Automated Assessment of Noisy Crowdsourced Free-text Answers for Hindi in Low Resource Setting

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

The requirement of performing assessments continually on a larger scale necessitates the implementation of automated systems for evaluation of the learners' responses to free-text questions. We target children of age group 8-14 years and use an ASR integrated assessment app to crowdsource learners' responses to free text questions in Hindi. The app helped collect 39641 user answers to 35 different questions of Science topics. Since the users are young children from rural India and may not be well-equipped with technology, it brings in various noise types in the answers. We describe these noise types and propose a preprocessing pipeline to denoise user's answers. We showcase the performance of different similarity metrics on the noisy and denoised versions of user and model answers. Our findings have large-scale applications for automated answer assessment for school children in India in low resource settings.

1 Introduction

Posing and assessing open-ended descriptive questions to children is crucial to evaluating their learning levels and improving their understanding of concepts. Unfortunately, this puts an enormous load on the classroom teacher who is faced with assessing and providing feedback to every child. As a result, teachers are not able to give writing assignments or do oral evaluation as often as they would wish. The problem is even more important in rural areas where availability of teachers is low, and children seldom get access to quality assessments. This highlights the importance of developing applications that automate assessments for children. Free-text questions allow a respondent to answer in open text format such that they can answer based on their complete knowledge and understanding. This means that response to this question is not limited to a set of options.

Though the evaluation of multiple-choice questions is straightforward and can be scaled, we need robust systems to assess the free text questions as well. This presents an interesting challenge for automated assessments as there are multiple versions of correct answers for the same question in free-text format. We show some examples in Table 1. Extracting information from the text in low resource languages such as Hindi is even more challenging for the NLP community. Further, crowdsourcing such free-text answers at scale brings another challenge of noise in the collected data.

This paper presents our experience with automated assessment of free-text answers by developing an automated assessment system for one of the low resource languages - Hindi which is the medium of instruction in government schools in Rajasthan and Uttar Pradesh states of India. The children were of age group 8-14 years from rural areas of India, and hence were not well equipped with technology like their urban counterparts. We developed an Android assessment app for these children to give assessments at any time they wanted to. One assessment has a mix of question types including

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

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 122–131
Online, Nov 19, 2020.

Question (Hindi / ETL)	Varieties of correct answers (Original / ETL)
आंखों में धूल चली जाने पर आंसू क्यों आ जाते हैं? / Why do we tear up when dirt enters our eyes?	aansu nikalne se dhul ke kan bahar aa jate hai / dust particles get released due to tears
	आंखों से धूल बाहर निकलने के लिए / to remove dirt from eyes
उत्तोलक कितने प्रकार के होते हैं? / How many types of levers are there?	तीन / three
	तीन प्रकार का / of three types
	3
	three
	प्रथम श्रेणी दितीय श्रेणी तृतीय श्रेणी यह तीन प्रकार के होते हैं / first class, second class, third class, these three types
	Teen / three

Table 1: Varieties of correct answers for the same question

MCQ and free-text questions. To facilitate the answer to free-text questions, we allowed the children to type the answers. We also enabled automated speech recognition services where the child could speak out the answer and then edit it. It also reinforced our objective to test the knowledge and understanding of a child in a particular subject and not their writing ability. The assessment app helped collect 39641 user answers to 35 different questions of Science topics.

We manually went through various user answers and found different varieties of noise that needed cleaning before it could be evaluated. With these noise types in answers, it is a challenging task even for a human evaluator to assess children's answers. It is also critical to enable assessment through denoising, as we do not wish to incorrectly mark an answer that can lead to a child getting discouraged from using the platform altogether. For example, a noisy answer which is correct is shown below:

User Answer: tan,k.g\n
Model Answer: १० किग्रा (ETL: 10 kg)

Though the test was in Hindi, we also found that around ~60% of the answers contained English alphabets. This may be for two reasons:

i. The child is not comfortable using Indic Keyboard and hence typed in English

ii. Language pack was not downloaded for Hindi in the phone and hence the Hindi text was transliterated to English while using Speech-To-Text (STT).

