A Review of an Information Extraction Technique Approach for Automatic Short Answer Grading

Uswatun Hasanah[1,2], Adhistya Erna Permanasari[1], Sri Suning Kusumawardani[1], Feddy Setio Pribadi[1]
[1]Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
[2]Department of Informatics Engineering
STMIK Amikom Purwokerto
Purwokerto, Indonesia
[1,2]uswatun.mti15@mail.ugm.ac.id, [1]adhistya@ugm.ac.id, [1]suning@ugm.ac.id, [1]feddy.setio.p@mail.ugm.ac.id

***Abstract*— The requirement for automatic short answer grading (ASAG) system brings researchers to discover more knowledge about this field. Many techniques have been developed to reach the highest accuration. It can be processed by following stages: creating data set, pre-processing, model building, grading, and model evaluation. One of the techniques which commonly used is information extraction technique. Information extraction is a technique that employing finding fact on the student answers as patterns and then matches these to the teacher answer. The accuration is pointed out in computer and human raters agreement. The goal of this paper is to present a review of several ASAG research which using information extraction technique. However, this paper does not conclude the best method which can be used for general cases.**

Keywords*— short answer; automatic grading; information extraction; pattern matching; natural language processing

I. INTRODUCTION

The successful of learning can be represented with various evaluation study and grading models. Evaluation models which mostly used by teacher is questions that require multiple choice or essay answers. Multiple choice question is evaluation model that consists of several wrong answers and one correct answer. In utilizing of e-learning, multiple choice grading is easier to be computed than essay grading. Essay grading need to be processed with Natural Language Processing (NLP) technique first. Although easy to be computed, multiple choice model considered has low reliability because student can chop off from alternative responses at hand [1]. Multiple choice model intuitively give hint that connected to the alternative response [1]. Different from multiple choices, essay requisite for response in natural language form. Essay response can be divided in two form: long-answer essay and short-answer essay [2]. Long-answer essay requires the student to give the response in two or more paragraphs, where short-answer essay contains of two or three sentences. Assessment point of long-answer essay can be considered by writing style and content [3] whereas short-answer essay only focus in content and ignoring the writing style [4]. Teachers are believed that short-answer model can strengthen learning and improving the cognitive skills, and they can make the students show their understanding effectively without show the prompts or clue [5].

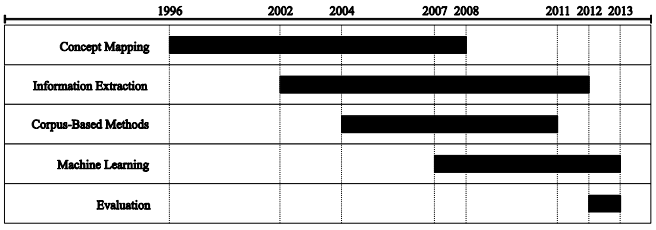
Question for short-answer grading can be considered as meeting at least five specific criteria [6]: (1) The question must require a response that recalls external knowledge instead of requiring the answer to be recognized from within the question. (2) The question must require a response given in natural language. (3) The answer length should be roughly between one phrase and one paragraph. (4) The main focus of assessment is answer's content, instead writing style. (5) An objective question design is needed to restrict the level of openness in open-ended versus close-ended responses.

TABLE I. PROPERTIES OF QUESTION TYPES [6]

Property	Question Type		
	Fill-the-gap	*Short answer*	*Essay*
Length	One word to a few words	One phrase to one paragraph	Two paragraphs to several pages
Focus	Words	Content	Style
Openness	Fixed	Closed	Open

Overall, the assessment for short answer has several disadvantages: it consuming times, not objective, and susceptible human errors. Some researchers have developed automatic grading system to solve those problems. Conole et al. [7] identified that automatic answer grading can record the interactions of student and analyze these to give a richer learning comprehension.

Fig. 1. The era of ASAG system.



Burrows [6] categorized ASAG system based on five eras: Concept Mapping, Information Extraction, Corpus-Based Methods, Machine Learning, and Evaluation. This paper will explain some various researches which are conducted in Information Extraction (IE) era. IE technique in ASAG system has to match facts between student answers and teacher answers. Based on the eras of ASAG system, IE technique is more well known and most commonly used (by range of its use) than other techniques. Information Extraction (IE) techniques were adopted for use in the application. Sukkarieh et al. [8] describe that the reasons for choice IE techniques were that these techniques do not require complete and accurate parsing, they are relatively robust in the face of ungrammatical and incomplete sentences and they are also easy to implement. Hence, the exploration of IE technique can be more deeply and give remedicals from the previous system.

