الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
ⵜⴰⵙⴷⴰⵡⵉⵜ ⴰⴽⵍⵉ ⵎⵓⵃⴻⵏⴷ ⵓⵍⵃⴰⵊ - ⵜⵓⴱⵉⵔⴻⵜ -
Faculté des Sciences et des Sciences Appliquées
وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
-البويرة-
كلية العلوم والعلوم التطبيقية

AUTOMATIC GRADING OF SHORT ANSWERS

Leila OUAHRANI

Thesis Supervisor: Pr. Djamal BENNOUAR

Submitted in fulfillment of the requirements for the degree of Doctor of Science in Computer Science

Computer Science Department
Faculty of Applied Sciences
Akli Mohand Oulhadj University – Bouira
2023/2024

The arbitration committee consists of:

Prof. Amad Mourad	President	University of Bouira
Pr. Boustia Narhimene	Examiner	University of Blida 1
Dr. Chouiref Zahira	Examiner	University of Bouira
Prof. Baadache Abderrahmane	Examiner	University of Algiers 1
Dr. Boufenar Chouki	Examiner	University of Algiers 1
Prof. OuldKhaoua Mohammed	Examiner	University of Blida 1
Prof. Djamal Bennouar	Supervisor	University of Bouira

In memory of my parents

To my family

Abstract

Developing effective Automatic Short Answer Grading (ASAG) for e-learning environments is challenging. We consider scoring a text-constructed student answer compared to a teacher-provided reference answer. In this thesis, we address three key issues. The first key issue involves addressing the challenge of managing diverse student answers, considering that the reference answer may not cover their full range and often includes only specific correct responses. Secondly, developing an accurate grading model that enhances sentence similarity computation without requiring a large number of manually marked student responses is essential. Thirdly, it is crucial to ensure seamless integration into the Learning Management System (LMS) to enhance the practicality and scalability of the proposed system. The proposed solution addresses these challenges through three key components: a sequence-to-sequence deep learning network paraphrase generator, a supervised grading model, and an extension to the LMS quiz system. First, we provide a sequence-to-sequence deep learning network aimed at producing plausibly paraphrased alternative reference answers based on the provided reference answer. Second, we develop a supervised grading model that enhances features with specific and general course domain information using computational distributional semantics. Finally, we extend the question engine of the LMS quiz system to our model as a plugin for the open-source Moodle platform. Templates for creating and grading short answer questions have been successfully established and shared. Conducted in Arabic and English, quantitative experiments show that the paraphrase generator produces accurate paraphrases. The grading model yields comparable results to state-of-the-art and is deployed with low computational complexity to support short answers in online assessment. Two case studies were conducted in a real educational environment. The first case study resulted in the creation of the AR-ASAG dataset, which is the first publicly available Arabic dataset for ASAG evaluation. The second case study involved conducting a qualitative evaluation of a controlled class of students through formative and summative assessments using the proposed solution. The discussion covers the findings and implications, emphasizing valuable insights to advance the e-assessment of free-text short answers in online higher education.

Keywords. Automatic Short Answers Grading, Automatic reference answer generation, automatic assessment tools, distributional semantics, Learning Management Systems, encoder-decoder, Paraphrase generation, supervised learning

الملخص

يلعب التعلم الإلكتروني دورًا مهمًا في التعليم العالي. التقييم هو أحد أهم أجزائه. يعد تطوير نظام تقييم آلي فعال للإجابات القصيرة (ASAG) لبيئات التعلم الإلكتروني مهمة صعبة. نأخذ في الاعتبار تسجيل إجابة الطالب التي تم إنشاؤها بواسطة النص مقارنة بالإجابة المرجعية المقدمة من المعلم. ونعتبر ثلاث قضايا في هذه الأطروحة. أولاً، إدارة تنوع إجابات الطلاب، حيث أن الإجابة المرجعية لا تغطي تنوعها وتحتوي فقط على أمثلة محددة للإجابات الصحيحة. ثانيًا، توفير نموذج درجات دقيق من شأنه تحسين حساب تشابه الجملة دون الحاجة الصعبة لعدد كبير من اجابات الطلاب المحددة يدويًا. ثالثًا، تحسين التكامل في نظام إدارة التعلم لجعل النظام المقترح ممكنًا عمليًا وعلى نطاق واسع. لذلك، فإن الحل المقترح لحل هذه القضايا يسلط الضوء على ثلاثة مكونات. أولاً، نقدم نموذجًا للتعلم العميق من تسلسل إلى تسلسل يستهدف إنشاء إجابات مرجعية بديلة معاد صياغتها بشكل معقول ومشروطة بالإجابة المرجعية المقدمة. ثانيًا، نقترح نموذج تصنيف خاضع للإشراف يثري الميزات بمعرفة مجال الدورة المحددة والعامة باستخدام دلالات التوزيع الحسابية. أخيرًا، قمنا بتوسيع نظام اختبار محرك الأسئلة الخاص بنظام إدارة التعلم ليشمل نموذجنا كمكون إضافي لمنصة موودل مفتوحة المصدر. لقد نجحنا في إنشاء نماذج ومشاركتها لإنشاء أسئلة ذات إجابات قصيرة وتقييمها. أظهرت التجارب الكمية، التي أجريت باللغتين العربية والإنجليزية، أن مولد إعادة الصياغة ينتج إعادة صياغة دقيقة. ويؤدي نموذج الدرجات إلى نتائج مماثلة لأحدث النتائج ويتم نشره بتعقيد حسابي منخفض لدعم الإجابات القصيرة في التقييم عبر الإنترنت. تم إجراء دراستي حالة في بيئة تعليمية حقيقية. أسفرت دراسة الحالة الأولى عن مجموعة بيانات AR-ASAG، وهي أول مجموعة بيانات عربية متاحة للجمهور لتقييم الإجابات القصيرة النصية. أجرت دراسة الحالة الثانية تقييمًا نوعيًا لفئة من الطلاب من خلال التقييمات التكوينية والختامية باستخدام الحل المقترح. تتم مناقشة النتائج والآثار المترتبة، ويتم تعلم الدروس لتعزيز التقييم الإلكتروني للإجابات القصيرة النصية في التعليم العالي عبر الإنترنت.

الكلمات المفتاحية. تصنيف الإجابات القصيرة تلقائيًا، إنشاء الإجابات المرجعية تلقائيًا، أدوات التقييم التلقائي، دلالات التوزيع، أنظمة إدارة التعلم، توليد إعادة الصياغة، التعلم الخاضع للإشراف.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AR-ASAG Dataset	ARabic Automatic Short Answer Grading dataset
ASAG	Automatic Short Answer Grading
AraT5	Arabic Text-to-Text Transfer Transformer
BERT	Bidirectional Encoder Representations using Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BLEU	Bilingual Evaluation Understudy
BOW	Bag Of Words
ELMo	Embeddings from Language Model
GPT	Generative Pre-trained Transformer
ICT	Information and Communication Technologies
QWK	Quadratic Weighted Kappa
LCMS	Learning Content Management System
LCS	Longest Common Subsequence
LMS	Learning Management System
LSTM	Long Short-Term Memory
ML	Machine Learning
MOODLE	Modular Object-Oriented Dynamic Learning Environment
NLP	Natural Language Processing
POS	Part-of-Speech
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation
SOTA	State-Of-The-Art
SVM	Support Vector Machine
SVR	Support Vector Regression
T5	Text-to-Text Transfer Transformer
TF-IDF	Term Frequency-Inverse Document Frequency
VLE	Virtual Learning environments

Acknowledgements

اللهم لك الحمد والشكر كما ينبغي لجلال وجهك الكريم وعظيم سلطانك وعلو مكانك.
اللهم لك الحمد والشكر في الأولى ولك الحمد والشكر في الأخرة ولك الحمد والشكر من قبل ولك الحمد والشكر من بعد وأناء الليل وأطراف النهار وفي كل حين ودائماً وأبداً.

It took a while to write this thesis. While there have been various challenges along the way, there have also been several advancements. It is now appropriate to express gratitude to the individuals who had an impact on the latter.

My greatest gratitude goes out to Professor Djamal Bennouar, my thesis supervisor, whose guidance, expertise, and hand unwavering support have been invaluable throughout this research endeavor.

I extend my sincere appreciation to the distinguished members of the thesis jury for their readiness to examine and comment on the thesis. Their expertise and thoughtful considerations will undoubtedly improve this research's thoroughness and quality.

I appreciate Pr. Wail Gomaa from Cairo University (Egypt) for giving us access to his Arabic dataset (the Cairo University dataset), which enabled us to conduct our initial tests very early on.

I would like to thank Mr. Cherif Zahar, assistant professor at the Computer Science Department of Blida 1 University, and Ms. Boukhatem Fawzia, an Arabic language teacher at Abd El-Hamid Mahri Ain Defla High School, for carrying out the tedious task of the manual annotation during the collection of the AR-ASAG dataset.

I will never be able to thank several people enough for their invaluable assistance in providing human expertise on our automatically generated Arabic and English paraphrased test sets. I gratefully name Mr. Taha Zerrouki, lecturer at the Computer Science Department of the University of Bouira; Ms. Zahra Fatma Zohra, assistant professor at the Computer Science Department of Blida 1 University; Mr. Maamar Missoum and Mr. Moustafa Messeded, lecturers at the English department of Blida 2 University; and Mr. Amirat Omar, official translator at the Bejaia course. Without their invaluable expertise and contribution, we would not have been able to carry out such an in-depth qualitative evaluation. Their support was truly remarkable.

I extend my sincere thanks to my master's students, whom I had the privilege of supervising as part of this thesis. Their diligence and enthusiasm enriched our collaborative efforts and catalyzed my progress in this thesis. I kindly name in the chronology: Mohamed Hamza Hannoufi & Adel Nassim Henniche, Garoudja Khadidja & Abdallah Amina, Atoub Yasmine & Benayad Asma, Madani Ahmed Abderraouf & Snoussi El Hareth, Boulhouache Houda & Abassi Selma, Oukina Faiza Radia & Amar Seti Imene, Hadjersi Mohamed & Benguergoura Oussama, Lamari Selena & Hamel Oussama, and Benali Ahmed Chouaib & Abdenour Ben Hamida.

I am grateful to my friends and colleagues at Blida 1 University's Computer Science Department for their encouragement and support.

A full mention for my family for their steadfast support, encouragement, and understanding throughout my academic journey. Their love, patience, and constant motivation have enabled me to pursue and accomplish this thesis with serenity.

To those whom I may have inadvertently forgotten to thank, please accept my sincere apologies. I am truly grateful for your presence in my academic journey.

Chapter 1: Introduction

The purpose of this chapter is to contextualize the planned research. We establish the study aims, define essential concepts, and provide an overview of the background, research environment, and methods.

1.1 BACKGROUND

The influence of globalization and technological progress in the 21st century has profoundly transformed the techniques by which students learn, strengthening the use of Information and Communication Technologies (ICT) for both teaching and evaluation. The transition has generated increasing interest among higher education institutions and instructors in connecting evaluation methodologies with current social requirements (Boitshwarelo et al., 2017). The shift from traditional assessment methods to modern approaches has led to the emergence of e-assessment, which relies solely on technological devices for assessment tasks (Jordan, 2013; Appiah & Van Tonder, 2019; Said et al., 2019).

E-learning, which has become increasingly popular in higher education, uses information and communication technologies (ICTs) to help the learning process. The Learning Management System (LMS), a form of ubiquitous computing, may be integrated into classroom pedagogy to facilitate learning both inside and outside of the classroom (Kumar & Sharma, 2016). The Learning Management System (LMS) may serve as a valuable tool for instructors and administrators in higher education, aiding them in their tasks and responsibilities. LMS is intended to assist educators in achieving their pedagogical objectives by facilitating the delivery of content and assessments to students (Machado & Tao, 2007).

Assessment plays a pivotal role in e-learning by evaluating the effectiveness of teaching methods and student learning outcomes. Nevertheless, online education frequently receives criticism for its elevated dropout rates, with some attributing this trend to a perceived deficiency in educational engagement (Qiu, 2019). Higher education institutions accommodate a significant volume of students, leading to distinct challenges for instructors, especially in devising equitable and efficient assessment methods across diverse learning settings. Instructors are tasked with

constructing tailored assessment strategies that suit the distinct needs of face-to-face interactions as well as the virtual nature of online learning environments. Online tests provide students with the flexibility to choose when and where they take assessments, leading to improved accessibility and timely feedback (Khan & Khan, 2019). The necessity for conducting extensive assessments at scale, coupled with the cost of manual grading, has driven the innovation and adoption of automated assessment systems (Jordan, 2013; Whitelock & Bektik, 2018).

Learner knowledge can be assessed through both objective and subjective tests. Objective tests typically include multiple-choice questions, true/false questions, and matching questions. In contrast, subjective tests focus on descriptive answers to open-ended questions, such as short-answer questions and essay questions (Ashton et al., 2005; Jordan, 2013). With the rise of technology in education and the shift towards more student-centered approaches, open-ended questions are becoming increasingly prevalent in higher education (Beckman et al., 2019). Open questions allow students to develop their own interpretations of the requirements, and online technologies provide greater flexibility and enable new types of interactions between teachers and students (Furst, 1981; Sychev et al., 2020).

Short answer questions require concise and focused written responses using appropriate vocabulary related to the subject. Short-answer questions are designed to evaluate students' understanding of key facts and fundamental concepts (Ziai et al., 2012). They are commonly used to assess students' comprehension, critical thinking, and ability to articulate their understanding of a topic. Education Bloom's Taxonomy states that short-answer questions effectively assess a learner's understanding and synthesis of information (Furst, 1981; Bloom, 1984). However, grading these short answers can be a daunting task for educators, especially in large classrooms or online courses where the volume of responses is high (Jordan, 2013).

Historically, short answer responses have been graded manually by teachers, involving reading each response individually, assessing its quality, and assigning a grade based on predefined criteria. This process can be time-consuming and subjective, leading to inconsistencies in grading across different teachers and instances. With the advancement of technology, particularly in the fields of natural language processing (NLP), machine learning, and artificial intelligence (AI), automated grading systems have emerged as a promising solution to the challenges of manual grading. These systems leverage algorithms to analyze and evaluate students'

short answer responses, providing quick and objective feedback to both students and teachers.

ASAG systems play a crucial role in scaling assessments to accommodate large numbers of students efficiently. They offer numerous advantages over manual grading methods. Firstly, they significantly reduce the time and effort required for grading, enabling teachers to focus more on providing personalized feedback and engaging with students. Secondly, they promote consistency and fairness in grading by applying predefined criteria consistently across all responses, thereby reducing the likelihood of bias or subjectivity. Finally, they facilitate timely feedback, allowing students to receive immediate insights into their performance and areas for improvement and more engagement. Formative assessment is becoming more and more necessary to assist students in assessing their learning (Hettiarachchi Enosha et al., 2015). This is especially useful if prompt and precise feedback is given. Both teachers and students would benefit from the LMS's integration of effective ASAG systems, which would also promote more participation in the assessment process.

Developing a robust scoring system for short answers in e-learning settings poses challenges due to the subjective nature of questions, diverse linguistic styles, and varying responses based on different topics. In recent years, automatic short answer grading has garnered significant attention from researchers across various academic institutions and research organizations worldwide.

The research environment surrounding this topic is characterized by interdisciplinary collaboration, technological advancements, and a commitment to addressing the challenges of traditional grading methods in educational settings by improving the grading accuracy and reliability. Research focused on educational technology, natural language processing (NLP), machine learning (ML), and artificial intelligence (AI) as well as techniques for analyzing the semantic and syntactic properties of text to infer meaning and relevance. Until recently, ASAG models have not achieved human-like performance in scoring answers (Shermis, 2015; Jayashankar & Sridaran, 2017; Schneider et al., 2023).

Automatic short answer grading faces several challenges and limitations. One of the main challenges is accurately assessing the quality and depth of student responses, particularly in subjective subjects or open-ended questions where there may be multiple valid interpretations. Additionally, automated grading systems may struggle with non-standard language, ambiguous wording, or unconventional

responses, leading to inaccuracies in grading. Moreover, few ASAG systems are integrated and made accessible on e-learning platforms, and the ones that are remain heavily reliant on human oversight. The scoring models are more sophisticated and more complex to implement at scale in practice. Currently, the transmission of course materials and objective tests is the main application of the LMS.

Beyond academic research, ASAG has practical applications in various educational settings, including higher education institutions, online learning platforms, and standardized testing environments. Motivated by the practical need for efficient and effective assessment solutions, we are driven to develop an ASAG approach that can address real-world challenges and meet the diverse needs of educators and learners.

This thesis aims to identify the critical components required for establishing an automatic short-answer grading system project in practice. The broad objective is to improve scalability, efficiency, pedagogical efficacy, assessment equality, educational impact, and easy and practical application. By addressing these goals, we want to contribute to the ongoing evolution of automated grading technology and its enormous potential for reform in educational settings.

1.2 RESEARCH ENVIRONMENT

This study was conducted at Algerian public higher education institutions, with a specific focus on the utilization of ASAG within them. In order to comprehend the context and extent of the suggested solution for ASAG development, it is necessary to examine the structure of the higher education sector in Algeria. Subsequently, a concise overview of the instances examined in this thesis is provided to situate the empirical context within the realm of higher education. This will provide a comprehensive background for the analysis and findings of this study.

1.2.1 E-learning in Algerian Higher Education Institutions

E-learning in Algerian higher education institutions is steadily gaining ground, although facing several challenges (Ghouali & Cecilia, 2021). Algerian higher education institutions have increasingly recognized the potential of e-learning to expand access to education, improve learning outcomes, and adapt to the demands of the digital age. There has been a growing interest in integrating e-learning technologies and methodologies into existing curricula.

The Algerian government has shown support for e-learning initiatives in higher education through policy frameworks and funding programs. Efforts have been made to develop infrastructure, provide training for educators, and promote the use of digital resources in teaching and learning (Benharzallah, 2020). One of the challenges facing e-learning in Algerian higher education is the need to localize and adapt digital content to align with the cultural, linguistic, and educational context of Algeria. Developing relevant and high-quality educational resources in Arabic and French languages is essential for ensuring the effectiveness of e-learning initiatives. Ensuring the quality and accreditation of e-learning programs and courses is another area of concern.

Algerian higher education institutions need to establish mechanisms for quality assurance, assessment, and accreditation of online learning offerings to maintain academic standards and credibility. Our university, like most public universities in our country, is equipped with an online learning environment that hosts the Moodle Learning Management System (LMS) as part of a government initiative to support in-person instruction and promote technology-enhanced learning in higher education. Courses and assessments are primarily conducted in Arabic, French, and English. Students generally follow a system of continuous formative assessments, which provide ongoing feedback to both students and teachers. Summative assessments are conducted at the end of the learning activities. Offering assessments as opportunities to practice is essential.

In a recent survey conducted to assess online learning in Algerian universities during COVID-19 (GUEMIDE & Maouche, 2020), online testing, comprising 5% of the assessment, has seen limited adoption among educators. Regrettably, only a small number of teachers have opted for online testing for their students. Those who have embraced this approach emphasized its application primarily in subjects where online features are deemed essential, such as oral skills and phonetics (GUEMIDE & Maouche, 2020). Generally, teachers in the majority of fields use short answer questions to evaluate their students, with the exception of medicine, where selection-type questions are the main technique (Bennouar, 2013). Particularly in technical domains like computer science, electronics, and civil engineering, short-answer questions are common. In contrast, social sciences such as psychology and legal science typically use essay-style questions (Bennouar, 2017). Because selection-type questions are inappropriate for gauging the breadth of students' learning in their courses, teachers prefer short answer questions. They contend that a written response's

content might include a variety of components that together provide a more realistic picture of a student's understanding of the material.

Responding to a short answer question involves more mental work than a selection-type inquiry since students have to come up with their own meaningful responses and arguments. When answering a selection-type question, the right answer is already included in the possibilities. In many disciplines, short-answer questions become the most used. Unfortunately, automated grading tools for short answers, provided in the LMS, are underused because not simple to use and require significant manual supervision.

The increasing number of students each year is making the task of grading manually their assessments tedious and time-consuming when considering short answer questions. More tutors are needed to carry out this task, which makes it difficult to reach an agreement on the evaluation criteria. In the realm of grading free-text short responses, the utilization of grading methods within Algerian universities faces considerable hurdles, primarily due to environmental constraints and unaddressed issues such as course specificity and language dependency. Recognizing these challenges, this thesis aims to identify the crucial factors influencing ASAG projects to enhance their development, implementation, and integration into higher education.

1.2.2 Cases in this Thesis

The goal for the ASAG project is to identify strategically important issues on which to focus in order to maximize the benefits of the ASAG system and positively contribute to enhancing learning and teaching. The prerequisites for obtaining and developing ASAG should be assessed in realistic settings. For these considerations, we conducted two case studies in the context of this thesis:

Case study 1 conducted at Blida 1 University. (Dataset collection).

The first case study was conducted at Blida 1 University in the computer Science Department. The objective was to collect data to train and evaluate our grading model. We created our own Arabic dataset in order to solve the issue of the scarcity of publicly available datasets for short response grading assignments and the requirement for realistic training data encompassing a range of learning objectives. This dataset includes answers from three master student courses and questions taken from the "cybercrimes" course. One hundred seventy high school students who are natural Arabic speakers took part in the course taught by the thesis author. An official test was

then used to authenticate the course. These tests were given in an outdoor setting. Initial system usage yielded satisfactory results, and baseline metrics were established. Details of the data collection process and the case study are outlined in Chapter 3.

Case study 2 conducted at Bouira University. *(ASAG Integration into the online system for evaluation on formative and summative assessments).*

For the purpose of qualitative evaluation, using the university's Moodle online platform, where students are enrolled in various courses, experiments were carried out with students. Both from home and at the institution, students can use the LMS. Formative and summative evaluations for the "cybercrimes" course were conducted using the proposed grading system incorporated into the LMS's Question Engine. Throughout the semester, students were assigned tasks as part of a formative evaluation. The tasks included both short answer and objective question types (multiple choice, fill in the gaps, etc.). In the student interface, a history feature allows students to revisit their past examinations. They can view detailed information about each question, including their own response, the grade they received, and the reference answers. More comprehensive evaluation details, results and discussions are provided in Chapters 3 and 4.

1.3 MOTIVATION FOR THE RESEARCH

Improving assessment methods may have a big influence on how well students are taught and learn. The rising use of Moodle Learning Management Systems (LMS) in Algerian higher education institutions provided the impetus to investigate ASAG systems in e-learning environments. This adoption was a component of a larger government program designed to support technology-enhanced learning and reinforce conventional in-person instruction.

Our motivation for studying ASAG systems is multifaceted. We aim to contribute to the ongoing evolution of automated grading technology and its transformative potential in higher education in several ways:

Scalability. The increasing adoption of e-learning platforms and online education has led to a surge in the demand for scalable assessment solutions. Manual grading of short answer responses in these contexts can be time-consuming and resource-intensive.

ASAG systems streamline the grading process for short answer questions in e-learning platforms, enabling instructors to handle a large volume of student responses efficiently. This scalability ensures that courses with a high enrollment can effectively manage assessments without overwhelming teachers.