Following are two examples of correct answers which are in English alphabets:

Question: हड्डियों के बींच जोड़ नहीं होते तो क्या होता?
(ETL: What would have happened if there were no joints between the bones?)
User Answer:

a. ham hil nahi pate (Transliterated to English)

b. We cannot be move and do action (English Translation)

In this paper, we describe various noise types identified in detail and propose a preprocessing pipeline for answers collected through automated speech recognition (ASR) enabled assessment app.

After the preprocessing, we compute similarity scores between user answers with their reference answer. We test different similarity metrics for both original noisy answers and their denoised versions. The idea is to measure the significance of denoising the answer before passing through assessment. Interestingly we find that denoising enables a simple word-matching based metric to perform as good as semantic similarity measures. This finding has promising implications for deployment of such solutions in low resource settings.

It is important to note that the focus of our study is for the rural children in remote areas of India with limited internet access. Hence, we need to find a solution which could be integrated with a system in low resource setting i.e., with low computing, memory and battery capacity. Thus, while more complex state-of the-art models like LSTMs, BERT may give higher performance for sentence similarity measurement, we choose similarity measures as our assessment methodology as the primary requirement we have is to keep the assessment model as simple as possible to ensure it does not add a lot of memory requirement to the app.

The main contributions in this paper are (1) listing the types of noise possible in a text-based ASR enabled assessment tool, (2) a preprocessing pipeline to handle those noises and transform the user answer to a format consumable by NLP systems, (3) comparison of various semantic similarity measures on noisy and denoised answers.

2 Previous Work

Research in the area of evaluation of descriptive free text answers has been in progress since a decade and a half. Burrows et al., (2015) did a comprehensive review of Automatic Short Answer Grading (ASAG) research and systems according to history and components. Their historical analysis identifies 35 ASAG systems within 5 temporal themes that mark advancement in methodology or evaluation.

Butcher et al., (2010) compared the marking accuracy of three separate computerized systems, one system (Intelligent Assessment Technologies FreeText Author) is based on computational linguistics whilst two (Regular Expressions and OpenMark) are based on the algorithmic manipulation of keywords. Patil et al. (2018) experimented with training a Naive Bayes classifier based on three parameters: Keywords, Grammar and Question Specific things. They proposed a system where students will have a certain degree of freedom while writing the answer as the system checks for the presence of keywords, synonyms, right word context and coverage of all concepts. But the experiment was conducted with only 20 students and 3 questions to each student.

Perez et al., (2005) presented a comparative evaluation between BLEU-inspired algorithm and a system based on Latent Semantic Analysis and proposed a combination schema. Despite the simplicity of these shallow NLP methods, they achieved state-of-the-art correlations to the teachers' scores while keeping the language-independence and without requiring any domain specific knowledge

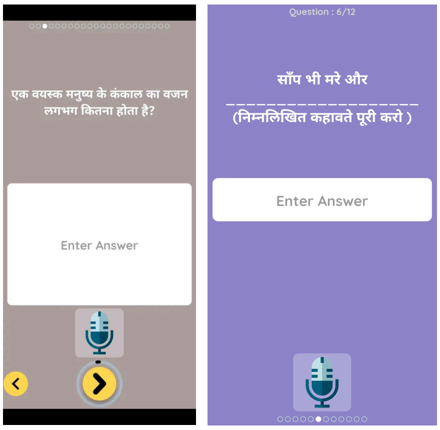
Lun et al., (2020) proposed multiple data augmentation strategies for improving performance on automatic short answer scoring. They combined it with the latest fine-tuned BERT model for the short answer scoring task, and show significant gain.

Bonadiman et al., (2019) discuss a new Question Paraphrase Retrieval (QPR) system that can be used to understand and answer rare and noisy reformulations of common questions by mapping them to a set of canonical forms.

To the best of our knowledge, this is the first study which handles noises in user answers while evaluating the short answers typed in by the children in Indian language. Also, focus is in a method with less computing and memory needs, so that it can be used for large scale implementation in Android devices.

3 Data and Attributes

We conducted our assessment in the Hybrid Learning program of Pratham which reaches more than 1000 villages and over 109,560 children. Every group of 5-6 children have one tablet provided by Pratham with digital learning content. An Android Assessment app (Figure 1) was developed and loaded in these tablets for children to give assessments anytime they wanted to. The assessment app helped collect 39641 user answers to 35 different questions of Science topics. Each answer was then manually evaluated as correct/incorrect by two different raters and their agreement score as calculated by Cohen's Kappa κ score is 0.74. Total Average number of unique correct answers per question is 34.