This paper consists of four sections. The first section contains explanation about automatic grading for short answer and it techniques. Afterwards, the second section will explain data sets and process that used in automatic short answer grading using information extraction techniques, such as collecting data set, pre-processing, grading models, and evaluation. The third section will discuss about possible future research directions in automatic short answer grading using information extraction techniques. Finally, conclusion will be drawn on the last section.

II. Materials and Techniques

A. Data Sets and Text Pre-processing

The first step to build an automatic grading system is creating data set. Data set for automatic short answer grading system usually consists of several questions, teacher answers and student answers. From the literatures, we identified some data set that using questions and answers from learning materials for school, such as Biology [8] and Science [9]. The other data sets using lecture learnings for college, like Computer Architecture [10], Science [11], and Biology [12]. Later, some paper does not describe their data set.

Text pre-processing is used to change the student answers into normalized form. Pre-processing text usually using natural language processing techniques, such as: tokenization, stemming or lemmatization, part of speech tagging, phrase recognition, parsing, stop words removal, sentence segmentation, punctuation removal, spelling correction, etc.

B. Information Extraction Technique

Information Extraction (IE) described as a process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts [13]. IE techniques applying a set of patterns to pull out pertinent information from syntactically analyzed pieces of text answers [14]. Evaluation is reached by comparing between the templates produced automatically by an extraction program with templates for the same texts produced by humans [13]. Appelt and Israel [15] define three crucial steps to write extraction rules by hand. First, determine all the ways in which the target information is expressed in a given corpus. Second, think of all the plausible variants of these ways. Last, write appropriate patterns for those ways. In order to write the patterns, Sukkarieh [8] devised a simple language shown in Fig. 2.

Fig. 2. Writing patterns using simple language [8].

```
Pattern         ->   Word | Word/Cat | Symbol
                     | Variable | Disjunction
                     | Sequence | k(N, Sequence)
                     (N is upper limit of length of Sequence)
                     | k(N, Sequence, Pattern)
                     (Sequence NOT containing Pattern)
Disjunction     ->   {Pattern, ..., Pattern}
Sequence        ->   [Pattern, ..., Pattern]
Word            ->   sequence of characters
Cat             ->   NN | VB | VG ...
Symbol          ->   & | % | $ ...
Variable        ->   X | Y | Z ...
```

Next, we will discuss the use of IE techniques based on IE era.

1) Parse Tree Matching

Mitchell et al. [9] built an system of AutoMark which is represent mark scheme answers as syntactic-semantic templates. A particular form of acceptable or unacceptable answer is specified by each template. First, student answers are parsed, and then matched against each mark scheme template. Finally, the system will compute a mark for each answer. The input text can contain multiple variations that robustly mapped. Overall, the rate of human-computer matched marking outcomes was achieved with 92,5%.

2) Regular Expression Matching

Bachman et al. [16] introduced WebLAS system that consists of three modules: task creation, task modification, and lexicon modification. The instructor can use task creation module and store it into database and preprocessing the task for automatic scoring. Instructor also can go back and make some modified tasks they have created. Later, the WebLAS lexicon modification is based on Wordnet. To score each element, WebLAS system simply match pattern from the regular expression.

Jordan [17] used PMatch system that matching simple sequences of words. Eleven short text questions were selected and for each answer were marked by a single human marker which is expert and comprehend with the question. Of course, the human marker will have made some mistakes, but this is disregard in the current study. It is also permitted for words of more than three letters, single incorrect, transposed, missing or extra letters. The system has feedback in order to inform students while a word they have used is not recognized by the spell-checker's dictionary, and offers suggestions for correct spelling. Finally, the human expert's marking was compared with PMatch's marking. The kappa inter-rater statistic was greater than 0,9 but one question has 0,84 because the question was answered by students very fine, so for a chance the computer and human were more probability agree.

3) Boolean Phrase Matching

Thomas et al. [18] employ a relatively simple algorithm based on matching key-phrases between a student's answer to a question and the corresponding specimen solution. It employed a thesaurus to cope natural language processing (NLP) task such as synonyms, plurals and verb tenses. Boolean expression is used to unite a set of phrases that have different allocated mark for each possible solution in every specimen. Boolean-AND expression define the required phrases and Boolean-OR expressions refer to acceptable alternative phrases. A significant relationship between the average human marker score and the electronic score showed 0,86 from analysis of the correlations.

