UKARA: A Fast and Simple Automatic Short Answer Scoring System for Bahasa Indonesia

Guntur Budi Herwanto[1], Yunita Sari[2], Bambang Nurcahyo Prastowo[3], Mardhani Riasetiawan[4], Isna Alfi Bustoni[5], and Indra Hidayatulloh[6]
[1,2,3,4,5] Department of Computer Science and Electonics
Universitas Gadjah Mada, Yogyakarta, Indonesia
[6] Department of Electronics and Informatics Engineering Education
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
gunturbudi@ugm.ac.id

Abstract. This paper presents UKARA, a fast and simple automatic short-answer scoring system for Bahasa Indonesia. Automatic short-answer scoring holds an important role in speeding up automatic assessment process. Although this area has been widely explored, only very limited number of previous work have studied Bahasa Indonesia. One of the major challenges in this field is the different type of questions which require different assessments. We are addressing this problem by implementing a combination of Natural Language Processing (NLP) and supervised machine learning techniques. Our system works by training a classifier model on human-labeled data. Using three different types of Programme for International Student Assessment (PISA) student responses, our system successfully produced the F1-score above 97% and 70% on dichotomous and polytomous scoring types respectively. Moreover, UKARA provides a user-friendly interface which is simple and easy to use. UKARA offers a flexibility for human grader to do re-scoring and re-training the model until the optimal performance is obtained.

Keywords: Automatic assessment, Bahasa Indonesia, Machine Learning, Natural Language Processing

INTRODUCTION

Research in the area of automatic short-answer scoring is getting more and more attention due to the need of making the assessment process faster. Approaches to this problem in general can be divided into two different groups: *reference* and *response*-based approaches (Sakaguchi, Heilman, & Madnani, 2015). The first approach works by comparing the student responses to the reference answers. Responses with higher semantic/syntactic similarity to the reference answers most likely to get higher score. Various text similarity methods (Agirre, Cer, Diab, Gonzalez-Agirre, & Guo, 2013) are commonly used in this approach. In contrast to the first method, *response*-based approach relies on the availability of pre-scored student responses in order to train good classification model. Features such as Bag-of-Words (BoW) and n-grams with powerful classifier like Support Vector Machine (SVM) have been proved to be effective (Magooda, Zahran, Rashwan, Raafat, & Fayek, 2016; Mohler, Bunescu, & Mihalcea, 2011; Pado, 2016) for this task.

Despite the rapid progress on automatic short-answer assessment, research in this area particularly for Bahasa Indonesia has been very limited and only recently emerged as a topic. According to Ratna et. Al (Ratna, Purnamasari, & Adhi, 2015), the main challenges on developing automatic short-answer grading for Bahasa Indonesia is the use of informal language in writing context. The diversity of local language of the student is one of many factors that influence the writing style and the vocabulary used by the students. In addition, students often include slang words which commonly used in daily conversation.

In this paper, we present UKARA, a fast and simple short-answer scoring system for Bahasa Indonesia. We implement a *response*-based approach to assign score to the student responses. BoW and character n-gram features with supervised classifiers are applied. The system is tested on three different Programme for International Student Assessment (PISA) student responses. Our system successfully obtained the F1-score above 97% and 70% on dichotomous and polytomous scoring types respectively. Moreover, UKARA provides a user-friendly interface which is simple and easy to use. UKARA offers a flexibility for human grader to do re-scoring and re-training the model until the optimal performance is obtained.

The remainder of this paper is structured as follows: in the next section, we provide a brief

review on related work. Section III presents the overview of the system. Section IV present the detail of datasets used in UKARA. Result and discussion are provided in Section V followed by conclusion and future work in Section VI.

RELATED WORK

Burrows et. al (Burrows, Gurevych, & Stein, 2015) provided a comprehensive review of automatic short-answer grading and divided the approaches into five different eras: concept-mapping, information extraction, corpus-based method, machine learning and evaluation. The historical review presented by Burrows et. al emphasized that there is a big movement towards reproducibility, standardized corpora and permanent evaluation in the area of automatic short-answer grading. The evaluation forum such as SemEval Semantic Textual Similarity (STS) (Agirre et al., 2014, 2015; Agirre, Cer, Diab, & Gonzalez-Agirre, 2012; Agirre et al., 2013) has made a major contribution to the fast progress of this task.