एक वयस्क मनुष्य के कंकाल का वजन लगभग कितना होता है?

Question : 6/12

साँप भी मरे और

____

(निम्नलिखित कहावते पूरी करो)

Figure 1: UI for free text question (space to type in the answer and a mic to use STT)

3.1 Challenges

Since any child could give the assessment whenever they like in an unsupervised way, this increased the chances of gibberish in the dataset as the child could use the app just for fun. Furthermore, the option to enter the answer through a speech recognition system was helpful for the child but also bought in its own unique challenges. Listed in Table 2 are the various types of noise we found in the data. Additionally, when we checked the evaluated data, the level of agreement between two annotators calculated by Cohen's Kappa κ score is 0.74. We can see that even with human raters there is a mismatch between the ratings. This puts a spotlight on the requirement of a standardized mechanism to evaluate children on the same scoring model. Another important finding from the evaluated dataset is that the same question can have many varieties of correct answers, a sample of such answers are highlighted in Table 1. On average, there are 34 unique correct answers per question in our crowdsourced dataset.

3.2 Noise Types in Crowdsourced Data

We begin by highlighting different types of noise and their prevalence in the data (Table 2) which we came across in the answers submitted by children. We have discussed them in detail below as it gives a fair idea about the noise types to expect in a text-based speech recognition enabled assessment app. This list is not exhaustive, but it covers the vast majority of cases observed in this paper's datasets.

Type of Noise	Percentage (%)
Punctuators	14.80
URL	0.12
Emoji	0.53
Subscripts/Superscripts	0.015
English Alphabets	59.4

Table 2: Different types of noise

1. **Punctuations:** Since our focus is on the semantic similarity of short answer questions, the significance of punctuation is less. We observed that there are unnecessary punctuations due to typos in ~15% of user answers. We removed the punctuators, and replaced comma with spaces. Example of such answers:

   10,kg
   हम,हिल,डुल,नही,सकते / ETL: we can't move
   Hilana,dulana,asmbhav,ho,jata

   These answers are correct but have unnecessary commas after every word. We replaced these commas with space. Also, we removed all other punctuations observed in the answers.

2. **Emojis:** Though Emoji can express an emotion for sentiment analysis it is irrelevant in our use case because our aim is to analyze knowledge of a child on a particular concept. Only 0.53% of user answers had some form of emoji and it seemed to be typo and thus can safely be removed. For example,
   😀रक्त का थक्का नही जमेगा (ETL: *There would be no blood clotting)* is a correct answer but has emoji as noise.

3. **Translated Text:** The assessment test was in Hindi and expected answers from children were in the same language. But we see that there are instances (~60%) where the answer is in English and some of which are correct too. Since, our main objective is to assess how well acquainted the child is with the concept irrespective of the language, we need to consider these answers and handle it accordingly. So, we used Google translate library to convert the answers back to Hindi before it can be used by the NLP systems. For example, BONES was converted to हड्डियों

4. **Transliterated Text:** We also found some scenarios where the answer was in native language (Hindi) but was transliterated in English. This happens when the device that the child is using to give assessment does not have the Language pack downloaded for their native language. In such scenarios, the Hindi text is transliterated to English while using STT. To handle the transliterated text, we use Indic Transliteration library to transliterate it back to Hindi for evaluation.

For example, *haddi* was converted to *हड्डी (ETL: Bone)*

5. **Digits:** While working on the solution, we realized there is a performance drop in the similarity metrics because of numbers. We need to ensure that both the ideal answer and user answer either specify numbers as digits or number names. Our ideal answers have numbers present as number names; hence, we converted the numbers to their corresponding word lexemes. For example, *4561* was converted to *चार हज़ार पांच सौ इकसठ* (ETL: *Four Thousand Five Hundred and Sixty-One*).

6. **URL:** There were answers which contained urls in them and are irrelevant in our context. These urls were also present sometimes along with the correct answers. Since, these urls are insignificant, they are considered noise and removed from the answers. Here's an example of a correct answer that has a url at the end which can safely be removed: *हड्डियोंसेhttps://faq.whatsapp.com/general/26000015?lg=en&lc=IN&eea=0* (Ideal answer here is *हड्डियों से*, ETL: *with bones*)

7. **Subscript & Superscript characters:** A few user answers had subscript and superscript characters in the answers. Although these subscripts and superscripts might be significant in Science answers, the question-answer set in our dataset had none. We were facing issues with these characters while using Google Translate and hence we considered this as noise and removed them from the answers. Going ahead, we will look into ways to handle them instead of removing.