Efficiency. Traditional manual grading methods are prone to human error and subjectivity, leading to inconsistencies in assessment outcomes. ASAG systems aim to improve the efficiency and reliability of grading by leveraging technology to analyze and evaluate short answer responses objectively and consistently.

Immediate Feedback. ASAG systems provide immediate feedback to students on their short answer responses. In e-learning environments where direct interaction with instructors may be limited, this instantaneous feedback is invaluable for promoting continuous learning and enabling students to identify and address areas of misunderstanding promptly facilitating a continuous feedback loop that enhances student engagement, comprehension, and retention of course materials.

Pedagogical Innovation. By automating the grading of short answer questions, instructors can focus on developing their students' thinking, critical analysis, and problem-solving skills.

Personalized Learning. By analyzing individual student responses, ASAG systems can offer personalized feedback and adaptive learning experiences tailored to each student's needs. This personalization enhances engagement and supports diverse learning styles, ultimately improving the effectiveness of e-learning platforms.

Time Savings for teachers. Automating the grading of short answer questions with ASAG systems frees up instructors' time, allowing them to focus on other aspects of course delivery, such as designing engaging content, facilitating discussions, and providing one-on-one support to students. This timesaving enhances instructors' ability to deliver high-quality instruction in e-learning environments.

Educational Impact. Our study in ASAG has the potential to have a significant impact on education by improving assessment practices, enhancing learning outcomes, and informing instructional decision-making. By advancing the state-of-the-art in automated grading technology, we may contribute to the broader goals of promoting educational equity, accessibility, and excellence.

Finally, Automated Short Answer Grading (ASAG) holds practical significance across diverse educational contexts, spanning higher education institutions, online learning platforms, and standardized testing environments. We are committed to designing

ASAG systems capable of tackling real-world hurdles and catering to the manifold requirements of educators and learners alike, fueled by the pragmatic demand for streamlined and impactful assessment methodologies.

1.4 AIM OF THE RESEARCH

The growth of e-learning has created several chances to enhance evaluation procedures. The majority of computer-assisted assessment systems offer multiple-choice, true/false, and matching question templates. A smaller number of systems only give rudimentary assistance with managing short answer questions, such as essays and short response questions.

Short response questions, which range from a few words to a few sentences and are written in plain language, are better at testing retained information by emphasizing recall and replication (Furst, 1981; Bloom, 1984). However, evaluating them is a difficult and essentially subjective procedure that calls for in-depth knowledge of material written in natural language.

The research environment surrounding this topic is characterized by interdisciplinary collaboration, technological advancements, and a commitment to addressing the challenges of traditional grading methods in educational settings. Automating the assessment of large numbers of students requires a multi-faceted challenging solution.

In this thesis, we may respond to three challenges managing especially: the lack of resources for low resourced languages considering the Arabic language as an example of under-resourced languages, the diversity of student responses, and the integration and the scalability of ASAG systems with educational technologies such as learning management systems and online assessment platforms

First, addressing the lack of resources challenge. Despite years of research, Automatic Short Answer Grading Systems (ASAG) are not widely used in practice because of their complexity (Liu et al., 2014). To address the ASAG task, various systems utilize machine learning, manually crafted patterns, or templates, in conjunction with information extraction techniques. Drawing on the works of Mohler and Mihalcea (2009), Mohler et al. (2011), Gomaa & Fahmy (2014a), Zahran et al. (2015), Magooda et al. (2016), (Sultan et al., 2016), Bennouar (2017), and Gomaa and Fahmy (2020), the approach taken involves treating the problem as a semantic similarity challenge between the Student's Answer (SA) and the teacher's Model

Answer (MA). The assessment of semantic similarity between two short texts primarily relies on two methods: topological similarity (knowledge-based) and statistical corpus-based similarity (Mihalcea et al., 2006).

Topological similarity utilizes data structures such as WordNet[1], thesauri, and dictionaries, which encompass information about concepts and their relationships. In contrast, statistical similarity employs vector space models to represent word correlations derived from text corpora. For English, numerous topological and statistical methods for determining semantic similarity already exist. However, due to their reliance on advanced, language-specific natural language processing techniques, only a few of these methods are adaptable to under-resourced languages like Arabic.

Researchers with the goal of automating language analysis, parsing, and annotation have created numerous NLP tools. In these activities, language resources are used primarily for two goals. First, they offer extensive annotated corpora that support statistical natural language processing methods. Second, they offer test collections (datasets) for assessing NLP systems against a gold standard. Initiatives like the Language Resources and Evaluation Map are among the attempts to catalog such NLP resources (Gratta et al., 2014).

These resources are restricted, nevertheless, for certain languages. Arabic is a relevant example. With very few significant exceptions, Arabic is known to have limited publicly available tools and resources despite its extensive usage (Mahmoud El-Haj et al., 2015). To be more precise, Arabic NLP lacks fully automated basic NLP tools, including tokenizers, part-of-speech taggers, parsers, stemmers, and semantic role labelers, in addition to corpora, lexicons, machine-readable dictionaries, and datasets (Ouahrani & Bennouar, 2019). Mahmoud El-Haj et al. (2015) state that practitioners of Arabic natural language processing have faced difficulties due to a lack of data and study. The Arabic WordNet (AWN)[2], which was created with an approach akin to WordNet, is deficient in concepts and semantic connections across synonym sets. Furthermore, privacy issues prevent the sharing of many datasets. The availability of high-quality datasets is paramount for training accurate ASAG models. However, biases within these datasets can affect the fairness and reliability of grading. For ASAG

[1] http://globalwordnet.org/
[2] http://globalwordnet.org/arabic-wordnet/

initiatives, academics frequently modify data from their own teaching experiences without taking into account the necessity of relevant comparisons with alternative techniques and literature works. With the exception of the Arabic Cairo dataset (Gomaa & Fahmy, 2014b), which is not accessible to the public, there are no additional datasets that fulfil the ASAG research requirements. As a result of the lack of a publicly accessible Arabic dataset, authors that concentrate on grading Arabic short answers frequently assess their models by examining individual samples (Wali et al., 2015; Al-Shalabi, 2016; Elghannam, 2016; Nababteh & Deri, 2017).

Thus, *to address the lack of resources challenge,* first, we provide the AR-ASAG (Ouahrani & Bennouar, 2020), a realistic Arabic Short Answer Grading for ASAG training and evaluation. Next, we investigate distributional semantics (Turney & Pantel, 2010; Higgins et al., 2014; Adams et al., 2016) for word distribution to overcome the topological resource scarcity . The word distribution semantic space is built on the COALS (Correlated Occurrence Analogue to Lexical Semantic) Algorithm (Rohde et al., 2004). Distributional word representations in text corpora are provided via the COALS space. Only a stemmer and an undifferentiated text corpus are the language-dependent resources needed. We enhance word distribution by term weighting. We use the collected dataset to address particularly the first aim. Using an incremental design process, we investigate the impact of word distribution, term weighting and stemming effect, as they are language-dependent aspects, on the grading process. The objective here is threefold. We aim to respond to these questions and identify:

(Question 1.1) How do the domain and dimension of semantic space distribution affect grading accuracy in an under-resourced language when using a semantic space approach for word distribution?

(Question 1.2) How can word weighting enhance grading quality, given that it has been rarely utilized in Arabic?

(Question 1.3) What impact do stemming techniques have on grading accuracy for a highly inflectional language like Arabic?

Second, addressing the diversity of student responses. When compared to a teacher reference answer, the ASAG forecasts similarity scores. The response that is most pertinently formulated is the reference answer. Accordingly, students are likely to receive a better score whenever their responses contain a high level of common text

that overlaps with the reference solution (Ramachandran et al., 2015). Short answers have many words in common; therefore, automatically evaluating them is troublesome since it depends on semantic similarity in the meaning (Ab Aziz et al., 2009). While a standard reference answer is regarded as the best option, it does not include all viable answers (Kumar et al., 2017). It leaves out certain possible ways to phrase the correct answer. Students blend synonyms, paraphrases, and other sentence patterns to create a response. Some student answers with a single reference answer could be correct since they do not resemble the reference answer at all. It may be possible to manage the variety of student answers and increase accuracy by providing several alternative reference solutions for the same question. There might be a few different ways to formulate the reference answer, but it would be challenging the instructor to do it manually.

The use of data augmentation in ASAG systems to enhance short answer grading has not gotten much attention. Because it requires human work and talent to generate paraphrases that accurately convey the content of the original reference response. Manually constructing alternative reference answers can be inefficient and time-consuming (Marvaniya et al., 2018). Moreover, it may not be a scalable or reliable approach for scoring short answers on a large scale. Consequently, there is a demand to automate the process of generating alternative reference answers.

In addressing this, two primary concerns emerge. Firstly, there is a necessity to generate automatically diverse reference answers capable of accommodating the variability in student answers. Secondly, enhancements to the grading model are required to ensure accurate grading.

One potential avenue for improvement involves enhancing sentence similarity computation through the utilization of multiple reference answers. Consequently, our suggested method for resolving both issues emphasizes two elements. Initially, we offer a deep learning sequence to sequence model that is designed to produce plausible paraphrases based on the provided reference answer. Second, we suggest enhancing the supervised grading model by utilizing characteristics from sentence embedding. The grading model enhances features to increase score accuracy. Although they employ different wording, paraphrases retain the same sense as the original sentences. By expanding the reference answer's language and writing style, the paraphrased answers will help it better encompass the range of student responses. While the English literature concerning the paraphrasing task is quite rich, limited works have dealt with

this task in Arabic. Our approach may respond to this challenging task in Arabic and establishes baselines.

Thus, *to address the challenging diversity of student responses*, the objective here is double. We aim to respond to these questions:

(Question 2.1) Can paraphrase generation enhance the ASAG system? How do multiple reference answers affect grading accuracy improvement?

(Question 2.2) How can we address the gap in Arabic paraphrase generation to enhance the quality of generated paraphrases in Arabic?

Third, addressing the integration of ASAG challenge into real and broader educational technologies. The integration of ASAG into broader educational technologies, such as learning management systems and online assessment platforms, is a growing challenge. Seamless integration enhances the user experience and facilitates the widespread adoption of ASAG in educational institutions. Despite extensive theoretical research, the practical adoption of ASAG systems within e-learning environments remains limited (Adams et al., 2016). Only a few ASAG tools are currently implemented and directly available on e-learning platforms, despite the advancement of grading models. There seems to be more emphasis on scoring accuracy than on seamlessly integrating grading systems into the e-learning environment.

Software in learning management systems (LMS) commonly employs regular expressions, templates, and logic expressions to match student responses with reference answers when grading short answers. While creating these manually can be time-consuming and beyond the capabilities of some teachers, they often yield high marks. The Open University-developed PMatch system from Open Mark (Butcher & Jordan, 2010; Jordan, 2012; Jordan, 2013) is regarded as the most advanced ASAG system for use in online learning contexts. It may yield extremely short responses, up to one phrase in length, and is based on matching keywords and their synonyms.

Training each question in the model necessitates a substantial number of student answers, which is a challenging. In order to determine all words, word stems, and synonyms required for appropriate responses against the reference answers, regular expressions use word-level pattern matching. Even with high scoring accuracy, LMS systems still do not use these types of questions enough. The time needed to

create answer matching and the need to collect multiple student answers for every question continue to be major obstacles.

In conclusion, the current state of e-learning systems' implementation and availability of ASAG tools is limited. They still need a lot of manual supervision, though. The scoring models have become increasingly complex and difficult to apply widely. One major challenge is that each question in the prompt requires hundreds of graded answers to train. The challenge in the LMS environments is to collect sufficiently manually labeled student answers for each question and the time needed to do it. Furthermore, the solutions developed for ASAGs do not envisage a harmonious integration with other types of questions and other activities assessment. Such an integration might use the potential of e-learning environments and make it easier for teachers to create a variety of tests that include designed and selected questions with the relevant feedback. We believe that the success of using Integrated Short Answer Grading on e-learning systems depends on how it is applied, even though the scoring is accurate.

Thus, *to address the integration and the scalability challenge in a real LMS environment,* the objective here is double. We aim to respond to these questions:

(Question 3.1) Can combining supervised learning with computational distributional semantics improve scoring accuracy while reducing the reliance on a large number of manually marked student responses?

(Question 3.2) How can integrating short answer assessment practices into e-learning environments enhance student achievement and support adaptive assessment design?

Aligned with these incremental goals and challenges, we have developed ISAGe (namely Integrated Short Answer Grader for e-learning environments) and made the following contributions:-learning environment and make the following contributions:

- We propose a supervised *general question model* that provides and enriches a listing of the most features that are important for the grading task and contributing to the score accuracy,
- The model uses computational compositional linguistics to integrate specific and general domain knowledge as features without the need for a huge number of labelled answers for each question,

- The model is trained on a realistic dataset (The AR-Arabic dataset) developed and evaluated in the context of the thesis,
- The developed dataset is publicly shared to be used in ASAG training and evaluation,
- Multiple reference answers are generated to improve the grading accuracy. Our offering is a sequence-to-sequence deep learning model with attention mechanism designed to generate credible paraphrased reference responses based on the given reference answer.
- The ASAG tool, which includes a grader and a paraphrase generator, is integrated seamlessly as a plugin into the online quiz system within the LMS. Our new ASAG Question Type Plugin expands the capabilities of the LMS's Question Engine to consider free-text short answer questions.
- Our proposed approach guarantees openness with the LMS and expands its quiz system's question engine to our ASAG model. This implies that it is feasible to make a quiz that smoothly combines established question types—such as essays and multiple-choice—with questions supported by our ASAG (free-text short answer questions).
- The grader in outsourced to run separately on a cloud when evaluating a student response to promote scaling in a real environment. This is especially interesting when a large number of students work at the same time, and on the other hand, encourages the design of incremental and adaptive ASAG assessment.
- Quantitative and qualitative experiments in Arabic and English are conducted to evaluate the proposed approach and its effectiveness in practice.

1.5 CONCEPTUAL FRAMEWORK AND METHODOLOGY

The research entails both the development of a conceptual framework (which serves as the theoretical foundation) and the definition of a methodology (which outlines the research techniques and methods) for assessing the effectiveness of the proposed Automated Short Answer Grading system. The methodology offers a systematic approach and specific techniques for executing tasks and achieving objectives within the framework. This section outlines the theoretical underpinnings guiding our research and the systematic approaches we have adopted to investigate responses to research issues. We introduce the selected methodology and its adequacy and application for this thesis.

1.5.1 The framework of the study

The framework of this research is shown in Figure 1, which also lists the different components of the study and how they relate to each other in order to help with decision-making and problem solving. Incremental design (Shore & Warden, 2008; Blokdyk, 2017) is used as a basis when developing the proposed ASAG solution in a systematic manner, gradually adding features and making improvements over time. Each design increment focuses on delivering a specific set of features, which are then tested, evaluated, and potentially refined before moving on to the next increment. Given that our research spans multiple domains, including ASAG (Automated Short Answer Grading) research, e-learning, educational research, and NLP (Natural Language Processing), the research action affects how incremental design should be focused on and which aspects should be emphasized.

To maximize the effectiveness of ASAG assessment practices, we should provide grounds for identifying critical features in the grading process by combining supervised learning, distributional semantics, and paraphrase generation. The evaluation of requirements for acquiring and creating ASAG has to be carried out in practical environments. The two case studies were conducted for this thesis with these factors in mind. The cases provided realistic data for quantitative evaluation, as well as feedback from practical experience for qualitative evaluation.

The Action Research within a qualitative evaluation was guided by continuous improvement, increasing the effectiveness and efficiency of short-answer grading practices. The action research entailed identifying research issues and concepts based on real-world needs and priorities, designing and implementing interventions to address identified challenges in short answer grading, collecting and analyzing data, and reflecting and iterating for continuous improvement.

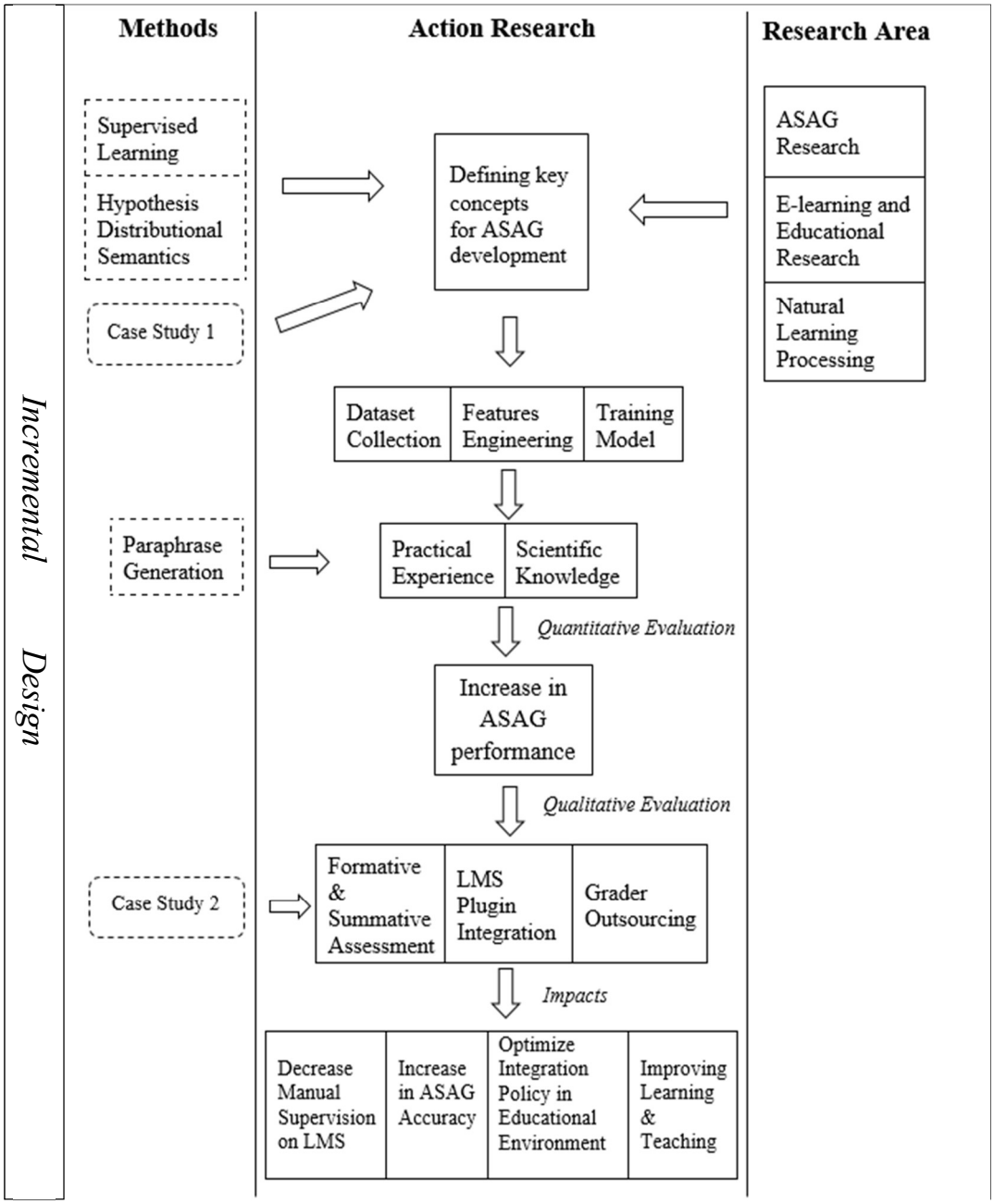

Figure 1 Framework of the thesis

1.5.2 Selection of Methodology

In this study, we employed a mixed-methods research methodology, which integrates quantitative and qualitative research approaches, to investigate research issues. This approach allowed us to gain a comprehensive knowledge of the ASAG development project by combining the strengths of both quantitative and qualitative methods according to our research questions. Furthermore, we incorporated an incremental design approach into our research methodology.

Incremental design involves breaking down the research process into iterative stages, each building upon the insights and findings from the previous stage (Winter, 2014). Combining mixed-methods research methodology with incremental design provided us with a structured approach to investigate systematically the influence of key factors (such as linguistic specifications, distributional semantics, paraphrase generation, feature engineering, and implementation practices) in real-world educational settings. Through each iteration, we aim to improve incrementally the accuracy, reliability, and usability of the ASAG system. We employed a multi-stage methodology that encompassed two case studies, feature engineering, model enhancement, and qualitative and quantitative evaluation techniques.

The first case study focused on data collection and model development. We collected the AR-ASAG Arabic dataset and evaluated it using an unsupervised grading model. Subsequently, we enriched feature engineering to propose a supervised model for improved accuracy. In the third stage, we enhanced the model by introducing paraphrase generation techniques to generate multiple alternate reference answers, thereby refining the accuracy of our ASAG system. Furthermore, we evaluated the generated alternate reference answers through intrinsic manual and automated metric evaluation. Additionally, paraphrase generation was evaluated extrinsically based on its impact on the ASAG task.

This comprehensive evaluation approach provided insights into the effectiveness of paraphrase generation techniques in improving the ASAG system's accuracy and reliability.

The second case study involved integrating our proposed ASAG system into an online evaluation system for formative and summative assignments. Experiments were conducted among students using the university's Moodle web platform to qualitatively assess the impact of integrating Automated Short Answer Grading (ASAG) into the learning and teaching processes. The ASAG system developed was seamlessly incorporated into the Question Engine Learning Management System for the 'cybercrimes' course, enabling both formative and summative assessments. Evaluations employed automated metrics, such as Pearson's correlation coefficient (r: higher values indicating superior performance), Root Mean Squared Error (RMSE: lower values indicating greater accuracy), as well as BLEU (Papineni et al., 2002), GLEU (Napoles et al., 2015) and METEOR (Lavie & Agarwal, 2007) scores, assessed against the mean of human-assigned ratings. The primary goal in both instances of this

thesis was to influence the educational environment in order to more effectively cultivate and employ ASAG to enhance the caliber of teaching and learning. Although the case studies were done in our academic e-learning environment, the findings have the potential to be relevant to a wider educational community. The results and consequences are presented and discussed in Chapter 4.

1.5.3 Rationalization of the Methodology

We discuss the rationale for the use of the selected methodology and how the selected methodology addresses the research objectives and how it aligns with the nature of the research topic and context. The selected methodology, which combines mixed-methods research methodology with incremental design, was chosen for its suitability in addressing the multifaceted nature of our research objectives and to adhere to the principles of validity and reliability, which are critical components of rigorous research inquiry.

The integration of mixed-methods research methodology allows us to leverage both quantitative and qualitative approaches, providing a comprehensive understanding of the research topic. Given the diverse aspects we aim to explore, ranging from distributional semantics to implementation practices, the inclusion of both quantitative and qualitative methods enables us to capture a wide range of perspectives and insights.

The aim of including a qualitative method is also based on the need to understand the environment and the impact of ASAG on it from a broader point of view as the research on e-assessment was absent at our university, and our knowledge of the implications was still modest. Additionally, the adoption of incremental design aligns with the iterative nature of our research process and the need for continuous improvement and optimization. By breaking down the research process into iterative stages, we can systematically refine our techniques, ensuring robustness and validity in our findings. We present a flexible and adaptive approach to developing and evaluating ASAG systems, which can ultimately result in more reliable and efficient solutions, by incorporating incremental design into a mixed-methods research methodology.