Current work on automatic short-answer scoring mostly falls into machine learning era in which natural language processing techniques are combined with classification or regression models. As an example, Sultan et. al (Zedan & Al-Sultan, 2017) applied some features such as term weight, length ratio, lexical, and semantic similarities with supervised learning algorithm. Their experiments, which were performed on SemEval-2013 (Agirre et al., 2013) and Mohler (Mohler et al., 2011) datasets concluded that the augmentation of key grading-specific construct (i.e question demoting) to text similarity features are proved to be effective by producing top results on multiple benchmarks. Similar to Sultan et. al, Pado (Pado, 2016) employed an array of features including n-gram, text similarity, dependency, abstract semantic representation and entailment votes. All features were computed in relation to the reference answer given for each question. His experiment results on two types of corpora: language skill and content assessment showed that each feature can have different levels of effectiveness depends on the corpus type.

We found very limited literature of automatic short-answer scoring for Bahasa Indonesia. One of them is a work by Ratna et. al (2015) who introduced a web-based essay grading system called SIMPLE-O. Their system utilized Latent Semantic Analysis (LSA) to estimate the similarity between student responses and reference answers. Their experiments which were conducted with 40 students and 3 lecturers as the grader demonstrated that SIMPLE-O obtained agreement with human raters above 86%.

SYSTEM OVERVIEW

UKARA is a web platform that automatically grades freely written text for the grader. The design of a web platform is to make the system easily accessible for people who want to use the system. The basic intuition of our automatic grading system is, the grader will grade about 10% amongst all of the answers. Currently, our system works best on dichotomous scoring types. In other words, the system works best to decide whether the answer is correct or incorrect. However, we were not limiting the user to provide more than two labels or polytomous scoring types. Based on this knowledge, the system will automatically grade the rest of the answers. Our platform consists of 5 modules: (1) Create Question (2) Upload Data (3) Explore Data (4) Training the Model (5) Predict and Model Tuning. The screenshot of the application can be seen in the appendix.

1. Create Question

   We believe every question have different characteristics and analysis to decide the grade. To catch this intuition, we decide that the basic unit of our system is question. The model will be produced independently on every question. On the security terms, the grader will only see the question that they have been created.
2. Upload Data

   We assume the grader has a different kind of source of data. In order to solve these varieties, we have provided a simple spreadsheet file consists of only 3 attributes, which is the student id, the free text answer, and also the human label or grade. As we have mentioned previously, the grader only requires to fill 10% of the label and leave the rest empty. The spreadsheet then can be uploaded to the system to be ready for the next process.
3. Explore Data

   Once the data have been uploaded to the system, it is essential to discover how the data looks like. We provide a simple search and filter to observe the condition of the data. The grader can filter by the label, or search based on keyword matching. In addition, if the users want to add a supplementary label to the dataset, they can add them directly under this module. The more label provided to the system, the more intelligent the system would be.
4. Training the Model

   The next step is to tell the machine to learn from the labeled data. We are using a machine learning algorithm called Adaptive Boosting (AdaBoost) (Freund & Schapire, 1997) to learn based on the human labeling provided by the grader. After a short period of time, the system will produce a grading model, tailored for a particular question. To ensure the quality of the

model, we show the accuracy of the model to the user. We calculate the accuracy by 5-fold cross-validation method. The aim of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias (Cawley & Talbot, 2010). The accuracy score can become the basis for deciding whether we are confident to use the model for a prediction on the next step. Based on our experience, if the accuracy score achieved more than 90%, we can be confident enough to predict the rest of the answer.

5. Predict and Model Tuning

   The final step of the system is to predict the rest of the data that has not been labeled. The prediction will utilize the model that has been created on the previous steps. The process will take time depends on the number of answers that will be predicted. In our experience, it will take around 5 minutes to complete 6.000 answers from the dataset. The final prediction can be seen on the system or can be downloaded in the form of a spreadsheet to be examined in offline by the grader.

The spreadsheet basically will fill the blank data in label column that has been imported previously. We are sure that there will be a misclassified grade that occurs, depends on the accuracy score. In order to improve the model, the grader can manually change the label and then import it back to the system. Hypothetically, the accuracy score will be improved based on the additional label from the user. This model then can be used for the future answer that employs the same question and grading criteria.

DATASET DETAILS

UKARA is tested on three different PISA datasets. These datasets represent a range of characteristics in term of response and scoring methods. All responses mostly consist of one or two sentences.

Machu Picchu

Machu Picchu consist of 6776 pre-scored student responses. This dataset is categorized as constructed response with dichotomous scoring. Students were given a 300-word reading passage that describes the offer from a travel agent to visit Machu Picchu. Table 1 shows the scoring guideline for this dataset.

Question:

Sebutkan dua cara yang disebutkan di brosur di mana calon pelanggan dapat memperoleh informasi lebih lanjut mengenai layanan yang ditawarkan oleh Biro Perjalanan Buana Wisata!

