INDONESIAN ESSAY GRADING MODULE USING NATURAL LANGUAGE PROCESSING

Try Ajitiono
Informatics Engineering
Bandung Institute of Technology
Bandung, Indonesia
13512052@std.stei.itb.ac.id

Yani Widyani, S.T., M.T.
Informatics Engineering
Bandung Institute of Technology
Bandung, Indonesia
yani@informatika.org

Abstract— **Moodle is a LMS application which serves as an online courses platform. As for now, Moodle doesn't have capability to grade essay question automatically, so teachers should grade all the submitted answers by students one by one. This is a problem for teachers because when two sentences are being compared, they don't only compare them syntactically, but also semantically. Our research proposes an essay grading module to help teachers in term of grading answers syntactically using common tasks in natural language processing such as formalization, sentence detection, stemming, POS (Part of Speech) tagging, and tokenizer. The developed module consists of one grader component and two plugins integrated to Moodle, one is "essayinagrader" plugin for the question type interface and the other is "igsuggestion" plugin for the quiz report interface. To grade an answer, "igsuggestion" plugin will use the developed grader component. This component utilizes INANLP, a NLP tool for Indonesian language, to process the chosen NLP tasks. In addition of INANLP, this component also uses WordNet thesaurus to compare synonyms and Jaro-Winkler algorithm to calculate string distance. Test results show that the developed module can grade answers with the same representation correctly. However, for answers with different representation, it will classify those differences as a false answer.**

Keywords—quiz; essay; INANLP; Moodle; plugin; grading

I. INTRODUCTION

Assessment in education is one of the methods to measure students' comprehension using natural language. In general, an assessment can be a multiple choice type or essay type. However, there are many other types such as matching, short answer, and true-false. Every type of question has its advantages and disadvantages. For example, in multiple choice question type, teachers don't know the students' thinking process, while in essay question type, they can see the process through the sentences students have written. This enhances the purpose of an assessment. However, teachers must spend more time to grade students' answers [1]. As technologies grow, CBT (Computer-Based Test) or online assessment becomes a common thing because it is available as a feature in a lot of LMS (Learning Management System). Although online assessment has the same purpose with written assessment, most of online assessments' question type are multiple choice [2]. Essay question type is rarely chosen as an assessment question type because the students' answer might not be the exact same as the teachers'. Multiple choice question type, to the contrary, is easily being compared because there is only one answer [3]. This is caused by sentences constructed by natural language might have only one meaning or vice versa.

One of the solutions to overcome that problem is the utilization of NLP (Natural Language Processing), string metric, and thesaurus. With these tools and algorithms, teachers can reduce their time to grade an answer syntactically. Currently, there are a lot of NLP libraries for certain languages, but in this case, the development will be using INANLP as Indonesian is the target language. String distance will be used to calculate the similarity of two words or sentences. Lastly, thesaurus will be used to find synonym sets of a word. With these three tools/algorithms, there is a chance to develop a grader for essay question type in an online assessment.

II. RELATED TOPICS

A. Online Assessment

In [3], an assessment question type can be multiple choice, true-false, matching, short answer, and essay. Every type of question has its own advantages and disadvantages. For example, multiple choice question type advantages are can be used to assess broad range of content in a brief period, skillfully written items can measure higher order cognitive skills, and can be scored quickly. However, it has several disadvantages, such as difficult and time consuming to write good items and some correct answers can be guesses. These disadvantages are contradictory to assessment's purpose which is to measure students' comprehension. Another example is essay question type. This type of question can be used to measure higher order cognitive skills with relatively easy to write questions. It is also difficult to guess the correct answer for respondents. Although these advantages cover the weaknesses of multiple choice question type, new problems arise such as time consuming to administer and score, difficult to identify reliable criteria for scoring, and the content is limited.

As technologies grew, online assessment has become a thing with the usage of LMS (Learning Management System). In [2], there are three advantages of online assessment compared to written assessment:

978-1-5090-5671-2/16/$31.00 ©2016 IEEE

1. No time spent revising answers.
2. Absence of human mistakes on correction (depending on the exam assembling).
3. Possibility of immediate feedback and grading.

Same research in [2] shows that the student's performance gap between written assessment and online assessment is not big in term of their averages. That research gave a conclusion that online assessment is a solution to current problems, such as increasing number of students, limited amount of time, lack of adjusted tests, and inadequate methods relating to new paradigms. In [4], each question type has its own way to be graded. To grade true-false and multiple choice question type, grader must check the selected radio buttons. To grade matching question type, grader must check the matching column pairs. Lastly, to grade short answer and essay question type, grader must do it manually.

