An Attempt to Create an Automatic Scoring Tool of Short Text Answer in Bahasa Indonesia

Husni Thamrin, Jan Wantoro
Dept. of Informatics
Universitas Muhammadiyah Surakarta
Sukoharjo, Indonesia
husni.thamrin@ums.ac.id

***Abstract*— Closed questions offer poor information on student's ability to manage and apply knowledge. On the other hand, open questions have advantages because it may be used to grasp students' conceptual maturity and ability of communication. However, scoring open question answer is not trivial and time-consuming so an automatic scoring tool becomes necessary. An attempt was made to create a scoring tool for open and short text question answer in Bahasa Indonesia that resembles the way school teachers do scoring. Automatic scoring of a student answer was based on the similarity between the answer and predefined key answers. The proposed automatic scoring tool has a form of correlation with human scoring so that the model may be used to predict teacher scoring.**

***Keywords*— automatic scoring; short text; bahasa Indonesia**

I. INTRODUCTION

Assessment in a teaching learning process is not a trivial work in all of its stages, including the scoring. Effort to make tools to help assess students work has been a long activity [1]. Currently, automatic scoring can be successfully carried out on closed questions [2], and they include multiple choice, with one or more correct answers, true or false, short answer, a word or simple phrase from a list, numerical, matching, and calculated questions.

Closed questions have been useful to check whether students have grasped the essentials. However, closed questions offer poor information on the student's ability to actively manage and apply their recently acquired knowledge [3]. Open questions are a common way to get an insight into the conceptual maturity a student may have achieved after a learning period. In open questions the student is asked to produce a piece of text using his/her own choice of words and form of expression (free-text answers). Questions can be very specific, with only a word or two for a correct answer, or require the student to write a brief essay on the subject.

Closed questions may be used cost effectively to verify comprehension. On the other hand, giving only closed questions in an assessment may cause student thought be filled with knowledge that is out of context. Knowledge obtained without proper or authentic context is not engaging or motivating, thus making it hard for students to comprehend and remember a concept [4]. Consequently, students will find it difficult to implement the concept to the real world. Generally, asking students to construct something in open questions provides better assessment than asking them to select among a small set of alternatives [5]. The use of open questions provides better benefits in at least two dimensions of assessment: conceptual maturity and ability of communication.

Even for short answer questions, the marking of open question tests proves much harder than that of closed questions tests [3]. The automatic marking of free text answers is still a field of research and researchers have been obtaining various results.

This writing describes an effort to develop automatic scoring of open short text question tests in Bahasa Indonesia. The mark is obtained by calculating the similarity of student answers to key answers. We have surveyed teachers to get insight into how they score open question tests. We found that teachers have a high variation in the way they mark student answers and that a similarity score is fairly indicative to predict the mark of teachers.

II. RELATED WORKS

A number of approaches have been proposed in the past for automatic scoring of short text answer. Among the best papers that present a result study for this topic is the one by [1]. They compare several knowledge-based and corpus-based methods of measuring semantic similarity of texts in auto grading system. They collected data from three student assignments and asked two human judges to score the answers by comparing the student answers with the golden key answers. Later on, similarity methods were applied to measure similarity between student answers to the golden answers. The researchers found that the corpus-based LSA (latent semantic similarity) methods give a measure with the best correlation compared to other methods.

Another research [6] develops a method called SynSemSim (syntactic-semantic-similarity) to measure the similarity of shot texts. They applied the method to an auto grading system and concluded that the method may successfully calculate the similarity of texts that have similar structure. The method fails to calculate similarity of compound sentences and sentences with a lot of stop word.

We have not found any researches in automatic scoring that took Bahasa Indonesia as the object. Most researches on Bahasa Indonesia have been about stemming [7, 8], clustering [9, 10, 11] and plagiarism detection [12].

III. Method

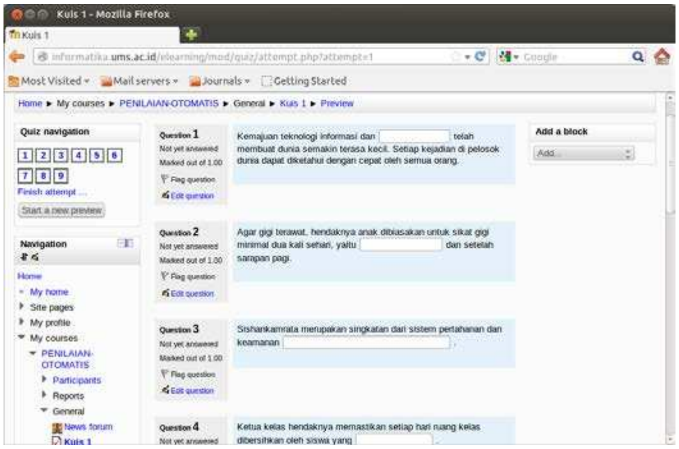
We made a forked version of Moodle learning management system so that it provides a short answer question type with automatic scoring. Moodle has already had a short answer question type and the scoring is based on the exact matching between a student answer and one or more key answers. Our version has an additional question type similar to short answer question but the scoring is calculated based on semantic similarity of a student answer and key answers. In case of misspelling, our version scores student answers by calculating the edit distance (such as Levenshtein distance).