8. **Human Reasoning:** Apart from the cases mentioned above, there are few other types of noise which require human reasoning to decode. For example,

   a. Missing space between words, ex: हमहिलनहिसकते (should have been *हम हिल नहि सकते (*ETL: *we cannot move)*

   b. Words replaced with phonetically similar sounding ones but with a completely different meaning. These might occur due to STT dialect/child speech issue.

      Example 1:
      Reference answer: 'ऊर्जा' (ETL: energy)
      Child's answer: 'उड जा' (ETL: fly)

      Example 2:
      Reference answer: *das gram hota hai*
      Child's answer: *that's gram hita hai*

   c. Answer hidden among other meaningful words which are irrelevant to the answer present, ex: *हड्डियों का बोल हड्डियों का* (ETL: *bones say with bones*)

4 Descriptive Answer Assessment

While including questions with descriptive answers is an important aspect of learning, it brings unique challenges in terms of assessment. In this section we describe our methodology for assessment of such descriptive free-text answers and the challenges that are unique to it, thus emphasizing the importance of more research in this area. One major challenge as described previously is the presence of various noise types in the crowdsourced answers data. In this section we describe a preprocessing pipeline to denoise the answers. We also provide description of the semantic similarity measures we use to assess the user answers at scale in low resource setting.

4.1 Data Preprocessing

The purpose of data preprocessing is to improve text quality for downstream evaluation methodologies. The quality of data can have a significant influence on assessment methods hence, removing the noise in the crowdsourced free-text data is essential. Our proposed preprocessing pipeline is shown in Figure 2.

First, we denoise the text data by removing the unwanted elements like the emoji, url, superscript/subscript characters and replace punctuations with space. For example,

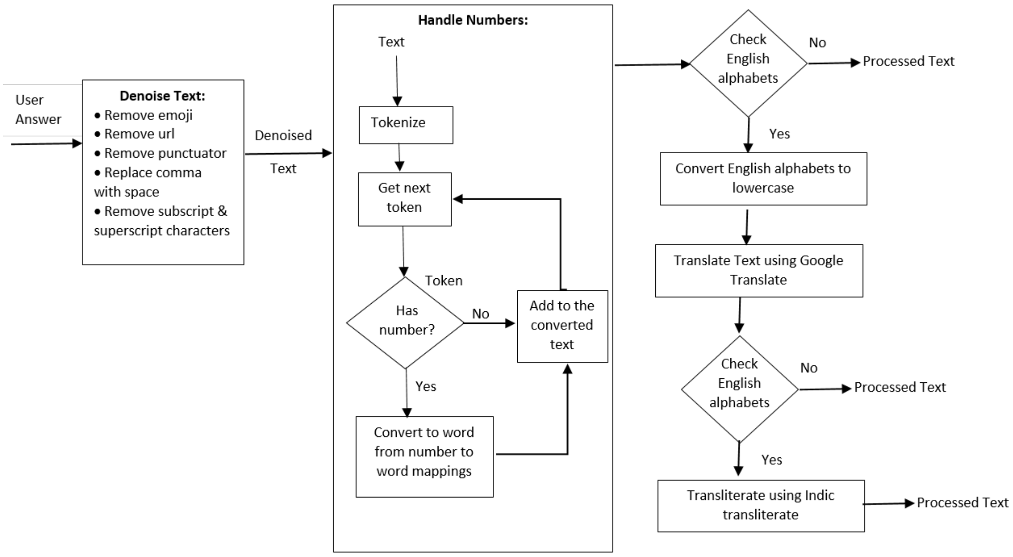

User Answer
Denoise Text:
• Remove emoji
• Remove url
• Remove punctuator
• Replace comma with space
• Remove subscript & superscript characters
Denoised Text
Handle Numbers:
Text
Tokenize
Get next token
Token
Has number?
No
Add to the converted text
Yes
Convert to word from number to word mappings
Check English alphabets
No
Processed Text
Yes
Convert English alphabets to lowercase
Translate Text using Google Translate
Check English alphabets
No
Processed Text
Yes
Transliterate using Indic transliterate
Processed Text