1.6 THE GENERALIZABILITY, RELIABILITY, VALIDITY, AND REPRODUCIBILITY OF THE RESEARCH

Generalizability refers to the applicability of our research beyond its original context. In the realm of Automated Short Answer Grading (ASAG), the ability to generalize research findings and theories is crucial for assessing their impact. Educational practitioners can use case studies to inform decision-making, while researchers can align their findings with quantified data to enrich the collective knowledge base of the field. Our study employs a mixed-methods research approach to investigate ASAG, allowing for a nuanced understanding of the phenomenon across diverse educational settings.

To achieve a comprehensive view of the research problem, we integrate both qualitative and quantitative findings considering the research questions at hand. The qualitative findings are often context-specific and emphasize depth of understanding over breadth. In contrast, quantitative findings offer a wider view and enable us to make general conclusions that extend beyond the specific case studies we carried out. By integrating quantitative and qualitative methodologies, we achieve a comprehensive examination of ASAG implementation, effectiveness, and implications. In today's society, research that directly influences practice is highly valued, underscoring the practical relevance of research endeavors.

Reliability refers to the degree to which several researchers working toward identical goals and examining the same issue produce consistent findings. As a result, our study was done in both Arabic and English. The results obtained from the English dataset indicate that the suggested study may be effectively generalized to other languages and datasets. The quantitative results of our study provided numerical data on the performance and accuracy of ASAG system, allowing for objective comparisons and analyses in the research field. In chapter 4, the implications of our research are presented and discussed.

Complementing the quantitative findings, the qualitative component of our study offered rich insights into the experiences, perceptions, and attitudes involved in ASAG implementation. While our research was conducted in a specific setting or population, the insights and conclusions drawn from our research can inform ASAG implementation and practice in other educational institutions, regions, or cultural contexts.

Validity means the extent to which the evaluation measures what it is supposed to measure. It refers to the extent to which the proposed models, solutions, or concepts exactly describe the phenomenon under investigation. We aimed to enhance the validity of our research by diversifying sources of data and perspectives:

Firstly, we addressed validity by carefully selecting the metrics that accurately capture the key constructs related to ASAG design, implementation, effectiveness, and impact. Through extensive literature review and pilot testing, we ensured that our quantitative measures (Pearson correlation, RMSE, Bleu, Gleu, Meteor, etc.) were aligned with the research objectives and the conceptual framework of ASAG systems.

Secondly, we focused on minimizing potential bias in automatic metrics that could influence the study outcomes in paraphrase generation. We undertook a manual assessment of the paraphrased responses, soliciting feedback from domain experts to evaluate the relevance and readability of the sample pairs. Furthermore, the proposed models underwent training and testing using the gold standard for human grades. To enhance the reliability of our findings, we employed systematic data collection procedures aimed at minimizing measurement errors and maintaining consistency in data collection and analysis, including manual correction and inter-annotator agreements.

Finally, qualitative methods were conducted to gain a more comprehensive understanding of the quantified results.

Reproducibility refers to the ability of other researchers to independently replicate our study methods and obtain similar results. In our thesis, we prioritized transparency and rigor to enhance the reproducibility of our research. We made our data and codes openly available to the research community, facilitating transparency and enabling other researchers to reproduce our findings. We shared the collected dataset publicly, allowing others to access and use the same data for further use and help overcome the problem of lack of ASAG datasets. Additionally, we shared codes for the trained models and LMS plugin templates to install the short answer plugin on a Moodle web platform, providing researchers with the necessary tools to replicate our design and implementation in their own contexts. By sharing our data and code through repositories or online platforms, we promoted the reproducibility of our research and encouraged collaborative inquiry and validation of our results.

For reproducibility, the created resources and codes are shared publicly during the development of the thesis:

- (AR-ASAG Dataset, 2020). The Arabic Dataset for Automatic Short Answer Grading and Evaluation
  - V. 1.0, ISLRN 529-005-230-448-6. https://www.islrn.org/resources/request/3582/
  - Shared on Mendeley (Elsevier): https://data.mendeley.com/datasets/dj95jh332j/1
  - Shared on Github: https://github.com/leilaouahrani/AR-ASAG-Dataset
- (Arabic Cyber Text Corpus, 2020). The Arabic In-Domain Cyber Text Corpus
  - V. 1.0, ISLRN 798-080-268-332-8. https://www.islrn.org/resources/request/2934/
- Plugin Moodle for ASAG (in Arabic and English): https://github.com/leilaouahrani
  - Arabic: https://github.com/leilaouahrani/ISAGe-Arabic
  - English: https://github.com/leilaouahrani/ISAGEe-English
- ARAG-ED: Alternative Reference Answer Generator Encoder-Decoder (in Arabic and English) :
  - Baseline Bi-Lstm : https://github.com/leilaouahrani/Bi-LstmPG
  - Encoder Decoder Without Attention Mechanism : https://github.com/leilaouahrani/ED
  - With Attention Mechanism : https://github.com/leilaouahrani/ARAG_ED
- COALS word distribution generator : https://github.com/leilaouahrani/COALS-Creator

1.7 RESEARCH SCOPE AND LIMITATIONS

This thesis focuses on the management aspects of ASAG development, such as grades, assessment activities, and feedback on the deployment in educational settings. Integrating ASAG systems into educational technologies and processes guarantees that they are compatible with learning management systems, assessment platforms, and institutional rules. This integration improves the educational experience for both students and educators. However, several more levels warrant effort, such as response interpretability or data privacy and security. As a result, ASAG systems may provide students and instructors with intelligible and practical explanations and recommendations for development.

ASAG systems may fail to protect the security and integrity of assessments in online settings, as automated grading systems are vulnerable to cheating or gaming methods. ASAG systems rely on collecting and analyzing student data, which raises privacy and security problems. Protecting sensitive information and adhering to data protection requirements are critical to upholding trust and ethical standards in ASAG initiatives. Addressing these additional problems will require interdisciplinary collaboration, rigorous research, and continual innovation in the application of ASAG systems *that are not considered in this thesis*.

1.8 THESIS OUTLINE

In the subsequent chapters of this thesis, there is a comprehensive exploration of the automatic short answer grading systems and the paraphrase generation task. Chapter 2 provides an in-depth analysis of existing literature, exploring the advancements, methodologies, and implications of automatic short answer grading systems and paraphrase generation tasks. Building upon this theoretical framework, Chapter 3 outlines the research design and data collection methodologies employed in this study. It provides insights into the systematic approach adopted for gathering data and designing the proposed solution. After executing the research design, Chapter 4 presents the outcomes and discussions derived from both quantitative and qualitative empirical evaluations. We provide a summary of the research findings, key insights, and overarching conclusions. Through a reflective lens, the practical implications of the research are examined. Furthermore, the contributions of this study to the field of automatic short answer grading are elucidated. Finally, Chapter 5 outlines critical pathways for future research and development in the field.

Chapter 2: Literature Review

The literature review presented herein delves into two domains within the realm of natural language processing and educational technology: automatic short answer grading (ASAG) and paraphrase generation used as data-augmentation strategy in this thesis. In the first section of the chapter, a landscape of ASAG systems is presented, examining from earlier models to advanced deep learning models and highlighting the most popular datasets and metrics used in ASAG evaluation. Additionally, the section discusses the challenges posed by Arabic NLP in the development of Arabic ASAG systems and highlights studies that have employed data augmentation for various reference answers in ASAG systems. E-assessment of short responses in Learning Management Systems with an emphasis on Moodle is presented since we foresee integrating ASAG into LMS. The subsequent section provides a comprehensive overview of paraphrase generation approaches.

2.1 AUTOMATIC SHORT ANSWER GRADING LANDSCAPE

The integration of Learning Management System (LMS) technology has become widespread in educational settings, facilitating the delivery of course materials and assessments. Within Virtual Learning environments (VLEs), assessment tools are commonly embedded, offering instructors and students a platform for evaluating learning progress and outcomes. In the development of e-learning environments, it is essential to incorporate authentic exercises drawn from real-world scenarios to enhance student engagement and satisfaction (Király et al., 2017). Well-designed assignments not only aid in reinforcing knowledge but also provide valuable opportunities for students to apply theoretical concepts in practical contexts. The concept of assessment can be broadly categorized into summative and formative assessment. Summative assessment typically occurs at the conclusion of a learning period or set of activities and serves to evaluate overall learning outcomes. Conversely, formative assessment is designed to provide ongoing feedback to both students and instructors, facilitating continuous improvement in teaching and learning practices. Offering assessments as opportunities for practice is crucial to promoting active learning and skill development among students.

Several types of questions can be automated, as demonstrated in Figure 2 (Burrows et al., 2015). The figure presents enough typical examples of questions to distinguish ASAG questions from other types of questions. Historically, automated assessment has primarily focused on objective question formats, which inherently limit the range of possible student responses. Most computer-assisted assessment systems predominantly offer templates for multiple-choice questions, true/false questions, and matching questions.

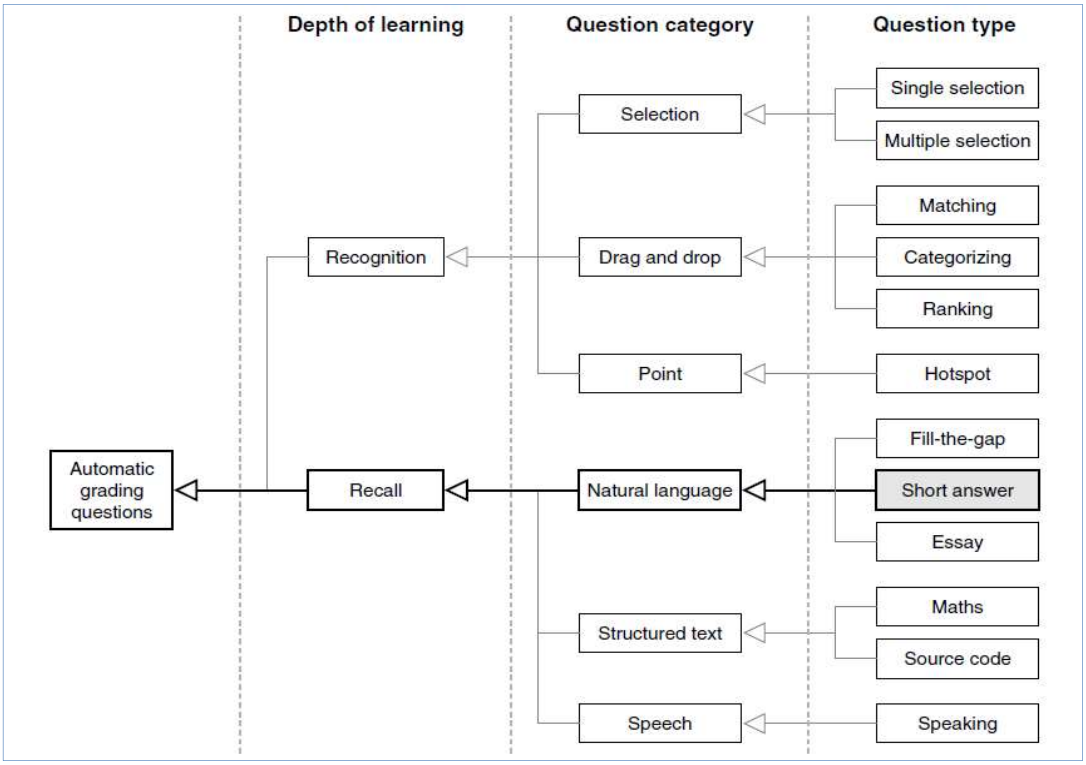

Figure 2 Overview of question types suitable for the application of automated grading techniques (Burrows et al., 2015).

In many examinations, short-answer questions, typically comprising a few words to a few sentences in natural language, are extensively employed (Bennouar, 2013). These questions often asses students with tasks like declaring, suggesting, describing, or explaining concepts, serving as vital components of assessments (Sukkarieh et al., 2003). Short answer questions are considered more effective for assessing acquired knowledge, with a focus on recall and reproduction (Furst, 1981). Automatic Short Answer Grading (ASAG) involves "assessing constructed short free natural language responses using computational methods" (Burrows et al., 2015). We identify a short answer as one that adheres to several key properties (Siddiqi et al., 2010; Burrows et al., 2015):

- It is written in natural language,
- It draws from knowledge beyond the scope of the question itself unlike selected questions,
- the length of the response should be about one phrase (few words) to one paragraph (maximum of 100 words),
- It prioritizes the content over the writing style, unlike essays, and
- It is both concise and precise, directly addressing the question or issue at hand.

The level of openness should be controlled through an objective question setting it apart from both open-ended and close-ended responses. An open-ended question cannot be answered with a simple "yes" or "no" response, or with a specific piece of information. Instead, it requires the student to provide a more detailed and elaborate answer. Open-ended questions are designed to encourage a full, meaningful answer using the subject's own knowledge and feelings. A close-ended question is a type of question that can be answered with a simple, direct response, often a single word or a short phrase. These questions usually have a limited set of possible answers, such as "yes" or "no," or a specific piece of information. Short answer questions, however, have posed challenges for e-assessment due to the diverse ways in which acceptable answers can be expressed, necessitating advanced automated natural language understanding capabilities (Sukkarieh et al., 2003; Ras and Brinke, 2015). However, only a limited number of systems provide basic support for managing short answer questions, such as short answer questions and essays (Conole & Warburton, 2005). However, with the increasing adoption of online courses and digital assessments, educators face new challenges in effectively grading student submissions, particularly short answer responses. However, evaluating short answer responses is a complex and inherently subjective task, necessitating thorough analysis and a deep comprehension of natural language texts.

Manual grading of short answer responses in online courses is time-consuming and labor-intensive, especially when dealing with large cohorts of students. Automated Short Answer Grading (ASAG) technology emerges as a promising solution to address these challenges, offering the potential to streamline the grading process, provide timely feedback to students, allows instructors to focus their attention on other aspects of course delivery and student support, and enhance the overall efficiency and effectiveness of assessment practices in education. Despite significant progress, challenges remain in the development and implementation of ASAG

technology (Butcher & Jordan, 2010; Jordan & Butcher, 2013). These challenges include the need for improved accuracy, scalability, and adaptability to different subject domains, and the incorporation of pedagogically assessment practices. Technical challenges associated with ASAG include the ability to handle linguistic variations, including differences in vocabulary, syntax, and writing styles across different languages and cultural contexts.

Addressing the complex task of Automated Short Answer Grading requires a multifaceted approach, encompassing a range of methodologies from statistical techniques to advanced natural language processing algorithms. Roy et al. (2015), Burrows et al. (2015), Galhardi and Brancher (2018), and Abbirah et al. (2022) conducted comprehensive surveys that offer valuable insights into the landscape of ASAG systems. Generally, the problem of short answer scoring is tackled through two primary approaches: response-based (rubrics-based) and reference-based methods (Sakaguchi et al., 2015). Response-based methods focus on extracting characteristics and features from student responses, such as lexical, syntactic, and semantic elements, to train models. These models can learn from a wide variety of student responses and accommodate different writing styles and expressions. However, they require a substantial amount of annotated data to train robust models effectively. Reference-based methods compare student responses to predefined reference answers provided by teachers, employing text similarity measures to assess how closely a student's response matches the reference answers. These methods can ensure consistency and alignment with expected answers but may struggle with correct responses that differ from the reference answers.

The approach of utilizing extracted rubrics rather than reference answers has only received limited recognition in ASAG literature. ASAG approaches draw upon a diverse range of techniques, including concept mapping, information extraction and pattern matching, document similarity, machine learning, attention-based and transformer-based with its own strengths and limitations. In this section, we provide a historical overview of the ASAG field and reveal the fundamental architecture of various systems.

2.1.1 Earlier approaches

The systems analyzed in this section took into account earlier approaches, including rule-based techniques like concept mapping and information extraction, as well as methods related to document similarity.

Concept mapping approaches analyze student answers by breaking them down into individual concepts. The grading process is based on the teacher and students' answers matching each other in terms of sentence-level concepts. Graders utilize concept-based lexicons, concept grammar (Nielsen et al., 2008), and textual entailment techniques (Levy et al., 2013) to identify the expression of concepts within the answers (Burstein et al., 1999). Various automated systems, such as Automatic Text Marker (ATM) (Callear et al., 2001), C-rater (Leacock & Chodorow, 2003; Sukkarieh & Blackmore, 2009), (Wang et al., 2008)' system, and others, aid in this process.

Information extraction and pattern matching approaches extract relevant information from parsed text chunks using predefined patterns. These patterns, whether manually crafted or generated automatically, are designed to pinpoint key concepts within text responses. Systems like Auto-Marking (Mitchell et al., 2002), WebLAS (Bachman et al., 2002), eMax (Sima, D. et al., 2009), FreeText Author (Jordan & Mitchell, 2009), IndusMarker (Siddiqi et al., 2010), and PMatch (Jordan, 2012) are examples of tools used for this purpose. However, creating patterns that cover all possible variations in student answers remains a challenge with these approaches. Sakaguchi et al. (2015) proposed a method that combines reference answers with rubrics. Their approach entailed presenting questions accompanied by aexemplary elements, to be discovered. This method employed a dual Support Vector Regression stack: the initial SVR aimed to align the student's response with a reference answer, while the second SVR analyzed shared key components in both the reference and student responses to assign points accordingly. Marvaniya et al. (2018) conducted a study to create scoring rubrics for student answers using clustering techniques. They selected sample responses from each cluster to use as benchmarks for scoring purposes. The incorporation of these exemplary responses with the reference answers significantly improved the accuracy of the scoring.

In learning management system (LMS) environments, the software employed for grading short-answer questions typically relies on various techniques, such as regular expressions, templates, and logic expressions, to identify specific terms or concepts within student responses. These systems can be categorized broadly into two

main groups: those utilizing computational linguistics for pattern matching and those employing keyword-based algorithms. Regular expressions, text templates, or predetermined patterns are frequently used to determine if a student's response fits specific words or phrases found in the reference answer. An example of this is the Moodle e-learning platform, which offers the Regular Expression Short Answer question type (Moodle, 2011). Teachers can input accurate answers using regular expressions. Similarly, systems such as PMatch and the Moodle Pattern Match question-type[3] used keyword matching and synonym identification. The process of accomplishing this involves using word-level pattern matching to compare all necessary terms, word stems, and synonyms with reference answers.

Although pattern matching offers a sophisticated alternative to traditional short-answer question types, its development necessitates the use of real student responses, which can be a time-consuming process. Despite achieving commendable scoring accuracy, such question types remain somewhat underutilized on LMS platforms. Research suggests that exploring machine learning for the development of response matching rules could address this challenge and facilitate wider adoption. Furthermore, investigating the consistency of student answers across different universities could enable the sharing of scored answers between ASAG systems

Document similarity techniques consider the ASAG problem as a matter of semantic similarity. Semantics play a crucial role in Natural Language Processing for Automated Short-Answer Grading Systems, as they enable the systems to understand and evaluate student responses accurately against a teacher-provided reference answer. ASAG system encounters semantic difficulties due to the diverse phrasing styles employed in responding, leading to challenges in accurately interpreting and evaluating student answers. In response to these challenges, the ASAG system employs various semantic similarity approaches. These include string-based, knowledge-based, and corpus-based methods. More recently, it has also incorporated deep learning techniques. These methods help to effectively tackle the semantic obstacles encountered in evaluating student answers (Mihalcea et al. 2006; Mohler and Mihalcea 2009; Mohler et al. 2011; Gomaa and Fahmy 2013; Cer et al. 2017; Kumar et al. 2017; Amur and Hooi 2022).

[3] https://docs.moodle.org/404/en/Regular_Expression_Short-Answer_question_type

String similarity measures assess how alike two text strings are for tasks like approximate string matching or comparison. These measures include various algorithms. The most used are described here. The Longest Common Substring (LCS) evaluates the length of the longest contiguous sequence of characters shared by two strings (Gusfield, 1997). The Damerau-Levenshtein (Hall & Dowling, 1980) distance counts the minimal number of operations required to transform one string into another. The Jaro–Winkler (Winkler, 1990) measure considers both the number and order of common characters, accounting for spelling variations. The Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) uses dynamic programming to perform global alignment across the entire sequence. Dice's similarity coefficient (Dice, 1945), often referred to as the Dice coefficient, evaluates the similarity between two texts by measuring the number of common terms (such as bigrams or other n-grams) relative to the total number of terms in both texts. The Jaccard similarity coefficient (Real & Vargas, 1996) is a metric for determining the similarity between two texts by quantifying the overlap between their sets of terms. It is calculated as the ratio of the count of shared terms (intersection) to the total count of unique terms (union) present in both texts. N-gram similarity (Barrón-Cedeño et al., 2010). Another measure, known as string-based similarity, was utilized at the beginning of the investigation to address short text similarity, specifically using cosine similarity. It calculates the cosine of the angle that the two vectors form, producing a similarity score that ranges from -1 to 1. Within the realm of text, vectors commonly denote the frequency or existence of terms in the two texts. Cosine similarity is extensively utilized in deep learning and neural networks to capture semantics through embeddings (Kenter & de Rijke, 2015; Lubis et al. 2021) as it is insensitive to the magnitude of the vectors, making it effective for comparing texts of different lengths.

Knowledge-Based Similarity leverages information from semantic networks to determine the degree of similarity between words. WordNet , a widely used lexical database for English, is frequently utilized for this purpose. WordNet comprises cognitive synonyms for nouns, verbs, adjectives, and adverbs, each representing a unique concept. There are six primary measures of semantic similarity: Resnik (Resnik, 1995), Lin (D. Lin, 1998), Leacock & Chodorow (Leacock & Chodorow, 1998), Wu & Palmer (Wu & Palmer, 1994). Popular packages for implementing

knowledge-based similarity measures include WordNet Similarity[4] and the Natural Language Toolkit[5] (NLTK).

Corpus-based similarity is a measure of semantic similarity that assesses the similarity between texts using information extracted from extensive corpora. The Hyperspace Analogue to Language (HAL) method (Lund & Burgess, 1996) creates a semantic space by looking at how often words appear together in a word-by-word matrix and giving each one a weight based on how close it is to the target word. Latent Semantic Analysis (LSA)(Deerwester et al., 1990) is a well-known technique that posits that words with similar meanings frequently appear in similar textual contexts. Through the statistical analysis of large document collections, LSA is largely used for its ability to reveal semantic relationships. The Extracting DIStributionally similar words using Co-occurrences (DISCO) (Kolb, 2008) method calculates distributional similarity between words by employing a simple context window of 3 words in the right and in the left to count co-occurrences. This approach has proven effective in various natural language processing tasks for capturing word similarities and relationships.

Hybrid methods utilize various similarity measures across different studies to achieve optimal performance by integrating multiple metrics. For example, (Li et al. (2006) presented a technique for assessing semantic similarity between sentences or very short texts, highlighting the impact of word order on sentence meaning. Similarly, Islam and Inkpen (2008) introduced the Semantic Text Similarity (STS) method, which evaluates the similarity of two texts by combining both semantic and syntactic information. Another approach that combines corpus-based semantic relatedness measures with knowledge-based scores has shown significant improvements in calculating semantic similarity between sentences. The UKP system, described by Bär et al. (2012) employs a log-linear regression model to integrate different methods for determining text similarities, including string similarity, semantic similarity, text expansion mechanisms, and measures of structure and style.

