ScAA: A Dataset for Automated Short Answer Grading of Children's Free-text Answers in Hindi and Marathi

Dolly Agarwal[1], Somya Gupta[2], Nishant Baghel[1]

[1]Pratham Education Foundation
[2]Pratham Volunteer*
[1]{dolly.agarwal, nishant.baghel}@pratham.org, [2]somya.gupta1@gmail.com

Abstract[1]

Automatic short answer grading (ASAG) techniques are designed to automatically assess short answers written in natural language. Apart from MCQs, evaluating free text answer is essential to assess the knowledge and understanding of children in the subject. But assessing descriptive answers in low resource languages in a linguistically diverse country like India poses significant hurdles. To solve this assessment problem and advance NLP research in regional Indian languages, we present the Science Answer Assessment (ScAA) dataset of children's answers in the age group of 8-14. ScAA dataset is a 2-way (correct/incorrect) labeled dataset and contains 10,988 and 1,955 pairs of natural answers along with model answers for Hindi and Marathi respectively for 32 questions. We benchmark various state-of-the-art ASAG methods, and show the data presents a strong challenge for future research.

1 Introduction

Answer assessment is a key component of teaching and learning process and automating it has many advantages including speed, availability, consistency and fairness of assessments. Though the evaluation of multiple-choice questions is straightforward and can be scaled, there is need for systems to assess free text answers as well. Prior research has shown that recognition questions (like MCQ) are deficient as they do not capture multiple aspects of acquired knowledge such as reasoning and self-explanation (Wang et al., 2008). The free text answers are important as they help measure the understanding of the student with respect to a particular concept. Grading responses to short answer questions is considered difficult as it requires deep understanding of natural language. There can be multiple versions in which a correct answer can be articulated for the same question (examples highlighted in Table 1). Any errors in assessment can affect students' learning engagement and feedback directly.

ASAG has been in research for many years now, but most of the work has been done primarily for English. In comparison, there has been far less research in similar areas for Indian languages, which are primary medium of instruction in a large proportion of schools of India. Thus, we present a dataset for ASAG for Indian languages - Hindi and Marathi to aid development of robust solutions.

An Android app[2] for assessment was developed by our team at Pratham and was piloted in our Hybrid Learning Program. The Hybrid Learning Program of Pratham is spread across 3 states of India - Rajasthan, Uttar Pradesh and Maharashtra. The Android Assessment app enabled collection of free text answers to questions in Hindi and Marathi essential for building systems for ASAG in Indian languages. These free-text answers are 2-

* Currently affiliated with LinkedIn. This work was done as a Pratham volunteer and is not connected with LinkedIn.

[1] Data is available at: https://github.com/PrathamOrg/ScAA-Dataset

[2] Pratham Online Assessment App: https://play.google.com/store/apps/details?id=com.pratham.assessment&hl=en_IN&gl=US

Question in Hindi	Question in Marathi	English Translation (ETL)	Correct Answers in Hindi	English Translation	Correct Answers in Marathi	English Translation
एक वयस्क मनुष्य के कंकाल का वजन लगभग कितना होता है?	एका प्रौढ व्यक्तीच्या सांगाड्याचे वजन अंदाजे किती असते?	What is the approximate weight of an adult human skeleton?	एक वेस्ट मनुष्य के कंकाल का वजन लगभग 10 किलोग्राम होता है	The skeleton of a n adult human we ighs about 10 kil ogram	10 किग्रॅ	10 kg
			१० किग्रा	10 Kilogram	दहा किलो ग्रॅम	Ten Kilogram
			लगभग 10 किलोग्राम	Approximately 10 Kilogram		
चाँद पर कोई किसी की आवाज क्यों नहीं सुन सकते?	अंतराळवीर चंद्रावर एकमेकांचे आवाज का ऐकू शकत नाहीत?	Why can't anyone hear someone's voice on the moon?	वायुमंडल नहीं है	No atmosphere	वातावरण न सते	There is no atmosphere
			क्योंकि वहां पर वायु नहीं है	Because there is no air	तिथे हवा नाही	There is no air
			वायु मंडल का अनुपस्थिती के कारण	Because of the absence of atmosphere	कारण तिथे हवा नाही	Because there is no air
			वायुमणडल ना होने से कारण	Due to lack of air		

Table 1: Samples demonstrating that same answer can be written in multiple ways

way labeled as correct/incorrect by human annotators.

