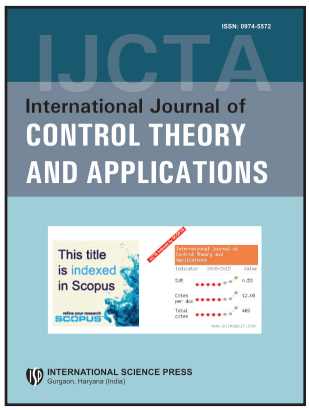

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 43 • 2016

Grading of Diagrams in Answer Scripts Using Support Vector Machine

Panchami K.S[a], Surekha Mariam Varghese[b] and Aby Abahai T[c]

[a,c]Department of Computer Science and Engineering, M.A College of Engineering, Kothamangalam, Kerala, India. Email: [a]panchami.ks1@gmail.com; [b]surekha.laju@gmail.com; [c]abytom@gmail.com

Abstract: Answer Script evaluation and grading is a tedious and time consuming process for many of teachers, due to nonavailability of expert teachers and volume of answer scripts to be dealt with complicates descriptive university student evaluation system. Automated answer script evaluation is remedy to this problem. This type of evaluation results in a faster, more accurate and unbiased way of valuation. Most of the descriptive type answers include figures that contributes certain percentage of marks to the answers. Due to the advances in Natural Language Processing, Currently there exists many automated text evaluation systems. These systems analyses a piece of text based on its semantics, context and spelling. Unfortunately there are very less number developments as of now in the field of diagram evaluation. The proposed system aims to be a stepping stone in this field.

Keyword: Automated Essay scoring (AES), Natural Language Processing (NLP), Optical Character Recognition (OCR), Machine learning and Support vector machine (SVM).

1. INTRODUCTION

Revision and feedback are essential aspects of the writing process, it's necessary to receive feedback from the teacher to improve the writing quality or ability of a student. In recent years, with the growing need of essay scoring in English writing skill, automated essay scoring (AES) has become a hot issue in the research of natural language processing. On one hand, the essay scoring task costs huge human resources but the efficiency is less. On other hand, the essay score given by human rater is mostly determined by rater's personal will, emotion and energy. An essay scored highly by one rater may receive a low score from another rater. Even the same rater probably gives different scores for the same essay at different times. Thus, the correctness of essay scoring cannot be guaranteed. So answer Script evaluation and grading can be tedious and time consuming process for many of teachers. It is a challenging task for many teachers to finish the essay scoring of all student essays in a short interval time. Thus, students cannot feedback on their essays in time, leading to the situation that it is hard for them to improve their writing skill.

In such requirement, researchers proposed automated essay scoring techniques, these techniques has the advantage of fairness, less human resource cost and timely feedback. These systems analyses a piece of text based

on its semantics, context and spelling. The research [1] on Automated Essay Scoring (AES) has revealed that computers have the capacity to function as a more effective cognitive tool. In general, automated essay scoring is a machine learning problem [2] and more specifically, it is a supervised learning problem. Scored essays can be seen as labelled training data and unscored essays as unlabelled test data. The main process of AES system is to learn a model from the training data and then use that model to score essays in the test data. Most of the descriptive type answers include figures that contributes certain percentage of marks to the answers, but there are very less developments as of now in the field of diagram evaluation. In this paper, we regard automated diagram evaluation as a ranking problem and plan to solve this problem by learning to rank algorithms [3]. Learning to rank is a family of supervised learning algorithms that automatically construct a ranking model or function to rank objects. The major advantage of learning to rank is its flexibility in incorporating diverse kinds of features into the process of ranking.

2. RELATED WORKS

A. Text Localization methods

Text localization is the process of detecting text region from whole region. According to the features utilized, it's classified into mainly three categories as shown in Figure 1 region-based and texture-based [4] and hybrid.