B. *Natural Language Processing*

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications [5]. In [6], there are two categories in NLP sub-problems. They are low-level and higher-level NLP tasks. Low-level NLP tasks used in this grader development are sentence boundary detection, tokenizing, stemming, and part-of-speech assignment (POS-tagging), while for the higher-level one, the used task is spelling/grammatical error identification and recovery. There are a lot of NLP libraries which can be used, such as Stanford NLP Parser, Apache OpenNLP, and INANLP.

1. Stanford NLP Parser [7] is capable to process a tokenizer, sentence splitter, part-of-speech, lemma, named-entities, constituency parsing, dependency parsing, sentiment analysis, mention detection, coreference, and open information extraction.
2. Apache OpenNLP [8] has features regarding text processing such as sentence detector, tokenizer, name finder, document classifier, part-of-speech tagger, chunker, parser, and coreference resolution. However, it is not maintained as last update was on 2014.
3. INANLP [9] has a lot of features such as sentence detector, stemming, tokenizer, formalization, part-of-speech tagger, chunker, coreference resolution, semantic analyzer, and name-entity tagger. It is exclusively made for Indonesian language.

C. *String Metric*

String metric, or string similarity, is a method to measure quantifies the similarity between two text strings for approximate string matching or comparison [10]. In [11, pp. 5], there are two matching techniques, phonetic encoding and pattern matching. Phonetic encoding is a technique that converts a name string into a code according to how a name is pronounced, while pattern matching is a technique that approximates the distance between two strings. As for this development, phonetic encoding won't be considered because of the context of online assessment. Experiments in [11, pp. 12] shows the experiment results with a lot of string similarity algorithms. In phonetic encoding, Phonex algorithm has bested three out of four categories, while in pattern matching, Jaro algorithm has bested two out of four. Its improvement, Winkler algorithm, also have relatively close score compared to Jaro. Winkler improves Jaro algorithm by adding the length of the common substring to the similarity calculation.

D. *Thesaurus*

In [12], thesaurus is a dictionary containing synonyms of words and concept correlation with these words. Synonym is words that represent the same thing. Thesaurus differs from language dictionary because language dictionary provides definition instead of synonyms. Used thesaurus in this development is Open Multilingual WordNet [13] – [17]. TABLE II shows some of the data provided by Open Multilingual WordNet.

TABLE I. Some of Open Multilingual WordNet Data [13] – [17]

Word ID and POS tag	Type	Word
00001740-a	ind:lemma	berdaya
00001740-a	ind:lemma	keahlian
00001740-a	ind:lemma	layak
00001740-a	ind:lemma	mahir
00001740-a	ind:def 0	memiliki sarana yang diperlukan

TABLE II shows that words with the same word ID and POS tag belong to the same synonym set. In addition to synonym set (row with type ind:lemma), there is definition (row with type ind:def N), where N is the number of definition.

E. *Moodle*

In [18], Moodle is a learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalized learning environments. With Moodle being an open source software, all developers can add plugin to their needs. Plugins that are related to Moodle quiz are activity modules are available things inside a course, quiz reports are different kind of interface showing students' attempts, quiz access rules are limitation on how a quiz can be accessed, question types are types of question in an assessment such as multiple choice, essay, matching, etc., question behaviors are how the quiz interact with a question, such as giving feedback, grading automatically or manually, and question import/export formats are how question bank can be utilized.

III. The Proposed Essay Grader

A. *Assessment Requirement*

The developed essay grader is capable of handling answers ranging from 1 word to 8 sentences. Essay question type can be classified into 2 categories, fill in the blank and free writing. Filling in the blank is a category which there is an exact answer, while in free writing, there are a lot of types, such as argumentative, descriptive, exposition, and persuasion. The scope of this development will be limited to fill in the blank with only one expected answer from the teachers as an input.

B. *Adding New Feature(s) to Moodle Quiz*

The problem currently is that the most used question type in online assessment is multiple choice. This is contradictory to the purpose of assessment, which is to measure student's comprehension, as the teacher can't know how the students reach an answer. In addition, according to TABLE I, multiple choice question type is relatively easy to guess. To increase the quality of the assessment, teachers can change the question type from multiple choice to essay, but it does come with some consequences. Teachers must spend a lot more time in grading these answers submitted by students. Moodle's essay question type has this problem, therefore a new question type must be added, as it is recommended to duplicate a template plugin and modify it rather than to modify it directly. Template plugin is any Moodle plugin that will be extended. For example, in this development, there are 2 developed plugins. The first is the extension of essay question type and the second is the extension of overview quiz report. Both of these plugins will be duplicated, then any needed changes will be made. For example, after duplicating a template plugin, the developer must change the plugin data such as plugin name, plugin tables in database, and plugin-specific strings. The next part is to modify or extend the features that plugin has. For example, to grade an answer, there must be a comparison inputted by the teacher. In this case, a new text input should be added to the form when the teacher is adding or editing a question.