The main contributions of this paper are:

- A dataset with 10,988 answers in Hindi and 1,955 in Marathi to 32 questions in 8 different topics of science (§ 3)
- Benchmark of various state-of-the-art methods for automated assessment of free-text answers by children (§ 4)

2 Prior Art

There are numerous standard ASAG datasets publicly available for the research community to experiment with: Beetle and SciEntsBank(SEB) - released as part of SRA corpus (Dzikovska et. al, 2013), CSD (Mohler and Mihalcea 2009), X-CSD (Mohler et. al, 2011), Powergrading (PG) (Basu et. al, 2013) and ASAP (Higgins et al., 2014). The total no of prompts in these Datasets are in the range of 10 (ASAP) to 135 (SEB). While some annotated dataset (PG; Beetle) have 2-way labels (correct/incorrect), a few (CSD; X-CSD; ASAP) have scores on an ordinal scale within a range, e.g. 0-5, SEB has a more complex 5-way labels. Though there are numerous ASAG datasets available, all these are in English and none are available for Indian languages. To the best of our knowledge, this is the first comprehensive corpus with reasonable size for Hindi and Marathi.

Numerous approaches have been tried before for ASAG. Burrows et al., (2015) and Roy et al., (2015) do a comprehensive review of ASAG systems. Mohler et al. (2011) show that answers can be accurately graded by using semantic measures. Rodrigues and Araújo (2012) propose word matching between (user answer, model answer) pair. Roy et al. (2016) propose an unsupervised ASAG technique using sequential pattern mining. Sultan et al. (2016) train a supervised model based on semantic similarity features. Supervised methods include neural architectures by Riodan et al. (2017) where performance and optimal parameter settings vary across prompts, a joint-multidomain deep learning architecture by Saha et al. (2019) which learns generic and domain-specific aspects. Lun et al. (2020) introduce data augmentation strategies and show that this combined with latest BERT model brings significant gain.

For benchmarking on our dataset, we prefer unsupervised methods like sentence level semantic

similarity and sequential pattern mining due to their generalizability and suitability for deployment in an unseen-question setting. We intentionally do not benchmark supervised ML approaches that require a large corpus of labeled answers for training as they get limited to a particular question pool and are hard to generalize.

3 ScAA Data Creation

We curate Science Answer Assessment (ScAA) dataset in Hindi and Marathi language comprising of 8 science topics with 4 questions per topic, i.e. a total of 32 parallel question-model answer pairs in Hindi and Marathi. The dataset is created via three stages: Question and model answer curation, User answer collection, User answer evaluation.

The questions were selected from Grade 8 level Science topics: Adaptation, Circulatory System, Eye and Vision, Heat, Simple Machine, Skeletal System, Sound, Water Chemistry. The users here are children in the age group of 8-14 years from 3 states of India: Uttar Pradesh, Rajasthan and Maharashtra.

3.1 Data Collection and Statistics

The data is crowdsourced via an Android app developed by Pratham to enable children to take assessments anytime they want. This Android Assessment App[3] is available on play store since November 2019 with 10,000+ downloads. Children could either type the answer directly using the phone keyboard or use Speech-to-Text (STT) service[4] and then edit it. We identify the issues in the data collected via this process and pre-processing methods in section 3.2.

Over a period of 8 months, the app helped collect ~50,000 answers from 11,476 children to 32 science questions from 3 states of India. The ScAA dataset was created by selecting a subset of these answers and getting each answer evaluated as correct or incorrect by two human annotators. The Cohen's Kappa κ score indicating level of agreement between two annotators was 0.75. ScAA consists of answers where both human evaluators matched in their markings. Detailed statistics are listed in Table 2.

	Hindi	Marathi
Total Questions	32	32
# total answers	10988	1955
# total unique answers	7205	1435
# total correct answers	3843	488
Average # unique correct answers per question	41	7
Average answer length (in words)	15	15

Table 2: Statistics of Evaluated Dataset

3.2 Noise Types and Data Processing

Since any child could give the assessment whenever they like without adult supervision through phone interface, this led to presence of noise in the dataset. The option of submitting the answer through STT service brought in its own errors as well. We preprocess the noise and clean it before benchmarking. The ScAA dataset that we present lists the original noisy as well as preprocessed answers for the benefit of NLP community. Table 3 lists various noise types we found in the data and how we processed them.