The text similarity approaches we mentioned are classified as non-deep learning techniques. Corpus-based measurements depend on the particular corpus they are used on, making use of its language and structure to compare two or more phrases.

[4] http://globalwordnet.org
[5] http://nltk.org/

On the other hand, knowledge-based measurements utilize ontologies like WordNet and dictionaries to construct semantic connections between words. Recent research indicate that deep learning methods have become increasingly popular, allowing computers to attain the highest level of performance in tasks related to text similarity (Abbirah et al. 2022; Amur and Hooi 2022). By combining deep learning approaches with corpus-based or string-based measurements, it becomes possible to conduct a more sophisticated analysis and achieve higher accuracy in assessing text similarity

2.1.2 Traditional Machine Learning (Hand-Engineered Features)

Machine learning systems for automated short answer grading leverage features extracted from natural language processing techniques, which are then integrated using classification or regression models, often through supervised learning methods (Galhardi & Brancher, 2018). These models typically rely on handcrafted features derived from dependency or constituency parsers to capture the structural and semantic nuances of both student and reference answers (lexical and semantic features based on document similarity methods).

Tools like Weka (M. Hall et al., 2009) can facilitate this process. Common features in this context include bag-of-words, n-grams and semantic text similarity, while Decision Tree, Support Vector Machine, Logistic Regression, Naive Bayes, Linear Regression, and K-Nearest Neighbors models are typical examples of the most learning algorithms used. For instance, researchers like Bailey and Meurers (2008), Gül (2007), and Cummins et al., (2016) have trained models on various questions within a specific domain using these features. Common features include similarity measures between student and reference answers, the length of the student's answer, question demotion, and overlap information.

In this context, Sultan et al. (2016) proposed a method for scoring short answers that combines several key features: term weighting, length ratios, question demotion, text alignment, and semantic similarity. This comprehensive approach aims to enhance the accuracy of ASAG by incorporating multiple dimensions of textual analysis, thereby providing a more robust evaluation of student responses. Although this method achieves good accuracy, it is limited by its reliance on high-quality dependency parsers, which are not widely available for all languages. This limitation restricts the applicability of such combined approaches to low-resource languages, such as Arabic, making them less suitable for large-scale real-world deployment.

Additionally, the development and maintenance of dependency parsers require significant linguistic expertise and resources, further complicating the extension of these methods to under-resourced languages.

Ramachandran et al. (2015) addressed this issue by utilizing a word-order graph approach to identify significant patterns from rubric texts and top-scoring student answers. This method involved constructing graphs that capture the sequential order of words, which can then be used to identify important syntactic and semantic patterns. They also incorporated semantic metrics to group related words, creating clusters of alternative answers that reflect different but correct ways of expressing the same idea. This approach allows for greater flexibility in recognizing correct answers, even when phrased differently from the reference answers.

2.1.3 Deep Learning Approaches

Since 2016, deep learning architectures have become increasingly popular for tasks involving text similarity, including Automatic Essay Scoring and Automatic Short Answer Grading (Abbirah et al., 2022). The progress in deep learning within the ASAG domain is closely intertwined with the methodological advancements in the field of NLP. The deep learning approaches used in Automated Short Answer Grading are a direct reflection of the historical evolution of Natural Language Processing and its techniques for text representation. We explore approaches involving in chronology three categories of models: word-embedding models, sequence-based models, and attention-based models and transformers.

Word embedding models use specific methods to generate vector representations of words, aligning related words to nearby vectors in a latent space for effective semantic representation. These models are trained on large, unannotated text corpora to enhance their understanding of word meanings (Mikolov, Chen, et al. 2013; Mikolov, Sutskever, et al. 2013; Mikolov et al. 2019). The creation of sentence embedding commonly entails aggregating individual word embedding through summation or averaging.

A wide range of publicly accessible word embedding models have been trained using extensive unlabeled data[6] ; the most used are Word2Vec, Glove, Fasttext, and Elmo. The Word2Vec model (Mikolov, Chen, et al., 2013) employs a neural network

[6] http://vectors.nlpl.eu/repository/

that acquires the ability to forecast a word by considering its neighboring context. After the training process, words that have similar meanings are placed in close proximity to each other in a vector space representation. This model is trained using a portion of the extensive Google News dataset, which contains 100 billion words. It is capable of generating vector representations for around 3 million words and phrases. Word2Vec has two main variations: skip-gram and CBOW (continuous bag of words). The skip-gram variation of the model predicts the context words surrounding a target word, whereas the CBOW variant predicts a target word based on a group of context words. GloVe (Global Vectors) (Pennington et al., 2014) takes parts from two main ways of making word embeddings: global matrix factorization and local context window, which is like word2vec. The model was trained using corpora that included Gigawords and the 2014 English Wikipedia dump. This training resulted in word vectors for 400,000 tokens. The fasttext model (Bojanowski et al., 2017) employs the skip-gram architecture, which represents each word as a set of its component character n-grams.

A vector representation is allocated to each letter n-gram, and the word vector is obtained by combining these n-gram vectors. The model provides a collection of 2 million word vectors that have been trained using data obtained from Common Crawl. Elmo (Embeddings from Language Models) (Peters et al., 2018) is a highly contextualized approach to word representation that captures differences in word usage across different language contexts (polysemy) as well as syntactic and semantic characteristics. Based on the internal state of a deep bidirectional language model trained on a dataset of one billion words, word vectors are produced.

Word embedding models have been the subject of multiple investigations in the ASAG fi.eld due to their ability to represent words as dense vectors in a continuous vector space, capturing semantic relationships and contextual meanings within text (Haller et al., 2022). These Approaches used embeddings of both student and reference responses to preserve the semantic and syntactic links between words. In reference (Magooda et al., 2016), authors compared various similarity metrics applied to pre-trained word vector representations from models like Word2Vec and GloVe to assess their effectiveness. Moreover, the researchers developed a grading system for short answers, showing similar effectiveness on the Cairo and Mohler datasets. Roy et al. (2016) introduce an ASAG method that overcomes the limitations of supervised approaches reliant on model answers and graded student responses. Their method incorporates an iterative ensemble approach that combines two classifiers: one

analyzes textual content in student answers, while the other utilizes numeric features from similarity metrics. Additionally, the method leverages transfer learning and canonical correlation analysis with embeddings to enhance the ensemble classifier's performance for scenarios lacking labeled data. The effectiveness and applicability of the approach are demonstrated through comprehensive evaluations on multiple ASAG datasets, highlighting its generalizability. Hassan et al. (2018) conducted a study examining various word embeddings, including Word2Vec, GloVe, and fasttext. They also explored paragraph embeddings like Doc2Vec, InferSent, and Skip-Thought.

To create paragraph embeddings for both student and model responses, they used the sum of word vectors from these word-embedding models. Among all the models studied, the Doc2Vec paragraph-embedding model performed the best, achieving a Pearson correlation value of 0.569. Gomaa and Fahmy (2020) employed an unsupervised learning method to create skip-thought vectors, transforming both student and reference responses into deep embedding vectors. This technique enabled the comparison of the similarity between the responses. Lubis et al. (2021) developed an automated grading method for short answers employing word embedding techniques and syntactic analysis to assess the precision of learners' responses by measuring semantic similarity.

Sequence-based models can capture the semantic characteristics of the text by considering word sequences of various lengths and the relationships between words in sentences that span larger distances. Recurrent neural networks (RNNs), specifically those utilizing long short-term memory (LSTM) networks, are employed to represent the sequential characteristics of textual data. Saha et al. (2018) advanced the field by blending handcrafted features with sentence embedding features to enhance the accuracy of automated short answer grading. They trained a neural network grading classifier using these combined features, which included traditional NLP metrics as well as modern embedding techniques. The handcrafted features provided domain-specific insights, while the sentence embeddings captured the broader semantic context of the student answers. Additionally, they trained an end-to-end deep neural network to learn embeddings, enabling the model to better generalize across different types of questions and answer styles.

Training embedding models requires vast amounts of unlabeled data, posing a significant challenge due to the resource-intensive nature of data collection and annotation. Moreover, incorporating question embeddings into sentence-level features

to capture the gap information between questions and answers is a novel approach that enhances the model's ability to understand the context of each question-answer pair. This method proves particularly beneficial when test questions have been encountered during training, as it allows the model to generalize better to unseen questions. However, learning question embeddings also present challenges, as they require a substantial amount of labeled data. This requirement can be a notable limitation for short answer grading tasks, especially in scenarios where obtaining labeled data is difficult or impractical. To address the same challenge, Kumar et al. (2017) proposed a solution involving a Siamese bidirectional LSTM applied to both student and reference answers. Their model leverages the Earth-mover distance similarity across all hidden states from both LSTMs, with a final regression layer used to produce grades. This approach offers a promising solution to the problem of incorporating question embeddings into the grading process.

Attention-based and transformer models are recent models that capture structural and semantic data using attention (Vaswani et al., 2017). Recent ASAG approaches have turned to these models. Attention-based mechanisms can detect long-term dependencies between words in a sentence. These models have demonstrated state-of-the-art performance across various natural language processing (NLP) tasks, including text summarization, sentiment analysis, and question answering. In the context of ASAG, Schneider et al. (2023) utilized a comprehensive dataset of 10 million question-answer pairs spanning multiple domains, categorized as either correct or incorrect. By fine-tuning the BERT transformer model (Devlin et al., 2019), they achieved an accuracy rate of 86% for automatic scoring. This study also emphasized trust and ethics, incorporating human oversight in the automatic scoring process. Their analysis of a human-graded sample of challenging questions revealed significant variability in BERT's performance on smaller datasets. This suggests that using a point grading system rather than the binary true/false classification could help identify substantial deviations from predicted accuracy, an aspect not addressed in their current dataset.

(Gaddipati et al., 2020) explored the impact of pre-trained embeddings from several transfer learning models, including ELMo (Embeddings from Language Models) (Peters et al., 2018), BERT (Devlin et al., 2019), GPT (Radford et al., 2018), and GPT-2 (Radford et al., 2019), on the ASAG task. Using Mohler's dataset (Mohler et al., 2011), they employed a regression model with cosine similarity features to

compare embeddings of reference and student answers. Their findings indicated that the ELMo model outperformed other transfer learning methods (BERT, GPT, and GPT-2) in the ASAG domain-specific. However, all pre-trained models underperformed compared to the Sultan system (Sultan et al., 2016) on the Mohler dataset. This discrepancy highlights the challenge: the pre-trained data for models like ELMo, BERT, GPT, and GPT-2 are extensive, yet the specific and small ASAG dataset domain limits their effectiveness. Schneider et al. (2023) concluded that while models like BERT excel on diverse datasets, they struggle with domain-specific datasets due to insufficient training data, which is a common issue in ASAG tasks. Agarwal et al., (2022) addressed this by using short text matching with a multi-relational graph transformer representation. By incorporating relation-enriched structural information, they aimed to capture more domain-specific properties. Their method achieved state-of-the-art performance on the Mohler dataset by embedding the semantic representation of relationships into token preparation. However, despite their power, transformer-based language models like BERT are computationally intensive (X. Huang et al., 2022), posing significant challenges for practitioners, especially in e-learning environments. The substantial memory and processing capacity required loading and storing model parameters could be prohibitive for extensive use in such contexts.

An alternative strategy to training various questions within the same model is to train a specific model for each question. This approach, demonstrated by Madnani et al. (2013), Luo et al. (2016), Zhang et al. (2019), and Kumar et al. (2019), requires a large number of labeled student answers for effective training of each question-specific grading model. However, collecting sufficient manually labeled student answers can be challenging, particularly in learning management system (LMS) environments where efforts to increase the volume of labeled student answers are often lacking. Despite these challenges, deep learning techniques using this approach have achieved good results in tasks such as textual similarity and textual entailment, as demonstrated by Zhang et al. (2019) and Kumar et al. (2019). However, the necessity of numerous graded student answers for each target question remains a significant drawback, limiting the applicability of such models to scenarios with access to a large volume of student answers.

Tulu et al. (2021) introduced an innovative deep neural network architecture designed to enhance text similarity analysis by integrating Manhattan LSTMs with

SemSpace vectors. The SemSpace vectors, sourced from the English lexical WordNet database, were utilized to accurately predict student grades from two ASAG datasets. The system has two identical LSTM networks. Each student and model answer pair are fed into the system as sense vectors, and the Manhattan distance is used to figure out how similar the vectors are at the output. Demonstrating state-of-the-art ASAG performance on the Mohler dataset underscored the system's effectiveness. Using individual training files for each question, the system achieved a Pearson's correlation of 0.95. This exceptionally high Pearson's correlation score of 0.95 signifies that the system's predictions closely align with human grading standards, underscoring the system's reliability and accuracy. However, when all questions, student answers, and reference answers are included in a single training file, the Pearson correlation drops significantly to 0.15. It takes a lot longer for the ASAG system to learn new words when there are a lot of words that are not in its vocabulary. This happens when the dataset has more words and the context training set is larger. This severely restricts the use of LSTM algorithms in this scenario.

2.1.4 Data augmentation for multiple reference answers in ASAG systems

The exploration of data augmentation strategies through the integration of multiple reference answers to enhance short-answer scoring represents a domain ripe for further investigation. Although traditional methods laid the groundwork for this field, as demonstrated by Leacock and Chodorow (2003), Sukkarieh and Blackmore (2009), Noorbehbahani and Kardan (2011), and Kumaran and Sankar (2015), recent studies have explored automated approaches.

Mohler and Mihalcea (2009) pioneered the adoption of the pseudo-relevance feedback method, originally derived from information retrieval techniques introduced by (Rocchio, 1971). Their method involved augmenting the reference answer with student responses that demonstrated the highest similarity scores according to a predefined metric. While this approach exhibited promising results in enhancing system quality, its implementation necessitated an intricate and time-consuming training process. According to Mohler and Mihalcea (2009), integrating top student responses with the teacher's answer helps to broaden the vocabulary of the teacher's answer, and they find this approach to be effective.

On a different front, Omran and Ab Aziz (2013) proposed the Alternative Sentence Generator Method, which relies on a comprehensive database of synonyms.

By replacing each word in the reference answer with its synonymous counterpart, this method aimed to encompass a broader spectrum of potential answers, thereby enriching the dataset and enhancing the scoring process for short-answer assessments. While producing a large number of sentences, this approach is not appropriate for low-resource languages like Arabic, where knowledge-based linguistic resources (dictionaries, thesaurus, lexicons, etc.) are either nonexistent or just partially available. It also does not make it easier to formulate the reference response in new ways.

In order to solve this, Dzikovska et al. (2014) used generalizable lexical representations and rules to include semantic information in the reference response. Similarly, by summarizing the highest-scoring student answers, (Ramachandran & Foltz, 2015) generated alternative reference texts using a graph-based cohesion approach. In place of conventional reference answers, these summarized responses showed how summarizing approaches may yield more representative and helpful phrases for grading.

In an effort to do away with the requirement for a training procedure, Pribadi et al., (2018) went over Mohler and Mihalcea's (2009) use of Rocchio's approach to provide alternative reference responses. In order to exclude non-contributive terms from the reference response, they applied the Maximum Marginal Relevance (MMR) technique, which was put out by (Carbonell & Goldstein, 1998), and they devised an unsupervised similarity method called Geometric Average Normalized-Longest Common Subsequence, or GAN-LCS. Their findings, however, did not demonstrate a noteworthy increase in correlation when compared to the gold standard on the English Mohler dataset (Mohler et al., 2011) which already produces results that are quite accurate.

Additionally, we attract interest from related fields that focus on semantic textual similarity and paraphrasing concepts (recognition and generation). Research in automated short answer grading has also been approached as a paraphrase recognition challenge.

Koleva et al. (2014) developed a system designed to grade reading comprehension tests for German foreign language learners leveraging paraphrase recognition techniques. The core of their approach involved aligning words from paraphrase fragments that were extracted from parallel corpora. This alignment process was crucial for identifying semantic entailment relationships between student responses and the reference answers provided by the instructors. The system-extracted

features from these aligned paraphrase fragments and fed them into a linear regression model. This model evaluated the strength of the semantic connections between the fragments, effectively distinguishing between correct and incorrect answers. False answers were characterized by the absence of paraphrase fragments, whereas correct answers demonstrated strong semantic links through the presence of these fragments. The system was tested using the German CREG corpus (Ott et al., 2012), a benchmark dataset for evaluating comprehension tests. The system achieved an accuracy rate of 86.8%, underscoring its effectiveness in correctly identifying semantically accurate answers. However, it is important to note that the system did not use a point-based grading approach. Instead, it employed a binary evaluation method, classifying answers as either true or false. While this method proved useful for identifying correct responses, it fell short of capturing the more nuanced differences that might exist between the scores assigned by human graders and those generated by the automated system. This limitation suggests that while the system is highly accurate in determining whether an answer is correct, it does not fully replicate the depth and subtlety of human grading, which often considers partial correctness and varying degrees of answer quality. This method represents a notable advancement in applying paraphrase detection for educational assessment. However, it also highlights the need for additional improvements to incorporate sophisticated scoring systems that captured better the nuances of human judgment. At present, there is no research linking automated short answer grading with the task of paraphrase generation.

2.1.5 Arabic Automatic Short Answer Grading and Challenges

2.1.5.1 Arabic ASAG Approaches

In terms of Arabic, some of the ASAG approaches that have been studied include a hybrid method that combines various methods for text similarity measurements; string-based, corpus-based, and knowledge-based text similarity (Gomaa & Fahmy, 2012). In this study (Gomaa & Fahmy, 2014a), the authors explored text similarity methods for automatic scoring of short answers in Arabic. They compared string-based and corpus-based measures, assessed their combined impact, and managed student responses. The research also aimed to provide immediate feedback and introduced a benchmark dataset for Arabic, known as the Cairo dataset.

Abbas and Al-qazaz (2015) proposed a vector space model approach with latent semantic indexing LSA. A combination of LSA and POS tagging for syntactic

analysis is proposed by Mezher and Omar (2016). The study of Magooda et al. (2016) investigated various methods for constructing vectorized space representations for Arabic and evaluated these models through both intrinsic and extrinsic assessments. The extrinsic evaluation measured the effectiveness of the models by examining their performance in Short Answer Grading tasks using the Cairo dataset. The study released a collection of Arabic standard word embeddings created using the Skip-gram, CBOW, and GloVe models for public use.

Using stemming strategies and Levenshtein edit operations, Al-Shalabi (2016) proposed an automated system for Arabic essay scoring in online tests. Text similarity measures that use both corpus-based and string-based methods are proposed by Shehab et al. (2018).

More recently, Abdeljaber (2021) examine the application of the longest common subsequence (LCS) string-based similarity for measuring the similarity of short answers to Arabic essay questions. The LCS algorithm is enhanced through weight-based measuring techniques using Arabic WordNet. The experiments on a dataset of 330 student-collected responses yielded positive outcomes; however, a comprehensive comparison with other studies in the field is necessary to establish the significance of the findings.

The proposed system in (Salam et al., 2022) employs a hybrid methodology to enhance the performance of Long Short Term Memory (LSTM) networks using the Grey Wolf Optimizer (GWO) algorithm in an Arabic ASAG. The simulation results indicate that the GWO-enhanced LSTM model surpasses traditional LSTM models in performance, though it necessitates a longer training period. Nael et al. (2022) fine-tuned the BERT and ELECTRA pre-trained models using a translated version of the ASAP Short Answer Scoring dataset, resulting in a QWK grading score of 0.78. The dataset was translated from English to Arabic via the Google Translate API.

2.1.5.2 Arabic NLP Challenges in the ASAG Field

The majority of ASAG study has focused on the English language, whereas there have been comparatively less studies conducted on Arabic. More than 400 million people speak Arabic, which is the official language of 22 nations. It is recognized as the fourth most frequently used language on the Internet (Guellil et al., 2021). Arabic is a highly prevalent and complex natural language. It is defined by a significant amount of extensive morphology, intricate morpho-syntactic agreement rules, and a considerable

number of irregular forms (Mustafa et al., 2017). Moreover, Arabic is considered one of the languages that has fewer resources available for its study and development (Mahmoud El-Haj et al., 2015).

In recent years, we have witnessed great research interest in the field of Arabic language applications (Guellil et al., 2021). However, automated evaluation in Arabic still represents a major challenge (Ditters, 2013; Ouahrani and Bennouar 2019). Applying natural language processing tasks in general and determining answer scores is a major challenge in the Arabic language. The Arabic language has many features, which are a challenge to the automatic answer estimates in Arabic. Some research on Arabic uses English translation to take advantage of the availability of resources and knowledge in English (Gomaa & Fahmy, 2014b; Nael et al. 2022).

Below, we review the main challenges that must be faced when designing an automated evaluation system for short text answers formulated in Arabic (Ouahrani & Bennouar, 2019):

The first challenge is that there are three types of Arabic: classical (standard), Modern, and Colloquial.

Standard Arabic, used in the Qur'an, is more complex in grammar and vocabulary than Modern Arabic. It contains a large number of diacritics that facilitate the pronunciation of words and reveal them in their grammatical cases. *Modern Arabic*, where all diacritics have been deleted to make the reading and writing process easier and faster, is considered the official language of the Arab countries and is used in daily life, education, and the media. Usually, Arabic-based computational research uses modern Arabic (Ditters, 2013; Al-Thubaity, 2015; Mustafa et al., 2017). In colloquial language, grammar and vocabulary are less developed compared to modern Arabic. However, most people use it in their daily spoken conversations and in written messages informally for its simplicity (Guellil et al., 2021).

Arabic speakers often make grammatical errors when using Standard Arabic and tend to mix it with their local colloquial dialects. Moreover, the dialects vary significantly across different Arab countries, which complicates the task of Arabic answer grading in recognizing user language. For instance, a person from Palestine might use different vocabulary and grammatical structures compared to someone from Algeria when responding to the same question, making it challenging automated systems to assess accurately the correctness of their answers. The use of slang and informal language in colloquial Arabic further complicates the grading process.

Additionally, variations in pronunciation and regional accents contribute to the difficulty of evaluating Arabic language responses accurately. Consequently, creating a reliable and effective automated grading system for Arabic language learners is a significant challenge that must address these linguistic complexities.

The second challenge is related to morphology, capitalization, diacritics and stemming.

First, the Arabic language's complexity arises from its morphological variations, where the shape of letters changes based on their position within a word. Additionally, words can include various combinations of prefixes, roots, and suffixes, making the morphological process highly intricate (Mustafa et al., 2017). For instance, the word "كتب" can mean different things depending on the context and letter placement, such as "he wrote" or "books." These characteristics add to the difficulty of determining the grammatical status of words in Arabic sentences.

Second, the Arabic language does not utilize capitalization for proper nouns like country names and personal names, unlike Latin languages where these names begin with capital letters (Abouenour et al., 2013) Consequently, computer-assisted Arabic language assessment programs may struggle to identify named entities, as they are treated like any other words. This makes it more challenging to recognize these nouns in responses. Researchers still developing specialized algorithms and tools to enhance named-entity recognition in Arabic texts, as this remains a complex task in the Arabic language (Guellil et al., 2021; Qu et al., 2024).