Figure 2: Data Preprocessing pipeline

User Answer: रक्त का थक्का नही जमेगा\n\ n😀 😀😀😀😀😀 \n
Processed Text: रक्त का थक्का नही जमेगा (ETL: blood clots will not form)

Next, this text is processed to convert numbers to their corresponding word lexemes. For example,

User Answer: 5 किलो ग्राम (ETL: *5 Kilogram*)
Processed Text: पांच किलो ग्राम (ETL: *Five Kilogram*)

We then check for presence of English alphabets in the processed text, which we identify and then pass through Google Translate service to convert it to Hindi. For example,

User Answer: We can not move , we can not work
Processed Text: हम हिल नहीं सकते हम काम नहीं कर सकते (ETL: *We can't move, we can't work*)

As mentioned before there are additionally a few instances where the Hindi text is transliterated to English which needs to be converted back to Hindi. These are not converted by the Google translation service and hence we check for English alphabets again and transliterate them to Hindi. We use the Indic Transliteration library to do the same. For example,

User Answer: hilna dulna asmbhav
Processed Text: हिलना दुलना असम्भव (ETL: *impossible to move*)

This final denoised data is now ready for assessment, methods for which are described in next section.

4.2 Answer Assessment Methodology

We model the assessment of user answers against ideal answers as a similarity task. It is important to note that the focus of our study is for the rural children in remote areas of India with limited internet access. Hence, we need to find a solution which could be integrated with a system in low resource setting. Thus, we choose similarity measures instead of paraphrase detection as our assessment methodology to ensure it does not add a lot of memory requirement or internet connectivity to the app. Additionally, training robust Paraphrase identification systems requires availability of large amounts of corpus that we do not have the luxury of for Hindi language, and education data. We thus lean towards using word embedding based similarity measures to capture semantic similarity among user and ideal answers. To compare performance of these similarity measures on our dataset, we need ground truth dataset of manually evaluated actual vs user answer pairs. The methodology used for creation of this ground truth is described below.

Ground Truth using Human Evaluation: A web portal was created to evaluate the answers of children with respect to questions and model answers. Every answer contains three options for the evaluator - correct, incorrect, can't say. An additional optional field Remarks was given to highlight any comments/irregularities that were seen in the data. The answers were evaluated by corporate volunteers of Pratham. Each answer was evaluated by two people, and we collect human evaluations on a total of 15479 user answers.

4.3 Semantic Similarity Measures

Each user answer is assigned a similarity score with ideal answer using the various scores described in this section. The intent is to measure semantic similarity and not syntactic similarity for the purpose of this task. Since this task is for Science subject, ensuring the user answer matches semantically to the ideal is more important than ensuring syntactic correctness. We use the following similarity scores to provide a benchmark on the dataset.

1. **Baseline:** We compare random score assignment with the described scores as baseline. This is generated as by marking the user answer as correct based on a coin toss.

2. **Jaccard Similarity:** It calculates the number of words from user answer appearing in the ideal answer sentence. This is normalized w.r.t the total number of words present in the given answers as described in equation (1), where J is the jaccard similarity score between C, the set of words in user answer and I, the set of words in ideal answer.

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

3. **Semantic Similarity Scores:** We segment the user and ideal answer into their constituent words. We then retrieve word embeddings for each of these words. User and ideal answer are represented as vectors by taking the average of these word embeddings. We then calculate the cosine similarity score between the answers. The Hindi word embeddings used are:

IndicNLP: 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).

5 Results and Analysis

We now compare the performance of various similarity scores on the human evaluated ground truth on answer assessment. We study the performance of these similarity measures before and after denoising and observe that denoising leads to considerable improvements in their performance.

5.1 Results

The results have been measured on 9055 user answer, ideal answer pairs out of 15479 answers where both human evaluators matched in their markings. Among these 9055 evaluated and matched answer pairs, 30% are correct and 70% are incorrect answers. The marking is converted to binary where we assign label 1 if the human evaluators marked the user answer as correct, and 0 if the human evaluator marked the answer as wrong.