Noise Type	Example	Processing
Transliterated Text	हड्डियों से appears as *Haddiyon se* in user answer	Transliterated it using Indic Transliterate*
Translated Text	हड्डियों से appears as *bones* in user answer	Translated it using Google Translate**
Code Mixed Language	bol our सॉकेट ke madat sa	Translated/Transliterated the English words
Special symbols and characters	|\£\%£`¡$°©$¶¶`=	Removed Special symbols & extra spaces
URLs	*हड्डियोंसेhttps://faq.whatsapp.com/general/26000015?lg=en&lc=IN&eea=0*	Removed these URLs
Emoji Characters	😹😹😹😹😹	Removed the emojis
Phonetically Similar words with different meanings	'ऊर्जा' (energy) recorded as 'उड जा' (fly) by STT service	Not processed

Table 3: Noise types in the data and processing

[3] Pratham Online Assessment App: https://play.google.com/store/apps/details?id=com.pratham.assessment&hl=en_IN&gl=US

[4] https://developer.android.com/reference/android/speech/package-summary
* https://pypi.org/project/indic-transliteration/
** https://pypi.org/project/googletrans/

4 Automated Short Answer Assessment

We model the assessment of user answers against reference answers as a similarity task. Each (user answer, model answer) pair is assigned a similarity score using the various state-of-the-art methods for assessment of free-text answers described in this section. We use random score assignment as baseline. User and model answers are tokenized into their constituent words using indicNLP tokenizer (Kunchukuttan et al., (2020)).

1. **Jaccard Similarity:** We calculate the number of words from user answer appearing in the model answer sentence. This is normalized w.r.t the total words present in the given answers (1), where J is the jaccard similarity score between C, the set of words in user answer and I, the set of words in model answer.

$$J = \frac{(C \cap I)}{(C \cup I)} \quad (1)$$

2. **Word based Semantic Similarity:** Answer sentences are represented by taking average of their word embeddings. We then calculate cosine similarity between them. The word embeddings used are:

   Indic NLP: Pre-trained word embeddings available for 1.1B Hindi tokens trained using FastText on corpus crawled from news websites (Kunchukuttan et al., (2020))

   fastText: Pre-trained word embeddings for Hindi, trained on Wikipedia and Common Crawl datasets consisting of 1.8B tokens (Grave et al., (2018))

3. **Sentence Similarity using S-BERT**: Sentence-BERT (Reimers and Gurevych, (2019)) finetunes a pre-trained BERT network using Siamese and triplet network and adds a pooling operation to the output of BERT to derive a sentence vector. Cosine similarity is used to compare the generated user and model answer vectors.

4. **Sequential Pattern Matching:** (Roy et. al (2016)) define a method to extract commonly occurring patterns p using support *sup(p)* to quantify the notion of commonalities from user answers and lexical diversity via type-token ratio *TTR* (eq 2). The score $Sc(s_i)$ for user answer s is calculated using this *TTR* and *sup(p)* as described in (eq 3). While this doesn't need a model answer, note that this method is most effective for batch mode as it banks on pattern mining from repeating answers and hence does not work well for real-time ASAG.

$$\text{TTR(d)} = \frac{\#distinct\ patterns\ of\ length\ d}{\#patterns\ of\ length\ d} \quad (2)$$

$$\text{Sc(s)} = \sum_{p \in si} \text{sup}\ (p)^{len(p)} * TTR(len(p)) \quad (3)$$

5 Results and Analysis

We now benchmark the ASAG methods described earlier on ScAA taking only the unique answers for evaluation (Table 4). The resulting data has 20% correct answers for Hindi and 16% for Marathi. We evaluate the models based on cost, the number of wrong assessments the similarity scores result in as compared to the actual ground truth (eq 4). We convert all scores to binary by selecting the best threshold t (table 5) for each method that minimizes this cost c and marking scores above t as 1 (correct), else 0. FP, FN, TP, TN are number of false positives, false negatives, true positives and true negatives in data.

$$c = (FP+FN) / (TP+FP+FN+TN) \quad (4)$$

Note that while on full ScAA with repeating answers PatternMatch-Repeat is comparable to S-BERT, its use is suitable in batch mode to extract answer patterns. It therefore renders itself unusable in apps, where the requirement is for real time evaluation for single (user answer, model answer) pair.