4) Syntactic Pattern Matching

Sukkarieh et al. [8] built a system namely auto-marking where the expert examiners give score for each answer: 0 for incorrect answer, 1 for partially correct or incomplete answer, or 2 for correct and complete answer. First, a chunk of answer has been identified that qualifies for a mark. The system using knowledge-engineering approach, so the grammar that describing the patterns is constructed by hand and the rules only can be modified by people who is familiar with the grammar and the system. Three sets of data from biology task are abstracted the pattern. First, the compact key answers which provided by the examiners are fleshed. Second, the own version of the answers are used. For the last set of data that abstracted patterns over are the training data from University of Cambridge Local Examinations Syndicate (UCLES). After checking out the variant ways answers could be written, a simple language in which to write the patterns is devised. This technique gives accuracy 88%. Sukkarieh et al. use different method from text classification, known as the k nearest neighbor (KNN) technique, which just has accuracy 67%. However, the result of KNN can be increased while adding more linguistic information represented in the vectors.

Sima et al. [10] designed eMax system to assessing short texts in the following three main steps: syntactic analysis, semantic analysis, and scoring. First, teachers enter questions and determine the initial teacher answer space of the question (ASQ), containing fully correct and expected answers. The system also supports multiple correct answers for a single question. The ASQ can be consisted of expected knowledge elements, logical and semantic relationships. Teacher setting up exam by entering relevant parameters such as length, score range limits for grading, etc. Afterwards, eMax system used to do examination. For evaluation, the initial teacher answer space can be escalated if necessary. Scoring is performed according to the scheme specified by the teacher with the return values delivered by the semantic analysis [19].

Siddiqi and Harrison [12][20] built a short-answer making system and identified two error from C-rater system [21]: misses and false positives. A miss refers to C-rater's inability to recognize a correct concept in a response. A false positive occurs when a C-rater assigns too much credit for a response while the concept(s) are not present in the response. In order to solve that problem, Siddiqi adopt an IE technique from Sukkarieh et al. [8]. The processing of the student's answer text is performed in three phases: spell checking and correction, parsing, and comparison. The average agreement rate is 96%. It's performed better than other short-answer marking systems like Sukkarieh et al. [8] where had an agreement rate of 93% and C-rater developed by Leacock and Chodorow [1] had an average accuracy of 84%. However, the data sets used are different. Standardized data set is absolutely needed to examine each system in order to get a more effective comparison.

5) Syntactic-semantic Pattern Matching

Jordan and Mitchell [11] developed the templates in a computerized mark scheme with offline process using FreeText Author software. Input of free-text answers are processed by a sentence analyzer and the output is matched against each mark scheme template. The mark score determined by the result of the matching process. Each model answer associate with appropriate feedback. FreeText Author have user-friendly interface and makes the user to focus only on the tasks. FreeText Author's main components are a mark scheme panel, a model answer list, a synonym editor and a response list. FreeText Author using a machine learning algorithm to generate templates from model answers. Before enter a new model, user must determine the keywords. User can add synonyms for each keyword and considering suggestion from thesaurus.

6) Semantic Word Matching

Cutrone et al. [22] developed Auto-Assessor system that employ Natural Language Processing tools including WordNet.NET and SharpNLP in order to evaluate student responses. An answer key contains a single correct answer and will be provided for two human graders. The single correct answer used to be a benchmark. Next, the system establish automatically between the range of acceptable answers and appropriate deductions at the far ends of the range. Finally, level of agreement between system and each human grader are determined. In addition, human graders also compared one to another to decide between two humans. The system is able to processing a single sentence but the grammar and spelling are ignored.

7) LRS Representation Matching

Hahn et al. [23] developed CoSeC-DE system using Lexical Resource Semantics (LRS) which incorporated capability of precisely semantic distinctions into the robustness and modularity needed to represent meaning in real-life applications. The system employed freely available corpus namely CREG corpus. LRS structures are derived in two steps. First, surface representations are mapped to syntax-semantics-interface representations, which abstract away from some form variation at the surface. In the second step, rules map these interface representations to LRS representations. CoSec-DE has accuracy up to 86,3% while the other project CoMiC-DE [24] has 84,6% accuracy with the same data set.

C. Model Evaluation

Evaluation model is used to assess effectiveness from grading model by comparing accountability between system

(computer) and human raters. The evaluation metrics are most commonly represented by accuracy agreement, different variants of kappa, and pearson correlation. Values greater than 0,75 can represent the excellent agreement [25].

III. DISCUSSION

Information extraction techniques have been used by previous researchers and show good performance. From the literature review, we can summarize the papers based on method, data set, and human-computer agreement (HCA).