We conducted a survey to learn how Indonesian teachers score students' answers. We made question and answer sheets with a total of 40 questions that were answered by students. We asked more than 150 local teachers to score the answers. Students had simulatively answered the questions in various ways, some with an exact match to the key answers, some with wrong spelling, and some others with synonymy or phrases that have related meaning to the key answer. Later on, teachers scoring were statistically analyzed and compared to methods of automatic scoring.

IV. Result and Discussion

Rather than creating our own software, we have chosen to modify Moodle to provide an open short text question type. Moodle is an open source software, which means that the software may be changed or re-engineered to suit one's needs. Creating a new question type involves several steps from copying and modifying a folder and several files. Closed short text question type is already available in Moodle. Therefore, creating the open short text question type is simplified by changing the scoring algorithm of the closed question type. Consequently, the look of the modified question type does not change (see Fig. 1).

Scoring of closed short text question type is conducted by comparing the student answer to one or more key answers. In the modified open version, scoring is conducted in three conditional steps. Firstly, if a student's answer has a match to one of the key answers, the student get a maximum score.



Question 1
Not yet answered
Marked out of 1.00

Kemajuan teknologi informasi dan ______ telah membuat dunia semakin terasa kecil. Setiap kejadian di pelosok dunia dapat diketahui dengan cepat oleh semua orang.

Question 2
Not yet answered
Marked out of 1.00

Agar gigi terawat, hendaknya anak dibiasakan untuk sikat gigi minimal dua kali sehari, yaitu ______ dan setelah sarapan pagi.

Question 3
Not yet answered
Marked out of 1.00

Sishankamrata merupakan singkatan dari sistem pertahanan dan keamanan ______.

Question 4
Not yet answered

Ketua kelas hendaknya memastikan setiap hari ruang kelas dibersihkan oleh siswa yang ______.

Fig. 1. Modified version of Moodle showing several short answer questions

Secondly, if the answer has a lexical meaning, scoring is calculated based on the lexical relationship such as synonym or hyponymy. Lastly, if a student answer is not in a standard dictionary, then scoring is based on the edit distance between the student answer and one of the key answers.

An attempt was made for the scoring algorithm to predict the scoring of human, who are the local teachers. They were asked to score question and answer pairs, as if the answers were made by students. Various answers were simulated. Of all 40 questions, one was answered with exact match to the key answer and 2 were answered using synonymy. Additionally, 3 answers were unrelated, 7 answers contain typo resulting words without meaning, while 3 answers contain typo having new meanings. An example below shows a question and a student answer that has typographical error causing it to have a new meaning:

> *Setelah sarapan pagi, Nina berkeras dengan memasukkan pakaian dan peralatan lain ke dalam ransel. Kepergiannya kali ini hanya untuk dua hari sehingga bawaannya tidak banyak.*
> (After breakfast, Nina _____ by putting her clothes and accessories into her backpack. Her travel this time was just for two days so she did not bring many items.)

Key answer for the above question is “berkemas” (to do packing), and a student had answered the question with “berkeras” (to insist).

Table 1 shows the average scores that teachers gave to various types of student answers. An answer that matches the key was scored nearly 100 points and the others were scored expectedly less depending on the similarity of the answers with the key. The higher the similarity of a student answer to the key answer, the higher the score.

TABLE I. Teachers' Scoring on Students' Answers

Relationship between students' answers and key answers	Average score (by teachers)	Standard Deviation
Exact match	99.6	3
Synonymy	76	23
Hyponymy	81	18
Hyponymy, 2 level	72	22
Hypernymy	72	21
Hypernymy, 2 level	57	26
Meronymy	62	27
Holonymy	62	29
1 letter typo, no lexical meaning	70	15
2 letter typo, no lexical meaning	59	23
1 letter typo, has lexical meaning	43	35
2 letter typo, has lexical meaning	34	30
No relationship	35	30

Table 1 presents several scoring data when a student’s answer and the key answer are related in the context of elements of a WordNet. WordNet element typically consists of word, word definition, and related words. Related words can be hyponyms/troponyms (words that have smaller scope in meaning), hypernyms (words that have wider scope in meaning), and synonyms (words that have similar meaning). One node in a WordNet is called a synset which contains the word, definition, synonyms and reference to the related words. So synonyms are in the same synset and two synsets are directly connected if the containing word of a synset is related to the word of the other synset.