Additionally, most contemporary written Arabic texts are devoid of diacritics, which increases the language's ambiguity since a single short vowel can be optional. Therefore, the form of a word may have different meanings depending on the context. This creates ambiguity while evaluating students' responses and affects the accuracy of calculating grades.

Finally, The Arabic language is part of the Semitic language family, which also includes Aramaic. It features a lexicon primarily built from trilateral and quadrilateral roots, uses a right-to-left writing system, and has an alphabet consisting mainly of consonants. A stemmer is an automatic process that maps different morphological variants of words to a single representative form called a stem (Lovins, 1963). Stemming techniques typically involve a list of affixes (prefixes and suffixes) and a set of predefined de-suffixation rules to determine the stem of a word. For the Arabic language, automating the identification of a word's root or stem is particularly

challenging. The root often has a very abstract meaning, making it less suitable for natural language processing (NLP). Additionally, Arabic words can be borrowed from various contexts, adding to the ambiguity and complicating mechanical interpretation. The two most effective approaches to Arabic stemming are root extraction and light stemming. Root stemming involves removing known prefixes and suffixes to isolate the root of a word and identifying the pattern that corresponds to the remaining word. Light stemming, a simpler process, stops at removing prefixes and suffixes without attempting to find the root word.

The third challenge is related to the lack of linguistic resources.

Overall, the availability of Arabic language resources for research purposes is limited. Compared to English, Arabic is under-resourced, lacking sufficient data and tools, which hinders natural language processing (NLP) research in the language (Mahmoud El-Haj et al., 2015; Guellil et al., 2021; Qu et al., 2024). In the realm of ASAG, two particularly challenging issues related to resources are *the scarcity of Arabic datasets* for training and evaluation, and *the limitations of the Arabic WordNet* (lexical resources in general). The latter, a lexical database, significantly affects the similarity calculations between reference answers and student responses.

Arabic WordNet (AWN) limitations. Knowledge-based similarity (Lin, 1998; Leacock and Chodorow 1998; Wu & Palmer, 1994) calculates the similarity between words by utilizing information obtained from semantic networks. The English WordNet (EWN) is the most often utilized semantic network. WordNet [15] is an English language lexical database that categorizes English words with similar meanings into groups of synonyms, each accompanied by a concise and broad description. It also elucidates several semantic links among groups of synonyms, such as the relationship of antonyms, the relationship of wholes and parts, and the relationship of inclusiveness. Initially, WordNet was created with the aim of assisting scientific research. However, over time, its purpose evolved to establish WordNet as a crucial tool in natural language processing. This was achieved by providing a wide range of features and functionalities.

- An amalgamation of a dictionary and thesaurus that is designed to be more user-friendly and easy to navigate,
- Support for automated text analysis, and
- Support for of artificial intelligence applications.

WordNet is utilized in several domains, including data retrieval, semantic similarity, and word meaning disambiguation. Based on the success of WordNet in English, WordNet International, a non-profit public organization, is currently implementing multiple projects to develop lexical databases for low-resourced languages such as Arabic, Persian, Albanian, African, and Indian languages.

In ASAG field, WordNet is instrumental in enhancing the accuracy of the system's scoring results. It serves as a resource for supplying synonyms of words found in the model answer. During the comparison process, if any of these synonyms match a word in the student's answer, it indicates that the student's word is synonymous with the model answer word. Consequently, the student's word is considered (or replaced) without altering its meaning. Arabic WordNet (AWN) has been created specifically for utilization in Arabic natural language processing (NLP) applications such as question answering, query expansion, and text disambiguation. Several iterations of AWN have been published; yet, its representation of the Arabic language still falls behind other comparable WordNets, hence restricting its efficacy (Abouenour et al., 2013; Regragui et al., 2016). The Arabic WordNet has over 18,925 Arabic word meanings that are divided into around 9,698 synsets. However, it is significantly smaller in comparison to the English WordNet. Table 1, obtained from (Abouenour et al., 2013), presents a juxtaposition of the contents of Arabic WordNet and English WordNet. Additionally, it calculates the proportion of word lemmas in each WordNet compared to the overall number of words in comprehensive lexical resources for both languages.

Table 1 demonstrates that the AWN, when compared to the English WordNet, only encompasses 9.7% of the anticipated number of word lemmas in the Arabic lexicon examined, whereas the English WordNet covers 67.5%. The Arabic lemmas account for approximately 7.5% of the lemmas found in the English WordNet. In addition, the number of synsets in AWN accounts for just 8.2% of the synsets in the English WordNet. The connection between word lemmas and synsets is made by means of word-sense pairings, which make up 9.1% of those identified in the English WordNet. Furthermore, AWN synsets employ just three semantic relations (hyponymy, synonymy, and equivalence), but the English WordNet has seven relations, which encompass antonymy and meronymy. AWN has several drawbacks, but it also has advantages such as its adherence to the WordNet structure, its connection with other ontologies, and its consideration of Arabic-specific traits.

However, the restricted coverage significantly limits its applicability to a few projects and poses challenges in the Automated Short Answer Grading domain.

Table 1 Comparison of AWN and EWN statistics (extracted from (Abouenour et al., 2013)

Figures	Arabic	English
Synsets	9,698	117,659
Word-Senses	18,925	206,941
Word Lemmas (WL)	11,634	155,287
Language Lemmas (LL)	119,693	230,000
Ratio lemmas (WL/LL)	9.7%	67.5%
Ratio Word-lemmas (WN/English WN)	7.5%	100.0%
Ratio Synsets (WN/English WN)	8.2%	100.0%
Ratio Word-senses (WN/English WN)	9.1%	100.0%

Scarcity of Arabic datasets for ASAG training and evaluation. The scarcity of Arabic datasets for Automatic Short Answer Grading (ASAG) systems significantly affects their development and effectiveness. This lack of resources leads to challenges such as data collection, limited model training, poor generalization, difficulty in benchmarking, bias issues, and limitations in language-specific features. Data collection and annotation involve collecting large amounts of student responses and meticulously annotating them with accurate grades. This process is time-consuming and resource-intensive, requiring manual effort from educators and linguistic experts. Limited training opportunities for supervised learning models, which require extensive training on large datasets, also hinder the development process. The scarcity of evaluation datasets limits comprehensive testing and validation. Standardized datasets provide a common reference point for evaluating and comparing ASAG systems, but in Arabic, each research team may use different data for training and evaluation, making it difficult to compare results across studies. Evaluation metrics, such as accuracy, precision, recall, and F1 score, are difficult to use due to the absence of standardized Arabic datasets. Reproducing results is difficult due to the scarcity of standardized datasets, as researchers cannot replicate experiments to verify findings or build upon existing work.

Limited comprehensive evaluations, which require datasets covering a wide range of topics, question types, and answer styles, can lead to overfitting and limited generalizability. Some research on Arabic uses English translation to take advantage of the availability of resources and knowledge in English (Gomaa and Fahmy 2014b; Nael et al. 2022).

While using English translations in Arabic research provides access to extensive resources and facilitates global collaboration, it also comes with significant challenges. Employing English translations in research can result in several adverse effects, such as the loss of linguistic nuances, the introduction of bias, reliance on translation quality, insufficient representation of Arabic-specific concerns, and potential issues in resource allocation. These problems can lead to misinterpretations, oversimplifications of the original content, ultimately compromising the accuracy and authenticity of the research. Consequently, it is essential to create native Arabic resources and tools.

In conclusion, while these challenges are particularly relevant for Arabic, they are also applicable to numerous other under-resourced languages in the ASAG field. This underscores the broader need to develop native resources and tools tailored to these languages.

2.1.6 E-assessment of short answers in Learning Management Systems

Several learning management systems (LMSs) may be used to create, oversee, and share digital materials for both in-person and online instruction. An LMS enables the incorporation of conventional teaching methods with digital learning materials, while also offering pupils customized e-learning possibilities (Boitshwarelo et al., 2017).

The discipline of e-learning has seen enormous expansion, especially since 2020, because of the COVID-19 pandemic. The COVID-19 pandemic has significantly restricted the feasibility of in-person teaching for many educational institutions globally (Ghouali & Cecilia, 2021; Dias et al., 2020; Raza et al., 2021). As a result, educational institutions have had to adapt by shifting to online teaching, implementing new assessment methods, adjusting research approaches, and altering scholarly discourse practices (Byrnes et al., 2021).

In recent years, the importance of Learning Management Systems (LMSs) has grown significantly, thanks to the improved availability of high-speed internet and developments in online education technology. Many educational institutions have successfully integrated Learning Management Systems (LMSs) into their curriculum, reaping benefits such as improved student engagement and streamlined administrative processes. These institutions are actively exploring the efficacy of various LMS types, including personalized learning platforms and collaborative tools. The platforms that are often used are Edmodo, Moodle, MOOCs, and Google Classroom (Setiadi et al.,

2021). Moodle[7], an acronym for Modular Object-Oriented Dynamic Learning Environment, functions as a Virtual Learning Environment (VLE) that enables online communication between teachers and students in an e-learning setting. In addition, Moodle serves as a Learning Content Management System (LCMS), enabling educators to develop and oversee customized online courses, exchange documents, engage in real-time interactions, award grades, give assignments, administer exams, and track student progress over time. Moodle, unlike commercial VLE systems, is an Open Source Software (OSS), allowing free access, customization, and community-driven development by users worldwide. The system is highly versatile and user-friendly, enabling students to access conveniently it from any location, at any time.

Moodle is a globally recognized and widely used platform. In a recent systematic review conducted by (Altinpulluk & Kesim, 2021), Moodle was identified as the leading open-source LMS in the field. The platform is extensively used by educational institutions, organizations, and individuals for the purpose of online learning and training (Wu, 2008; Kumar & Sharma, 2016; Florjancic, 2016; Boitshwarelo et al., 2017; Gamage et al., 2022). It provides a diverse range of interactive courses, spanning various subjects and available in multiple languages. The platform has been deployed in 241 countries globally, with an estimated user base exceeding 417 million users. Additionally, there have been over 2.3 billion course enrollments and about 9,123,664,464 exam questions (Moodle-Stats, 2024)[8] .

Moreover, recent research highlights the positive impact of instructors acquiring knowledge and utilizing LMSs like Moodle on student performance in educational assessment and evaluation (Oguguo et al., 2021). Out of the 155 papers analyzed in a recent survey on the use of Moodle for teaching and learning (Gamage et al., 2022), only 33% of the research examined assessment, which included both summative and formative evaluation conducted on the platform. Most of these publications concentrated on assessing students' performance at the course conclusion, mainly using multiple-choice questions. The element of "luck" present in multiple-choice questions is commonly considered equitable. The quiz and lesson modules in Moodle have a diverse range of question types. The available question types include calculating, multiple choice, true/false, short answer, matching, and essay questions,

[7] https://moodle.org/
[8] https://stats.moodle.org/

etc. Various question behaviors can be utilized when generating a quiz with these questions. (Question types and behaviors are detailed in Appendix A).

The Short answer question type in Moodle uses regular expression (Moodle, 2011), allowing users to input correct responses as regular expressions. A short answer question requires the student to input a single word or phrase as a response to the question. Responses may or may not differentiate between uppercase and lowercase letters. The response may consist of either a single word or a phrase, but it must precisely correspond to one of the permissible answers provided. It is advisable to minimize the necessary answer's length to prevent overlooking a valid response that may be expressed in a different manner. Students are subjected to several limitations while formulating their responses.

There are two challenges facing tutors. Initially, it pertains to the manual creation of grammatical templates that indicate the reference answer. The second is about how well students follow the guidelines in the template. Any extra space, misspelled words, etc., result in penalties for the students. These factors make it uncommon for teachers to employ this type of question.

2.1.7 ASAG Evaluation

In the assessment of Automated Short Answer Grading, it is crucial to consider not only the grading algorithms but also the datasets and evaluation metrics used to determine its effectiveness. We present the most used datasets and conduct a comparative analysis of the evaluation metrics.

2.1.7.1 Datasets

There are several commonly used standard datasets for most NLP tasks. These datasets are frequently used to assess how well novel approaches for these tasks perform. A primary challenge to ASAG's aim is the need for more diverse and extensive datasets that cover a wide range of student writing scenarios. Several publicly available datasets have been used to evaluate the effectiveness of different ASAG systems. These datasets show notable differences in subjects covered, dataset sizes, and the grading criteria employed. Competitions have been used as platforms to introduce specific datasets, like the ASAP and SemEval-2013 datasets, fostering innovation and progress in ASAG systems. These contests have facilitated the increase of interest and engagement in the advancement of ASAG systems. Nevertheless, there is still a pressing need for more diverse and extensive datasets that accurately capture the

intricate nuances involved in evaluating student writing. The majority of the datasets are in English. Additionally, the other languages on the list are Chinese (Wang et al., 2008), Arabic (Gomaa and Fahmy, 2014b), Japanese (Takano & Ichikawa, 2022), Hindi (Roy et al., 2015; Agarwal et al., 2020), and German (Ott et al., 2012). We present in the following the dataset most used in the ASAG literature evaluation.

Mohler et al. (2011)[9] (Mohler et al., 2011)**.** It is commonly employed to evaluate performance on English ASAGs. The dataset is obtained from initial computer science projects in a Data Structures course at the University of North Texas. A group of undergraduate students from the Computer Science program provided responses. The authors employed a fusion of graph-based alignment and lexical similarity metrics to evaluate brief replies. The dataset comprises 81 questions, accompanied by a cumulative count of 2273 responses. Two human judges evaluated the dataset based on criteria including relevance, coherence, and technical accuracy, using a rating scale from 0 to 5. The judges reached a consensus (assigning the same rating) in 57.7% of the cases. This dataset is an expanded version of the one used by Mohler & Mihalcea's (2009)'s study, demonstrating a 64.43% agreement rate among annotators for each question assessment.

ASAP-SAS Dataset (2012)[10]**.** The Hewlett Foundation introduced the Automated Student Assessment Prize: Short Answer Scoring Corpus on Kaggle. This dataset features responses from students in grades 8 through 10, each comprising fewer than 50 words. It includes 10 different prompts, each covering a distinct topic and corresponding to separate questions, with 17,204 responses. These responses are evaluated using two distinct scoring scales, one ranging from 0 to 2, and the other ranging from 0 to 3. Additionally, the dataset provides a marking rubric for each prompt.

The SemEval 2013 Dataset (Dzikovska et al., 2013)[11]. Introduced in 2013 by the SemEval workshop, this dataset includes two subsets: BEETLE and SCIENTSBANK. The BEETLE subset, derived from transcripts of the BEETLE II tutorial dialogue system by Dzikovska et al., (2010), presents 56 questions related to basic electricity and electronics. The responses, gathered from about 3,000 students, are typically one

[9] https://web.eecs.umich.edu/~mihalcea/downloads.html
[10] https://www.kaggle.com/competitions/asap-sas/data
[11] https://www.kaggle.com/datasets/azzouza2018/semevaldatadets

or two sentences long. The SCIENTSBANK subset originates from a collection of student responses to assessment questions compiled by Nielsen et al., (2008). It contains approximately 10,000 answers addressing 197 questions spanning 15 diverse science domains. The responses are classified into three labeling schemes: 2-way (correct and incorrect), 3-way (correct, contradictory, and incorrect), and 5-way (correct, partially correct or incomplete, contradictory, irrelevant, and not in domain).

Cairo University Arabic Dataset (Gomaa & Fahmy, 2014b)**.** The dataset encompasses questions derived from a chapter of the official Egyptian Environmental Science curriculum. It includes 61 questions, each accompanied by 10 responses, amounting to 610 answers along with their English translations. Each student response is evaluated on a scale from 0 to 5 by two expert annotators. A Pearson correlation coefficient of 0.86 and a Root Mean Square Error (RMSE) of 0.69 demonstrate the agreement between the annotators.

2.1.7.2 Evaluation Metrics

Automatic Short Answer Grading (ASAG) models are designed to automatically assess student responses and assign scores that ideally align with those of human graders. Depending on whether an ASAG system is developed as a classification or regression model, different assessment metrics are used. This section provides a comprehensive review of popular metrics used in the evaluation of ASAG models, examining their applicability, benefits, and limitations.

Pearson's r correlation. Pearson's correlation is a metric used in the ASAG context to measure the relationship between instructor marks and model predictions. It ranges from -1 to 1, with values indicating a perfect positive linear correlation, a negative linear correlation, or a perfect negative linear correlation. In Automatic Short Answer Grading (ASAG), a high positive r value indicates a strong positive relationship between the model's predicted scores and human-assigned scores, while a low r value suggests minimal to no linear correlation, indicating poor alignment. A negative r value indicates an inverse relationship, which is typically undesirable in this setting.

Root Mean Square Error (RMSE). It is a commonly used metric to measure the differences between values predicted by a model and the actual values observed. It is particularly useful for regression tasks. RMSE is calculated as the square root of the average of the squared differences between predicted and observed values. Lower

RMSE values indicate a closer match between predicted and actual values, signifying better model performance. Higher RMSE values indicate poor model performance due to a significant difference between the predicted and actual values. In the context of Automatic Short Answer Grading (ASAG), low RMSE indicates that the model's predicted scores closely match the human-assigned scores, demonstrating high accuracy and reliability in grading. A high RMSE indicates significant discrepancies between the model's predicted scores and the human-assigned scores, suggesting poor performance and potential inaccuracies in grading.

F1 score. The F1 score measures a classification model's balance between precision and recall, ensuring accurate assessment of performance.

It is calculated as 2× (Precision×Recall)/(Precision+Recall).

True Positives (TP) indicate accurate predictions, false Positives (FP) indicate incorrect predictions, and false Negatives (FN) indicate failures. A score of 1 indicates perfect precision and recall, while a score of 0 indicates poor performance. Balancing false positives and false negatives is crucial for accurate assessment. Automatic Short Answer Grading (ASAG) uses a F1 score to assess the trade-off between precision and recall in classification models. A high F1 score ensures accurate prediction of correct scores and high recall, while a low F1 score may lead to incorrect classifications or missing important details.

Quadratic Weighted Kappa (QWK). It is a statistical measure that assesses the agreement between two sets of categorical ratings, with a focus on ordinal contexts such as educational assessments. QWK considers chance agreement and assigns higher weights to larger discrepancies between the ratings. In interpreting QWK scores, a value of 1 signifies perfect agreement, 0 indicates agreement no better than chance, and negative values denote worse than random agreement, highlighting systematic disagreement. In the context of Automatic Short Answer Grading (ASAG), a high QWK score signifies strong agreement between the human-assigned grades and the automated system, whereas a low QWK score indicates significant disagreement.

It ensures that larger errors are penalized more heavily, thus providing a more nuanced and accurate assessment of the model's performance.

Discussion on metrics suitability. Table 2 provides a quick overview of each metric's applicability, benefits, and limitations according to whether the ASAG model is performing classification or regression tasks. Pearson Correlation Coefficient and

Root Mean Square Error (RMSE) are two commonly used metrics for regression model evaluation. Pearson's correlation coefficient measures the linear relationship between predicted and actual scores, offering benefits such as trend alignment and interpretability, which are valuable for assessing model performance.

RMSE measures the average magnitude of errors between predicted and actual scores, offering a clear indication of the model's accuracy in the context of ASAG systems. RMSE is a standard metric for regression tasks, enabling direct comparisons with other models and state-of-the-art methods, making it a valuable tool for assessing model performance. The squaring of differences in RMSE means that significant errors or outliers can have a disproportionately negative impact on the overall evaluation of the model.

Quadratic Weighted Kappa (QWK) measures the agreement between predicted and actual scores by taking into account the ordinal nature of the data and penalizing larger discrepancies more heavily, making it a suitable metric for tasks like short answer grading. It is highly suitable for tasks involving ordinal data, such as short answer grading. Despite its utility, QWK is less commonly adopted in ASAG evaluations, presenting challenges in benchmarking against state-of-the-art models and comparing results with existing research. In conclusion, Pearson Correlation Coefficient and RMSE are highly effective for regression models, providing insights into trend alignment and error magnitude.

The F1-Score is crucial for classification models, especially when balancing precision and recall is important. Quadratic Weighted Kappa (QWK), although less commonly used in current literature, provides a sophisticated measure of agreement for ordinal data, making it particularly relevant for grading tasks. Finally, the Pearson correlation coefficient and RMSE are appropriate for evaluating regression models, ensuring that findings are relevant and interpretable within the context of state-of-the-art ASAG models.

Table 2 Overview of ASAG metrics applicability, benefits and limitations.

Metric	Applicability	Benefits	Limitations
Pearson	- Classification - Regression	- Measures linear relationship between predicted and manual scores. - Useful for trend analysis.	- Ignores magnitude of errors. - Assumes linearity, which may not always hold true.
RMSE	- Regression	- Measures average magnitude of errors. - Standard metric for regression tasks.	- Does not indicate direction of errors (overestimation or underestimation).
F1-Score	- Classification	- Provides balance between precision and recall. - Relevant for classification tasks.	- Less suitable for continuous or ordinal scoring.
QWK	- Classification - Regression	- Handles ordinal data (grades) - Considering severity of discrepancies (in errors).	- Limited use in existing literature.

2.2 PARAPHRASE GENERATION OVERVIEW

2.2.1 Approaches

Our investigation focuses on the integration of paraphrase generation techniques into the Automated Short Answer Grading task. Paraphrase generation focuses on transforming a given text into equivalent or semantically similar text. Constructing high-quality paraphrases poses a significant challenge within the field of natural language processing. The primary aim is to automatically produce alternative answers to a given reference answer.

Paraphrase generation methods can be broadly categorized into two main approaches: controlled methods and deep learning methods. Paraphrasing can be done using manual rules and alignments with a thesaurus or by using statistical machine translation (SMT) methods. SMT treats paraphrasing as a form of machine translation that is limited to a single language (Wubben et al., 2010). These strategies use many methods, such as phrasal and lexical dictionaries (Huang et al., 2019) keyword-based approaches (Zeng et al., 2019), sentential exemplars (Chen et al., 2020), syntactic trees and tree encoders (Kumar et al., 2020), retriever-editors (Kazemnejad et al., 2020), syntactic transformations (Goyal & Durrett, 2020), and retrieval target syntax selection (Sun et al., 2021).

Modern paraphrase systems utilise existing parallel corpora to train sequence-to-sequence models, aiming to improve performance by drawing inspiration from the

success of deep learning networks. In order to enhance the performance of deep neural networks, Prakash et al. (2016) suggested the use of a technique called stacking, which involves combining many layers of Long Short-Term Memory (LSTM) with residual connections. Gupta et al. (2018) developed the VAE-SVG-eq (Variational Auto-Encoder for Sentence Variant Generation). It is a paraphrase generator that utilises the LSTM architecture (Hochreiter & Schmidhuber, 1997) in conjunction with the Variational Auto-Encoder (Kingma & Wellivng, 2013). The system has exhibited exceptional performance compared to existing approaches for generating paraphrases. Yang et al., (2020) created the Generative Adversarial Paraphrase Model (GAP), which is an end-to-end conditional generative architecture. This model is capable of generating paraphrases through adversarial training and is particularly noteworthy because it does not require any extra language signals to work.

