ATHENA PUBLISHING

0 ’1- ! & 1) 2 !
+

9 C @A ! , 0 =

% $!
9 C @A + - .

%
9 C @A ! , 0 = $ < . ? ,
D . # ’ 2
) ! 3 46 ;5! ; + !

& #
. 77 .

’ ! ’ "

Natural Language Processing Research
Vol. **1(3-4)**, 2021, *pp.* 46–55
DOI: https://doi.org/10.2991/nlpr.d.210212.001; eISSN: 2666-0512
https://www.atlantis-press.com/journals/nlpr/

Research Article

Web-Based Turkish Automatic Short-Answer Grading System

Ebru Yilmaz Ince[1,*], Akif Kutlu[2]

[1]*Isparta University of Applied Sciences, Isparta, Turkey*
[2]*Süleyman Demirel University, Isparta, Turkey*

ARTICLE INFO

Article History
Received 01 Dec 2020
Accepted 10 Jan 2021

Keywords
Engineering education
Educational technologies
Automatic short answer grading
Latent semantic analysis
Natural language processing

ABSTRACT

In this paper, a web-based Turkish automatic short answer grading system (TASAG) is developed to score exam questions and to generate online exams for automatic scoring of short-answer questions. The novelty of this study is that TASAG is the first software of its kind for the Turkish language. The algorithm of the TASAG software is a hybrid that determines which method will be used at runtime based on the word number dimensions to achieve accurate scoring. To measure the software's accuracy, a case study is performed for computer engineering. Instructors' scoring results and the TASAG software scoring results were compared. Two instructors prepared different answer keys for the same exam to increase the accuracy of the scoring. The scoring results, which are compared in the figures, are very close to each other, which indicate the effectiveness of the TASAG software. Moreover, the TASAG arithmetic mean scores for each answer key are calculated and given as the final score for the exam. According to the results, the TASAG software can be used for automated short answer grading system in Turkish with a 92% success rate.

1. INTRODUCTION

Technological developments have resulted in new educational software such as learning management systems (LMSs) to utilize distance education and web-based learning. LMSs help instructors to manage online courses and to prepare exams. Students take exams and the results are scored and entered into the LMS by instructors. Scoring an exam may take a long time when there are a large number of students, questions, and long answers. To create a smart LMS, several new technological developments can be used [1]. For example, exam results can be scored automatically by computers based on the instructors' answer key. Automatic short answer grading system (ASAG) software have been developed to automatically assign a score to an answer through a comparison with one or more correct answers using a computer [2].

ASAG software must be designed for the language in which the test is administered. In the literature, there are several ASAG systems for different languages (mostly English); however, at present, there is no ASAG for Turkish. So, it is not possible to compare them with our proposed web-based Turkish automatic short answer grading system (TASAG) software (the first ASAG software for Turkish). The new ASAG software must be designed in Turkish that is being in accordance with grammar rules of Turkish. Also, short-answer question type as one-word answer is used in university entrance exam in 2017 at Turkey. For these reasons, this paper proposes a web-based automated grading system called TASAG system for open-ended short-answer question. The TASAG has a user-friendly web interface. Instructors can use the TASAG for web-based assessment and for data recording of exam questions and results. It is expected that the TASAG will facilitate students e-learning because it is enhanced with artificial intelligence software, human–computer interaction, feedback, and personalized learning methods.

This paper is organized into four sections. In Section 2, related works are given. In Section 3, technical and computational background of the TASAG software is presented. In Section 4, proposed TASAG system is applied to a real case study and results are evaluated. Finally, conclusions and thoughts on future studies are given.

2. RELATED WORK

There are many researches about ASAG systems [3,4] which are given in chronological order (Table 1). Burrows, Gurevych, and Stein [5] categorized ASAG systems into five eras that are thematically consistent set of activities as following: concept mapping, corpus-based method, machine learning, information extraction, and evaluation.

*Corresponding author. Email: ebruince@isparta.edu.tr

3. PROPOSED TASAG SYSTEM

The TASAG was developed for administering online exams and scoring open-ended short-answer questions automatically in Turkish language (Figure 1). In this system, instructors can create exam and prepare questions. Students can get exam on the system by using online exam module. When student answer a question, answer key of the question is gained from the system. Then, answer key and student's answer are compared with similarity methods such as Cosine, ILSA, and LSK. Obtained question score and also total exam score is shown to the students instantly at the end of the exam. Also, exam score is saved to the system and instructor can analyze the exam results of each student.

In this part, the detailed information about the technical background and computational background of the TASAG system is given to clarify the mechanism of the system.

3.1. Technical Background of the TASAG Software

The developed TASAG software is based on a n-tier software architecture. The software has three layers: the presentation layer, the business layer, and the data layer, as shown in Figure 2. These layers make the software flexible, modular, and scalable.

In the presentation layer, web pages are prepared using ASP.NET, AJAX, and C#.NET programming languages. The instructors can use these web interfaces to add, update, and view exam questions and the exam scores of students. Student affairs includes course assignments and course progress; additionally, it can list and view instructor processes and student processes. Students can register for courses, take exams, and view their own exam scores. The TASAG software uses user-friendly web 2.0 pages. Instructors can use TASAG for web-based assessment and data recording for exam questions and results. Also, the exam preparation page for instructors is shown in Figure 3. A sample student page for viewing the automated scoring results of TASAG is shown in Figure 4.

The business layer includes the most important modules: automatic assessment, the similarity engine with the Turkish WordNet framework, the Zemberek framework, and preprocessing. The Extended Turkish WordNet framework finds semantically related words. The Zemberek framework performs morphological analysis of the Turkish text. The preprocessing module includes the natural language processing functions (i.e., pruning stop words, white space tokenization, stemming, trimming, and lower case). The

Table 1 | Examples of ASAG systems.