C. *NLP Library Requirement*

In II.B, there are three NLP libraries that can be used. Apache OpenNLP has the least priority because it is no longer maintained. With Apache OpenNLP out of the options, the remaining choices are Stanford NLP Parser and INANLP. However, the scope of development is to build a grader for Indonesian language and Stanford NLP Parser doesn't have its model, so INANLP is the chosen library. INANLP is also capable of handling the grading processes such as formalization, sentence detector, and part-of-speech tagger.

D. *The Usage of String Metric to Compare Sentences*

According to [11], Jaro has the best average results in pattern matching technique category. However, Jaro doesn't count the common substring in the compared sentences. This element is very important because it will help in giving some tolerance when it comes to typos at the beginning or middle of a sentence.

E. *The Usage of Thesaurus to Handle Synonyms*

A single word can have different structure with the same meaning, for example, in Indonesian, "fauna" (fauna) has the same meaning with "hewan" (animal). Therefore, synonyms must be put into consideration to prevent synonym words being marked as a wrong answer. TABLE II shows some of the data inside Open Multilingual WordNet for Indonesian language. There is a thing to consider as this thesaurus is delivered as a .txt file. Searching through 107.042 words will take a pretty long time if it doesn't being indexed. In that case, this thesaurus' data structure will be modified to word ID, part-of-speech tag, type, and word. Then, it will be inserted into MySQL database to optimize searching time.

F. *The Grading Process*

These are the processes when the grader is about to generate grade recommendation for teachers.

1. The grader will receive two arguments, the first is expected answer, the second is students' answer.
2. If there is only one word, then the grader will do Winkler algorithm straight away, because INANLP can't process part-of-speech of one word. If there is one sentence, then proceed to process 3. If there are two sentences or more, compare each of the sentence from expected answer to students' answer using process 3, forming a matrix of sentences comparison.

   TABLE II. Comparison for Answers with More Than Two Sentences

	A	B	C	D	E	F
B'	60%	**100%**	50%	40%	45%	65%
A'	**100%**	60%	30%	15%	60%	65%
C'	30%	50%	**100%**	25%	40%	45%
E'	60%	45%	40%	20%	**100%**	20%
D'	15%	40%	25%	**100%**	20%	15%
F'	65%	65%	45%	15%	20%	**100%**

   TABLE IV shows that every row has a highest percentage. When the matrix has finished filling the value of each sentence's comparison, the highest value of each row will be taken. That cell's column and row is removed and the value is added to the total value. Repeat this until the final row. The final result is total value divided by total number of sentences in expected answer.
3. Split sentences using INANLP's sentence detector
   a. For every sentence, use INANLP's formalization, so every informal word will be converted to formal word.
   b. Compare expected answer to students' answer with processes as below.
      i. Use tokenizer to split words from expected answer and students' answer.
      ii. Use stemmer and part-of-speech tagger respectively to the tokenized words.
      iii. Eliminate matching pairs and add word pair count by 1 for each eliminated pair. Final value of this result is pair count divided by total expected answer's words. Any exceeding words will be contained in a variable and considered at the end of the whole process.
      iv. If there are remaining words from the previous process, the remaining processes will be done. If all words are eliminated already, skip the remaining processes.
      v. Check synonym sets of the expected answer's words. If the grader doesn't found any synonym, take the synonym with the highest

Winkler value compared to the nearest students' answer. Save it in some variable. For every synonym-students' word answer elimination, add 1 to the synonym pair count. Final value of this process is synonym pair count divided by total expected answer's words.

vi. If there are remaining words from the previous process, concatenate remaining words from both expected answer and students' answer and compare both with Winkler algorithm. If all words are eliminated already, skip the remaining processes. Final value of this process is the maximum of Winkler value of remaining words, or Winkler value of remaining words, with the students' words being replaced by its synonyms' container from the previous process.

vii. Sum all final value from these processes. Exceeding words will cause point subtraction in this process.

4. Final grade recommendation is a decimal number between 0 and 1. That number represents similarity between the expected answer and students' answer. Grade recommendation with number "0" means the students' answer doesn't fit at all syntactically, while "1" means it fits perfectly to the expected answer.

