Towards Automated Evaluation of Handwritten Assessments

Vijay Rowtula
CVIT, IIIT Hyderabad
vijay.rowtula@research.iiit.ac.in

Subba Reddy Oota
IIIT Hyderabad
oota.subba@students.iiit.ac.in

C.V. Jawahar
CVIT, IIIT Hyderabad
jawahar@iiit.ac.in

***Abstract*—Automated evaluation of handwritten answers has been a challenging problem for scaling the education system for many years. Speeding up the evaluation remains as the major bottleneck for enhancing the throughput of instructors. This paper describes an effective method for automatically evaluating the short descriptive handwritten answers from the digitized images. Our goal is to evaluate a student's handwritten answer by assigning an evaluation score that is comparable to the human-assigned scores. Existing works in this domain mainly focused on evaluating handwritten essays with handcrafted, non-semantic features. Our contribution is two-fold: 1) we model this problem as a self-supervised, feature-based classification problem, which can fine-tune itself for each question without any explicit supervision. 2) We introduce the usage of semantic analysis for auto-evaluation in handwritten text space using the combination of Information Retrieval and Extraction (IRE) and, Natural Language Processing (NLP) methods to derive a set of useful features. We tested our method on three datasets created from various domains, using the help of students of different age groups. Experiments show that our method performs comparably to that of human evaluators.**

***Keywords*-automated evaluation; handwritten answers; self-supervised learning; deep learning;**

I. Introduction

Handwritten document analysis is explored in various ways in computer vision. Recent advances in deep learning and NLP has facilitated such tasks [1]. One such need is to automate the evaluation of students handwritten answers in schools and colleges. Scalable and reliable methods for evaluating the student performances critically lack in today's massive virtual as well as sizable real classrooms. As a result, instructors have to resort to simple boolean or multiple-choice questions. It is a known fact that the handwritten responses are a reliable means to check the comprehension levels and the expressive skills of the students [2]. They also reflect the student's traits (e.g., concentration, logical organization of the thoughts, etc.) in a useful way. Evaluating large numbers of handwritten answers from student tests is a time-consuming, monotonous and costly task. An effective automatic evaluation system can contribute a lot to the teaching/learning process in different ways. Such a solution can prune the answers from a large class to a smaller number, and use the limited human resources judiciously. Even minimal support from the automatic evaluation system, like keyword highlighting, can speed up the evaluation task. Instructors can also provide a quick glimpse of evaluation and feed-backs from the students before giving a final evaluation score.

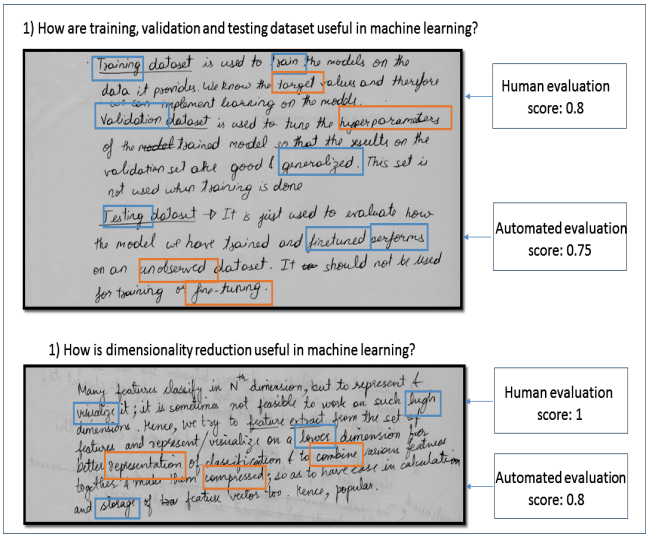

Figure 1: We try to assign a quantitative score to a handwritten answer that matches with the score assigned by a human evaluator. The figure depicts two examples where answers from datasets evaluated by a human evaluator and by our assistive evaluation framework. To evaluate automatically, we match the keywords that are present in the textual reference answer (blue box) as well as those that are not directly provided (orange).

Solving a similar problem such as evaluation of handwritten essays was earlier attempted using scoring features derived from reading comprehension research [3]. However, such features may not be effective when dealing with handwritten short answers, where students answers are usually limited to a few keywords. In this work, we tried to design a solution that helps in automatic evaluation of the answers as illustrated in Figure 1. Unlike text-based automated short answer evaluation, where performance of automated evaluation has seen a great boost using keyword recognition by word embedding, this task has been rarely pursued. As highlighted in the figure, we detect the keywords which include (a) keywords from the textual reference answer and (b) semantically relevant keywords obtained using information retrieval and (NLP) methods, to derive the evaluation

score. Our results show that our method score predictions are on par to human evaluation scores. We believe that such an automatic evaluation solution can help the large-scale evaluation that the modern educational systems demand.