Example System	Method	Developer (Reference)
c-rater	Concept mapping	Leacock and Chodorow [6]
Atenea	Corpus based method	Alfonseca and Perez [7]
SAMText	Corpus based method	Bukai *et al.* [8]
e-Examiner	Machine learning	Gutl [9]
CAM	Machine learning	Bailey and Meurers [10]
FreeText Author	Information extraction	Jordan and Mitchell [11]
eMax	Information extraction	Sima *et al.* [12]
CoMiC-EN	Machine learning	Meurers *et al.* [13]
Auto-Assessor	Information extraction	Cutrone *et al.* [14]
PMatch	Information extraction	Jordan [15]
Willow	Corpus based method	Perez-Marin and Pascual-Nieto [16]
ETS	Evaluation	Heilman and Madnani [17]
UKP-BIU	Evaluation	Zesch *et al.* [18]
SoftCardinality	Evaluation	Jimenez *et al.* [19]

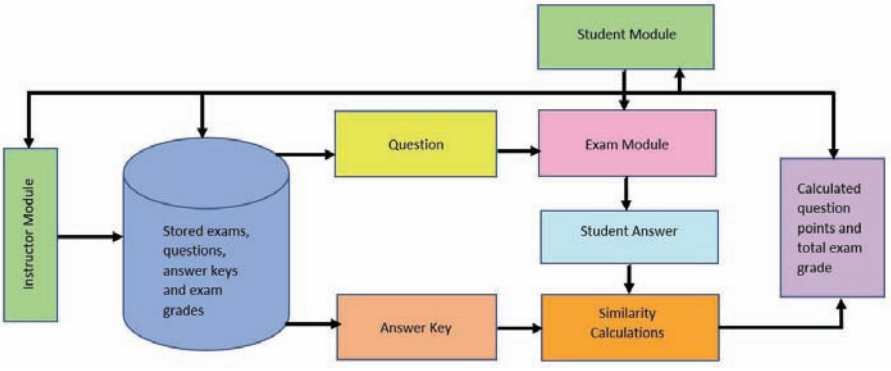

Figure 1 | The schema of proposed TASAG system.

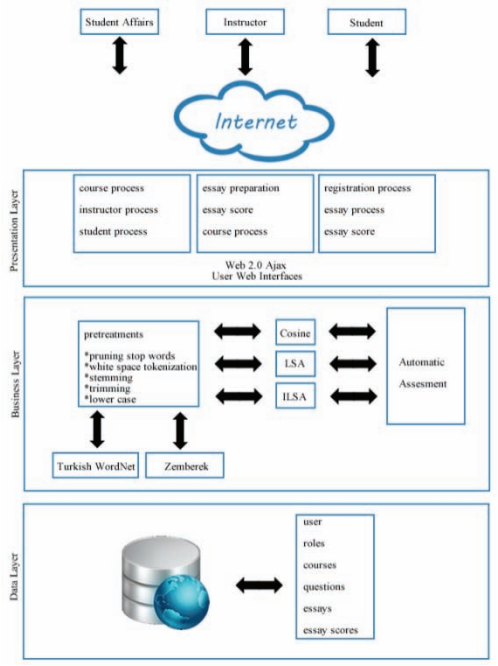

Figure 2 The TASAG software architecture.

Figure 3 Exam preparation page for instructor.

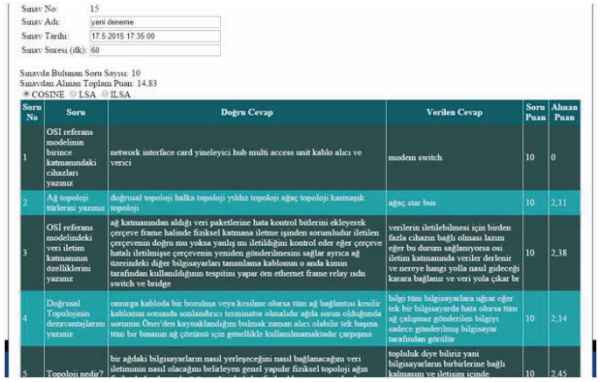

Figure 4 Student page for automated scoring results of the TASAG software.

similarity engine calculates the cosine, latent semantic analysis (LSA), and integrated latent semantic analysis (ILSA). Automatic assessment determines the exam score. In this layer, Zemberek [20] is used as a generic NLP framework, WordNet [21] is used to automatically extract synonyms and hyponyms, and Visdic [22] is utilized for viewing and editing the Turkish WordNet [23], which is stored in XML format. These are requirements for morphological analysis and determining semantic relations of words for automatic short-answer question scoring. A cross-platform numerical analysis and data processing library, ALGLIB2, is used to compute the SVD, which is needed for the LSA algorithm [24].

The data layer includes all of the data used by the TASAG software such as users, roles, courses, questions, exams, and exam scores. Using the data layer, instructors can perform adding, updating, and viewing functions. An MS SQL Server 2012 database management system is used in TASAG system.

3.2. Computational Background of the TASAG Software

Language is the most important communication tool; it allows problems to be solved using NLP algorithms in computer environments. NLP is a combination of artificial intelligence, computational linguistics, and computer science, which is concerned with the interaction between computers and human language.

Most NLP systems use classical linguistic levels of preprocessing, morphology, syntax, semantics, and pragmatics in a sequential architecture [25]. The preprocessing module used in this research includes natural language processing functions (i.e., pruning stop words, white space tokenization, letter tokenization, trimming, lower case and stemming). Pruning stop words removes stop-words from token streams, a Turkish stop-words list is used. Letter tokenization is done with Regex. Replace function by changing the nonword characters to space, then white space tokenization splits text based only on whitespace. Trimming is done with the Trim function. Lower case normalizes token text to lower case and done with ToLower function. Stemming attaches stem words to their root forms, which is vital for Turkish because it is an agglutinative language. Stemming is done with the help of Zemberek library.