Table 1. Scoring Guidelines for Machu Picchu

Score Code	Scoring Guideline
1	if both responses are true. Responses are considered true if students mentioned the phone number and website URL OR the students mentioned the method to get the information (call the travel agent and visit the travel agent's website)
0	if only one response is true, or the responses are irrelevant to the question.

Sample real responses:

- telepon (021)73811111 website: www.buanawisata.com
- berbagai perjalanan wisata yang akan membantu mehnguba kota yang sunyiini menjadi hidup

Jaket

This dataset was constructed from 3125 pre-scored student responses and falls into constructed response with polytomous scoring category. Students were given a study case in order to evaluate their financial literacy skills. The scoring guideline for this dataset is presented in Table 2.

Table 2. Scoring Guidelines for Jaket

Score Code	Scoring Guideline
11	if the response related to the idea of spending money unwisely by buying unnecessary thing.
12	if the response related to the idea of spending money lavishly.
00	if the response is irrelevant to the question.

Sample real responses:

- Maman baru saja membeli jaket dan juga dia memiliki jaket yang serupa jadi akan terjadi pemborosan jika maman membelinya juga.
- karena dia sudah memiliki jaket yang sama

Sepeda

This dataset was constructed from 3144 pre-scored student responses and falls into constructed response with polytomous scoring category. There 4 score codes for this dataset which are: 11, 12, 13 and 00. Here are the sample responses for dataset sepeda:

- kirana dapat memutuskan untuk menyewa sepeda seharga yg lebih mahal agar kirana dapat membeli sepeda lain

- keuntungannya adalah 8 minggu biayanya 560 lebih kecil daripada membeli

RESULT AND DISCUSSION

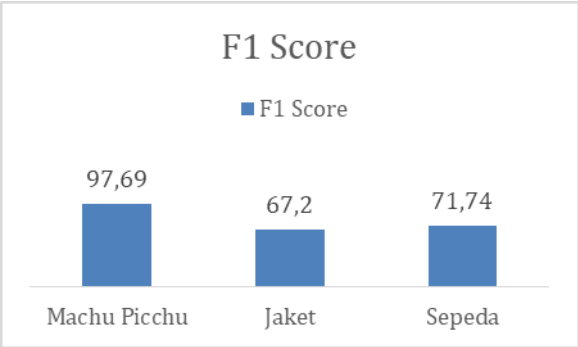
The main concern of our platform is to help the grader obtain an accurate result of the prediction. We evaluate the model using F1 Score. The formula for F1 score can be seen in equation (1):

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (1)$$

Precision and recall are the underlying measure for F1 Score. The precision will tell that the true answer prediction is actually true, and the recall will tell how the model able to find the correct answer on the data. The F1 score gives fair weight to both measures. The formula for F1
We applied the same training model on three datasets that we have previously defined. The detail result can be seen on Table 3:

Table 3. System performance

Dataset	F1 Score
Machu Picchu	97.69
Jaket	67.20
Sepeda	71.74



From the result, we can see the best F1 score obtained from the Macchu Picchu dataset, which has 97,69 score. Followed by Sepeda dataset and Jaket dataset. This score shows the superiority of dichotomous prediction vs polytomous prediction.

CONCLUSION AND FUTURE WORK

In this paper, we present UKARA, a fast and simple automatic short-answer scoring for Bahasa Indonesia. Based on our experiment results, our system works best on dichotomous type of question with F1-score of 97.69 %. There is, however, immense scope for improvement. We find that polytomous type of question is challenging and needs further exploration. In addition to that, automatic typo correction is necessary to be applied, since our system tends to suffer performance drop when dealing with typo and informal language.

ACKNOWLEDGEMENT

The authors would like to acknowledge Pusat Penilaian Pendidikan (PUSPENDIK), Badan Penelitian dan Pengembangan, Kementerian Pendidikan dan Kebudayaan, Republic of Indonesia for providing datasets and funding support.

*****

REFERENCES

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2014). SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 81–91). Association for Computational Linguistics. https://doi.org/10.3115/v1/S14-2010

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., … Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 252–263). Association for Computational Linguistics. https://doi.org/10.18653/v1/S15-2045

Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics -- Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 385–393). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/S12-1051

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., & Guo, W. (2013). *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (pp. 32–43). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/S13-1004

Burrows, S., Gurevych, I., & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, *11*, 2079–2107.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting *, *139*, 119–139.