We limit our attention to evaluating the handwritten answers digitized as images. Our use case is an online system where students upload the handwritten answers as images digitized by their mobile phones or a scanner. As a preprocessing step, we segmented the handwritten datasets into word images and annotated them with their equivalent text. It helped us to concentrate only on the aspect of the automated evaluation of students answers and not dealing with issues of automated word segmentation from handwritten documents. We borrowed ideas from information retrieval, document image analysis and NLP for automated assessment. In the rest of the paper, we present (i) a word spotting based automatic evaluation solution based on the deep learned features (Section II). (ii) A self-supervised enhancement of the word spotting (Section III-A). (iii) A set of features in the image space that captures the semantics and scores computable in the image space (Section III-D) and, (iv) experimental validation on a set of student answers from a real classroom (Section IV).

A. Related Works

Automated evaluation of assessments is an active area of research in the text domain. A multitude of measures were proposed for computing similarity between the reference answer and the candidate answer in the past, based on semantic content features [4], [5]. Various linguistic aspects of the sentences were covered using knowledge-based features [6], corpus-based features [7], alignment-based features [8] and, literal-based features [4]. Though the text-based automatic evaluation is nearing the reliable deployment in the university education system, handwritten answers are not yet amenable for their processing. Evaluation of handwritten answers needs a significant advance in computer vision algorithms (e.g., word segmentation and recognition). A natural direction to evaluate the handwritten answers is to convert into textual content and then exploit the advances in the text-based automatic evaluation. While the optical character recognizer (OCR) can reliably recognize printed text, offline handwritten text recognizer for unconstrained vocabulary are not robust enough for the practical use due to the inherent complexity of a handwritten word image.

However, one can resort to image-based matching methods (popularly known as word spotting [9]) for matching with the textual content. More recently, with popularization of deep architectures [10], [11] and introduction of synthetic data [1] for training, there has been a significant improvement in both recognition and word spotting in multi-writer handwritten documents. In this work, we capitalize on this success of deep features and develop our automatic handwritten evaluation framework. There have been only fewer attempts to address the problem of handwritten text assessments. Srihari [12] proposed a method for automatic scoring of short essays from reading comprehension tests. They presented an end to end pipeline with handwriting recognition, tri-grams based contextual post processing and scoring methods using a latent semantic analyzer and a neural network. This semi-supervised evaluation approach with handwriting recognition can have errors in transcription thereby reducing the accuracy of the system. Other attempts [13] in this space are also restricted to handwritten comprehension with semi-supervised evaluation and does not discuss much on context analysis.

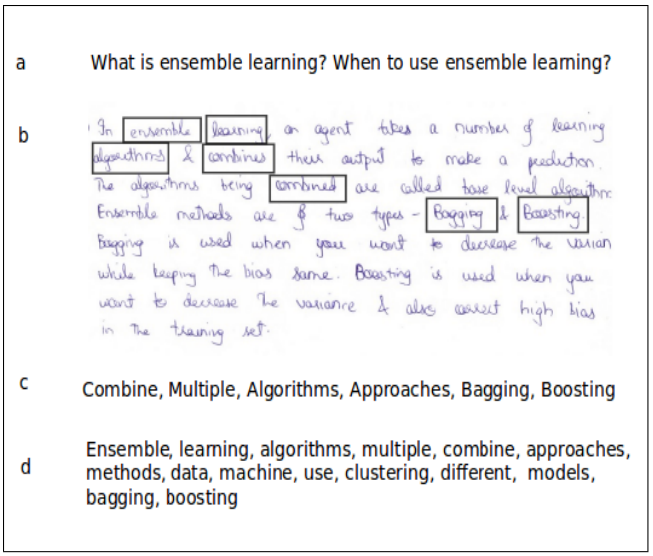

Figure 2: A sample answer. a) Question from university exam, b) student's handwritten answer, with word spotting, c) keywords from textual reference answer and, d) keywords after query expansion.