Similarity Measure	Hindi Data	Marathi Data
Baseline	0.50	0.50
Jaccard	0.76	0.75
indicNLP	0.80	0.80
fastText	0.78	0.69
S-BERT	**0.86**	**0.82**
PatternMatch-Unique	0.80	0.73
PatternMatch-Repeat	0.87	0.81

Table 4: ROC AUC for various approaches

Similarity Measure	Hindi Data	Marathi Data
Baseline	0.999	0.999
Jaccard	0.158	0.251
indicNLP	0.747	0.788
fastText	0.793	0.801
S-BERT	0.700	0.905
PatternMatch	0.245	0.539

Table 5: Threshold t for various approaches

Child's Answer / ETL	Model Answer / ETL	Human	Jaccard	indicNLP	S-BERT
मुखिया तरंग / Head waves	विद्युत् चुम्बकीय तरंग / electromagnetic waves	0	**1**	0	0
पचास किलो / Fifty Kilo	दस किग्र / Ten Kg	0	0	**1**	0
संवहन / Convection	वाष्पन / Evaporation	0	0	0	**1**
धूल के कण आंसुओं के साथ बाहर निकल आते है / Dirt particles get released with tears	धूल कण आशु के साथ में बाहर चले आते है / Dirt particles get released with tears	1	1	1	**0**
कारण तेथे वातावरण नाही / Because there's no atmosphere	कारण चंद्रावर वातावरण नाही / Because there's no atmosphere on moon	1	1	1	**0**
बल / force	बलभुजा / arm	0	0	0	**1**
प्रकाश साचे परावर्तीत / reflecting light mold	प्रकाशाचे परावर्तन / reflection of light	0	0	0	**1**

Table 6: Errors by Jaccard, indicNLP and S-BERT in marking. Columns 3,4,5,6 show marking by various methods

5.1 Error Analysis and Discussion

We now show some examples where S-BERT, IndicNLP and Jaccard based similarity measures make assessment errors in Table 6. Row 1 shows an example where the child's answer is incorrect, but Jaccard Similarity assigns it a high score due to matching word "तरंग / waves". IndicNLP incorrectly assigns high similarity among number names while S-BERT incorrectly assigns a high similarity score to word pair ("संवहन/convection", "वाष्पन/Evaporation") perhaps because they appear in similar context (rows 2 and 3).

S-BERT additionally marks a correct answer as incorrect for Hindi and Marathi in rows 4 and 5, while all the other methods mark the answer correctly. More S-BERT errors on Marathi answers are shown in rows 6 and 7. Rows 5 and 7 depict a contrast in evaluation by S-BERT where the model answer and child's answer are similar in words, but not in meaning, which it fails to capture properly. This shows its sensitivity to the input training data and absence of generalization to sentences that may unseen earlier.

The error analysis and examples showcase that while state-of-the-art models like S-BERT give best performance, they are far from being fit for deployment as the errors in assessment can directly affect students' learning engagement. Additionally, they need high latency and good compute power for assessment. A critical requirement for us is to keep the methods simple for low resource settings to cater to rural children in remote areas with limited internet access and Hindi/Marathi as the primary medium of instruction.

6 Conclusion

In this paper we present ScAA, a dataset of children's free-text answers to 32 questions of grade 8 level Science topics in Hindi and Marathi along with their user answers. This dataset is intended to facilitate research in automatic assessment of short answers in Indian languages. We benchmark the performance of various state-of-the-art ASAG methods on ScAA and observe that even though BERT based model performs best, it makes errors in assessment that can affect students' learning engagement, thus leaving scope for improvement before such techniques can be deployed in real world and presenting a strong case for more research in this area. We believe that this dataset will be useful for the research community working on automated short answer assessment for Indian languages and aid in solving a very practical problem for society at scale.

References

Basu, Sumit, Chuck Jacobs, and Lucy Vanderwende. "Powergrading: a clustering approach to amplify human effort for short answer grading." *Transactions of the Association for Computational Linguistics* 1 (2013): 391-402.

Burrows, Steven, Iryna Gurevych, and Benno Stein. "The eras and trends of automatic short answer grading." *International Journal of Artificial Intelligence in Education* 25, no. 1 (2015): 60-117.

Dzikovska, Myroslava O., Rodney D. Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. "SemEval-2013 Task 7: The Joint Student

Response Analysis and 8th Recognizing Textual Embodiment Challenge." In *Second Joint Conference on Lexical and Computational Semantics (* SEM): Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol. 2. Association for Computational Linguistics*, 2013.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning Word Vectors for 157 Languages. *In Proceedings of LREC 2018, 15th conference on International Language Resources and Evaluation*.