G. *Module Integration with Moodle*

The grader was developed in Java language, while Moodle is running in Apache PHP server. To bridge them, there is a PHP script to execute command-line, so the Moodle plugin will execute the grader through that script.

IV. Development Results

After deployed to Moodle, the plugins and grader were tested to ensure its quality. The tests will consist of quiz question and expected answer for every possible answer.

TABLE III. Test Cases

Test Case Number	Quiz Question	Students' Answer	Grade
1	Sebutkan ibu kota dari provinsi Jawa Barat! Jawab dengan satu kata langsung.	Bandung.	1.0
2		Bandungh.	0.98
3	Sebutkan ibu kota dari provinsi Jawa Barat! Jawab dengan kalimat panjang, "Ibu kota dari provinsi Jawa Barat adalah..."	Ibu kota dari provinsi Jawa Barat adalah Bandung.	1.0
4		Ibu kota dari provinsi Jawa Barat adalah Bandungh.	0.9975
5		Ibu kota daripada provinsi Jawa Barat adalah Bandung.	1.0
6		Sepertinya ibu kota dari provinsi Jawa Barat adalah Bandung.	0.9375
7		Ibu kota dari Jawa Barat adalah Bandung.	0.9791
8	Sebutkan ibu kota dari provinsi Jawa Barat, Timur, dan Tengah berturut-turut! Jawab dengan jawaban panjang.	Ibu kota dari provinsi Jawa Barat adalah Bandung. Ibu kota dari provinsi Jawa Timur adalah Surabaya. Ibu kota dari provinsi Jawa Tengah adalah Semarang.	1.0
9		Ibu kota dari provinsi Jawa Barat adalah Bandungh. Ibu kota dari provinsi Jawa Timur adalah Surabaya. Ibu kota dari provinsi Jawa Tengah adalah Semarang.	0.9991
10		Ibu kota daripada provinsi Jawa Barat adalah Bandung. Ibu kota daripada provinsi Jawa Timur adalah Surabaya. Ibu kota daripada provinsi Jawa Tengah adalah Semarang.	1.0
11		Ibu kota dari provinsi Jawa Tengah adalah Semarang. Ibu kota dari Provinsi Jawa Timur adalah Surabaya. Ibu kota dari provinsi Jawa Barat adalah Bandung.	1.0
12		Sepertinya ibu kota dari provinsi Jawa Barat adalah Bandung. Ibu kota dari provinsi Jawa Timur adalah Surabaya. Ibu kota dari provinsi Jawa Tengah adalah Semarang.	0.875
13		Ibu kota dari Jawa Barat adalah Bandung. Ibu kota dari provinsi Jawa Timur adalah Surabaya. Ibu kota dari provinsi Jawa Tengah adalah Semarang.	0.872
14	Tuliskan kalimat ini dengan mengganti "organisme" dengan sinonimnya: "Kelinci adalah organisme."	Kelinci adalah makhluk.	1.0
15	Kapan Harambe, gorila di Kebun Binatang dan Taman Botani Cincinnati, ditembak? Jawab dengan jawaban panjang.	Harambe ditembak pada tanggal 28 Mei 2016.	1.0
16		Harambe ditembak beberapa bulan yang lalu.	0.6914
17		Harambe ditembak 4 bulan yang lalu.	0.72
18	Siapa yang menembak Harambe di Kebun Binatang dan Taman Botani Cincinnati? Jawab dengan jawaban panjang.	Harambe ditembak oleh penjaga Kebun Binatang dan Taman Botani Cincinnati.	1.0
19		Harambe ditembak oleh pegawai resmi Kebun Binatang dan Taman Botani Cincinnati.	0.8511
20		Harambe ditembak oleh petugas yang bekerja sebagai penjaga kandang disana.	0.7666
21		Penjaga kandang di Kebun	0.8181

Test Case Number	Quiz Question	Students' Answer	Grade
		Binatang dan Taman Botani Cincinnati adalah adalah penembak Harambe.	
22	Pada umur berapa Harambe mati? Jawab dengan jawaban panjang.	Harambe mati pada umur 17 tahun.	1.0
23		Harambe mati pada umur tujuh belas tahun.	0.6904
24	Siapa pemain Argentina yang pernah mencetak gol dengan tangan pada Piala Dunia 1986? Jawab dengan jawaban panjang.	Gol dengan tangan pada Piala Dunia 1986 dicetak oleh Diego Maradona.	1.0
25		Diego Maradona mencetak gol dengan tangan pada Piala Dunia 1986.	1.0