Considerable advancements have been achieved in the domain of natural language processing (NLP) in recent years, namely via the development of comprehensive language models that utilise transformer topologies. These models have revolutionized various NLP tasks by leveraging massive amounts of data and sophisticated algorithms. With the availability of implementation codes and datasets on popular platforms like Hugging Face and GitHub, these models have become increasingly accessible to researchers and practitioners alike. GPT-2[12] (Radford et al., 2020) and BART[13] (Lewis et al., 2020) are two well-known pre-trained language models that serve as effective encoder-decoder frameworks for a variety of NLP applications. Furthermore, the "Text-to-Text Transformer" T5[14] (Raffel et al., 2020) has received attention for its capacity to transform a wide range of text-based language difficulties into a single text-to-text format. These models have not only demonstrated cutting-edge performance across a wide range of NLP tasks but have also enabled fast experimentation and advancement within the NLP community.

Despite the widespread adoption of these advanced models in NLP, there remains a notable gap in their application to the Arabic language, particularly in tasks such as paraphrase generation. While considerable progress has been made in English and other widely spoken languages, the unique characteristics of Arabic pose distinct

[12] https://huggingface.co/gpt2
[13] https://huggingface.co/docs/transformers/model_doc/bart
[14] https://github.com/google-research/text-to-text-transfer-transformer

challenges for NLP tasks. Consequently, research efforts in this area have been relatively limited, with few studies exploring paraphrase generation specifically for Arabic. One such approach, proposed by (Alkhatib & Shaalan, 2018), focuses on paraphrasing Arabic metaphors using neural machine translation techniques. Their method involves translating metaphors into a pivot language before converting them into English using a bilingual corpus. Similarly, Al-Raisi, Bourai, et al. (2018) developed a bidirectional LSTM neural network trained on their Arabic dataset (Al-Raisi, Lin, et al., 2018) for generating paraphrases. However, the evaluation of their system relied solely on cosine similarity, lacking results based on standard automatic metrics commonly used in paraphrase generation research.

Moving forward, there is a pressing need for further research and development in Arabic NLP, particularly in tasks like paraphrase generation. By leveraging the capabilities of advanced language models and adapting them to suit the complexities of Arabic language processing, researchers can pave the way for significant advancements in this field and address the unique challenges posed by Arabic text. As the authors concluded, "*the neural model has* learned *interesting linguistic constructs like phrases used for sentence opening but the output is still far from practical applicability*".

In our pursuit to bridge this gap and elevate the standard of paraphrase generation in Arabic, we embarked on an investigation centered on sequence-to-sequence deep learning models. By leveraging the capabilities of these advanced models, we aim to generate paraphrases that capture the nuances and intricacies of the Arabic language. The generated paraphrases hold significant potential beyond mere linguistic variation. They serve as invaluable resources for enhancing Automated Short Answer Grading (ASAG) systems. By providing alternative reference texts, they offer a diversified perspective for evaluating student responses, thereby contributing to the refinement and improvement of ASAG methodologies. Through our research endeavors, we aspire not only to advance the field of Arabic natural language processing but also to empower educators and institutions with innovative tools for enhancing the assessment process. By harnessing the power of deep learning and the rich linguistic heritage of the Arabic language, we endeavor to pave the way for more accurate, efficient, and culturally relevant ASAG systems.

2.2.2 Common Evaluation Metrics

To evaluate the quality of generated paraphrases during paraphrasing, we can employ established machine translation metrics such as BLEU (Papineni et al., 2002), GLEU (Napoles et al., 2015), and METEOR (Lavie & Agarwal, 2007). An observation frequently made in the machine translation field is that automatic measurements exhibit a strong correlation with human evaluations at the system level (Wubben et al., 2010 ; Shen et al., 2022). This indicates that the correlation analysis between automatic assessment measures and human ratings remains stable throughout the entire translation system.

BLEU is known for its rapid computation speed and language independence, which are crucial factors in efficiently evaluating translation quality. The evaluation of the paraphrase involves counting the n-grams it shares with a set of reference paraphrases, known as 'ground-truth paraphrases.' This assessment considers two critical aspects of translation: its faithfulness to the original text and its overall coherence and fluency.

GLEU (Google-BLEU) is a modified version of the BLEU metric. It is designed to precisely assess the accuracy of grammatical mistake correction in n-grams produced by comparing them to all the reference texts.

METEOR relies on evaluating the accuracy and recall of paraphrases through the analysis of unigrams, contributing to a comprehensive assessment approach. This significantly strengthens the alignment with human evaluations, highlighting METEOR's effectiveness in capturing human evaluation criteria. METEOR determines the similarity score between two texts by combining measures of unigram accuracy and unigram recall, as well as additional metrics like stemming and synonym matching, offering a comprehensive evaluation approach.

Unlike the BLEU measure, which considers precision, ROUGE-L takes into account both precision and recall in its score calculation. ROUGE-L(Lin, 2004) is based on the notion of the Longest Common Subsequence (LCS), which identifies the longest sequence of words that appear in the same order, even if they are not consecutive

2.3 SUMMARY

ASAG systems have transitioned from relying on designed hand-engineered text features to utilizing feature learning architectures powered by deep learning. Natural

language processing has been instrumental in driving this transition by enabling more sophisticated feature learning techniques through deep learning. To contribute to the domain, our study aims to fill in some gaps in the existing ASAG literature as we observe:

Scarcity of resources. ASAG systems encounter difficulties in accurately assessing primarily due to the scarcity of reliable data. Although deep learning has excelled in NLP tasks, it introduces challenges for ASAG, such as the high demand for computing resources and extensive annotated datasets and linguistic resources. The scarcity of datasets evaluation hampers comprehensive testing and validation. Standardized datasets are used as a universal reference point for evaluating and contrasting ASAG systems.

Research teams frequently employ diverse datasets for training and assessment, which makes it challenging to compare outcomes across different research in the field. On the other hand, the limited scope of lexical natural language processing resources such as Arabic WordNet presents substantial difficulties in the ASAG field.

Finally, the necessity of a huge number of graded, annotated answers to train each question is a significant problem that is time-consuming and costly to have. This problem may also have an unfavorable effect on the e-learning system's server performance. As a result, ASAG studies need to tackle basic issues like data scarcity and domains specificity.

Linguistic variations and diversity of student responses. The exploration of data augmentation for improving short response grading in ASAG systems has been limited. Manually generating alternative reference responses is a time-consuming process that demands substantial expertise and exertion (Marvaniya et al., 2018). Moreover, it is not feasible for assessing short answers on a big scale. Therefore, it is necessary to automate the process of generating alternate reference responses.

By including numerous alternative reference solutions for a given question, it is possible to account for the variations in student answers and improve the grading accuracy.

Integration into educational sittings and scalability. In practice, very few ASAG tools are implemented and are made available directly on the e-learning system even though the grader models are more sophisticated. The emphasis in the research field is much more on the score accuracy rather than on the practical integration of the grading system in the e-learning environment. Solutions developed for ASAGs do not

envisage a harmonious integration with other types of questions and other assessment activities. This may put the teacher at ease in developing various quizzes combining constructed and selected questions with the corresponding feedback in the same quiz. The effectiveness of implementing Integrated Short Answer Scoring in e-learning systems might rely more heavily on its application method than solely on the precision of the scoring itself.

To tackle these challenges, we developed ISAGe (Integrated Short Answer Grader for e-learning environments) trained on a realistic dataset collected and evaluated to effectively bridge these gaps. Chapter 3 provides a detailed description of the proposed system, highlighting its features and design.

Chapter 3: Research Design and DATA

In this chapter, we outline our research design, which includes the following components: (1) Proposed Approach, (2) Data Collection, (3) Proposed Features, (4) Proposed Scoring Model, (5) Paraphrase Generation for Alternative Reference Answers, and (6) Technical Design and Integration of the Solution into the Learning Management System.

3.1 PROPOSED APPROACH

Automating the assessment of large numbers of students requires a multi-faceted solution. To deal with these our goals, we develop ISAGe namely Integrated Short Answer Grader for e-learning environments. The research system design covers and combines approaches from computational distributional semantics, supervised learning, paraphrase generation, and LMS technology.

In this thesis, we need key concepts to facilitate fast computation and large-scale deployment. The main requirements for the ISAGe system focused on the formative and summative assessment scenarios into the LMS include:

- Accuracy, scalability, and easy to use.
- Flexible design to be integrated into existing e-learning environment by extending the LMS quiz system to the proposed ASAG.
- Standard-conform interfaces with the LMS environment.
- Incremental design for updating system models to new specifications.

The proposed approach is driven by a feature engineering strategy that integrates and enhances text similarity metrics, term weighting, answer length statistics, and difficulty features. Instead of necessitating separate training for each question and having access to hundreds of student responses for each one, we utilize features to train a supervised regression model that is question-general. This model learns from both the specific domain of the course and the broader general domain.

Our approach to feature engineering involves the integration of both specific domain knowledge and general domain knowledge as features, utilizing distributional semantics. Specifically, we employ semantic space distribution to capture domain-specific knowledge within the subject area, while leveraging Word Embeddings

trained on extensive general domain corpora to encompass broader, general domain knowledge. The aim is to simulate the behaviour of human teachers. Therefore, while the precise methods of human grading remain unclear in general, it is understood that teachers can acquire grading skills through their extensive domain-general knowledge. They can assimilate grading criteria by applying this knowledge to the specific subject matter at hand. Additionally, we enhanced our feature set by proposing two novel additions: the word vector alignment similarity and the knowledge gap between question and answer features. The word vector alignment similarity aims to improve semantic similarity by identifying the closest matching word in the reference answer for each word in the student's answer. The knowledge gap between the question and answer measures how well the student's answer aligns with what is expected in the question, as defined by the reference answer. Unlike rarely previous work such as that by (Saha et al., 2018), which uses deep trained sentence embeddings to capture this information gap, our approach is manually crafted. This ensures that the calculation of these features remains easy and fast.

In a second phase, we propose a deep learning model to generate automatically multiple alternate reference answers, thereby refining the accuracy of our model by considering different formulation of the reference answer.

Finally, a cloud-based LMS integration of the tool is proposed *to promote scaling* and *to favor an adaptable scoring model* to new assessment specifications or new features.

Training neural models necessitates the availability of data. However, public datasets specifically for short answer grading tasks are scarce. Ideally, we would train our model on a realistic dataset that includes diverse questions aligned with learning objectives and then provide a comprehensive evaluation of the approach. To address these dual challenges, we have conducted a case study to develop our own dataset for the Arabic Language, tailored for ASAG training and evaluation. By focusing our efforts on the Arabic language, we aim to address the challenges faced by under-resourced languages and offer our contribution to their development in the ASAG field.

Two factors influenced our design choices throughout the thesis: enhancing notation precision considering low resources and ensuring the system remains feasible and scalable in practice.

When implementing ASAG into LMS, it is essential to consider factors beyond just accuracy, such as server performance and practical efficiency. Due to its integration with online learning platforms, the scoring model must maintain low computational complexity to ensure optimal resources.

3.2 DATA COLLECTION (COURSE, PARTICIPANTS, AND DATA ANNOTATION)

The availability of high-quality training data, including labeled and domain-specific information, poses a significant challenge for Automated Short Answer Grading. For Arabic, we developed the AR-ASAG dataset (Ouahrani & Bennouar, 2020). This dataset is the first Arabic dataset available to the public[15] for use in ASAG training and evaluation. It is validated and registered under the International Standard Language Resource Number (ISLRN= 529-005-230-448-6)[16]. The dataset will be useful for other studies evaluating research related to automatic short answer grading and Arabic semantic similarity. Given that, data is fundamental we focused on collecting authentic data from students through a case study.

The Case study 1 was conducted with 170 students selected from the first-year master's degree program in the computer science department at the University of Blida 1. The students, all native Arabic speakers, came from three distinct areas of study (Software Engineering, Computer Systems & Networks, and Information Systems Security). They were enrolled in the "Cybercrimes" course, a mandatory component of their academic programs. Information on the course may be seen in Appendix D, and the course materials were accessible for students online[17]. The course was delivered in person during the first semester of the 2018–2019 academic year in Arabic by the author of the thesis, an experienced educator in the field, under standard teaching conditions. The evaluation of the course's instruction involved a meticulously designed final examination using short answer questions, custom-made to facilitate the collection of the dataset. The evaluation of the course's instruction included a meticulously designed final examination with short answer questions. This exam was custom-made to facilitate the collection of data for evaluating the students' comprehension and learning outcomes.

[15] https://data.mendeley.com/datasets/dj95jh332j/1
[16] https://www.islrn.org/resources/request/3582/
[17] https://elearning.univ-blida.dz/course/view.php?id=484

Creating a high-quality dataset for Automated Short Answer Grading involves multiple steps:

- *Collection of Responses:* This involves gathering a diverse array of student answers across various questions, grade levels, and question types. This diversity is essential to ensure the ASAG models are robust and accurately reflect students' understanding.
- *Annotation:* Expert educators manually grade these responses, providing both scores and feedback. This process is labor-intensive and costly.
- *Quality Evaluation:* Maintaining consistency and accuracy in annotations is critical. The dataset underwent several evaluations to ensure its reliability and effectiveness in the ASAG context.

3.2.1 Data Collection

The dataset includes grades based on student responses from three distinct classrooms, each taking three separate tests. Each test comprised 16 short answer questions, resulting in 48 questions across all tests. The number of responses per question varies, reflecting the natural variability found in classroom assessments. To maintain data integrity, identical student responses are reported only once in the dataset. 2133 distinct student responses were obtained as a result of this meticulous curation. For each question, the instructor provided a reference response, serving as a benchmark for grading. The inclusion of these reference responses allows for a thorough and consistent evaluation of the student's answers. The dataset encompasses five types of questions, ensuring a diverse and comprehensive collection that can be used for various analytical and training purposes. These question types cover a broad spectrum of cognitive skills, from simple recall to more complex analytical thinking, providing a robust foundation for training neural models.

1	**" عرف"**	Define?
2	**"إشرح"**	Explain?
3	**"ما النتائج المترتبة على"**	What consequences?
4	**"علل"**	Justify?
5	**"ما الفرق"**	What is the difference?

Figure 3 depicts the distribution of answers by type of question. The question types are designated as 1, 2, 3, 4, and 5. Two human experts, both experienced computer science instructors, independently rated the student answers using a grading scale ranging from zero (indicating a totally inaccurate response) to five (indicating an excellent answer). To establish a reliable benchmark, we use the average of the grades assigned by these two experts as the gold standard. This average serves as the reference point against which we evaluate the outputs of the grading model. Table 3 illustrates this process by presenting a question-and-answer pair along with three sample student responses. For each response, the table shows the grades awarded by the two human experts.

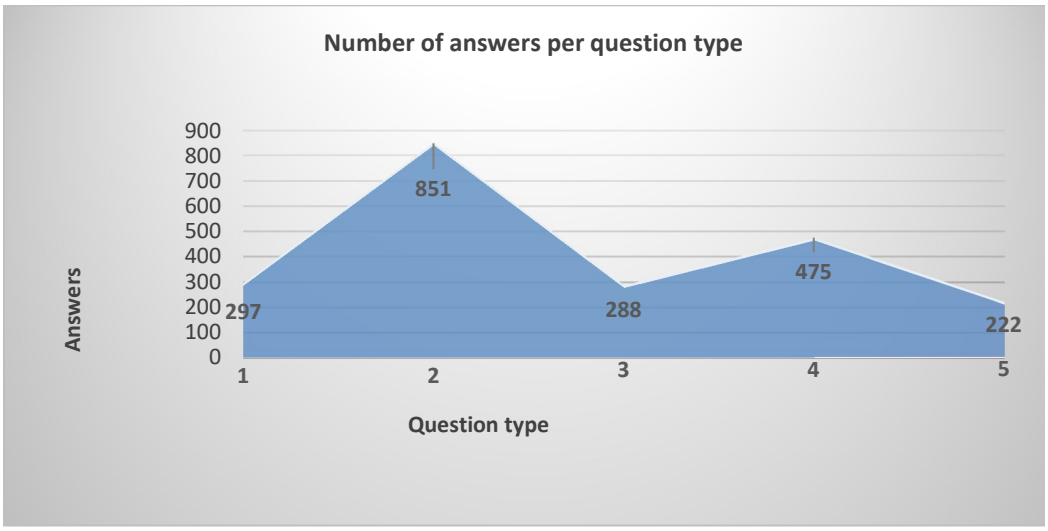

Figure 3 : Number of answers per question type (Ouahrani & Bennouar, 2020)

Table 3 Sample question, Reference Answer, Student Answers and the two Manual Grades

Sample Questions, Reference Answers, Student Answers and the two Manual Grades (AR-ASAG Dataset) (Ouahrani & Bennouar, 2020)			
Question	عرف مصطلح الجريمة المنظمة على الانترنت *Define: Online Organized Crime*		
Reference answer	عنف منظم تقوم به جماعات ترتكب أفعالا تخترق بها القانون للحصول على مكاسب مالية، بطرق وأساليب غير مشروعة تنفذ بعد تدبير وتنظيم *It is an organized violence by groups committing acts to gain financial gain, in unlawful ways using measure and organization.*		
Student answer 1	هي عنف منظم تقوم به جماعات من اجل كسب الأموال وتعتمد على التنظيم وكسب الأموال. *It is an organized violence organized by groups to make money and depends on organization and making money.*	5	5
Student answer 2	هي سلوك غير قانوني تقوم على التنظيم بهدف سرقة المعلومات او تغييرها. *It is an illegal behavior based on the purpose of stealing or changing information.*	4.5	4
Student answer 3	عنف منظم يسعى من خلاله تحقيق مطالب مالية غير شرعية تقع على الأنترنت *Organized violence which seeks to achieve illegal financial requests using Internet*	2.5	3

3.2.2 Inter-Annotator Agreement

The process of manually assigning grades posed significant challenges for the annotators. Their difficulty stemmed not from determining whether answers exhibited semantic similarity, but rather from accurately gauging the nuanced degrees of similarity between two responses and subsequently assigning grades accordingly. This task required meticulous attention to detail and objectivity to ensure a fair and consistent evaluation across all responses. This task required careful consideration and nuanced judgment, making the grading process inherently complex.

In order to assess the agreement among the annotators, we used two statistical measures: Root Mean Squared Error (RMSE) and Pearson's correlation coefficient (r). A lower value of RMSE indicates greater consistency, whereas a higher value of r indicates better consistency. RMSE provided a measure of the average magnitude of error in the grades assigned by the annotators, with lower values indicating closer alignment with the average grades. Pearson's correlation coefficient, on the other hand, measured the strength and direction of the linear relationship between the grades assigned by the two annotators, with higher values indicating greater consistency. The evaluations were performed by comparing these measures with the mean of the grades awarded by humans on a question-by-question basis.

To achieve a thorough and precise investigation of grading consistency, each question and its related student response were considered individual data points. This approach allowed us to examine the precision of the grades assigned to each individual response, providing a comprehensive understanding of the annotators' grading behavior. This approach underscored the paramount importance of precisely determining the grade for each individual answer. The results revealed a notable correlation coefficient of r = 0.8384 and a Root Mean Squared Error (RMSE) of 0.8381 between the assessments of the two experts. Automated Short Answer Assessment inherently entails subjectivity, given its focus on evaluating content.

Since subjectivity is an intrinsic aspect of any evaluative process (Brown et al., 1999), scrutinizing the grades assigned by the two annotators brings to light the inherent subjectivity involved in grading short-answer questions. This subjectivity can be influenced by a variety of factors, such as personal bias, prior knowledge, and even mood at the time of evaluation. As depicted in Table 4, both annotators assigned matching grades to 34.83% (743 answers) of cases. In the majority of cases, 54.14% (1155 answers), the grade discrepancy was minimal, not exceeding one point.

However, in a significant portion, 11.01% (235 answers), the discrepancy exceeded one point, and in a smaller percentage, 2.15% (46 responses), the difference surpassed two points on a five-point scale. Notably, in instances of disagreement, the second annotator tended to assign a higher grade, doing so 38.2% of the time. The average grade given by the first grader was 2.86, while the second grader's average grade was slightly higher at 2.94 for the entire dataset. The presence of subjectivity becomes apparent when examining the distribution of grade discrepancies between the two annotators, as depicted in Figure 4. In addition to presenting the RMSE error for the overall dataset, we also provide the median RMSE error for each individual question. The average RMSE for the entire dataset (Av (RMSE)) was calculated to be 0.5629, offering insight into the level of agreement between annotators on a question-by-question basis. This examination highlights how annotators may interpret student responses differently, resulting in discrepancies in grading. The observed subjectivity can be attributed to the diverse evaluation criteria employed by different annotators, who may adhere to distinct frameworks for assessing student answers, even when model answers are available. This variability underscores the complex nature of grading and emphasizes the importance of understanding and addressing these differences to ensure consistency and fairness in assessment. The lack of explicit instructions on grading beyond the [0..5] scale further contributes to these differences. The variability in manual grading underscores the challenges inherent in developing automated grading systems that aim to replicate human assessment. The automated systems must account for the nuances of human judgment and provide consistent, fair evaluations. The inclusion of manual grades in the dataset allows for a thorough comparison between human and automated grading, highlighting areas where the automated system may need adjustment to better align with human evaluators. Understanding these discrepancies is crucial for improving automated assessment technologies and ensuring they can handle the subjectivity inherent in grading short-answer questions. Table 5 positions the AR-ASAG dataset in relation to other frequently utilized datasets for ASAG assessment and evaluation, as discussed in section 3.2.2. Initially, the dataset was evaluated using an unsupervised grading model, with the outcomes presented as baseline metrics for our current study. Subsequently, the dataset was employed to train and evaluate the supervised grading model. A comprehensive evaluation of these results is provided and discussed in Chapter 4 (Results and Evaluation).

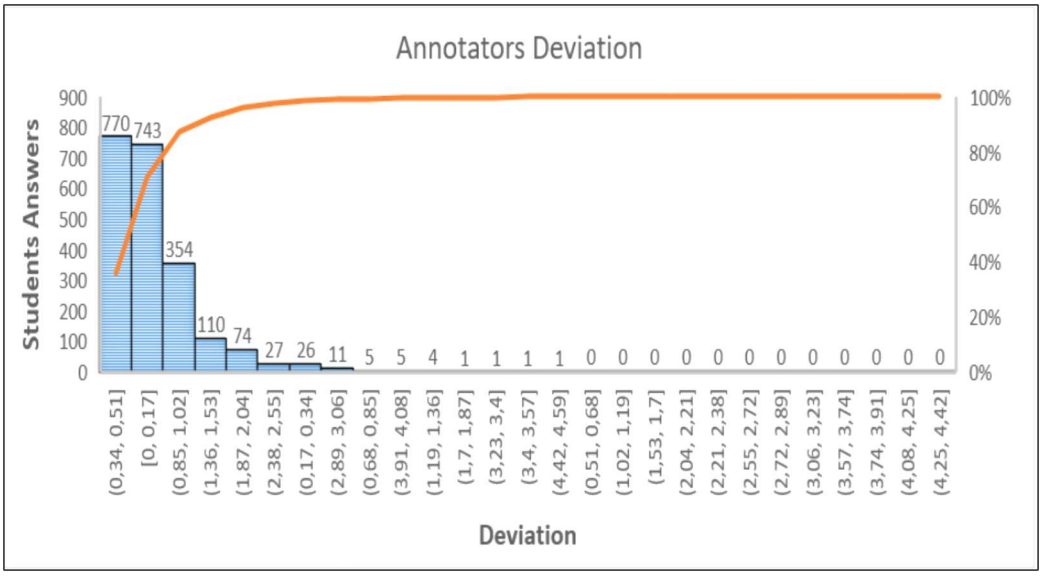

Figure 4 Inter-Annotator Agreement between human experts

Table 4 Annotators Analysis

Difference	Number of answers	%
0	743	34.83
0 <D <= 1	1155	**54.14**
1 <D <= 2	189	8.86
2 <D <= 3	38	1.78
D > 3	8	0.37

Table 5 AR-ASAG Dataset vs. ASAG Datasets

Dataset	Lang.	Answers	Domain	Availability
Texas (extended)	English	2273	Data structures Course	Yes
Cairo University	Arabic	610	Environmental Science Course	No
The SemEval 2013 Dataset	English	17,204	Multi-topic	Yes
ASAP-SAS Dataset	English	10,000	Scientific domain	Yes
AR-ASAG	Arabic	2133	Cybercrimes	Yes

3.2.3 Dataset Versions

The AR-ASAG Dataset is accessible in many forms, including TXT, XML, XML-MOODLE, and Database (.DB). The ".DB" format allows for data exports adapted to different analytical needs, as well as the efficient and continual extension of the dataset

through the incorporation of new tests and examinations. The dataset management application automates all exports to various versions. The XML-MOODLE format is especially effective for analyzing short response grading systems on the MOODLE platform, a popular e-learning system. This dataset serves as a question bank, allowing for performance comparisons using the MOODLE platform's short response system, which is based on grammar and pattern matching. Notably, the dataset contains manual grades, allowing for a thorough investigation of the automated system's performance in comparison to human annotators, which is critical in this subjective arena with no established assessment standards.