Zemberek [20] is used for the morphological analysis of the Turkish text in the TASAG software. Firstly, Zemberek loads the binary root file during the initialization of the library. Then a root word is read, related special cases are attached to the root object, and into a special direct acyclic word graph (DAWG) tree (shown in Figure 5) the resulting object is stored to provide fast access and ease of extensibility. The structure of Zemberek contains letters and alphabets, suffix information, and suffix special cases for Turkic languages.

An Extended Turkish WordNet framework is developed to find semantically related words [26]. The Extended Turkish WortNet contains computer network terms and semantic relations as being synonymous and hyponymous were determined from Computer communications and network technologies book [27] that is written in Turkish. Also, the Extended Turkish WordNet contains words and relations from The Balkan Turkish WordNet [23,28] and Turkish Synonymous Hypernymous Dictionary [29]. Extended Turkish Wordnet helps similarity calculations to be more accurate. For example, suppose a question given as “OSI katmanlarını yazınız?” (In English; “Write the OSI layers?”) and answer key given as “uygulama, sunum, oturum, ulaşım, ağ, data bağlantı ve fiziksel” (In English;”application, presentation, session, transport, network,

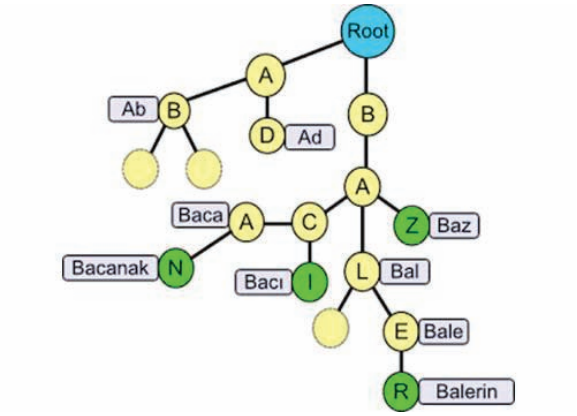

Figure 5 | Simplified structure of such a tree graph [32].

data link and physical"). If student answer as "uygulama, sunum, oturum, taşıma, ağ, veri bağlantısı ve fiziksel katman," although answer is completely correct student cannot exact (full) point for this answer because in Turkish "ulaşım (in answer key)" and "taşıma (in student answer)" are synonym. Moreover, "data (in answer key)" and "veri (in student answer)" are synonym. For these reasons, while similarity between the answer key and student answer is calculated, these words are analyzed whether there is a synonym or not. If synonym is found from Extended Turkish Wordnet, it is replaced so the student gets the correct score for the question more accurately. Extended Turkish Wordnet is prepared with the concept map method which provides deeper relations for words to get synonyms and hyponyms.

The TASAG uses a hybrid model combining cosine and ILSA methods. If the answer key from the instructor has less than ten words, which is an inadequate dimension for LSA [30], the cosine similarity algorithm is used for scoring student answers. Otherwise, ILSA is used, i.e., if the answer key contains more than ten words. The cosine method in the business layer controls synonymy, abbreviation, and digit cases of the texts.

ILSA method in the business layer controls synonymy, hypernym, abbreviation, and digit cases of the texts between the corresponding row vectors which, in this study, are the instructor answer key word vector, the student answer word vector, synonym WordNet list, hypernym WordNet list, and the index of the vectors. Then, the following steps are performed;

- Check if answerKeyVector and studentAnswerVector words are synonymous with the help of synonymWordNetList, when condition is true copy the synonym word found from answerKeyVector to studentAnswerVector. This creates the matchup for two words that have the same meaning. The algorithm of synonymous condition of the ILSA is given (see Algorithm 1).

Algorithm 1: Synonymous condition of the words

answerKeyVector a1,……. am ← array of answerkey words
studentAnswerVector b1,…… bn ← array of studentanswer words
synonymWordNetList s1,……, sp ← array of synonym words in WordNet
foreach words in answerKeyVector, do
foreach words in studentAnswerVector, do
foreach words in synonymWordNetList, do
if Synonymous condition is true ← control answerKeyVectorWord and studentAnswerVectorWord with the help of synonymWordNetList
studentAnswerVectorWord = answerKeyVectorWord
end
end
end
end

- Check if answerKeyVector and studentAnswerVector words are hypernymous, a value μ is set. μ is a weight value based on the WordNet node distance between the words. This step finds the relations for two similar meanings by using Extended Turkish WordNet. Additionally, if studentAnswerVector has two or more hypernyms, they will be added to the similarity with μ. In Figure 6, there is an Extended Turkish WordNet component that contains "Bilgisayar Ağı," which means "computer network," for the first node. "Bilgisayar Ağı" and "OSI" have a one-node distance. Additionally, "Bilgisayar Ağı" and "Fiziksel" have a two-node distance ("Fiziksel" means 'physical'). In this study, one-word-node distance is taken into account and the μ value, set at 0.5, has a one-word node distance. Therefore, if we took a two-word-node distance for two nodes, the μ value would be set at 0.25, as shown in Eq. (1). The algorithm of hypernymous condition of the ILSA is given (see Algorithm 2).

$$\mu_{\text{value}} = \frac{1}{2^{distance}}$$

Algorithm 2: Hypernymous condition of the words

answerKeyVector a1,……. am ← array of answerkey words
studentAnswerVector b1,…… bn ← array of studentanswer words
synonymWordNetList s1,……, sp ← array of hypernym words in WordNet
foreach words in answerKeyVector, do
foreach words in studentAnswerVector, do
foreach words in hypernymWordNetList, do
if hypernymous condition is true ← control answerKeyVectorWord and studentAnswerVectorWord with the help of hypernymWordNetList
studentAnswerVectorWord = "|"+answerKeyVectorWord;
end
end
end
end

- Check if answerKeyVector or studentAnswerVector contains an abbreviation, then replace the abbreviation with the long form. This converts "SDU" to "Süleyman Demirel University" for both answerKeyVector and studentAnswerVector.
- Check if answerKeyVector or studentAnswerVector contains a digit, then replace the digit with the word form. This converts "6" into "altı" ("altı" means "six") for both answerKeyVector and studentAnswerVector.