We compare the performance of various similarity measures described earlier on this evaluation data. Figure 3 and Figure 4 show the ROC curves for each of these similarity measures along with the area under the ROC curve (ROC AUC). In Figure 3, we plot the performance of similarity scores on the original noisy user data. In Figure 4, we plot the performance of similarity scores on denoised user data (output from our preprocessing pipeline). It can be observed that embeddings based semantic similarity scores have similar performance on our answer assessment dataset and considerably outperform the simple jaccard similarity method on noisy data.

It is very interesting to observe that though the performance of embedding based methods sees an ~8.9% lift upon denoising, the performance of Jaccard Similarity improves considerably with a 23.1% lift making it comparable to the embedding based semantic similarity scores, fastText and indicNLP on this dataset.

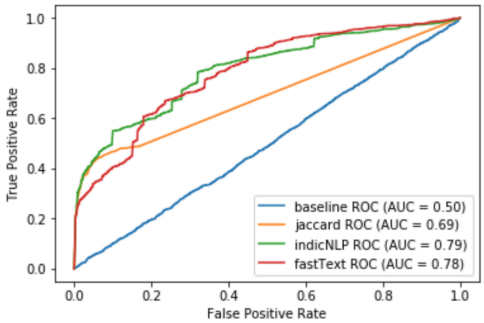


Figure 3: ROC Curves for various similarity scores on original noisy data

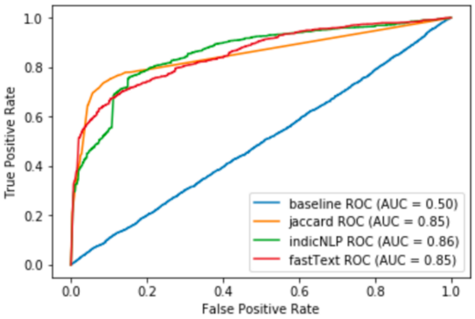


Figure 4: ROC Curves for various similarity scores on denoised data

Since we want to focus on the errors made by these similarity measures, we additionally evaluate the model based on cost, i.e. the number of wrong assessments the similarity scores result in as compared to the actual ground truth. We select the threshold to convert a given similarity score to binary that minimizes this cost. In the equation below, FP is false positives, FN is false negatives, TP is the number of true positives and TN, the number of true negatives in evaluation data.

$$\text{cost} = (FP + FN) / (TP + FN + FP + TN) \qquad (2)$$

Table 3 and 4 show the comparison of various metrics for similarity scores and the minimum cost associated with each on noisy and denoised data respectively.

Similarity	Cost	Precision	Recall	F1 Score
baseline	0.267	0.43	0.00	0.00
jaccard	**0.185**	**0.86**	**0.36**	**0.51**
indicNLP	**0.183**	**0.85**	**0.37**	**0.52**
fastText	0.205	0.86	0.27	0.41

Table 3: Performance of similarity measures on original noisy user answers

Similarity	Cost	Precision	Recall	F1 Score
baseline	0.281	0.27	0.00	0.00
jaccard	**0.126**	**0.82**	**0.69**	**0.75**
indicNLP	0.169	0.70	0.69	0.69
fastText	0.149	0.87	0.54	0.67

Table 4: Performance of similarity measures on denoised user answers

5.2 Error Analysis

We show a few examples comparing Jaccard Similarity errors and indicNLP Similarity errors on the denoised data. Row 1 in Table 5 shows an example where the answer is incorrect, but Jaccard Similarity assigns it a high score due to matching word "तरंग / waves". Row 2 shows error made by indicNLP as it assigns a high similarity score to word pair ("वायुमंडल / atmosphere", "गुरुत्वाकर्षण / gravitation") perhaps because they appear in similar context often. Additionally, row 3 shows error made by indicNLP as it assigns a high similarity score among numbers and number names.

User Answer / ETL	Ideal Answer / ETL	Ideal	Jaccard	indic NLP
मुखिया तरंग / Head waves	विद्युत् चुम्बकीय तरंग / electromagnetic waves	0	1	0
क्योंकि वहां पर गुरुत्वाकर्षण कम होना है / Because there's less gravitation	वायुमंडल ना होने के कारण / Due the absence of atmosphere	0	0	1
तीन / three	पांच प्रकार / five types	0	0	1

Table 5: Errors by Jaccard and indicNLP in assessment

5.3 Ablation Study

We have demonstrated performance of various similarity measures on both original (noisy) and denoised evaluation data. In this section, we perform an ablation study of different noise types and effect of their removal on the performance of each of these similarity measures to get a better understanding of their relative importance.