II. Scoring by Word Spotting in Images

We developed our scoring model based on a word spotting. Here, our interest lies in finding the matching score between the keywords associated with the Textual Reference Answer (TRA) and Handwritten (HW) document images, written by different writers in an unconstrained setting. Word spotting is typically formulated as a retrieval problem where the query is an exemplar image (query-by-example), and the task is to retrieve all word images with similar content. It uses a holistic word image representation which does not demand character level segmentation. Many of the popular features [14] are limited for the multiple-writer scenarios due to high intra-class variations. Such a problem is now successfully addressed using CNN features [1], [15] for handwritten word images. In this work, we used architecture inspired by HWNet-v2 [1] which is pre-trained on a large corpus of synthetic handwritten word images and later fine-tuned on IAM dataset [16]. The HWNet-v2 is a ResNet34

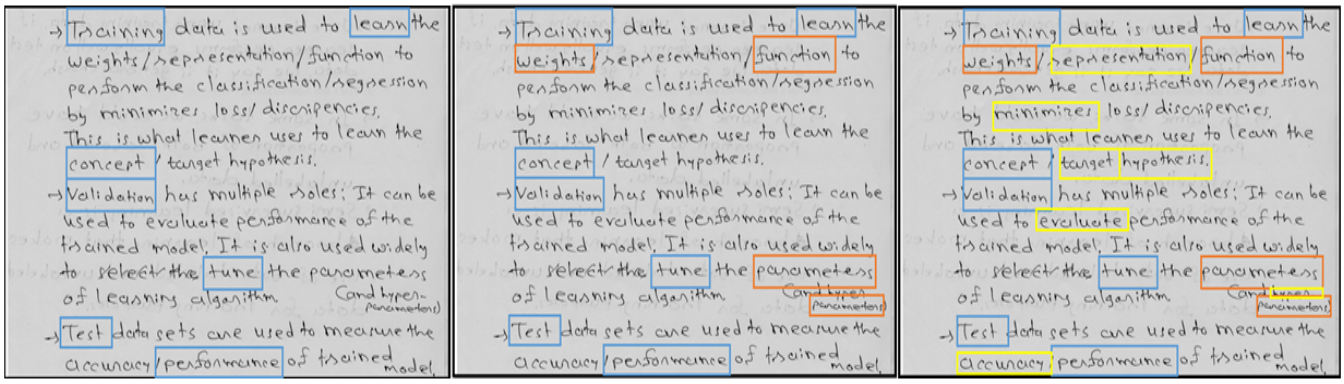

Figure 3: Samples of word spotting improvements with our context retrieval enhancements. i) Word spotting with ground truth keywords, ii) with query expansion and, iii) LDA with query expansion. We observe an improvement in the number of keywords detected for question "How are training, validation and testing datasets useful in machine learning?"

network with 4 ResNet blocks and two fully connected (FC) layers as penultimate layers instead of global average pooling, as proposed in original ResNet architecture [17]. The model is further fine-tuned on the training datasets created by us (Section IV-A) to learn the natural variations in writer styles.

A. Keyword Extraction

A primary source of keywords for word spotting is the Textual Reference Answer (TRA) provided by instructors for each question. Keywords are either manually annotated by the examiner from TRA or extracted from TRA using NLP techniques. From the linguistic aspect, the building blocks of a sentence is a noun phrase (NP) and a verb phrase (VP). NP represents topics or subjects/objects in a sentence, while VP describe some action between the subject/objects in a sentence. We used the keywords from both NP and VP since they can sufficiently describe the topic and hence the context is derived from them. We used Stanford core NLP tools [18] like POS tagger and sentence parser to extract keywords from textual reference answer. The keywords thus extracted are further filtered by the examiner by intuition and experience, if required.

We match the keywords in the image space. For image matching, we synthesize the images from keywords of TRA using multiple synthetic fonts. Given the keyword images, we extract the corresponding features from a model trained on word spotting task. Later, we do word spotting on segmented answer images from answer sheets using nearest neighbor search with a threshold for image matching set empirically. We observed that our model performs with an accuracy of 82% on word-spotting task on our dataset (more details later).

Although the performance seems reasonable, we show in the next section that given the nature of our problem, we can further improve the word spotting performance by restricting the vocabulary to a particular domain. A grading framework solely dependent on keywords from textual reference answer would be unable to detect semantically relevant keywords, thus marking multiple answers invalid. Figure 2 demonstrates an example of a handwritten answer with just the reference answer based keywords and semantically related keywords. In the next section, we present our enhancements to address these issues.