In the next step of ILSA calculation, LSA is applied to the corresponding row vectors; similarity is calculated according to the cosine similarity that LSA uses. ILSA does not use latent semantic kernels (LSKs) [31]; instead, ILSA directly controls synonymy and hypernym using the above rules and equations as an a priori process.

4. APPLICATION AND EVALUATION OF TASAG SOFTWARE

The aim of the TASAG software is to prepare online exams for Suleyman Demirel University, Faculty of Engineering, Computer Engineering Technology Program (Distance Education). Instructors use TASAG to prepare online exams, homework, and quizzes that have short-answer questions stored in the system and to automatically assess them.

The TASAG software can be used by other departments after adding technical terms for their specialized field to the Extended Turkish WordNet and Zemberek. Additionally, in the presentation layer, AJAX is used for web user interfaces. When a request is sent in the presentation layer, it does not influence the entire webpage because AJAX overhauls the site page data without needing to reload. In this manner, website pages are stacked and the reaction to client requests is fast. In addition, TASAG uses separate servers for the data layer and the business layer. Therefore, the workloads of the two servers do not influence each other. The TASAG software is created for the web environment. However, the n-tier design provides extensibility, scalability, and adaptability. Therefore, it can be utilized as a part of different environments, including the entirety of Süleyman Demirel University. The application server and database server can be scaled up by adding additional servers if needed.

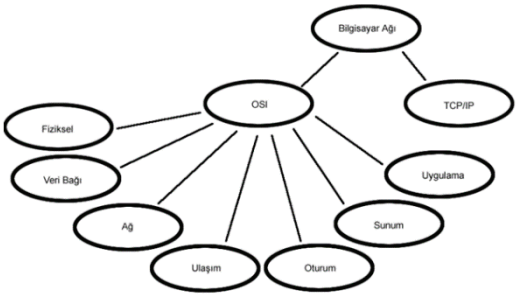

Figure 6 The extended Turkish WordNet nodes.

There are many studies in the literature which automatic grade short-answer questions such as c-rater [6] project. However, no studies have been published for Turkish automatic short answer grading. The TASAG software uses a hybrid model combining cosine and ILSA methods. The TASAG software defines which method will be used for evaluation of scores according to the number of words in the answer keys. The number of words in the answer keys is considered, by the reason of choosing the correct dimensionality for LSA is important to success [32]. Dimensionality of LSA is composed with word number which is used in the answer key of the question and document number which is equal to number of students that have the test. LSA is applicable to longer texts [30,32,33] and also ILSA contains a LSA calculation step. For this purposes, 15 tests are performed with different answer key word count to calculate Cosine and ILSA similarity. If word counts are less, LSA and LSA-based ILSA give worst results when compared with Cosine similarity. Experimental results (Table 2) shows that Cosine similarity gives better results up to 10 words rather than LSA and ILSA for the same question. When the keyword counts increased, Cosine similarity decreases, however, ILSA similarity increases accordingly with the increase of word count in the keyword. There is a rapid change in Cosine and ILSA result when word count is reached to 10. So, proposed system is developed according to these experimental results.

To measure the accuracy of the TASAG software, a case study is performed at Suleyman Demirel University, Faculty of Engineering, Department of Computer Engineering (Distance Education). Forty-one students took an exam which has 10 short-answer questions for a computer networks course (e.g., "Ağ Topolojisi nedir?" which means "What is topology of network?"; "Doğrusal Topolojinin dezavantajlarını yazınız" which means "Write disadvantage of the bus topology." The questions answers have no anaphora references to the question.

Table 2 Cosine and ILSA comparison test results.

Test Number	Answer Key Word Count	Cosine Result	ILSA Result
1	4	0.956	0.682
2	5	0.948	0.695
3	6	0.935	0.718
4	7	0.921	0.724
5	8	0.914	0.741
6	9	0.896	0.763
7	10	0.839	0.817
8	13	0.821	0.822
9	15	0.816	0.831
10	16	0.792	0.839
11	18	0.787	0.846
12	22	0.773	0.862
13	26	0.750	0.874
14	30	0.729	0.889
15	32	0.704	0.917

Example of the question, answer key 1, and answer key 2:

- Question: Yineleyicinin görevleri nelerdir?
- Answer Key 1: Ağ kablosunun erişebileceği maksimum mesafeyi uzatırlar. Ağdaki maksimum düğüm sayısını arttırır. Kablo arızalarının etkisini azaltabilir. Farklı kablo tipleri kullanan ağları birleştirebilir.
- Answer Key 2: Ağın maksimum mesafesini uzatır. Ağdaki bilgisayar sayısını arttırır. Kablo arızalarını azaltır ve farklı kablo tiplerindeki ağlar birleştirebilir.

Two instructors performed question and answer key preparation to assess the validity and reliability of the exam scoring. Instructors can prepare and enter different amount of answer key for each question without any limitation. Firstly, 10 questions were prepared for the exam by both instructors working together. The instructors entered the correct answers (answer key) for the questions with the question point (each question is out of 10 points for this case study) into TASAG for each question. After the students had taken the exam, student answer and answer key are compared with similarity methods. For example, if the similarity between student answer and answer key is 0.712, questions score is found as 0.712 x 10 (question point defined and entered to the system by instructor) which is equal to 7.12 points. After the calculations are done for each question, both instructors and students could see the exam scores, which were automatically evaluated by the TASAG software. Each answers score can be seen separately; the total score for the exam can be seen by both instructors and students. To analyze and evaluate the TASAG software's accuracy, the instructors scored the students answers according to their answer key. Instructors prepare answer key according to the keywords. Then, while scoring student answer of a question, instructor looks if student answer consists keywords from answer key or not. So, instructor gives a point (score) to the question subjectively. However, while instructors may be biased for any student (e.g., regarding classroom performance grade of student), TASAG system score each student answer objectively according to the same answer key so there are no subjective criteria. Answer key 1, prepared by instructor 1, and was entered into the TASAG software. The TASAG 1 score was calculated by TASAG based on answer key 1. A comparison of instructor 1's score and TASAG1's score shows similar results, as shown in Figure 7.