We consider three buckets of noise types: (1) Punctuations, urls and emojis (2) Presence of numbers (3) Translated or transliterated text. These buckets are created based on prevalence from Table 2.

Table 6 shows the performance of similarity measures on removing each of these noise buckets individually and compared with the original (noisy) answers plus the fully denoised answers. We observe that while removal of each noise type leads to improvement in the similarity measure AUCs, it is the cumulative effect of removing all the noise types that boosts their performance overall.

Similarity Score →	**jaccard**	**indicNLP**	**fastText**
Original Answers	0.69	0.79	0.78
Punctuations, urls, emojis removed	0.69	0.80	0.79
Numbers processed	0.84	0.82	0.83
Translated, Transliterated Answers	0.70	0.82	0.80
Fully Denoised Answers	**0.85**	**0.86**	**0.85**

Table 6: ROC AUC of similarity measures upon each individual noise type removal

5.4 Discussion

Even though Jaccard similarity is a crude measure of similarity, it outperformed other semantic similarity scores for answer assessment task upon denoising. Lift of 23.1% in performance of a simple, feasible measure like Jaccard due to denoising is promising for us as our requirement is to deploy the solution in low resource setting where fast and accurate computation of a similarity score for assessment is very critical in absence of internet connectivity. Jaccard similarity, however, fails to capture the synonym relations between words, and while intuitively it seems that it would underperform compared to embeddings, this benchmark shows that in order to outperform a simple score, there is considerable scope for improvement in embeddings, especially for education domain. We also observed that embeddings for words like *ऑक्सिज़न* (ETL: *Oxygen*), *भारहीनता (ETL: Weightlessness), तारत्व(ETL: String), रक्तकणिका (ETL: Blood cell)* etc. are absent from fastText embeddings leaving scope for improvement through fine tuning. We believe that using science textbook data for fine tuning the word embeddings (trained on generic data), can help alleviate some of the mentioned concerns.

6 Conclusion

In this paper we have described various noise types and proposed a preprocessing pipeline to denoise crowdsourced free-text answers provided by children for grade 8 level Science topics in Hindi. This work is intended to facilitate research in automatic assessment of student's free text answers in regional India languages in low resource settings. We have compared the performance of various semantic similarity scores using human evaluated ground truth on their original noisy and denoised versions. We see that denoising helps Jaccard Similarity outperform semantic similarity measures thus presenting a strong case for feasibility of automated assessment in low resource settings. In the next phase of this work, we will fine tune the existing embeddings on education domain. With Pratham's reach into 22 states and up to 15 million children in India, we can scale our crowdsourcing easily to include more responses in other regional languages. We hope that our findings mutually benefit the research community working in the area of descriptive answer assessments for regional languages meanwhile solving a very practical problem for society at scale.

Acknowledgements

This work has been supported by the Pratham Education Foundation. We are thankful to the corporate volunteers who helped us create the evaluation data for this work. We are also thankful to the web and android application team of Pratham for enabling the collection and recording of this data.

References

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar, 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv preprint* arXiv:2005.00085.

Daniele Bonadiman, Anjishnu Kumar, and Arpit Mittal, 2019. Large Scale Question Paraphrase Retrieval with Smoothed Deep Metric Learning. *arXiv preprint* arXiv:1905.12786.

Diana Pérez, Alfio Massimiliano Gliozzo, Carlo Strapparava, Enrique Alfonseca, Pilar Rodríguez, and Bernardo Magnini, 2005. 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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov, 2018. Learning word vectors for 157 languages. *arXiv preprint* arXiv:1802.06893.

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

Philip G. Butcher, and Sally E. Jordan, 2010. A comparison of human and computer marking of short free-text student responses. In *Computers & Education* 55, no. 2 (2010): 489-499.

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

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

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, and K. Toutanova, 2019. Natural questions: a benchmark for question answering research. In *Transactions of the Association for Computational Linguistics*, *7*, pp.453-466.

Transliteration tools to convert text in one indic script encoding to another.
https://pypi.org/project/indic-transliteration/

Translator: Google Translate API for Python.

https://pypi.org/project/googletrans/