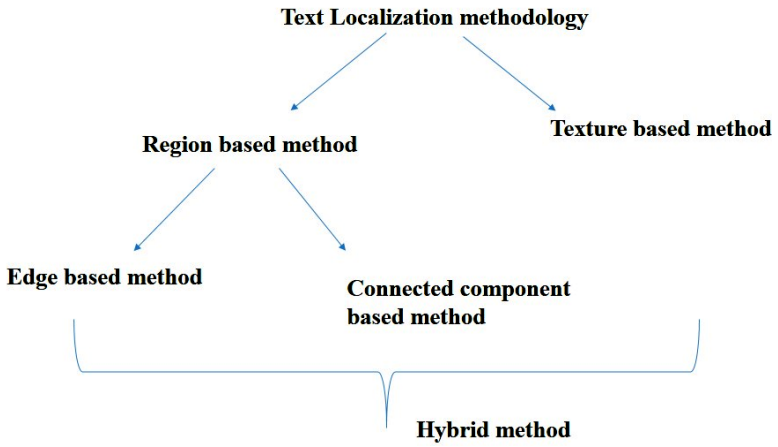

Figure 1: Text localization methods

In Region based Text Localization method, pixels exhibiting certain properties are grouped together. It uses the properties of edges, colors or intensity similarity. This method can be further classified into two: edge-based and connected component (CC)-based methods. Edge based method uses the information of edge to detect text and it finds the edges of the text boundary and merged them together, then alter out the non-text regions. Connected component method uses bottom up approach in which grouping of smaller components are done into larger components until all regions are identified in the image. It focuses similarity criterions of text, such as edge, color, stroke width, size and gradient information, which gather pixels together and form a connected components and filtered out non-texts CCs with geometric hypothesis or conditional random fields (CRFs). Texture-based methods are based on that text in images is different from background in textural characteristic. Hybrid approaches introduce textural property of text regions into region-based approach, take advantages of both region-based approaches which can closely cover text regions and texture-based approaches which can estimate of the coarse text location in cluttered scenes.

B. Automated Essay Scoring

In 1960's research on automated essay scoring system started, after 40 years of research four mature AES systems are commercially available. These systems uses large number of essay features and machine learning algorithms to predict and grade essays. In 1966 the first automated essay scoring system, Project Essay Grading (PEG), is developed by Ellis Page. It is widely used in testing companies, universities, and public schools. The PEG system uses shallow text features of essays and multiple linear regression to learn the scoring function. But scores given by the PEG system and scores given by human raters is proved to be high [3].This is because, PEG system only considers shallow text features of essay while ignores the essay content, leading that it is easy to be cheated by students. In the last of 1990s two automated essay scoring system are developed, one of them is Intelligent Essay Assessor (IEA), is developed. The IEA systems scores essay by measuring the semantic features and computes the semantic matrix of scored essays and unscored essays by a semantic text analysis method named Latent Semantic Analysis (LSA) [3]. Another one is E-rater, developed by Educational Testing Services (ETS) in America, and has been currently used for essay scoring in the Graduate Management Admissions Test (GMAT). In 2003 another mature commercial automated essay scoring system is developed based on artificial intelligence, IntelliMetric by Vantage Learning Company. These system extracts more than 300 text features, including both shallow text features and deep text features, feature extraction is complicated.

C. Introduction to Learning to Rank

Learning to rank is also called machine learned ranking is introduced to solve ranking problems in information retrieval [5]. In recent years, due to rapid growth of information searching for the information we need in the Internet has become more and more difficult. The central issue in modern information retrieval system is to rank the retrieved documents. The initial solution to this problem, a scoring function or model is build based on measurement. Recently, with the development of machine learning algorithms, researchers try to apply these techniques to solve the ranking problem and have introduced many innovative and effective ranking models. This research area is known as learning to rank. It is a type of supervised or semi-supervised machine learning algorithm that automatically construct a ranking model or function from training data. Current learning to rank algorithms fall into three categories, that is, the point-wise, pair-wise, list-wise approaches. Point-wise approach takes individual documents as training examples for learning a scoring function. In fact point wise approach can be modeled as both multiple linear regression and support vector regression (Vapnik et. al., 1996). Pairwise approaches process a pair of documents each time and modeled as ranking as a pairwise classification problem and the loss function is always a classification loss. Representative algorithms are ranking SVM (Joachims, 2006), RankNet (Li et. al., 2007), (Yannakoudakis et. al., 2011) apply pairwise approach and ranking SVM, to automated essay scoring achieve better performance than support vector regression. In list wise approaches, ranking algorithms process a list of documents each time and the loss function aims at measuring the accordance between predicted ranking list and the ground truth label. Representative algorithms are LambdaMart (Wu et. al., 2008), RankCosine (Qin et. al., 2008), etc.

3. PROPOSED WORK

In this section, we will firstly give architecture of proposed system. Then, we discussed the details of the key step in automated diagram evaluation system and of how to evaluate diagrams present in answer. The system uses Machine learning, Image processing and Natural language processing to effectively evaluate diagrams present in answer scripts. Implementation of the system includes different modules from text localization to final valuation of the answer that include scoring of the diagram. These modules are executed in a sequential manner for the proper implementation. Proposed system architecture is shown in Figure 2 as given below. Here system modules are executed in mainly two phases, training and evaluation. During training time outputs from all modules are saved in a file, which is the data used for evaluating new diagram, this is done in evaluation time.