The curves are similar; generally, there is little difference between the curves. Therefore, the scoring values of Instructor 1 and TASAG 1 are close. Similarly, answer key 2 (prepared by instructor 2) for the same questions were entered into the TASAG software. TASAG 2's scores were calculated based on answer key 2. Instructor 2's score and TASAG 2's score was compared; the results are shown in Figure 8.

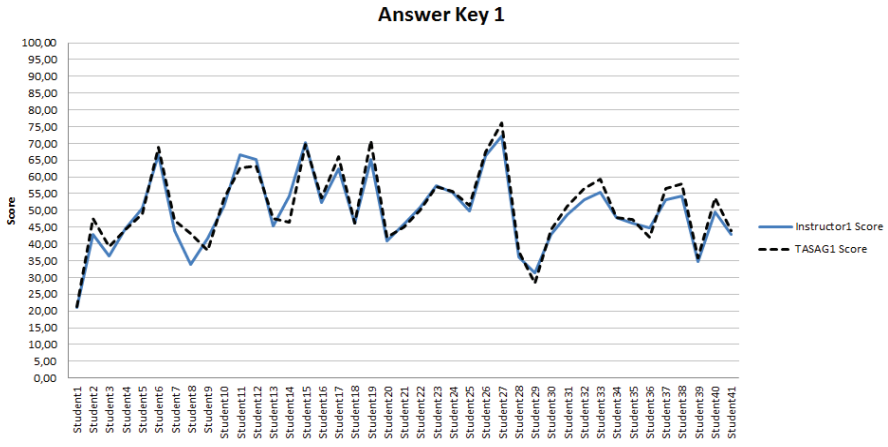

Figure 7 Comparison graphics of Instructor1 score and TASAG1 score.

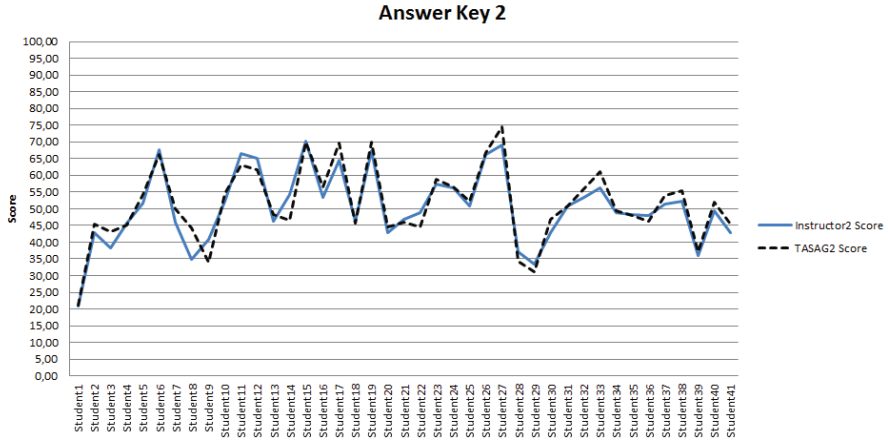

Figure 8 Comparison graphics of Instructor2 score and TASAG2 score.

A comparison of instructor 2's score and TASAG 2's score shows similar results, as shown in the figure. Finally, the TASAG arithmetic mean score (TASAGAMS) is calculated using the arithmetic means of TASAG 1's score and TASAG 2's score to achieve optimal concordance. Comparisons of this score with the instructors' scores are shown in Figure 9. The TASAGAMS score is very close to both instructor 1's score and instructor 2's scores. Therefore, TASAG gives the TASAGAMS score as the final score for the exam result to both instructors and students.

Kendall's W, is coefficient of concordance, can be used for assessing agreement between 3 or more raters. In this study, Kendall's W was calculated between 3 rankers' instructor 1, instructor 2, and TASAGAMS, which is used for assessing agreement among instructor and system scores. Kendall's W ranges from 0 (no agreement) to 1 (complete agreement) [34]. The Kendall's W result was 0.9254 among instructor 1, instructor 2, and TASAGAMS; the coefficient of 0.9254 indicates an excellent degree of agreement (Table 3). Additionally, Kendall's W was calculated among instructor 1, instructor 2, and TASAG 1 (and then among instructor 1, instructor 2, and TASAG 2). The Kendall's W result was 0.9204 among instructor 1, instructor 2, and TASAG 1. The Kendall's W result was 0.9174 among instructor 1, instructor 2, and TASAG 2.