III. Enhancements

A. Self Supervised Word Spotting

It is a well-known fact that CNN trained for a related task could be adapted or fine-tuned to get reasonable and even state-of-the-art performance for new tasks [19]. In our case, we use a similar strategy where we reformulate the problem of word spotting from generic vocabulary to word classification limited to question/reference answer specific keywords. While grading a specific question, we are interested in doing accurate word spotting only on a set of words that are semantically related to the TRA (discussed in Section III-B). Since the domain of keywords for a specific question is limited (approximately 5-25 words), we fine-tune the model to spot these limited keywords more accurately. We froze all the layers of the model (discussed in Section II) except the FC layers, replacing softmax layer to match the number of new keywords and fine-tune the model with very low learning rate. For generating the training data automatically from the keywords of TRA, we use synthetic handwritten fonts as suggested in [1]. This process repeats for every new question and its reference answer (TRA). We refer this as self-supervised word spotting where the entire process happens without any external human supervision.

B. Contextual Query Expansion

Word spotting using keywords from TRA provides baseline scores for the evaluation. However, students are likely to use paraphrasing with synonyms and acronyms in answers

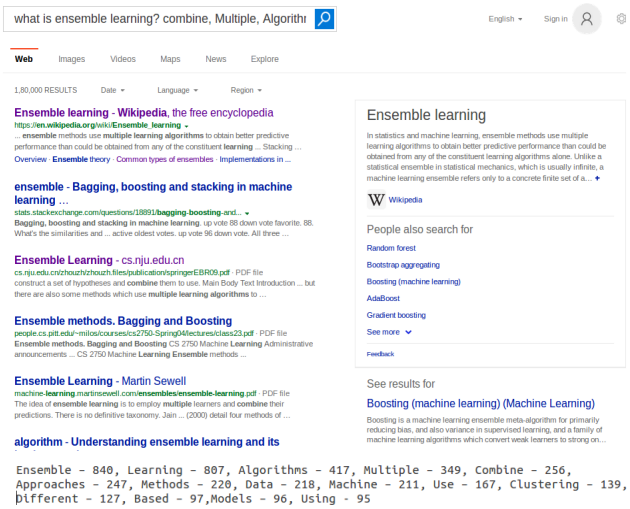

Figure 4: Example of results obtained from querying the search engine. We can observe contextually relevant terms in definitions along with query terms.

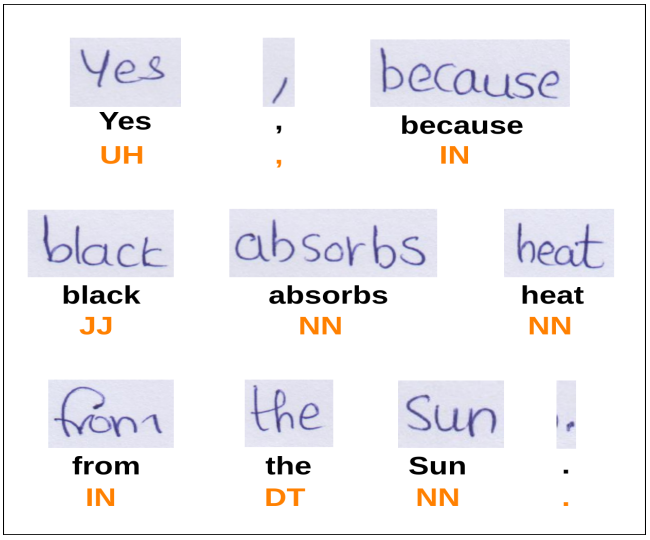

Figure 5: The figure shows an example from the SE dataset where words are classified into POS tags. Word images with the transcribed text and their POS tags are available during train and testing.

which can make automatic evaluations difficult. Alternatively, we can expand keywords using knowledge-based sources like WordNet and Thesauri but can result in false positives due to underlying ambiguity in word senses which could be only resolved by understanding the context. Other sources like Wikipedia articles, query reformulation logs and search results obtained from the web (together called as corpus-based sources) provides a set of contextual texts that are used to expand the original sparse keyword representation [20]. In our experiments, we use web search results to expand our query representations.

Query expansion is formulating a given query to retrieve a relevant document or information retrieval. It involves finding various semantically related words from words in a query such as synonyms, antonyms, meronyms, hyponyms, and hypernyms. It also involves a pre-processing step of stemming the queried words and automatically fixing the spelling errors. We observed that the keywords embedded in a question and textual reference answer could help in understanding the context and hence narrow down extraction of contextually relevant information significantly. We run constructed query of words against a Bing search engine's index and retrieve the top 500 documents [21]. The titles and descriptions from results are then concatenated and used as our expanded keyword representation.

In Figure 4, we show portion of the expanded representation for the short text segment "ensemble learning". As we see, this expanded representation has many contextually relevant terms, such as "Bagging", "Boosting" and "AdaBoost" that are not present in the surface keyword representation. To pick the most informative keywords from these results, we first weight each expanded keywords using TF-IDF scores and select only the top-N words. In another approach, we considered a query as "topic" and Latent Dirichlet Allocation (LDA) [22] is used on query results (documents) to form a cluster of words that often occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. We used MALLET framework [23] for topic modeling.