Teachers' average score has a kind of positive correlation with the distance of two synsets. Distance of two synsets is measured by counting the number of synset nodes connecting both synset in the WordNet. When the two words are in the same synset (i.e. they are synonymous and the distance is zero), teachers’ score is high. On the other hand, when the distance is 2 or both words are 2 level apart, lower score was given. This result agrees with what was proposed by previous researchers [13, 14] who pointed out that shorter path from one node to another in a WordNet resembles higher similarity of the containing words.

Assessment to answers with typographical errors results to different score depending on the number of mistyped letters and whether the errors cause the word to have different meaning. One letter typo is regarded acceptable as long as it does not make a new word. However, more than one letter typo is not tolerable so that teachers gave more severe penalty. On the other hand, answers with typos that make a new word is perceived similar to unrelated answers.

Looking at the table, we suggest that the most suitable model to predict teachers' score is a set of logical rules; such as if answer is synonymy, then score is 80. We could not establish a mathematical formula that relates teacher score and the distance between a student answer and the key answer. Our survey shows, for instance, that synonymous answer is scored similar to the first level hyponym answer. In contrast, Resnik's statement suggests that two synonymous words have higher similarity than two hyponymous words, so the former should have higher score than the latter. The use of the rule should mind the condition that the score is predictive and the standard deviation is large. Hence an individual teacher may give score quite differently from the score of the model. Large standard deviation means that teachers’ scoring varies considerably and they are quite subjective when they score open short text answer.

V. CONCLUSION

Automatic assessment of open short text answer can be implemented as a tool to help teachers score students' answers. Teachers score has a positive correlation to the similarity of answers and the keys. It is suggested that a set of logical rules may be implemented to predict teachers scoring.

REFERENCES

[1] Mohler, M., & Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, (2009) 567-575.

[2] Cole, J.: Using Moodle. Moodle. http://download.moodle.org/docs/using_moodle/ ch5_quizzes.pdf (2006)

[3] Gonzalez-Barbone, V., Llamas-Nistal , M.: eAssessment of Open Questions: an Educator's Perspective , 38th ASEE/IEEE Frontiers in Education Conference (2008) F2B1-F2B6

[4] Jona, K.: Rethinking the Design of Online Courses, in Proc. ASCILITE 2000, New South Wales, Coffs Harbour, (2000)

[5] Sargeant, J., McGee Wood, M., Anderson, S. M.: A Human Computer Collaborative Approach to the Marking of Free Text Answers, Proceedings of the 8th CAA Conference, Loughborough University (2004)

[6] Winarsono, D., Siahaan D.D., Yuhana, U.: Sistem Penilaian Otomatis Kemiripan Kalimat Menggunakan Syntactic-Semantic Similarity pada Sistem E-Learning, Kursor vol. 5 no. 2 (2009) 75-82.

[7] Talla, F.: A study of stemming effects on information retrieval in Bahasa Indonesia, (2003)

[8] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Williams, H. E.: Stemming Indonesian: A confix-stripping approach, ACM Transactions on Asian Language Information Processing (TALIP), vol. 6, no. 4, (2007) 1-33.

[9] Hamzah, A., Soesianto, F., Susanto, A., Istiyanto, J.E.: Studi Kinerja Fungsi-Fungsi Jarak dan Similaritas dalam Clustering Dokumen Teks Berbahasa Indonesia, Prosiding Seminar Nasional Informatika 2008, Yogyakarta (2008)

[10] Asy'arie, A. D., & Pribadi, A. W.: Automatic news articles classification in indonesian language by using naive bayes classifier method, Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, (2009) 658-662.

[11] Hamzah, A.: Temu Kembali Informasi Berbasis Kluster untuk Sistem Temu Kembali Informasi Teks Bahasa Indonesia, Jurnal Teknologi, vol.2, no.1, (2009) 1-7.

[12] Hamzah, A.: Aplikasi N-Gram Untuk Deteksi Plagiat Pada Dokumen Teks, Prosiding Seminar Nasional Teknoin (2011).

[13] Resnik, P.: Using information content to evaluate semantic similarity, Proceedings of the 14th International Joint Confer-ence on Artificial Intelligence, (1995)

[14] Wu, Z., and Palmer, M.: Verb semantics and lexical selection, Proceedings of the Annual Meeting of the Association for Computational Linguistics., (1994)