Then, paired t test was performed. According to the t-test results, there is no significant difference between TASAGAMS and either instructor 1 or instructor 2 ($p > 0.05$), t and p values are shown Table 4. The TASAG uses a hybrid model combining cosine and ILSA methods. For the test of ILSA, a comparison is examined for ILSA that is used in the TASAG. Kim and Kim [35] developed autonomous assessment system based on combined a LSK [30], first of all, a term-sync set matrix was created to transform the WordNet hierarchical structure. The term-conjugate set matrix was created using the terms t_i and N pairs of synonyms obtained from the text as in Formula (2). According to the WordNet hierarchy, synonyms have the value 2 for synonyms, 1 for words with 1 distance, and 0.5 for words with 2 distances. While creating synonyms, words with 0, 1, and 2 distances were taken into consideration, words with a distance greater than 2 were not used.

$$t_i, = s_1, s_2, \ldots \ldots, s_i, \ldots ..s_N \tag{2}$$

The term-co-cluster matrix created in formula (3) was subjected to singular value decomposition and the value obtained when the obtained values were reduced to the k dimension was assigned as the P matrix. Identity calculation between two texts was calculated according to Formula (3).

$$\text{similarity}\left(d_1, d_2\right) = \cos\left(P^T d_1, P^T d_2\right) = \frac{Td_1{}^T PP\, d_2}{\|P^T d_1\| \|P^T d_2\|} \tag{3}$$

ILSA and the LSK method is compared. Sixty-three students took an exam which has 10 open-ended, short- answer questions for a computer networks course. The results are shown in Table 5. Instructor, ILSA and LSK total score mean correlation results are shown in Table 6, according to the Pearson's correlation. According to the obtained data, it was determined that instructor and ILSA total score mean is closer, the instructor and LSK total score mean.

In this case study, two instructors gave scores for the same questions to increase the accuracy of the TASAG software's scoring results. In practice, more instructors can provide answer keys for the same exam so that the accuracy of the software can be increased. Consequently, the TASAG software can be used effectively for the evaluation of shortanswer question scoring, instead of the instructors. Moreover, although the case study performed was for a computer networks course, the TASAG can be used for other courses

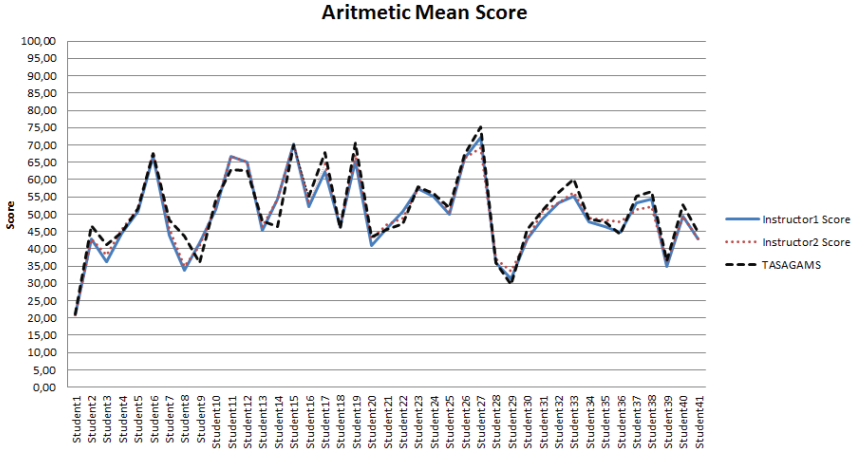

Figure 9 | Comparison graphics of Instructor1, Instructor2, and TASAGAMS scores.

Table 3 | Kendalls' W values.

Raters Comparison	W	Chi-square	df	p
instructor1, instructor2, TASAGAMS	0.9254	117.0	40	0.0001
instructor1, instructor2, TASAG1	0.9204	117.0	40	0.0001
instructor1, instructor2, TASAG2	0.9174	116.4	40	0.0001

Table 4 t and *p* values for instructor 1 and TASAGAMS, instructor 2, and TASAGAMS.

Question	Rater 1	Mean1	t1/p1	Rater2	Mean2	t2/p2
Question 1	Instructor 1	3.9317	−0.862	Instructor 2	3.9317	−0.862
	TASAGAMS	4.2973	0.391	TASAGAMS	4.2973	0.391
Question 2	Instructor 1	6.9268	0.684	Instructor 2	6.9268	0.684
	TASAGAMS	6.5024	0.496	TASAGAMS	6.5024	0.496
Question 3	Instructor 1	6.2683	−0.791	Instructor 2	6.2683	−0.791
	TASAGAMS	6.8320	0.431	TASAGAMS	6.8320	0.431
Question 4	Instructor 1	6.0976	0.137	Instructor 2	6.0976	0.137
	TASAGAMS	5.9907	0.891	TASAGAMS	5.9907	0.891
Question 5	Instructor 1	6.1463	−0.65	Instructor 2	6.9512	0.772
	TASAGAMS	6.5062	0.517	TASAGAMS	6.5062	0.442
Question 6	Instructor 1	5.9512	−0.22	Instructor 2	5.9512	−0.22
	TASAGAMS	6.0727	0.826	TASAGAMS	6.0727	0.826
Question 7	Instructor 1	1.2683	−0.094	Instructor 2	1.2683	−0.094
	TASAGAMS	1.3112	0.925	TASAGAMS	1.3112	0.925
Question 8	Instructor 1	1.7531	−0.201	Instructor 2	1.6341	−0.471
	TASAGAMS	1.8617	0.841	TASAGAMS	1.8617	0.639
Question 9	Instructor 1	3.2927	0.147	Instructor 2	3.1707	−0.046
	TASAGAMS	3.1988	0.884	TASAGAMS	3.1988	0.963
Question 10	Instructor 1	8.1220	−0.692	Instructor 2	8.1220	−0.695
	TASAGAMS	8.4806	0.491	TASAGAMS	8.4806	0.489

Table 5 Instructor, ILSA , and LSK total score mean.

Question	Instructor Score	ILSA Score	LSK Score
Question 1	100	100	91
Question 2	22	21	10
Question 3	65	79	100
Question 4	45	40	39
Question 5	85	90	70
Question 6	30	30	30
Question 7	12	13	23
Question 8	67	62	63
Question 9	45	50	50
Question 10	85	70	80

Table 6 Correlation between instructor, ILSA, and LSK total score mean.