C. POS Tagging and NER

Despite the usage of a good semantic query expansion methods, we may not to retrieve the necessary keywords every time. Since we are working on automated short answer evaluation, keywords are not always relevant. Boolean answers are not uncommon in assessments and at times an adverb like "not" can change the meaning of the answer despite presence of keywords. Hence, parts of speech (POS) tagging and named entity recognition (NER) on handwritten document images are helpful as an extra set of features for automated evaluation. POS and NER tagging is a NLP problem, to parse a sentence and assign parts-of-speech tags per word and classify the words into pre-defined entity categories such as the names of people, streets, organizations, dates, etc. POS tagging and key phase detection (Figure 5) from document image is quite difficult without transcription to text. However, such detection is essential since handwritten text recognition is not yet perfected and hence NLP tools cannot be used directly [24]. We used POS tags and named entities spotted from the student's answers as additional features to model and automatically evaluate the student handwritten answers. For this, we used

	CRD dataset			CD dataset			SE dataset		
Experiments	P	R	F1	P	R	F1	P	R	F1
Base Keywords	0.61	0.78	0.68	0.84	0.80	**0.82**	0.72	0.63	0.67
QE on Question	0.65	0.70	0.67	0.71	0.54	0.61	0.70	0.62	0.66
QE on Question & TRA	0.70	0.75	0.72	0.73	0.55	0.63	0.68	0.60	0.64
TF-IDF based QE	0.71	0.76	0.73	0.71	0.48	0.57	0.70	0.68	**0.69**
LDA based QE	0.71	0.78	**0.74**	0.70	0.51	0.59	0.71	0.65	0.68

Table I: The table show results for all experiment methods using **base features** on CRD, CD and SE datasets. The experiments are listed on the left. QE stands for query expansion, P for precision, R for recall and F1 for F1-score.

	CRD dataset			CD dataset			SE dataset		
Experiments	P	R	F1	P	R	F1	P	R	F1
Base Keywords	0.67	0.79	0.72	0.64	0.75	0.69	0.63	0.72	0.67
QE on Question	0.65	0.69	0.67	0.64	0.73	0.68	0.64	0.75	0.69
QE on Question & TRA	0.70	0.72	0.71	0.63	0.79	**0.70**	0.66	0.64	0.65
TF-IDF based QE	0.69	0.85	**0.76**	0.71	0.60	0.65	0.66	0.75	0.70
LDA based QE	0.71	0.77	0.74	0.62	0.78	0.69	0.70	0.74	**0.72**

Table II: The table show results for all experiment methods using **semantic features** on CRD, CD and SE datasets. The experiments list is on the left. QE stands for query expansion, P for precision, R for recall and F1 for F1-score.

a method described in [25]–[27] where, a CNN + RNN model architecture is used to take the advantage of sequential knowledge in successive word images. We trained a similar architecture on IAM dataset to detect POS tags and named entities directly from word images segmented from the handwritten text without transcribing word images to text. We used 58 unique POS tags and 6 named entities obtained using python based NLP tool named Spacy for tagging on the datasets.

D. Features for Grading

Our aim is to design a solution that assigns a quantitative score that is very similar to the score assigned by a human instructor. We do this by training a neural network in a supervised way on a set of features described below.

Base Features: The keywords spotted from TRA in a student's handwritten answer is the essential clue of its proximity to the textual reference answer. We capture this with (i) ***unique terms***: the count of unique keywords from TRA spotted in the students answer. (ii) ***keyword recall***: the ratio of *unique terms* spotted to count of actual keywords in TRA and (iii) ***word count***: the number of words segmented from the text. We refer to these three features as the BASE FEATURES.

Lexical Features: We also capture the features related to the lexical complexity. They are (iv) ***tokens***: the total number of terms from the ground truth keywords (from TRA) spotted, including term repetitions This feature characterizes the student's domain vocabulary knowledge (and not the common words). These features are like noun phrases and repeated n-grams [5] captured by a parser on a transcribed text. (v) ***unique terms - token ratio***: the ratio of the number of the unique terms spotted, to that of *tokens* [28]. The purpose of this feature is to capture the excessive use of keywords to enlarge the answer artificially instead of the precise description.