TABLE II. LIST OF PREVIOUS IE PAPERS

Information Extraction Techniques				
Paper	***Year***	***Method***	***Data Set***	***HCA***
Mitchell et al. [9]	2002	Parse Tree Matching	Science	92,5%
Bachman et al. [16]	2002	Regular Expression Matching	-	-
Thomas et al. [18]	2003	Boolean Phrase Matching	-	86,04%
Sukkarieh et al. [8]	2003	Syntactic Pattern Matching	Biology	88%
Sima et al. [10][19]	2007	Syntactic Pattern Matching	Computer Architecture	-
Siddiqi and Harrison [12][20]	2008	Syntactic Pattern Matching	Biology	96%
Jordan and Mitchell [11]	2009	Syntactic-semantic Pattern Matching	Science	-
Cutrone et al. [22]	2011	Semantic Word Matching	-	-
Hahn et al. [23]	2012	LRS Representation Matching	CREG Corpus	86,3%
Jordan [17]	2012	Regular Expression Matching	Science	84%

Thomas et al. [18] provide a simply assessment using Boolean phrase matching method with notation AND and OR. However, the system needs to be integrated for all users, such as students, examiners, and administrators. Hahn et al. [23] presented an approach for aligning underspecified semantic representations using LRS representation matching for reading comprehension questions. The system (CoSeC-DE) show a good performance, better than CoMiC-DE [24] with the same data set. Mitchell et al. [9] see that the template-based approach can makes the various types and complexity items into accurate and robust computerized marking. Jordan et al. [17] perform ASAG system which using a large number of responses so it can achieve remarkable marking accuracy. Bachman et al. [16] using regular expression matching method that suited to NLP tasks, and the instructor can modify it anytime. Cutrone et al. [22] interpreting the semantic meaning of the response using semantic word matching method. The challenge is how to grade multiple sentences based on set of synonym and adding more NLP tasks such as spell checker and grammar checker. Sukkarieh et al. [8] demonstrate that syntactic pattern matching method can perform better than KNN method. However, in the next research they want to try use machine learning algorithm to reach the same accuracy in IE technique. Sima et al. [10][19] employ syntactic analysis, semantic analysis and scoring using an approach of answer space. It restricted by single sentence on α version of system, but the β version is available to assess the answers that consist of two or three sentences. Siddiqi and Harrison [12][20] perform a good HCA and give some reparation to c-rater system [21]. Jordan and Mitchell [11] employ seven questions for various sum of student answers. Each question has different HCA (about 89,4% to 99,5%). The system has a good interface and has feedback for student answers.

From methods that explained, Siddiqi [20] has the highest accuracy to the number of 96%, but this result cannot be expressed to the best method because the other researchers use different data set. A valid accuracy of HCA just can be measured if they use the same data set. For the future works, standardization of data set is needful. Despite of the data set usually collected from beginning of researcher's institution, they can use two or more methods to test the same data set. IE technique which is applicated to ASAG system can be one of the alternative computer assessments. However, it is likely to add more technique to increase IE performance, such as employing material learning as a corpus in order to set some key answers. In addition, thesaurus can be employed to make out variation of synonym which is very possible turned up in student answers.

The system's interface is also important to well-designed. The system must be easy to use by people who are not familiar within the system. They don't need to understand everything inside the system and how it works. Setting of input text must provide some NLP tasks as needed, such as spell checker, grammar checker, etc. Later, the system must provide some linguistic library to make out the alternatives of words given, such as synonym, metonym, etc. FreeText Author system by Jordan and Mitchell [11] have a good interface between all.

Next, human graders normally consist of more than one grader because it can influence the objectivity of human grading. Grading just by one human grader is possible to make human error occurrence. On the other hand, grading by humans also compared to report agreement between human graders.

IV. CONCLUSION

By applying of ASAG system, a teacher can be the main assessor, but student answer can be assessed by assistance of computer. The result of assessment between system and human can be considered to make decision for final grading. In this context, computer can help to detect error and dereliction of human graders. From the previous methods, we can conclude that IE technique is a rule-based method. IE technique extracts the facts from student's answer as patterns and then matches these to the teacher answer. It must be a simpler technique, but the results show satisfactory.

From the previous research, auto-marking system by Sukkarieh [8] has a good method and the method most commonly used (based on IE era) namely syntactic pattern matching. Although their system only has 88% accuracy, but this method is reusable by Siddiqi [12][20] with adding some tool such as spell checking and correction, and the result can

reach 96% accuracy. We believe this method can be improve for a better system like adding some language processing tool. However, FreeText Author by Jordan and Mitchell [17] has the best system interface because it's facilitating user with a mark scheme panel, a model answer list, a synonym editor and a response list.

