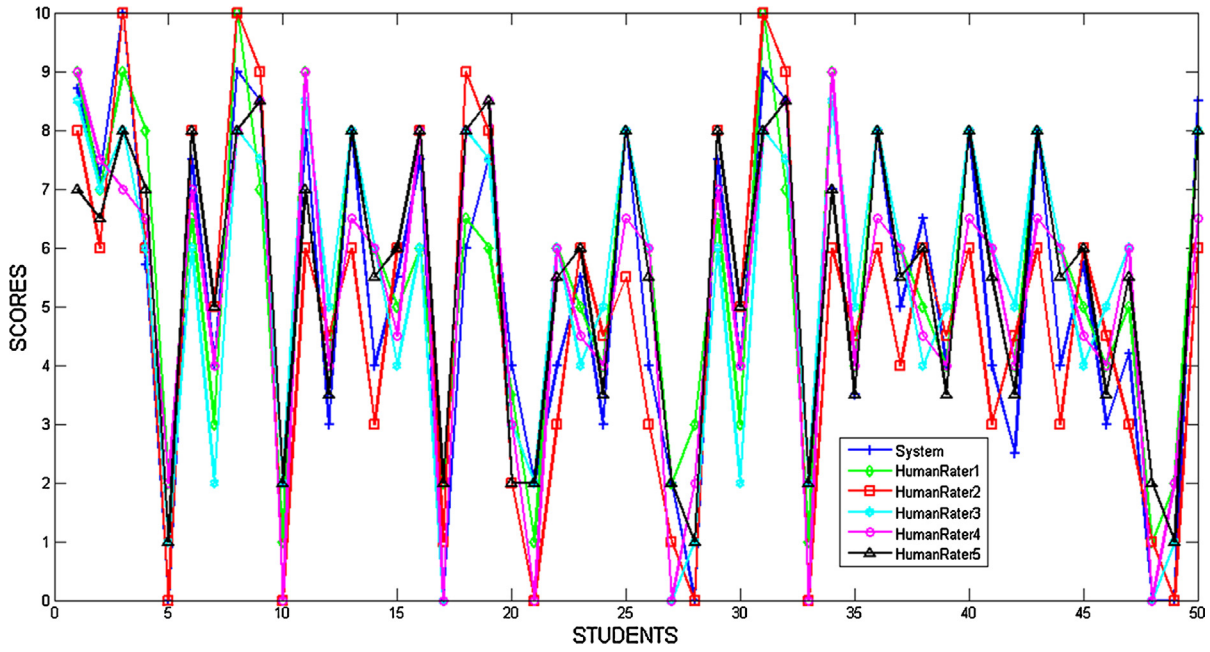

Fig. 6. Comparison among system score of 50 student answer and 5 human raters.

5. Conclusion

In this paper, we augmented the input by concatenating it with the first-order difference of the semitone filtered spectrogram. We observed a significant increase of the transcription performance with this addition. Compared to feedforward neural networks, RNN are capable of learning temporal dependency of sequential data, which is the property found in music answer scoring. Also, LSTM unit has a memory block updated only when an input or forget gate is open, and the gradients can propagate through memory cells without being multiplied each time step. This property enables LSTM to learn long-term dependency. In this paper, LSTM is expected to learn the continuity of onset, sustain and offset within a note as well as the relation among notes. The LSTM units are also set to be bidirectional, indicating that the input sequence is not only presented in order but also in the opposite direction. Throughout backward and forward layers together, the networks can access to both history and future of the given time frame. Comparative analyses show that the proposed technique outperforms existing techniques.