Syntactic Features: We use the following features to capture the **syntactic clues** from the images using word spotting. (vi) ***words length***: a simple word count obtained after segmentation of handwritten answer image after filtering out anomalies based on word image size. (vii) ***term strength***: the purpose of this feature is to count the number of unique terms in the answer and standardize this count with the total number of words in the essay. (viii) ***token strength***: the purpose of this feature is to count the tokens in the answer and standardize this count with the total number of words in the essay. It captures the strength of prioritized usage of the contextual words instead of simple words.

NLP Features: We capture the semantic clues by measuring the organization of the answers in terms of the presence of named entities and its supporting keywords in phrase or sentence. We used the method described in Section III-C to classify the words into their respective POS and named entity tags. We used the following features to capture the semantic clues. (ix) ***nouns phrase ratio***: ratio of nouns and adjectives spotted in students answer, with respect to nouns and adjectives in textual reference answer (TRA). (x) ***verb phrase ratio***: ratio of verbs and adverbs spotted in students answer, with respect to verbs and adverbs in textual reference answer (TRA). (xi) ***named entities match count***: total count of named entities matched between students answer and textual reference answer. Features described from (iv) to (xi) are together referred as SEMANTIC FEATURES.

With all these features computed from the student handwritten answers, we train a simple multi-layered neural network to predict the human score. We trained the network using mean squared error (MSE) loss and stochastic gradient descent (SGD optimizer to predict a score in the range $[0, 1]$.

IV. Experiment Results and Discussion

A. Datasets

To validate our method, we collected handwritten answers to a set of questions from school and college students. We selected questions from three domains: machine learning, operating systems, and basic science. We choose these domains due to the matured vocabulary of these areas and presence of enough Internet resources. The questions are mostly descriptive, listing or differences based. Typical answers are one to four sentences long. Examples of questions in our dataset are: (a)"What are the roles of training, validation and test datasets in machine learning?" (b)"Why is dimensionality reduction is very popular in many machine learning solutions as a pre-processing step?" Examples of handwritten answers is shown in Figure 3. In all these cases, a human evaluated the answer first, and the human score is normalized to $[0, 1]$, and used as a signal for the supervision or the evaluation. We created corresponding textual reference answer and textual students answers separately for validation.

*Class Room Dataset (*CRD*):* This dataset consists of answers from an actual university examination. We describe the details in Table III. This dataset consists of a set of 6 questions answered by 96 students in an examination. The total number of answers extracted is 576. An independent human evaluator HE provided a score $[0, 1]$ based on the correctness of the answer.

*Controlled Dataset (*CD*):* We created this dataset in an artificial class environment wherein 15 students participated to answer 10 questions. This dataset has simple questions, to imitate complexity of questions in high schools and colleges. As described in Table III, we obtained a limited dataset of 150 answers from this exercise. This dataset have images, their corresponding text and the human scores.

*SciEntsBank Dataset (*SE*):* The textual corpus was created as a part of Joint Student Response Analysis and Recognizing Textual Entailment Challenge in text domain [29]. The task is to develop models for automating the assessment of student responses to questions in the science domain. Of the two datasets provided, we used SciEntsBank Dataset (SE) for our third experiment, since this dataset contains a single reference answer provided by an expert instructor to every question and a clear demarcation in answer evaluation. The evaluation of datasets are given in three formats: i)2-way, ii)3-way and iii)5-way evaluation schemes where labels focused on correctness and completeness of the response content. We evaluated student answers against the reference answer, using the 2-way evaluation scheme which classifies the answer either as "correct" or "incorrect".

The SciEntsBank test corpus has about 5835 responses to 196 assessment questions in 15 different science domains. The test corpus is further divided into Unseen Answers (UA), Unseen Questions (UQ) and Unseen Domains (UD). We selected a subset of 69 questions from complete test corpus based on simplicity of answers and converted the corresponding multiple textual answers provided per question in the dataset, into 3152 handwritten student answers with the help of 12 students. We chose this dataset due to its relevance in the research community for ASAG task. This dataset also covers a broader domain of science and not just subject based question answers as in our earlier datasets.

Controlled	**Count**
No. of Students	15
No. of Questions	10
Total Answers	150
Class Room	
No. of Students	96
No. of Questions	6
Total Answers	576
SciEntsBank Handwritten	
No. of Students	12
No. of Questions	69
Total Answers	3152

Table III: Details about the datasets used in our experiments - Controlled, Class Room and SciEntsBank.