Although research of ASAG system using IE techniques only popular until 2012, in future works IE techniques can employ some technique in outside of IE scope, such as corpus-based methods. Corpus is needfull to handling variation of synonym within student answers. Corpus can be taken from Wikipedia or another language tools. With the help of the corpus, the system is expected to find the semantic meaning of words on the student answers and matched with the keywords provided.

REFERENCES

[1] M. K. Singley and H. L. Taft, "Open-ended approaches to science assessment using computers," *J. Sci. Educ. Technol.*, vol. 4, no. 1, pp. 7–20, 1995.

[2] D. Callear, J. Jerrams-smith, V. Soh, J. Jerrams-smith, and V. Soh, "CAA OF SHORT NON-MCQ ANSWERS," 2001.

[3] M. Juzaiddin, A. Aziz, F. D. Ahmad, A. Azim, and A. Ghani, "Automated Marking System for Short Answer Examination (AMS-SAE)," no. Isiea, pp. 47–51, 2009.

[4] S. G. Pulman, W. St, J. Z. Sukkarieh, and W. St, "Automatic Short Answer Marking," 2005.

[5] Intelligent Assessment Technologies Ltd., "E-Assessment of Short-Answer Questions On-Screen Assessment," 2009.

[6] S. Burrows, I. Gurevych, and B. Stein, *The Eras and Trends of Automatic Short Answer Grading*. 2015.

[7] G. Conole and B. Warburton, "A review of computer-assisted assessment," *Alt-J*, vol. 13, no. 1, pp. 17–31, 2005.

[8] J. Z. Sukkarieh, S. G. Pulman, and N. Raikes, "Auto-marking : using computational linguistics to score short , free text responses," pp. 1–15, 2003.

[9] T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge, "TOWARDS ROBUST FREE-TEXT RESPONSES," 2002.

[10] D. Sima, B. Schmuck, S. Szöll, and Á. Miklós, "Intelligent Short Text Assessment in eMax," pp. 435–445, 2009.

[11] S. Jordan and T. Mitchell, "e-Assessment for learning ? The potential of short-answer free-text questions with tailored feedback questions with tailored feedback," 2009.

[12] R. Siddiqi and C. Harrison, "A Systematic Approach to the Automated Marking of Short-Answer Questions," pp. 329–332, 2008.

[13] J. Cowie and Y. Wilks, "Information Extraction," 2000.

[14] S. R. B, Y. Narahari, and O. D. Deshmukh, "A Perspective on Computer Assisted Assessment Techniques for Short Free-Text Answers," vol. 1, pp. 96–109, 2015.

[15] Douglas E. Appelt; David J. Israel, "Introduction to Information Extraction Technology," *IJCAI-99 Tutor.*, pp. 1–41, 1999.

[16] L. F. Bachman, N. Carr, G. Kamei, M. Kim, M. J. Pan, C. Salvador, and Y. Sawaki, "A reliable approach to automatic assessment of short answer free responses," pp. 1–4, 2002.

[17] S. Jordan, "Short-answer e-assessment questions : five years on," 2012.

[18] P. Thomas, "The Evaluation of Electronic Marking of Examinations," pp. 1–7, 2003.

[19] D. Sima, B. Schmuck, S. Szöll, Á. Miklós, and A. Motivation, "Intelligent Short Text Assessment in eMax," 2007.

[20] R. Siddiqi and C. J. Harrison, "On the Automated Assessment of Short Free-Text Responses," pp. 1–11, 2008.

[21] C. Leacock and M. Chodorow, "C-rater : Automated Scoring of Short-Answer Questions," pp. 389–405, 2003.

[22] L. Cutrone and M. Chang, "Auto-Assessor : Computerized Assessment System for Marking Student ' s Short-Answers Automatically," pp. 81–88, 2011.

[23] M. Hahn and D. Meurers, "Evaluating the Meaning of Answers to Reading Comprehension Questions A Semantics-Based Approach," pp. 326–336, 2012.

[24] D. Meurers, R. Ziai, N. Ott, and J. Kopp, "Evaluating Answers to Reading Comprehension Questions in Context : Results for German and the Role of Information Structure," pp. 1–9, 2011.

[25] J. Fleiss, B. Levin, and M. Cho Paik, *Statistical Methods for Rates and Proportions*. 2003.