As shown in TABLE IV, there are 25 test cases for the grader. The first part of test is the one-word case. In this case, the Jaro-Winkler algorithm is used and the only factor that affects the grade is typo and less/more characters. For example, in test case number 2, the answer should be “Bandung.”, but the students' answer is “Bandungh.”, so its grade is slightly lower than test case number 1. The second part of test is the one-sentence case. In this case, each word in a sentence will have its POS tag generated by INANLP. The total number of words in expected answer will be the comparison whether the students' answer inputted less or more words. Things that affect the grade in this case are typo, less/more characters, and less/more words. Modifying conjunction words such as “dari” doesn't affect the grade as the grader removes all conjunction words before comparing words and their POS tags. The third part of test is more than one-sentence case. In this part of test, things that affect the grade is typo, less/more characters, less/more words, and less/more sentences. If the sentences are swapped, the grader will still mark the answer as correct. Last part is synonym testing. The fourth part of the test shows the grader accepts synonyms of a word, as long as INANLP and the thesaurus classify it with the same POS tag. The fifth, sixth, and seventh part of the test show that answers with different representation having less grades than they should be. The last part of the test shows that if two answers have same representation of words, they will have the same grade.

CONCLUSION

The developed module can help teachers to analyze students' answer syntactically. This module is capable of handling small errors with high tolerance rate. However, it still can't grade answers with different representation correctly as those differences will be classified as a false answer.

FUTURE RESEARCH

A method must be developed to recognize the similarity between words with the different representation but have the same meaning. By using that method, automated semantical grading in online assessment will be possible.

ACKNOWLEDGMENT

This work is based on my undergraduate thesis project. I would like to thank Mrs. Yani Widyani, S.T., M.T. as the supervisor to my work.

REFERENCES

[1] John K. Lewis. (2013). Ethical Implementation of an Automated Essay Scoring (AES) System: A Case Study of Student and Instructor Use, Satisfactions, and Perceptrons of AES in a Business Law Course. Salve Regina. Salve Regina University.

[2] Margarida Amaral & Hugo Riberio. (2010). Computer-Based Assessment: Sounds Easy, Is It Really? Porto. Universidade do Porto.

[3] Mary E. Piontek. (2008). Best Practices for Designing and Grading Exams, Michigan, The University of Michigan.

[4] University of Oregon TEP. (2013). *Online Assessment with Blackboard*. Oregon, University of Oregon.

[5] E. D. Liddy. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. New York. Marcel Decker, Incorporated.

[6] Prakash M. Nadkarni, Lucila Ohno-Machado, Wendy W. Chapman. (2011). Natural Language Processing: An Introduction. New Haven. Yale Center for Medical Informatics.

[7] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[8] Apache OpenNLP Documentation. (2014). https://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html. Apache OpenNLP Development Community.

[9] Ayu Purwarianti. (2015). INANLP. Bandung. Bandung Institute of Technology.

[10] Jiaheng Lu, Chunbin Lin, Wei Wang, Chen Li, Haiyong Wang. (2013). String Similarity Measures and Joins with Synonyms. Renmin University of China, China.

[11] Peter Christen. (2006). A Comparison of Personal Name Matching Technical and Practical Issues. Canberra. The Australian National University.

[12] Nigel A. Caplan (2011). *Using a Thesaurus*. http://writingcenter.unc.edu/files/2011/12/thesaurus.pdf.

[13] Francis Bond, et. al. (2014). The combined Wordnet Bahasa NUSA: Linguistic studies of languages in and around Indonesia 57: pp 83–100 (URI: http://repository.tufs.ac.jp/handle/10108/79286)

[14] Nurril Hirfana, Suerya and Bond. (2011). Creating the Open Wordnet Bahasa in Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25), Singapore.

[15] Hammam Riza, Budiono, and Chairil Hakim. (2010). Collaborative work on Indonesian wordnet through Asian wordnet (AWN) In Proceedings of the 8th Workshop on Asian Language Resources, pages 9–13. Beijing.

[16] Francis Bond, Hitoshi Isahara, Kyoko Kanzaki and Kiyotaka Uchimoto. (2008). Boot-strapping a WordNet using Multiple Existing WordNets. In LREC-2008, Marrakech.

[17] Lian Tze Lim & Nur Hussein. (2006). Fast prototyping of a Malay wordnet system In Proceedings of the Language, Artificial Intelligence and Computer Science for Natural Language Processing (LAICS-NLP) Summer School Workshop, pages 13–16.

[18] Moodle Dev Docs. (2016). Moodle Developer Documentation. https://docs.moodle.org/dev.