B. Evaluation Methodology and Metrics

We quantitatively evaluated performance of the automatic evaluation (AE) exhaustively. The experiments do not consider the accuracy of segmentation in reporting evaluation metrics. We compare performance of our solution with that of human evaluation (HE) in the following way. First, we normalize the AE and the HE scores to a binary $[0, 1]$ value to reflect notation of "correct" and "incorrect" answers. Note that the AE and the HE scores are in the range of $[0, 1]$. Since even a low score from HE reflects certain degree of correctness in students answer, we lowered the threshold θ to 0.25 from 0.5 when converting to a binary range. Automatic evaluation is valid, if both the human and algorithm scores match. Otherwise, we consider AE as incorrect. We then compute, precision, recall and F1-score for the automatic evaluation.

C. Qualitative Results

We conducted 5 different experiments based on the keywords from TRA and keywords using different query expansion methods described in Section III-B. These experiments were conducted first with BASE FEATURES and then with SEMANTIC FEATURES, as shown in Table I & II. The first experiment **Base Keywords** was with keywords from TRA. In the second, both verb phrases and noun phrases extracted from the question are used in **Query Expansion on Question** experiment. This experiment sets the platform for unsupervised evaluation where keywords are from the question but not TRA, and therefore human intervention is not required fro creating a TRA. We used bing search API to query the keywords from the question. The results were tokenized, converted to lower case, stop words were removed,

and top 15 most repeating words were extracted and used as query words for word spotting. The model is trained on features obtained from the expanded representation. In the **Query Expansion on Question &** TRA experiment, the relevant query words are extracted from web, based on the keywords from both question and TRA. An example of query expansion is seen in Figure 4. Not all keywords are equally important in the context of a question. Hence, we performed a **Weighted Query Expansion** experiment with top-N keyword weights calculated from search result documents using TF-IDF scores. We conducted another experiment using LDA **based Query Expansion** from search results.

Base Features based Evaluation: We demonstrate the assessment performance using just BASE FEATURES obtained using the method described in Section II. We used total word count, unique keyword count and keyword recall as features for training and testing the model. Each dataset is split into training and testing sets, and we use the prediction from trained model to evaluate the answer as valid or invalid. Prediction probability, which is in the range of 0 and 1 is used as our grading score, as described in Section IV-B. From Table I, we observe high precision scores across most of the experiments. We observed better performance using query expansion methods on CRD (using LDA) and SE (using TF-IDF) datasets, but CD dataset has a better score with base keywords. We attribute this due to presence of more definition and list-based questions in CD dataset, where keywords from TRA are sufficient and may not need query expansion. We observed that the baseline method perform poorly on the dataset of higher complexity (SE).

Semantic Features based Evaluation: In the second set of experiments, we added semantic features mentioned in Section III-D in addition to baseline features. From Table II, it is evident that the accuracy of semantic features is better than base features for the complex CRD dataset. We also observed high recall scores across most of the experiments. We argue that this is probably due to combination of an increase in the number of features and keyword coverage by query expansion methods. From the Table II, we observe better performance using query expansion (LDA specifically) methods on all the datasets. These experiments prove that topic modeller trained on search query documents and weighted query expansion methods (TF-IDF) has better key terms for word spotting.

We observed from above experiments proves that the automation (semi-supervised) in keyword extraction from the question and TRA using query expansion can help instructor with evaluation and grading. The results in Table I & II in general show that models trained on semantic features perform better than the base features and query expanded keywords provide better coverage of keywords for word spotting based evaluation.

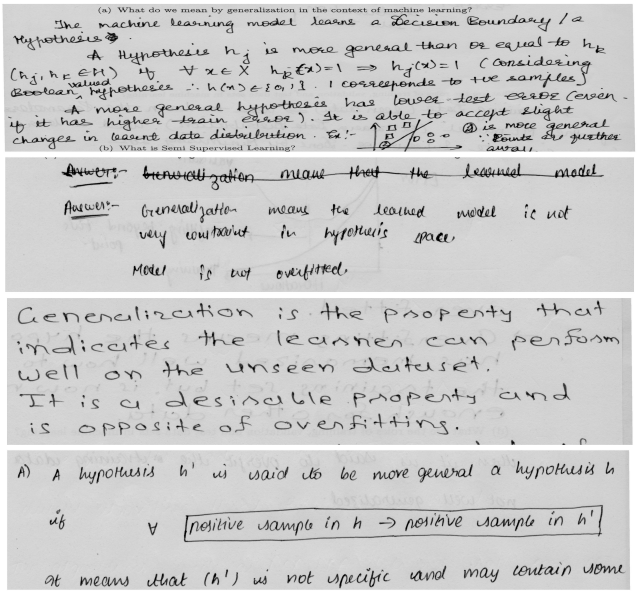

Figure 6: List of failure scenarios due to i) figures and equations, ii) scratched lines, iii) improper word, character spacing and, iv) text highlighting using boxes.