		Instructor Score	ILSA Score	LSK Score
Instructor score	Pearson correlation	1	.967*	.882*
	Sig. (2-tailed)		.000	.001
	N	10	10	10
ILSA score	Pearson correlation	.967*	1	.923*
	Sig. (2-tailed)	.000		.000
	N	10	10	10
LSK score	Pearson correlation	.882*	.923*	1
	Sig.(2-tailed)	.001	.000	
	N	10	10	10

*$p < 0.01$.

by adding terms to Turkish WordNet [23,28] that are specific to the course content.

ASAG systems are important for distance education and web-based learning systems. Moreover, the recent increase in distance education programs increases the need for automatic scoring systems. In an exam, if there are very many questions and students, scoring of the exams becomes time consuming. Moreover, in the digital age, there is no need to write paper exams. For these reasons,

using the TASAG, instructors only need to provide questions and correct answers using user-friendly web interfaces. After students take the exams, both students and instructors can see automated scorings from the TASAG, within a few minutes, without any interference. Therefore, time and resources are saved by employing the TASAG, software.

The software is beneficial for education in terms of saving both instructors' time and economic costs. The evaluation results suggest that the TASAG, software could provide effective scoring systems for Turkish educational institutes such as schools, colleges, and universities.

5. CONCLUSIONS AND FUTURE WORK

Researchers have developed ASAG software for their own languages, typically English. In this paper, a web-based Turkish automatic short answer grading software was developed and employed for a real exam. The novelty of this study is that TASAG is the first software of its kind for the Turkish language. The algorithm of the TASAG software is a hybrid that determines which method will be used at runtime based on the word number dimensions to achieve accurate scoring. In a case study, instructors scoring results and the TASAG software scoring results were compared. Two instructors prepared different answer keys for the same exam to increase the accuracy of the scoring. The scoring results, which are compared in the figures, are very close to each other, which indicate the effectiveness of the TASAG software. Moreover, TASAGAMS scores for each answer key are calculated and given as the final score for the exam. Therefore, high score accuracy is achieved.

In real-world situations, the instructor can prepare different answer keys for the same exam as if there were two different instructor answer keys to create an arithmetic mean of the TASAG for optimal concordance of scoring. The TASAG can be used for automated short-answer question scoring in Turkish with a 92% success rate according to Kendalls' W concordance. Consequently, the TASAG, which uses natural language processing methods, is used effectively in computer engineering education for a computer network course exam which has short-answer, open-ended questions. If the Extended Turkish WordNet used by the systems is enriched, the TASAG software can be used for other courses and by other educational institutes.

In future, the TASAG can be opened to other institutes for educational purposes by using web services. Moreover, other NLP methods such as the ontological approach [36], or multidimensional assessment method (Hoang and Ngamnij [37], or LSA derivatives [38], or Wikipedia-based explicit semantic analysis (ESA) [39,40], or ESA derivatives [41] can be used in future research. In this research, ESA was not utilized because ESA uses in a high-dimensional space of natural concepts derived from Wikipedia, or, for Turkish, Vikipedi [42].

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

The corresponding author worked in writing the paper, all authors worked collaboratively to write the literature review and discussion. All authors read and approved the final manuscript.

Funding Statement

The project was funded by The Scientific and Technological Research Council of Turkey (project number EEEAG-114E952).

ACKNOWLEDGMENTS

The authors wish to thank The Scientific and Technological Research Council of Turkey, who financially supported this project (project number EEEAG-114E952).

REFERENCES

[1] E. Hettiarachchi, E. Mor, M.A. Huertas, A.E. Guerrero-Roldn, Introducing a formative E-assessment system to improve online learning experience and performance, J. Univ. Comput. Sci. 21 (2015), 1001–1021.

[2] M. Mohler, R. Mihalcea, Text-to-text semantic similarity for automatic short answer grading, in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Athens, Greece, 2009, pp. 567–575.

[3] M.O. Dzikovska, R.D. Nielsen, C. Leacock, The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications, Lang. Resour. Eval. 50 (2016), 67–93.

[4] E. Agirre, A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, L. Uria, Semeval-2016 task 2: interpretable semantic textual similarity, in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, CA, USA, 2016.

[5] S. Burrows, I. Gurevych, B. Stein, The eras and trends of automatic short answer grading, Int. J. Artif. Intell. Educ. 25 (2015), 60–117.

[6] C. Leacock, M. Chodorow, C-rater: automated scoring of short-answer questions, Comput. Humanit. 37 (2003), 389–405.

[7] E. Alfonseca, D. Perez, Automatic assessment of open ended questions with a BLEU-inspired algorithm and shallow NLP, in: J. Vicedo, P. Martinez-Barco, P. Munoz, M. Saiz Noeda (Eds.), Advances in Natural Language Processing, vol. 3230 of Lacture Notes in Computer Science, Springer, Berlin, Germany, 2004, pp. 25–35.

[8] O. Bukai, R. Pokorny, J. Haynes, An automated short-free-text scoring system: development and assessment, in Proceedings of the 20th Interservice Industry Training, Simulation, and Education Conference, Association for Computational Linguistics, Orlando, Florida, USA, 2006, pp. 1–11.