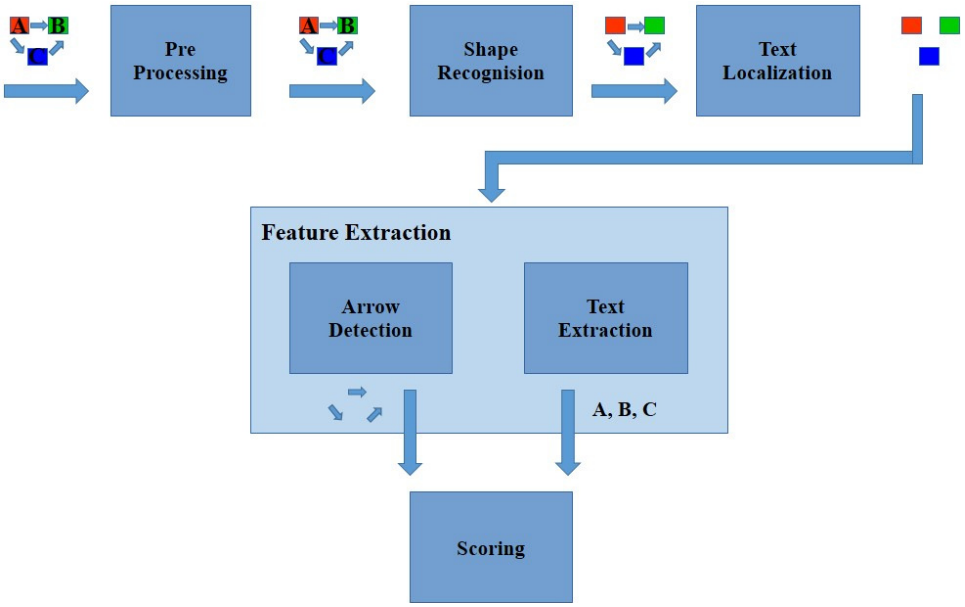

Figure 2: System Architecture

A. Pre-processing

Pre-processing is the initial step, it's the process of removing deficiencies in raw data. Here this is applied for image enhancement for future purpose. There are many pre-processing steps are exists, among that we choose adaptive thresholding [6], which typically takes gray scale or colour image as input and outputs a binary image. For each pixel in the image, a threshold has to be calculated and if the pixel value is below the threshold it is set to the background value, otherwise it assumes the foreground value. The pre-processed image will be the input to further processing.

B. Shape Recognition

In a diagram image different block shapes may be present triangle, square etc., which is an important factor needs to be considered in diagram evaluation. Many shape recognition algorithms have been proposed earliest and detailed survey of shape recognition algorithms can be found in [7], [8]. Here we choose contour detection algorithm using opencv as for identifying different shapes associated with diagram. Contours can be explained simply as a curve joining all the continuous points (boundary), having same colour, and intensity or pixel value. Contours are a key property for shape analysis and object detection and recognition. Contour approximation method is the key property to recognize block shapes. In which initially determines boundary of shape with same intensity, then it stores the (x, y) coordinates of the boundary of a shape. All boundary points are need not to be saved, it removes all redundant points and compresses the contour, thereby saving memory. Below image of a rectangle demonstrate this technique.

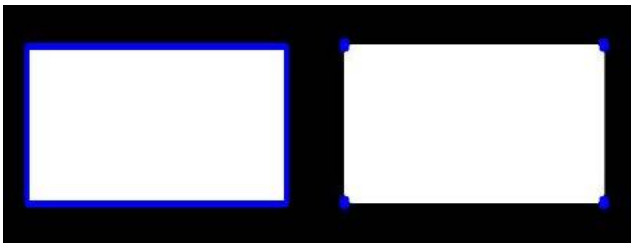

Figure 3: Shape recognition

C. Text Localization

The second key step is text localization, determines text region where the text present. Here input will be the output from shape recognition. Text present diagrams is present inside the blocks, so block boundary is need to be identified first. That will be done during shape recognition. The detected text regions is saved into separate image for further processing. And also we determines centroid point of the each detected text region. This is done by using image moments property. Using image moments we can calculate some features like center of mass of the object, area of the object etc. Image moments are calculated by using given equation,

$$Cx = M10/M00 \text{ and } Cy = M01/M00 \quad (1)$$

D. Text Extraction

Text extraction is the stage where the text components are segmented from the background, in order to facilitate its recognition. In this phase text content is extracted from each text region detected. For text extraction purpose there exists a technology named Optical Character Recognition Engine (OCR). It takes as input an image or image file and outputs a string. The system uses google tesseract OCR for this purpose [9]. Here the input will be detected text region saved as separate image, each image processed through tesseract and outputs text present in image. Extracted words saved along with appropriate contour, i.e. text corresponds to a contour saved in itself. This will be used in future purpose.