Hao-Chuan Wang, Chun-Yen Chang, and Tsai-Yen Li. 2008. Assessing Creative Problem-solving with Automated Text Grading. *Computers and Education*, 51(4):1450–1466.

Higgins, Derrick, Chris Brew, Michael Heilman, Ramon Ziai, Lei Chen, Aoife Cahill, Michael Flor et al. "Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring." *arXiv preprint* arXiv:1403.0801 (2014).

Horbach, Andrea, and Manfred Pinkal. "Semi-supervised clustering for short answer scoring." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

Kunchukuttan, A., Kakwani, D., Golla, S., Bhattacharyya, A., Khapra, M.M. and Kumar, P., 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv preprint* arXiv:2005.00085.

Kumar, Sachin, Soumen Chakrabarti, and Shourya Roy. "Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading." In *IJCAI*, pp. 2046-2052. 2017.

Lun, Jiaqi, Jia Zhu, Yong Tang, and Min Yang. "Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring." In *AAAI*, pp. 13389-13396. 2020.

Mieskes, Margot, and Ulrike Pado. "Work Smart-Reducing Effort in Short-Answer Grading." In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018) at SLTC, Stockholm, 7th November 2018*, no. 152, pp. 57-68. Linköping University Electronic Press, 2018.

Mohler, Michael, and Rada Mihalcea. "Text-to-text semantic similarity for automatic short answer grading." In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 567-575. 2009.

Mohler, Michael, Razvan Bunescu, and Rada Mihalcea. "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments." *In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 752-762. 2011.

Pérez, Diana, Alfio Massimiliano Gliozzo, Carlo Strapparava, Enrique Alfonseca, Pilar Rodríguez, and Bernardo Magnini. "Automatic Assessment of Students' Free-Text Answers Underpinned by the Combination of a BLEU-Inspired Algorithm and Latent Semantic Analysis." In *FLAIRS conference*, pp. 358-363. 2005.

Piyush Patil, Sachin Patil, Vaibhav Miniyar and Amol Bandal.2018. Subjective Answer Evaluation Using Machine Learning in *International Journal of Pure and Applied Mathematics*, Volume 118 No. 24 2018

Ramachandran, Lakshmi, Jian Cheng, and Peter Foltz. "Identifying patterns for short answer scoring using graph-based lexico-semantic text matching." In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 97-106. 2015.

Rodrigues, Fátima & Araújo, Lília. (2012). Automatic Assessment of Short Free Text Answers. *4th International Conference on Computer Supported Education*. 2.

Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint* arXiv:1908.10084 (2019).

Riordan, Brian, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chungmin Lee. "Investigating neural architectures for short answer scoring." In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 159-168. 2017.

Roy, Shourya, Sandipan Dandapat, Ajay Nagesh, and Yadati Narahari. "Wisdom of students: A consistent automatic short answer grading technique." In *Proceedings of the 13th International Conference on Natural Language Processing*, pp. 178-187. 2016.

Roy, Shourya, Y. Narahari, and Om D. Deshmukh. 2015. A Perspective on Computer Assisted Assessment Techniques for Short Free-Text Answers. *In Proceedings of the International Conference on Computer Assisted Assessment (CAA)*, pages 96–109. Springer.

Sultan, Md Arafat, Cristobal Salazar, and Tamara Sumner. "Fast and easy short answer grading with high accuracy." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1070-1075. 2016.

Sahu, Archana, and Plaban Kumar Bhowmick. "Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance." *IEEE Transactions on Learning Technologies* 13, no. 1 (2019): 77-90.

Saha, Swarnadeep, Tejas I. Dhamecha, Smit Marvaniya, Peter Foltz, Renuka Sindhgatta, and Bikram Sengupta. "Joint Multi-Domain Learning for Automatic Short Answer Grading." *arXiv preprint* arXiv:1902.09183 (2019).

Wang, Tianqi, Tomoya Mizumoto, Naoya Inoue, and Kentaro Inui. "Identifying Current Issues in Short Answer Grading." *ANLP-2018* (2018).

Zhang, Yuan, Rajat Shah, and Min Chi. "Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading." *International Educational Data Mining Society* (2016).