[9] C. Gutl, e-Examiner: towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems, in Proceedings of the 2nd International Conference on Interactive

Mobile and Computer Aided Learning, Amman, Jordan, 2007, pp. 1–10.
[10] S. Bailey, D. Meurers, Diagnosing meaning errors in short answers to reading comprehension questions, in Proceedings of the 3rd ACL Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Columbus, OH, USA, 2008, pp. 107–115.
[11] S. Jordan, T. Mitchell, e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback, Br. J. Educ. Technol. 40 (2009), 371–385.
[12] D. Sima, B. Schmuck, S. Szöll, A. Miklos, Intelligent short text assessment in eMax, in: I.J. Rudas, J. Fodor, J. Kacprzyk (Eds.), Towards Intelligent Engineering and Information Technology, vol. 243 of Studies in Computational Intelligence, Springer, Berlin, Heidelberg, Germany, 2009, pp. 435–445.
[13] D. Meurers, R. Ziai, N. Ott, S.M. Bailey, Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions, Int. J. Contin. Eng. Educ. Life Long Learn. 21 (2011), 355–369.
[14] L. Cutrone, M. Chang, Kinshuk: Auto-assessor: computerized assessment system for marking student's short-answers automatically. In: N.S. Narayanaswamy, M.S. Krishnan, Kinshuk, R. Srinivasan (Eds.), Proceedings of the 3rd IEEE International Conference on Technology for Education, IEEE, Chennai, India, 2011, pp. 81–88.
[15] S. Jordan, Short-answer e-assessment questions: five years on, in Proceedings of the 15th International Computer Assisted Assessment Conference, Southampton, UK, 2012, p. 1.
[16] D. Perez-Marin, I. Pascual-Nieto, Willow: a system to automatically assess students free-text answers by using a combination of shallow nlp techniques, Int. J. Contin. Eng. Educ. Life Long Learn. 21 (2011), 155–169.
[17] M. Heilman, N. Madnani, ETS: domain adaptation and stacking for short answer scoring, in Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics, Atlanta, GA, USA, 2013, pp. 275–279.
[18] T. Zesch, O. Levy, I. Gurevych, I. Dagan, UKP-BIU: similarity and entailment metrics for student response analysis, in Proceedings of the 17th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Atlanta, GA, USA, 2013, pp. 285–289.
[19] S. Jimenez, C.J. Becerra, A.F. Gelbukh, A.J.D. Bátiz, A. Mendizábal, SOFTCARDINALITY: hierarchical text overlap for student response analysis, in SemEval@ NAACL-HLT, Georgia, USA, 2013, pp. 280–284.
[20] M.D. Akin, A.A. Akin, Zemberek an open source NLP framework for Turkic languages structure, Structure. 10 (2007), 1–5.
[21] G.A. Miller, Dictionaries in the mind, Lang. Cognit. Process. 1 (1986), 171–185.
[22] A. Horak, P. Smrz, New features of wordnet editor VisDic, Rom. J. Inf. Sci. Technol. 7 (2004), 1–13.
[23] S. Stamou, K. Oflazer, K. Pala, D. Christodoulakis, D. Cristea, D. Tufis, *et al.*, BALKANET: a multilingual semantic network for the balkan languages, in Proceedings of the International Wordnet Conference, Mysore, India, 2002, pp. 21–25.
[24] S. Bochkanov, V. Bystritsky, ALGLIB-a cross-platform numerical analysis and data processing library ALGLIB Project. Novgorod, Russia, 2011. [Online]. Available: http://www.alglib.net/
[25] M.H. Stefanini, Y. Demazeau, TALISMAN: a multi-agent system for natural language processing, in: J. Wainer., A. Carvalho (Eds.), Advances in Artificial Intelligence, Springer, Berlin, Heidelberg, Germany, 1995, pp. 312–322.
[26] Y. Aktaş, E.Y. İnce, A. Çakır, Wordnet ontology based creation of computer network terms by using natural language processing, J. Tech. Sci. 7 (2019), 1–9.
[27] R. Çölkesen, B. Örencik, Bilgisayar haberleşmesi ve ağ teknolojileri, Papatya Press, İstanbul, Turkey, 2003.
[28] O. Bilgin, Ö. Çetinoğlu, K. Oflazer, Building a wordnet for Turkish, Rom. J. Inf. Sci. Technol. 7 (2004), 163–172.
[29] Türkçe Eş Anlamlar Sözlüğü, Mythes-tr. 2018, https://github.com/maidis/mythes-tr.
[30] A. Kontostathis, Essential dimensions of latent semantic indexing (LSI), in 40th Annual Hawaii International Conference on System Sciences (HICSS 2007), IEEE, Waikoloa, HI, USA, 2007, p. 73.
[31] N. Cristianini, J. Shawe-Taylor, H. Lodhi, Latent semantic kernels, J. Intell. Inf. Syst. 18 (2002), 127–152.
[32] T.K. Landauer, S.T. Dumais, A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychol. Rev. 104 (1997), 211.
[33] P. Foltz, W. Kintsch, T. Landauer, The measurement of textual coherence with latent semantic analysis, Discourse Process. 25 (1998), 285–307.
[34] M.G. Kendall, B. Babington-Smith, The problem of m rankings, Ann. Math. Stat. 10 (1939), 275–287.
[35] Y.B. Kim, Y.S. Kim, Latent semantic kernels for WordNet: transforming a tree-like structure into a matrix, in International Conference on Advanced Language Processing and Web Information Technology (ALPIT'08), IEEE, Dalian, China, 2008, pp. 76–80.
[36] Y. Zhang, R. Witte, J. Rilling, V. Haarslev, Ontological approach for the semantic recovery of traceability links between software artefacts, IET Softw. 2 (2008), 185–203.
[37] L.P. Hoang, A. Ngamnij, Assessment of open-ended questions using a multidimensional approach for the interaction and collaboration of learners in E-learning environments, J. Univ. Comput. Sci. 19 (2013), 932–949.
[38] S. Hao, Y. Xu, D. Ke, K. Su, SCESS: a WFSA-based automated simplied chinese essay scoring system with incremental latent semantic analysis, Nat. Lang. Eng. 22 (2016), 291–319.
[39] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, J. Artif. Intell. Res. 34 (2009), 443–498.
[40] F. Rahutomo, M. Aritsugi, Econo-ESA in semantic text similarity, SpringerPlus. 3 (2014), 149.
[41] Y. Haralambous, V. Klyuev, Thematically reinforced explicit semantic analysis, Int. J. Comput. Linguist. Appl. 4 (2013), 79–94.
[42] Vikipedi, 2016. https://tr.wikipedia.org/wiki/Vikipedi