E. Arrow Detection

The input in arrow detection stage is arrow present in image only, remaining portion present erased for better output. Here also contour properties are used for intended purpose. Shape detection is started by detecting contours present in the input, then we found the area of each contour detected. Our aim is to determine head and tail portion of each arrow. And we use a mathematical concept that contour area <90 obviously the head area and remaining portion is the line portion. From the remaining part we have to find out the appropriate line portion correspond to the head portion. For that we find four extreme point's extreme left, extreme right, extreme bottom and extreme top. Then find out the smallest distance between head and these extreme points using Eucledian distance formula.

$$d(X_i, X_j) = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + ... + (X_{im} - X_{jm})^2} \quad (2)$$

where, $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. $x_j = (x_{j1}, x_{j2}, \ldots, x_{jm})$ x_i and x_j are data objects with m dimensions

Line with smallest distance from head will be line corresponding to the head. Next step is to detect the tail point in line, for that same method is used but instead of smallest distance we consider the largest distance. Output from this phase is head and tail point corresponding to an arrow. `And also find out contour corresponding to head and tail in an arrow.

F. Feature Extraction

In this phase explains how to extract features that is used as key points to evaluate diagrams. During training time training samples will passed through above phases and output will be saved. After training, a new diagram is passed to evaluate. Then we will compare outputs from evaluation phase to training phase. Comparison is carried out in both arrow detected and text extracted. In text comparison, check the text extracted in training is present in text extracted in evaluation, and if text is present information will be saved. Arrow comparison is done by comparing text correspondence to head and tail of the arrow of training sample is same as the question diagram. If it's same info will be saved. All these extracted info passed to next phase ranking/scoring.

G. Scoring

In this phase an appropriate score is given to diagram, using machine learning technique support vector machine. "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges [10]. Firstly, we take training set as a list of features of diagrams in training. Then, we apply the ranking model to the test set and get a list of scores of diagrams in test set. Here we call score given by ranking model predicted score. Then we calculate the actual score depends upon the percentage of error threshold.

4. CONCLUSION

There exist many text evaluation, unfortunately there are no developments as of now in the field of diagram evaluation. Proposed system efficiently evaluate diagram by considering text present in diagram, semantics of the text, Blocks shape and arrow direction. The system aims to be a stepping stone in the field of diagram evaluation

REFERENCES

[1] Mark D. Shermis and Jill C. Burstein, "Automated Essay Scoring: A Cross-disciplinary Perspective", 1st ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2002.

[2] Larkey and Leah S., Automatic essay grading using text*categorization techniques*. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 90-95, 1998.

[3] Hongbo Chen, Ben He, Tiejian Luo and Baobin Li A Ranked-based Learning Approach to Automated Essay Scoring. 2012 Second International Conference on Cloud and Green Computing, November 01 - 03, 2012.

[4] Pooja B, Archana Ghotkar,"A Survey on Text Localization Method in Natural Scene Image", *International Journal of Computer Applications (0975 – 8887),* Volume 112 _ No 13, February 2015.

[5] Hang Li. "A short introduction to learning to rank" IEICE Transactions on Information and Systems 94-D (10):1854-1862 · October 2011.

[6] Gerhard Roth, Drerek bradley, "Adaptive thresholding using integral image" Journal of Graphics Tools, Volume 12, 2007 - Issue 2.

[7] M. Hagedoorn, "Pattern Matching Using Similarity Measures", PhD thesis, Universiteit Utrecht, 2000.

[8] R. C. Veltkamp and M. Hagedoorn, "State of the Art in Shape Matching", Technical Report, Utrecht, 1999.

[9] Ray Smith Google Inc,"An Overview of the Tesseract OCR Engine"

[10] Simon Tong, Daphne Koller,"Support vector machine active learning with applications to text classifications", Journal of Machine Learning Research (Nov):45-66, 2001.