Funding

This study was not funded by any organization.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Arzt G, Widmer S, Dixon. Automatic Page Turning for Musicians via Real-Time Machine Listening. In: Proceedings of the conference on ECAI 2008: 18th European Conference on Artificial Intelligence, 1. p. 241–5. 1.
[2] Alzahrani Ahmed, Alzahrani Abdulkareem, Al Arfaj Fawaz K, Almohammadi Khalid. AutoScor: An Automated System for Essay Questions Scoring. Int J Hum Soc Sci Educ (IJHSSE) 2015;2(5).
[3] Ali Muftah Ben, MohdJuzaiddin. AES for Short Answers in English Language. J Comput Sci 2013; 9(10) ISSN:1549-3636.
[4] HSSINA Badr, Bouikhalene Belaid, Mebouha Abdelkrim. Evaluation of semantic similarity using vector space model based on textual corpus. 13th International Conference Computer Graphics, Imaging and Visualization. IEEE; 2016.
[5] Galitsky Boris. Matching parse thickets for open domain question answering. Data & Knowledge Engineering, 107. Elsevier; 2017. p. 24–50.
[6] Chen, Yen-Yu, Liu Chien-Liang, Chang Tao-Hsing, Lee Chia-Hoang. An unsupervised automated essay scoring system. IEEE Intelligent Syst 2010.
[7] Kerr Deirde, Hamid. Automatic Short Essay scoring using NLP to extract semantic information. J Natl Center Res Eval, Standards, Stud Test (CRESST) 2013.
[8] Perera Govinnage R, Perera Deenuka N, Weerasinghe AR. A dynamic semantic space modelling approach for short essay grading. International Conference on Advances in ICT for Emerging Regions (ICTer). p. 043–9.
[9] Chen Hongbo, He Ben. AES by maximizing human-machine agreement. Proceeding of National/International Conference on Empirical Methods in Natural Language Processing. p. 1741–52.
[10] Jovita, Linda, Hartawan Andrei, Suhartono Derwin. Using vector space model in question answering system. Procedia Comput Sci 2015;59:305–11. Published by ICCSCI 2015 Elsevier.
[11] Zupan Kaja, Bosni Zoran. Automated essay evaluation with semantic analysis. Knowledge-Based Systems, 120. Elsevier; 2017. p. 118–32.
[12] Crowston Kevin, Liu Xiaozhong, Allen Eileen, Heckman Robert. Machine learning and rule-based automated coding of qualitative data. Pittsburgh, PA, USA: ASIST; 2010.
[13] Lee Kong Joo, Choi Yong-Seok, Kim JeeEun. Building an automated English sentence evaluation system for students learning English as a second language. Computer Speech and Language, 25. Elsevier; 2011. p. 246–60.
[14] Kosseim Leila, Yousefi Jamileh. Improving the performance of question answering with semantically equivalent answer patterns. In: Data & Knowledge Engineering, 66. Elsevier; 2008. p. 53–67.
[15] Bin Li, Jun Lu, Jian Yao. Automated essay scoring using the KNN Algorithm. IEEE Comput Soc 2008;1.
[16] Tandalla Luis. Scoring short answer essays. Int J Comp Sci Tech (IJCST) 2012;2 (1).
[17] Heie Matthias H, Whittaker Edward WD, Furui Sadaoki. Question answering using statistical language modelling. In: Computer Speech and Language, 26. Elsevier; 2012. p. 193–209.
[18] Yang Min-Chul, Lee Do-Gil, Park So-Young, Rim Hae-Chang. Knowledge-based question answering using the semantic embedding space. In: Expert Systems With Applications, 42. Elsevier; 2015. p. 9086–104.
[19] Ab Aziz MohdJuzaiddin, Dato' Ahmad Fatimah. Automated Marking System for Short Answer Examination (AMS-SAE). Int J Adv Comput Sci Appl(ISIEA), IEEE 2009;3(11).
[20] Kolomiyets Oleksandr, Moens Marie-Francine. A survey on question answering technology from an information retrieval perspective. In: Information Sciences. Elsevier; 2011. p. 5412–34.
[21] Moreda Paloma, Llorens Hector, Saquete Estela, Palomar Manuel. Combining semantic information in question answering systems. In: Information Processing and Management, 47. Elsevier; 2011. p. 870–85.
[22] Sravanthi Pantulkar, Srinivasu B. Semantic similarity between sentences. Int Res J Eng Technol (IRJET) 2017;04(01).
[23] Williams RJ, Peng J. An efficient gradient- based algorithm for on-line training of recurrent network trajectories. Appears Neural Comput 1990;2:490–501.
[24] Semire DIKLI. Automated essay scoring. Turk Online J Distance Educ-TOJDE 2006;7(1). ISSN 1302-6488, Article: 5.
[25] Hazrina Sofian, MohdSharef Nurfadhlina, Ibrahim Hamidah, Azmi Murad Masrah Azrifah, Mohd Noah Shahrul Azman. Review on the advancements of disambiguation in semantic question answering system. In: Information Processing and Management. Elsevier; 2016. p. 1–18.
[26] Hu Xinming, Xia Huosong. Automated assessment system for subjective questions based on LSI. In: Third International Symposium on Intelligent Information Technology and Security Informatics. IEEE; 2010. doi: https://doi.org/10.1109/IITSI.2010.76.
[27] Li Yali, Yan Yonghong. An effective automated essay scoring system using support vector regression. In: Proceeding of International Conference of Intelligent Computation Technology and Automation(ICICTA). Hunan: IEEE Computer Society; 2012.
[28] Yang. A review of strategies for validating computer-automated scoring, 15. Lawrence Eribaum Assocaites, Inc.; 2002. 4.