3.3 PROPOSED FEATURES

Selecting the most informative features for the regression model is crucial for achieving optimal performance in automatic short answer grading. Proper feature selection can enhance the accuracy and generalizability of the regression model. The process of the feature selection highlights the identification of features, creation of derived features, evaluation of the features and their combination using metrics and datasets and the refinement of selected features (combine, test and refine).

We focus on new features like specific and general domain knowledge, word alignment matching, answer length, question difficulty, and knowledge gaps between questions and answers.

3.3.1 Specific and general domain knowledge as features

In recent years, *computational distributional semantics techniques* have been abundant (Higgins et al. 2014; Adams et al. 2016). Distributional methods for meaning acquisition rely on a set of assumptions known as the distributional hypothesis (Harris, 1968), which forms the foundation of statistical semantics. This theory is sometimes articulated as the notion that "words that appear in similar contexts have similar meanings" (Turney & Pantel, 2010). The fundamental method involves gathering distributional data in vectors with a large number of dimensions and defining distributional semantic similarity based on the similarity of these vectors.

Generating distributional representations from word space models follows a consistent set of foundational steps. Initially, word features are extracted from a corpus, and then the semantic similarity between words is assessed based on the distribution of these features (Turney & Pantel, 2010). In these models, words are

represented typically as vectors, where semantically similar words are positioned close to each other in the semantic space. Compositional distributional semantic models extend this approach by describing the meaning of entire phrases or sentences, building upon the principles of traditional distributional semantic models. To achieve this, it is necessary to build distributional representations of the words within sentences. We utilize two distinct models to encapsulate domain knowledge: one that concentrates on specific details and another that encompasses general concepts.

For specialized knowledge in a particular field, we implement the COALS (Correlated Occurrence Analogue to Lexical Semantic) model (Rohde et al., 2004) to develop the semantic representation of a specific domain-related corpus. To gather extensive knowledge across various domains, we employ the sophisticated deep learning word embedding model known as Skip-gram (Mikolov, Sutskever, et al., 2013). This model is trained on a vast corpus of wide-domain texts to produce word vector representations.

3.3.1.1 Semantic Space Model for learning domain-specific features

We used the COALS method (Rohde et al., 2004) for semantic space processing for two main reasons. Initially, it offers a higher level of accuracy in predicting human similarity assessments compared to prior algorithms like Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996), Latent Semantic Analysis (LSA) (Deerwester et al., 1990), and Random Indexing (Sahlgren, 2005). Furthermore, unlike algorithms like LSA that work with groups of input documents, COALS uses a single text corpus and a moving window method to find word pairs that go together. The dimensions of a COALS co-occurrence matrix are relatively constant, in contrast to an LSA matrix, which scales proportionally with the number of documents. Thus, COALS is demonstrating more scalability and simplicity in its implementation. This is appropriate for Arabic, as the scarcity of Arabic resources can restrict the options available when attempting to locate a corpus (Al-Thubaity, 2015). The conceptual space development pipeline utilizing the COALS method comprises several key components, as illustrated in Figure 5. The key components are as follows:

- *Corpus Processing:* This initial phase entails cleaning and organizing the text corpus to ready it for analysis.

- *Context Selection:* During this stage, pertinent contexts within the corpus are identified, with an emphasis on the co-occurrence of word pairs within a specified text window.

- *Feature Extraction:* Here, features are derived from the identified word pairs and their contexts, serving as the foundation for building the semantic space.

- *Semantic Space Normalization:* Finally, the extracted features are normalized to establish a coherent and scalable semantic space.

These components align with the framework detailed in the work of Ouahrani and Bennouar (2018), ensuring a systematic approach to developing semantic spaces using the COALS method.

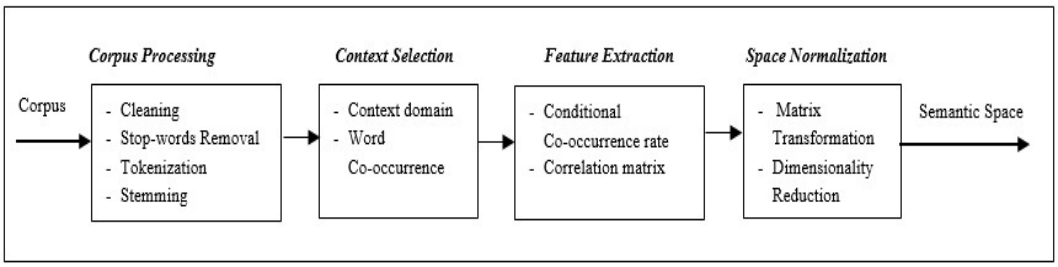

Figure 5 Semantic Space Pipeline

1) ***Corpus Processing.*** The inputs are normalized using pre-processing techniques like cleaning, stop-word removal, tokenization, and stemming. The methodology we suggest may be used on any textual dataset. In order to conduct Arabic experiments, it seems most suitable to utilize a varied corpus of ordinary Arabic. In order to extract the semantic space, the corpus has to undergo preprocessing, which involves cleansing, normalization, stop-word removal, tokenization, and stemming.

   a) *The cleaning task focuses on eliminating extraneous letters, punctuation, and non-native language elements.*

   b) *The normalization task seeks to standardize the many forms of expressing a term.* Exclusively for the Arabic language:

      - Suppression of diacritics
      - Normalize characters: (آ - إ - أ) to (ا) and (ة) to (ه) and (ى) to (ي).
      - Remove aesthetic additions. e.g. الأعمـــــــــــال à الأعمال
      - Normalize all numeric digits to the "NUM" token.

   c) *The tokenization task involves the identification of atomic units, such as words or tokens.*

d) The removal of stop-words task. Stop-words are functional words like prepositions, pronouns, and conjunctions that are significant due to their linguistic role. They possess minimal semantic information. By eliminating these terms, we reduce the overall count of unique words in the corpus. These terms can be identified by applying a frequency threshold to a large corpus. However, there are numerous existing compilations of stop-words that have already been generated and made available, and these are often quite similar.

e) The stemming task is to reduce all words with the same root to a single canonical form (by removing prefixes and suffixes, one may determine a word's true root and find the pattern that corresponds with the remaining word).

2) *Context Selection.* The process entails gathering co-occurrence counts by incrementally increasing the window size to four and constructing a matrix to store these co-occurrence data. The elements of the matrix represent the cumulative weights of the occurrences of the row term with the column term, as defined by the distributional hypothesis. According to this hypothesis, it is not necessary for identical words to be adjacent; instead, they should co-occur with the same set of other words. Specifically, two words are considered co-occurrents if their neighborhood size is less than the window size. With a window size of four, only the four neighboring elements on both the left and right sides are considered. For illustration, we will demonstrate this process using a simple text collection referred to as Corpus-Example, as shown in Table 6.

Table 6 Sample text corpus (corpus-example)

Arabic Corpus	شركات امن المعلومات رصدت ارتفاعا في عدد المكالمات الوهمية الصادرة عن أجهزة الهاتف النقال الذكية
English Corpus	Information security companies spotted a rise in the number of phantom calls from mobile smart devices

When examining the word "ارتفاعا", its immediate adjacent words will have a weight of 4, the words next to those will have a weight of 3, and so on.

1	2	3	4	**0**	4	3	2	1
الوهمية	المكالمات	عدد	في	**ارتفاعا**	رصدت	المعلومات	امن	شركات

Table 7 presents the initial symmetric word-by-word co-occurrence matrix, which was generated from the "Corpus-Example" sample corpus. Essentially, the dimension of the context vector aligns with the most commonly occurring canonical phrases in the corpus following the stemming process. To optimize memory usage, dynamic hash tables are employed to construct the resulting sparse matrix.

3) ***Word Feature Extraction.*** The process involves identifying the characteristics that most effectively differentiate word meanings and then quantifying the rate at which they occur together in a certain area. The conditional co-occurrence rate seeks to determine if a word (w_i) has a higher or lower frequency of occurrence near another word (v_j) compared to its overall frequency. In order to measure the likelihood for words to co-occur, the occurrences of words are analyzed using Pearson's correlation. The correlation matrix is computed by applying formula (1) to each member of the co-occurrence matrix.

$$\begin{cases} w'_{a,b} = \dfrac{T w_{a,b} - \sum j\, w_{a,j} \cdot \sum i\, w_{i,b}}{\left(\sum j\, w_{a,j} - (T - \sum j\, w_{a,j}) \cdot \sum i\, w_{i,b} \cdot (T - \sum i\, w_{i,b})\right)^{1/2}} \\ T = \sum_i \sum_j w_{i,j} \end{cases} \quad (1)$$

a & b : Terms of the co-occurrence matrix (row i and column j).
$w_{a,b}$: Element of the co-occurrence matrix of the terms a and b
$\sum j\, w_{a,j}$: Sum of the columns of the term row a.
$\sum i\, w_{i,b}$: Sum of the rows in the column of the term b
$T = \sum_i \sum_j w_{i,j}$: Sum of all elements of the co-occurrence matrix.

By applying this correlation, the newly calculated cell values will span from -1 to 1. A correlation of zero indicates that term w_i and term v_j are statistically independent, and the occurrence of word w_i is equally probable near v_j as in any other context. A positive correlation indicates that the occurrence of term w_i is more probable when term v_j is present compared to when it is not. In a sizable corpus, the correlation values are often diminutive, making it uncommon for the correlation value to surpass 0.01. Furthermore, the bulk of correlations have a negative relationship. Table 8 shows the correlation matrix calculated for the "Corpus-Example" sample corpus.

4) ***Normalization.*** Negative results, they are assigned a value of zero. This is done since negative correlations convey less information. Conversely, positive values undergo a square root alteration to amplify the importance of several small values relative to larger ones. The matrix's rows and columns represent the semantic context vectors of the associated row words and column terms, respectively, because of its symmetry. The vectors of all words constitute the semantic space.

The correlation between the vectors of two words indicates the extent to which their semantic meanings are comparable. Table 9 presents the normalized correlation matrix computed for the "Corpus-Example" sample corpus.

Table 7 Semantic Space Algorithm (Context Selection): The symmetric word-by-word co-occurrence matrix with a ramped, 4-word window.

	شرك	امن	علم	رصد	رفع	عدد	كلم	وهم	صدر	جهز	هتف	نقل	ذكي
شرك	0	4	3	2	1	0	0	0	0	0	0	0	0
امن	4	0	4	3	2	1	0	0	0	0	0	0	0
علم	3	4	0	4	3	2	1	0	0	0	0	0	0
رصد	2	3	4	0	4	3	2	1	0	0	0	0	0
رفع	1	2	3	4	0	4	3	2	1	0	0	0	0
عدد	0	1	2	3	4	0	4	3	2	1	0	0	0
كلم	0	0	1	2	3	4	0	4	3	2	1	0	0
وهم	0	0	0	1	2	3	4	0	4	3	2	1	0
صدر	0	0	0	0	1	2	3	4	0	4	3	2	1
جهز	0	0	0	0	0	1	2	3	4	0	4	3	2
هتف	0	0	0	0	0	0	1	2	3	4	0	4	3
نقل	0	0	0	0	0	0	0	1	2	3	4	0	4
ذكي	0	0	0	0	0	0	0	0	1	2	3	4	0

Table 8 Semantic space Algorithm (Word Feature Extraction): Raw counts are converted to correlations.

	شرك	امن	علم	رصد	رفع	عدد	كلم	وهم	صدر	جهز	هتف	نقل	ذكي
شرك	-0,048	0,301	0,182	0,088	0,007	-0,069	-0,069	-0,069	-0,069	-0,067	-0,063	-0,057	-0,048
امن	0,301	-0,068	0,204	0,119	0,047	-0,018	-0,082	-0,082	-0,082	-0,08	-0,075	-0,068	-0,057
علم	0,182	0,204	-0,084	0,153	0,086	0,027	-0,032	-0,092	-0,092	-0,089	-0,084	-0,075	-0,063
رصد	0,088	0,119	0,153	-0,095	0,128	0,072	0,015	-0,041	-0,097	-0,095	-0,089	-0,08	-0,067
رفع	0,007	0,047	0,086	0,128	-0,1	0,12	0,065	0,01	-0,045	-0,097	-0,092	-0,082	-0,069
عدد	-0,069	-0,018	0,027	0,072	0,12	-0,1	0,12	0,065	0,01	-0,041	-0,092	-0,082	-0,069
كلم	-0,069	-0,082	-0,032	0,015	0,065	0,12	-0,1	0,12	0,065	0,015	-0,032	-0,082	-0,069
وهم	-0,069	-0,082	-0,092	-0,041	0,01	0,065	0,12	-0,1	0,12	0,072	0,027	-0,018	-0,069
صدر	-0,069	-0,082	-0,092	-0,097	-0,045	0,01	0,065	0,12	-0,1	0,128	0,086	0,047	0,007
جهز	-0,067	-0,08	-0,089	-0,095	-0,097	-0,041	0,015	0,072	0,128	0,095	0,153	0,119	0,088
هتف	-0,063	-0,075	0,084	-0,089	-0,092	-0,092	-0,032	0,027	0,086	0,153	-0,084	0,204	0,182
نقل	-0,057	-0,068	-0,075	-0,08	-0,082	-0,082	-0,082	-0,018	0,047	0,119	0,204	-0,068	0,301
ذكي	-0,048	-0,057	-0,063	-0,067	-0,069	-0,069	-0,069	-0,069	0,007	0,088	0,182	0,301	-0,048

Table 9 Step 3-Semantic Space Algorithm: Negative values discarded and the positive values square rooted.

	شرك	امن	علم	رصد	رفع	عدد	كلم	وهم	صدر	جهز	هتف	نقل	ذكي
شرك	0	0.549	0.427	0.297	0.084	0	0	0	0	0	0	0	0
امن	0.549	0	0.452	0.345	0.217	0	0	0	0	0	0	0	0
علم	0.427	0.452	0	0.391	0.293	0.164	0	0	0	0	0	0	0
رصد	0.297	0.345	0.391	0	0.358	0.268	0.122	0	0	0	0	0	0
رفع	0.084	0.217	0.293	0.358	0	0.346	0.255	0.1	0	0	0	0	0
عدد	0	0	0.164	0.268	0.346	0	0.346	0.255	0.1	0	0	0	0
كلم	0	0	0	0.122	0.255	0.346	0	0.346	0.255	0.122	0	0	0
وهم	0	0	0	0	0.1	0.255	0.346	0	0.346	0.268	0.164	0	0
صدر	0	0	0	0	0	0.1	0.255	0.346	0	0.358	0.293	0.217	0.084
جهز	0	0	0	0	0	0	0.122	0.268	0.358	0	0.391	0.345	0.297
هتف	0	0	0	0	0	0	0	0.164	0.293	0.391	0	0.452	0.427
نقل	0	0	0	0	0	0	0	0	0.217	0.345	0.452	0	0.549
ذكي	0	0	0	0	0	0	0	0	0.084	0.297	0.427	0.549	0

3.3.1.2 Word Embedding Model for learning domain-general features

We employ distributed vector representations derived from the Word2vec embedding model (Mikolov, Chen, et al., 2013) to acquire the relationships between words within sentences, facilitating analysis. These representations enhance our understanding of semantic connections between words.

The Word2vec model consists of two main architectures: the continuous bag of words (CBOW) model and the skip-gram model. The CBOW model infers the adjacent context of the target word, while the skip-gram model infers the adjacent context of an input word. We use the skip-gram model for word general domain knowledge acquisition. The Skip-Gram model is a key component of word embedding techniques, enabling the acquisition of general domain knowledge (Mikolov, Sutskever, et al., 2013). It captures semantic relationships between words, enabling the transfer of knowledge across tasks and domains. The model's core function is semantic representation, which predicts surrounding words based on a target word, allowing it to identify and encode contextual relationships between words.

The model's ability to learn generalizable representations extends beyond a single domain, allowing for knowledge transfer between domains. Unsupervised learning is another advantage, allowing the model to learn from unlabeled text data, enhancing its understanding of the domain. However, the model has a major limitation: potentially overlooking long-range dependencies as focusing on local context (Mikolov, Sutskever, et al., 2013). Long-term dependencies usually correlate with the specific task domain (ASAG in our context). To overcome this limit, we combine it with the COALS model.

3.3.1.3 Leveraging COALS and Skip-Gram for Comprehensive Domain Knowledge Representation

The integration of COALS and Skip-Gram models could potentially improve the quality of knowledge representation, thereby increasing the effectiveness of the grading model.

COALS for Domain-Specific Knowledge. COALS specializes in capturing domain-specific knowledge by analyzing co-occurrence patterns within a designated corpus, adapting well to variations within the domain. This capability makes it particularly suitable for representing the terminology, concepts, and interrelationships specific to the domain of study. By examining the corpus, COALS identifies words that

frequently appear together, indicating strong semantic connections between them. For example, in the realm of cybercrimes, terms like "crime" and "internet" demonstrate a more robust association compared to a more generalized text corpus. Answers are aligned by associating them with a domain-specific vocabulary corpus that is relevant to the subject area of the course. They are then transformed into vector representations that capture the precise meanings of each word.

Skip-Gram for General Language Understanding. In contrast, the Skip-Gram model focuses on understanding broad semantic relationships among words across diverse contexts. By training on extensive and varied text data, Skip-Gram generates word embeddings that encode these general linguistic relationships, facilitating comprehension of language beyond specific domains, encompassing common word meanings, grammatical structures, and universal knowledge.

To elucidate the meaning of entire sentences, we generate distributional representations for words in both the reference response and the student answer. This involves creating vector distributions by summing the individual word vectors retrieved from the skip-gram model and the COALS model. As illustrated in Figure 7, the initial step involves tokenization to extract a list of words from the answer. For a given sentence answer, a pair of vectors (Vs, Vg) that encapsulate both specific and general knowledge domains is computed.

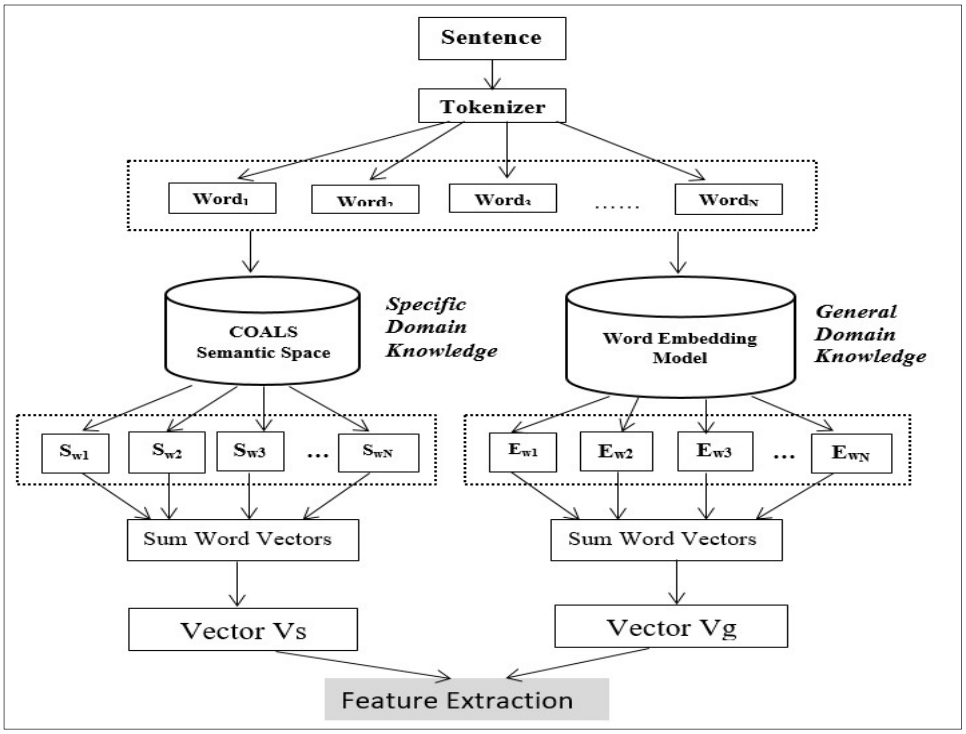

Figure 6 Generating Sentence Vectors from Pre-Trained and Semantic Space Word Vector Models

3.3.2 Text Similarities Features

In the case of ASAG, our primary interest is to obtain a measure of similarity between a given reference answer R = (r_1,... r_n) and a student response S = (s_1, ..., s_m) (where each r and s is a word token). We focus on statistical corpus-based measures learned from large text collections because of their large type of coverage. These similarities do not require any word data models. It significantly covers more tokens than any dictionary-based approach. The text similarity features are derived by combining lexical and semantic similarities. Selecting appropriate measures of similarity of two texts involves several steps. These steps ensure that the selected measures capture the diverse aspects of text similarity relevant to the specific ASAG context. Since detailing all the explored similarity measures is not feasible, we will instead outline the workflow used to select the retained similarities.

Workflow to Select and Combine Text Similarity Measures:

1. Identify Potential Similarity Measures and Word Distributions:

- *Lexical Measures:* We implemented 14 existing lexical similarity measures.
- *Semantic Measures:* We explored several semantic similarity measures, including cosine similarity, corpus-based and the impact of Word Embedding models like CBOW and Skip-Gram embeddings.
- *Custom Measures:* Custom domain-specific measures tailored to our specific needs. We proposed : The word vector alignment similarity model

2. Evaluate Similarity Measures:

- *Datasets:* We used the Cairo dataset and the STS SEMeval 2017 dataset for evaluation.
- *Metrics:* The performance of each similarity measure was validated using Pearson correlation and Root Mean Square Error (RMSE).