D. Discussion

Our method is a pipeline integrating information retrieval and NLP based feature analysis. Errors in initial stages of document image analysis gets propagated and impact evaluation scores to a certain extent. A primary limitation of our work is the lack of comprehension of complex mathematical equations and inferences, as shown in Figure 6. Tidiness and organized answers also matter. Our prototype fails to segment text with less spacing between words, high skew and excessive word scribbling which are add up in word count thereby effecting scores. Answers paraphrased with simple non-technical terms were also found relatively hard to evaluate. However, we hope that our approach with some changes can address the grading requirements in a variety of subjects across domains.

V. Conclusion

We demonstrate an automatic evaluation scheme for handwritten answers with performance comparable to the human evaluation. As a first step towards fully automating the grading schemes, we believe our method can be an assistance to the instructors, leveraging on the recent developments in handwritten document processing space. Our framework integrates ideas from information retrieval, natural language processing, and feature-based word spotting for this task. On real answers from a classroom, it provides scores that correlate highly with the human evaluators. The method is aimed at short descriptive answers, and it meets this purpose.

REFERENCES

[1] P. Krishnan and C. Jawahar, "Hwnet v2: An efficient word image representation for handwritten documents," *arXiv preprint arXiv:1802.06194*, 2018.

[2] M. Kiefer, S. Schuler, C. Mayer, N. M. Trumpp, K. Hille, and S. Sachse, "Handwriting or typewriting? the influence of pen-or keyboard-based writing training on reading and writing performance in preschool children," *Advances in cognitive psychology*, 2015.

[3] S. Srihari, J. Collins, R. Srihari, H. Srinivasan, S. Shetty, and J. Brutt-Griffler, "Automatic scoring of short handwritten essays in reading comprehension tests," 2008.

[4] D. Kanejiya, A. Kumar, and S. Prasad, "Automatic evaluation of students' answers using syntactically enhanced lsa," in *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2*, 2003.

[5] Y. Liu, C. Sun, L. Lin, X. Wang, and Y. Zhao, "Computing semantic text similarity using rich features," in *PACLIC*, 2015.

[6] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using wordnet and lexical chains," *Expert Systems with Applications*, 2015.

[7] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2008.

[8] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments." Association for Computational Linguistics, 2011.

[9] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, 2007.

[10] A. Poznanski and L. Wolf, "Cnn-n-gram for handwriting word recognition," in *CVPR*, 2016.

[11] S. Sudholt and G. A. Fink, "Phocnet: A deep convolutional neural network for word spotting in handwritten documents," *ICFHR*, 2016.

[12] S. N. Srihari, R. K. Srihari, P. Babu, and H. Srinivasan, "On the automatic scoring of handwritten essays." in *IJCAI*, 2007.

[13] E. A. Kozak, R. S. Dittus, W. R. Smith, J. F. Fitzgerald, and C. D. Langfeld, "Deciphering the physician note," *Journal of general internal medicine*, 1994.

[14] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Efficient segmentation-free keyword spotting in historical document collections," *PR*, 2015.

[15] H. Wei, H. Zhang, and G. Gao, "Word image representation based on visual embeddings and spatial constraints for keyword spotting on historical documents," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018.

[16] U.-V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, 2002.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[18] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014.

[19] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[20] A. Sordoni, Y. Bengio, and J.-Y. Nie, "Learning concept embeddings for query expansion by quantum entropy minimization," in *AAAI*, 2014.

[21] D. Metzler, S. Dumais, and C. Meek, "Similarity measures for short segments of text," in *European Conference on Information Retrieval*, 2007.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, 2003.

[23] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[24] D. Bär, C. Biemann, I. Gurevych, and T. Zesch, "Ukp: Computing semantic textual similarity by combining multiple content similarity measures." Association for Computational Linguistics, 2012.

[25] V. Rowtula, P. Krishnan, and C. Jawahar, "Pos tagging and named entity recognition on handwritten documents," in *Proceedings of the 15th International Conference on Natural Language Processing*, 2018.

[26] M. Carbonell, M. Villegas, A. Fornés, and J. Lladós, "Joint recognition of handwritten text and named entities with a neural end-to-end model," *arXiv preprint arXiv:1803.06252*, 2018.

[27] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[28] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® v. 2," *The Journal of Technology, Learning and Assessment*, 2006.

[29] M. O. Dzikovska, R. D. Nielsen, and C. Leacock, "The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications," *Language Resources and Evaluation*, 2016.