Magooda, A. E., Zahran, M. A., Rashwan, M. A., Raafat, H. M., & Fayek, M. B. (2016). Vector Based Techniques for Short Answer Grading. In {*FLAIRS*} *Conference* (pp. 238–243). {AAAI} Press.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). 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). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/P11-1076

Pado, U. (2016). Get Semantic With Me! The Usefulness of Different Feature Types for Short-Answer Grading. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2186–2195). The COLING 2016 Organizing Committee. Retrieved from http://aclweb.org/anthology/C16-1206

Ratna, A. A. P., Purnamasari, P. D., & Adhi, B. A. (2015). SIMPLE-O, the Essay Grading System for Indonesian Language Using LSA Method with Multi-Level Keywords. *The Asian Conference on Society, Education & Technology 2015*. Retrieved from http://papers.iafor.org/wp-content/uploads/papers/acset2015/ACSET2015_18875.pdf

Sakaguchi, K., Heilman, M., & Madnani, N. (2015). Effective Feature Integration for Automated Short Answer Scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1049–1054). Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1111

Zedan, H., & Al-Sultan, S. (2017). The specification and design of secure context-aware workflows. *Expert Systems with Applications*, *86*, 1339–1351. https://doi.org/10.1016/j.eswa.2017.05.078

APPENDIX

Here, we present the user interface for UKARA. Figure 1 shows the homepage of UKARA which consists of list of question set. Figure 2 shows the explore data module of one question set.

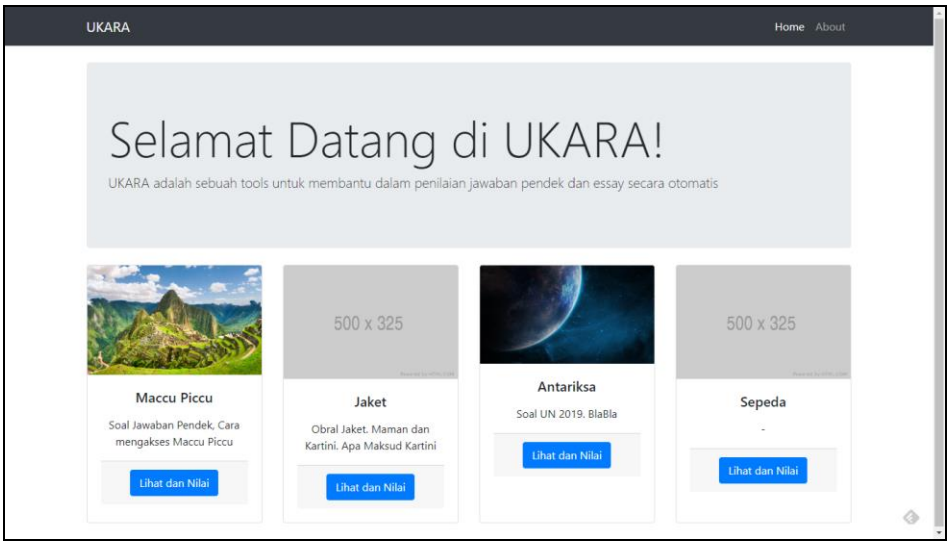


Fig 1. UKARA Home Page

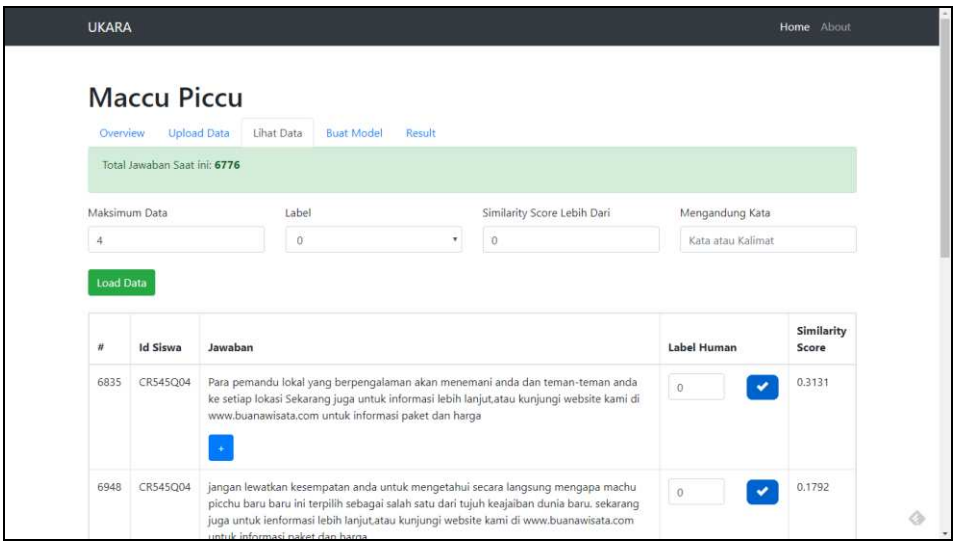


Fig 2. Explore Data