3. Combine, Test, and Refine:

- *Assign Weights:* Based on performance on the validation set, weights were assigned to each similarity measure.
- *Compute Weighted Similarity:* The weighted average similarity was computed, incorporating term weighting and other relevant factors.
- *Adjust Weights:* Weights were adjusted based on error analysis to improve performance.

3.3.2.1 Lexical Similarity

To avoid the limitation on word presentation's ability to capture the meaning of a sentence, which depends upon both syntax and semantics, we use syntactical features to improve the grading accuracy. We use Jaccard (Real & Vargas, 1996) and Dice coefficients (Dice, 1945) to compute word overlap features. We use the *Normalised Longest Common Subsequence* (LCS) introduced by Allison and Dix (1986) and modified by Islam and Inkpen (2008) to consider the length of both the shorter and the longer answer and the maximal consecutive longest common subsequence starting at any character.

3.3.2.2 Semantic Similarity

We define the similarity between R and S using the cosine similarity of the component word vectors in two ways:

- ***Vector Summation Model using cosine similarity.*** When measuring the semantic similarity of two answers, all word vectors appearing in the answer are summed. Thus, we can get vector representations of the two answers VR and VS. The similarity of the two answers can be measured with cosine similarity as follows:

$$\text{Similarity}(R, S) = \text{Cosine}(VR, VS)$$

$$\text{Where} \quad VR = \sum (ri) \text{ for } (i=1, m) \;\&\; VS = \sum (si) \text{ for } (i=1, n)$$

With this model, we consider symmetrically the student and reference answers and capture semantic similarity without considering the length or the implication of concepts.

- ***The proposed word vector alignment similarity model***

We proposed to a similarity measure based on an aggregate of word-level semantic similarity. The measure is customised based on domain-specific needs. The objective here is to assess whether the student's answer implies the reference answer taking in consideration the length of the answers. As presented below in algorithm 1, the similarity is based on one-to-many word vector alignment using the cosine distance between each aligned word vector of the student answer with all aligned word vectors of the reference answer. Thus, it captures the nuances and implications of the student's answer, even if it is not an exact match to the reference answer.

The word-to-word similarities are summed to give the overall score similarity using a joint similarity matrix (Mat (n × m)) in which every cell presents the cosine similarity between the column-word and the row-word vector distribution. This similarity is applied when the reference answer is more concise than student answer. Usually it is the case since reference answers generally contain only the important concepts, while student responses are often less relevant, without necessarily being less correct. The rows of the matrix are used for the words from the shorter answer (let be n words), while the columns represent the words from the longer answer (let be m). The similarity is calculated by aggregating the maximum of similarities for each word vector from an answer with all aligned word vectors of the other answer. The obtained score is multiplied by the reciprocal harmonic mean of m and n to obtain a balanced similarity between 0 and 1. The main idea here is to find, for each word in the student answer, the most similar matching in the reference answer.

Algorithm 1. Answer-to-Answer Semantic similarity using one-to-many word vector alignment

Input Mat (n, m), R, S, n, m
Output: sim

Sim=0;
Repeat
Find the maximum-valued matrix-element, M_{ij}
Sim ← *Sim*+ Mat_{ij}
Remove matrix elements of i^{th} row and j^{th} column from M
Until Mat empty
// *Reciprocal harmonic mean to put score between 0 and 1.*
Sim ← (*Sim*× (m+n))/2×n×m)

Return Sim
End Algorithm.

3.3.3 Word Weighting features

Term weighting enables the differentiation of the significant terms in the corpus (or sentences) from the less significant ones, hence enhancing similarity. In order to obtain pairs of words with their frequency weighted, a pre-processed corpus is required. These similarities allow the measurement of word overlap between the two answers at the expense of word frequency.

To improve similarity scores, different weights features are considered:

- ***IDF (Inverse Document Frequency).*** widely used in information retrieval (Salton & Buckley, 1988), it aims to give greater weight to the less frequent terms, considered more discriminating.

$IDF_i = \log \frac{\lvert D \rvert}{\lvert \{d_j : t_i \in d_j\} \rvert}$ $\lvert D \rvert$: Number of documents in the corpus $\lvert \{d_j : t_i \in d_j\} \rvert$: Number of documents where term t_i appears.	(2)

- ***NTFlog (Normalized TFlog).*** Unlike IDF, we propose NTFlog that measures the importance of the term in the complete corpus independently of the documents in which the term appears. The weight calculation is done in two steps:

  - Calculate TFlogs of corpus terms applying (2):

    $$TFlog(W) = -\log(Wc/N) \quad (2)$$

    Wc : Number of times the term W appears in the corpus
    N: Total number of words in the corpus.

  - Normalize TFlogs on [0..1] applying (3)

    $$NTFlog(w) = TFlog(w)/Max(TFlog) \quad (3)$$

Equation (2) assigns lower TFlog values to frequently occurring words and higher TFlog values to infrequent words. Dividing the TFlogs by the highest TFlog value (using formula (3)) ensures that less common terms are weighted closer to 1, while very common words are weighted closer to 0 due to normalization. This is because less frequent phrases are considered to be more distinctive.

- ***Part-Of-Speech(POS) tagging*** (Toutanova et al., 2003) weights to consider the syntactic aspect of the word's vector space. POS tagging is a sense disambiguation method that aims to assign each word in a text with a fixed set of parts of speech (verb, noun, adjective, adverb …) since words belonging to different categories contribute differently to the sentence meaning. As sentence meaning unfolds from the verb, we assign the highest weight to the verb followed by nouns, adjectives, and adverbs.

3.3.4 Answer Length statistics features

The length of words is a significant factor to consider when assessing the similarity between sentences (Zhao et al., 2014). Comprehending longer phrases is a greater challenge. In order to examine the impact of word length characteristics on sentence similarity, we incorporate three length features during the training phase:

- *The length of the student answer*.

- *The difference length between student answer and the reference answer.*
- *The redundancy frequency of terms in answers* represents the ratio of the number of words that repeat more than once in the student's answer to the total number of words in the answer. The redundancy frequency is considered an input feature since the calculation of similarity is biased by word repetition. By taking into account the redundancy frequency of terms in answers, the training phase can more accurately assess the similarity between a student's answer and the reference answer. This feature allows for a more nuanced evaluation of student responses, considering the impact of repeated terms on the overall length and structure of the answer. Ultimately, the inclusion of redundancy frequency as an input feature enhances the precision and fairness of the assessment process during training.

3.3.5 Question Difficulty Level Features

Since the performance of the students also depends on the difficulty level of the questions, we have introduced the difficulty level as a feature during the training process. Initially, the dataset did not contain this information. We calculated it based on the average score obtained for each question. If the average of the grades of all its student's answers is less than or equal to 1,5 the question is considered "difficult", otherwise if the average is between (1,5 and 3,5), the question is considered "average", otherwise, the question is considered "easy".

3.3.6 The information gap between question and answer

The gap between questions and answers expresses how much of what is expected by the question might be targeted in the answers. (Saha et al., 2018) is the only work, to our knowledge, that considered the question in the scoring model where the question gap is learned from sentence embeddings. We particularly propose this gap as similarities to move from vectors to tokens. This keeps features calculation easy and fast. For each set (question (Q), reference answer (R), student answer (S)), we generate the corresponding sentence vectors (Q, R, and S). The gap between expected knowledge in the question and the information in the reference and student answers is computed. We use the Hadamard product (Horn & Johnson Frontmatter, 2012) (noted ⊙ here) and the vector difference(noted -) to capture this gap (in the same way, that

information is collected at the gates in an LSTM neural network). Six features are proposed:

- {Cosine [(R⊙Q), (S⊙Q)], Cosine [(R-Q), (S-Q)]}: Capture the gap between what is expected in the reference answer and what is expressed in the student's answer concerning the question.
- {Cosine [S⊙ (S⊙Q), R], Cosine [[S⊙(S-Q), R]}: Capture how much of what is expected in the question by the reference answer could be expressed in the student answer.
- {Cosine [S⊙(R⊙S), R⊙(R⊙S)], S⊙ (R-S), R⊙(R-S)]}: capture the similarity between R & S knowing the gap between them.

To avoid influencing the calculation of the gaps, we proceeded to a question demoting of the dataset. We removed, from the student answers, the words used in the question's formulation. This avoids rewarding (in terms of similarity) the answer of a student that reproduced the words of the question.

3.4 PROPOSED GRADING MODEL

The proposed approach of combining semantic space and word embedding models for automatic short answer grading incorporates several key elements for achieving accurate and effective assessments:

- Generating COALS and Skip-Gram Vectors.
- Calculating Similarity Features using the COALS vectors.
- Calculating Similarity Features using the skip-gram vectors.
- Extracting Additional Features:
  - Answer Length: length of the student answer, the difference length between student answer and the reference answer, and the redundancy frequency of terms in answers.
  - Question Difficulty.
  - Gap between Question and Answer.
- Training the Regression Model using the extracted features and the dataset to initialize Weigh features in the trained model.
- Scoring the student answer using the trained model to predict the grading score.

The ASAG system requirements emphasize two primary processes, as depicted in Figure 7: the training process and the scoring process:

The training process is executed once. The system acquires knowledge from the given features and constructs a model that may be employed for further forecasting. Conversely, the continuous scoring process entails using the trained model to generate predictions on fresh data inputs:

1. Pre-processing: This stage comprises normalizing the dataset by cleaning, removing stop words, tokenizing, and stemming.
2. Feature Extraction: Features are extracted and used as inputs to train the model.
3. Model Training: The model is trained to determine the weights of its features.

The scoring process takes into account the question, the reference answer, and the student response to predict a grade.

4. Preprocessing: Answers and questions undergo normalization using pre-processing techniques such as cleaning, stop-word removal, tokenization, and stemming. This ensures that the text data is standardized and prepared for further analysis.
5. Feature extraction involves extracting features from the inputs and incorporating them into the trained model to determine the grade and generate matching feedback.

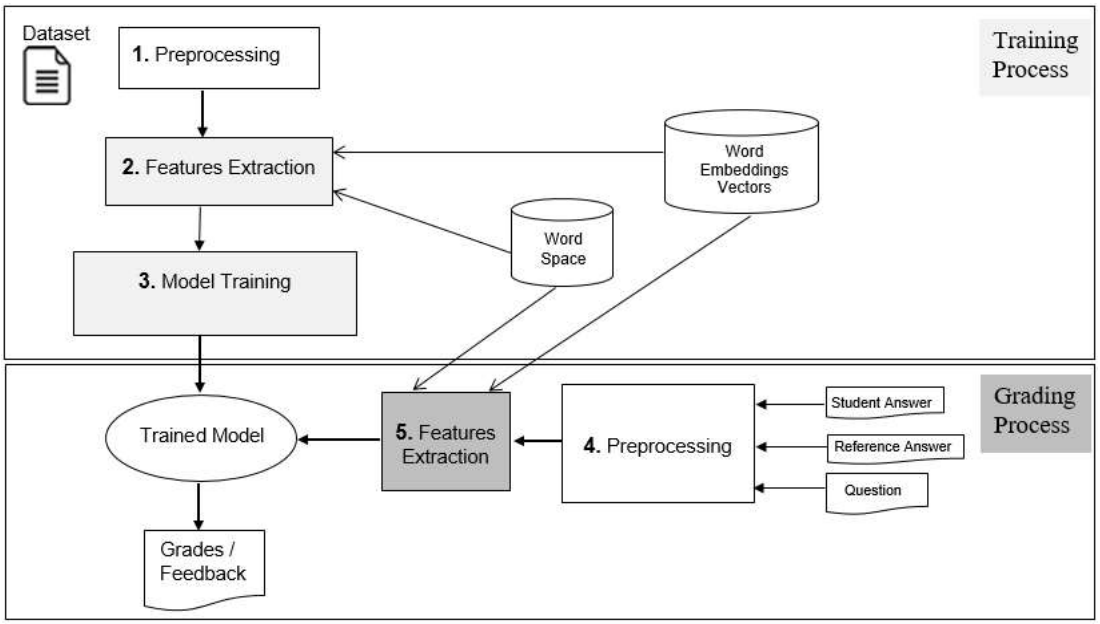

Figure 7 Automatic Short Answer Grading Framework Overview.

The proposed model is trained on the dataset's gold standard human grades using three linear regression methods implemented using the Scikit-learn module (Scikit-learn, 2019):

- ***Support Vector Regression (SVR)*** utilises a technique of mapping data points into a space with a high number of dimensions and then identifying the hyperplane that most accurately aligns with the data. SVR, unlike conventional regression techniques, is adept at dealing with non-linear associations and anomalies in the data. Through the process of optimising the margin of error, Support Vector Regression (SVR) is capable of generating precise forecasts. The algorithm selects the line that best matches within a specified threshold to decrease the discrepancy between the actual and anticipated grades. Additionally, SVR can handle large datasets efficiently, making it suitable for real-world applications in various industries. The model is trained with a regularization of 1. SVR works with the epsilon-insensitive hinge loss, which defines the tolerance margin where no penalty is given to errors. We fixed it at (ϵ-SVR=0,1).

- ***Multi-layer perceptron regression Artificial Neural Networks (Mlpregressor-ANN***) are powerful tools for predicting continuous values based on input data. By utilizing multiple layers of interconnected nodes, the MLPregressor-ANN is able to capture complex relationships within the data and make accurate predictionsThe algorithm operates by optimising the squared loss function through the use of stochastic gradient descent. The model is trained using an "identity" activation function, a regularization term of 0.001, a learning rate of 0.001, and the limited-memory to calculate the descent direction by conditioning the gradient with curvature information.

- ***Ridge Linear Regression (Ridge-LR)*** is a linear regression model that uses a regularization term to avoid over fitting. This regularization factor penalizes big coefficients, encouraging the model to perform better with fresh, previously unknown data. Ridge-LR achieves a compromise between reducing error on training data and keeping the model simple by adding this element to the standard least squares objective function. To handle a greater range of data, the model is trained via a second-degree "polynomial features" technique. This entails building a feature matrix that includes all possible polynomial combinations from the dataset.

The dataset is used to extract features. The dataset is divided into three sets: a training set (70% of the data), an evaluation set (10% of the data), and a test set (20% of the data). The dataset's features and grades are imported into a data frame, which is subsequently used for regression analysis.

Multiple iterations of each model are performed to determine the ideal parameters that have the most impact on the model's efficiency, resulting in the highest level of accuracy. In the subsequent section, we provide a comprehensive explanation of the features that are used as inputs for the training procedure.

3.5 PARAPHRASE GENERATION FOR ALTERNATIVE REFRENCE ANSWERS

3.5.1 Problem formulation

In paraphrase generation, we are interested in models that take in a source sentence (S) containing the words (s_1, s_2, s_3… s_n) and generate an output sentence (G) with the same meaning but a different surface containing the words (g_1, g_2, g_3… g_m). Formally, given an input sentence (source sentence) S where S = {s_1, s_2, s_3… s_n}, the aim is to generate one or more sentences G = {g_1, g_2, g_3… g_m} where the sentence length of the generated sentence and the input sentence may vary. Our goal is to find the sentence G such that the conditional probability $p(\mathrm{G}|\mathrm{S})$ is maximized. We model $p(\mathrm{G}|\mathrm{S})$ as a product of word predictions (formula (4)):

$$\mathrm{p}(\mathrm{G}|\mathrm{S}) = \prod_{1}^{M} p(\mathrm{g_t}\,|\mathrm{g_{1:t-1}}, \mathrm{S}) \qquad (4)$$

This indicates that the probability of generating each current word (g_t) relies on the previously generated words ($g_{1:\,t-1} = g_1, g_2 \ldots g_{t-1}$) and the source sentence S. This many-to-many sequence prediction problem predicts a sequence of words {g_1, g_2, g_3… g_m} from a sequence of words (s_1, s_2, s_3… s_n). Paraphrasing in writing involves rephrasing a text using different words while maintaining the same meaning. This broader view includes permissible paraphrases achieved through word substitution, lexical changes, grammatical adjustments, verb-noun conversion, and semantic implications.

The focus is on capturing the essence and main ideas rather than fine-grained linguistic distinctions of meaning between sentences. This exclusion is because the focus is on capturing the essence and main ideas rather than linguistic nuances. For

instance, in paraphrasing a scientific article, we could alter "The experiment yielded statistically significant results" to "The study produced findings considered statistically important." Despite the different wording, the main idea of the sentence remains unchanged.

3.5.2 Proposed Paraphrase Generation Model

For modeling the conditional probability p(G|S), we propose ARAG-ED namely (Alternative Reference Answer Generator Encoder-Decoder) (Ouahrani & Bennouar, 2024), an Encoder-decoder which targets generating plausible alternative reference answers conditioned on the provided reference answer. The Encoder-Decoder model, proposed by Cho et al. (2014), is a recurrent neural network that produces a sequence of outputs based on a given sequence of inputs.

ARAG-ED consists of two components: the encoder and the decoder. Gated Recurrent Unit (GRU) deep neural networks (Chung et al., 2014)) are used by each component to effectively handle input sequences of different lengths. This model offers several key benefits, such as the ability to train a unified end-to-end model that directly processes both the source and target phrases, leading to more efficient and accurate results. Additionally, it is capable of effectively handling input and output text sequences of varying lengths. Figure 9 shows that constructing paraphrases involves two steps: encoding and decoding. The encoder is used to transform the input sentence into an encoder vector that corresponds to the final hidden state. During decoding, the encoder vector is used as the initial hidden state of the decoder. This allows the decoder to anticipate words sequentially in order to construct a sentence.

The attention mechanism improves the model's performance by directing its focus towards key elements in the input text, allowing for more accurate word predictions during decoding.

Transfer learning using a unified Transformer framework like T5 (Raffel et al., 2020) was specifically done for the English language. While using T5 as a transfer learning strategy may seem appealing, we made the decision to construct an encoder-decoder model from scratch for two specific reasons. First, the T5 model has a multilingual variant known as mT5(Xue et al., 2021). Nevertheless, the performance of the model on non-English tasks remains uncertain, as it has not been benchmarked against monolingual language models that have been specifically trained on diverse non-English situations. In addition, Kreutzer et al. (2022) identified systematic issues

with the multilingual corpora used for training language models. Furthermore, our research focuses on enhancing the precision of the ASAG by exploring paraphrase-generating techniques specifically for the Arabic language.

Our objective is to develop a model that accurately represents common approaches to paraphrase production. We refrained from doing research on using parallel mT5 for generating Arabic paraphrases and examining the influence of paraphrases on the accuracy of ASAG. Transformers have a tendency to over fit, especially when trained on smaller datasets (Xu et al., 2021).

Undoubtedly, the self-attention method necessitates a substantial amount of computing, making it more computationally demanding compared to conventional encoder-decoder models (X. Huang et al., 2022). Training efficiently necessitates a substantial quantity of data. This might provide difficulty with the Arabic language, as there is a lack of available data, which may restrict their usefulness in contexts with limited resources. We conducted training on an attention-based standard encoder-decoder model and evaluated its performance on datasets in both Arabic and English languages.

3.5.2.1 Sentence Encoder

The encoder is composed of a gating layer that embeds the input information and an embedding layer that represents words.

Embedding layer. The objective is to construct embeddings from one-hot word representations. A one-hot vector is a 1×N vector where all elements are 0 except for a single 1 at the position corresponding to the word in the vocabulary. The vector consists of 0s except for a single 1 in a cell used to index the word in the vocabulary. Once the dataset is encoded, the one-hot vector for each word can be generated based on its index in the vocabulary (of size N), with the 1 positioned at the word's index. As an input to the embedding layer, each word of the input reference text is introduced by its one-hot vector to the neural network to generate a reduced representation (the embedding) while keeping the semantic links between the words in the text. We choose to retain a dimension of 256 for the vector space to enhance the representation of words in the embeddings. In figure 9, the words of the source sentence $S = \{s_1, s_2, s_3 \ldots s_n\}$ are represented by One-hot vectors which are coded into embeddings $\{E_{S1}, E_{S2} \ldots E_{Sn}\}$. The generation of embeddings in the model aims to capture similar word distributions, including synonyms, to enhance the representation of words.

Gated Recurrent Unit Layer (GRU). Gated Recurrent Neural Networks (GRUs)) (Chung et al., 2014) offer a solution to the gradient vanishing problem. GRU is a simplified version of LSTM with only two gates: a reset gate and an update gate for resetting and updating the cell's hidden state. LSTM uses three gates (forget gate, input gate, and output gate). GRU is less computationally expensive than LSTM and requires fewer parameters to train. GRUs can outperform LSTM networks on low-complexity sequences (Chung et al., 2014; Cahuantzi et al., 2021). We selected GRU cells due to their superior performance in terms of convergence time and iterative efficiency compared to other cell types. The number of GRU nodes used in the model is fixed at 1024. Hyper-parameters are defined as:

- *For optimization,* we employ the ADAM (Adaptive Moment Estimation) stochastic optimizer (Kingma & Ba, 2015)
- *Loss function.* The "sparse Categorical cross-entropy" function is used (formula (5)) to calculate the loss during the training process:

$$-\frac{1}{N}\sum_{i=1}^{n}\sum_{c=1}^{n} 1_{y_i \in C_c} \log P_{model}[y_i \in C_c] \qquad (5)$$

- Where n, C, and P are respectively the number of observations, the number of classes corresponding to the number of different words in the vocabulary dataset used in the one-hot representation, and the probability of the observation "i" relative to the class "c".
- *Batch size. The model trains progressively on dataset batches of the same size (64 pairs of sentences).*

Attention mechanism. The encoder-decoder network's performance deteriorates dramatically as the length of the input phrase rises (Cho et al., 2014). The issue is that the decoder stage only uses the most recent hidden state that the encoder generated for context. In the case of long input sequences, the encoder struggles to retain all essential information required for output generation until the final hidden state. To overcome this limitation, an attention mechanism is integrated into the encoder (Bahdanau et al., 2015).

The attention mechanism takes into account the entirety of the information contained in the concealed states at various time steps. During each time step, the attention mechanism leverages all hidden states of the encoder to generate a context vector. Alignment scores are computed by comparing the decoder's preceding hidden

state with all the hidden states of the encoder. In order to produce a word, careful consideration is given to each word in the input sequence. The weights given to the encoder, which provide a score for each hidden state, indicate attention. This guarantees that hidden states necessitating attention receive high scores for prioritization. The SoftMax function creates attention weights, which the encoder then applies to the scores it provides. Once all the attention weights have been computed, a context vector is derived using formula (6):

$$Context\ Vector = \sum_{i=1}^{n} P_i h_i \qquad (6)$$

h_i: Hidden state at time-step i,
n: represents the number of words in the source sentence, and
P_i: Weight of hidden state h_i

Each node (GRU_i) of the decoder, except the first, has as input, the output of the previous node(g_{i-1}), and the context vector generated by the attention mechanism (C_t) at time step t. The first node of the decoder (GRU_1) receives as input the last hidden state of the encoder with the first context vector (C_t) generated by the attention mechanism.

Figure 7 demonstrates that the model takes into account important terms in the original text in order to create new words. In the resulting paraphrase, the model may focus on the terms "most" and "appreciated" in the original sentence, which have significant attention weights in the context vector, to create the word "friendliest". The attention mechanism allows the decoder to selectively concentrate on terms that are very relevant when creating a word ("most", "appreciated", "employee", "firm"). The importance of each word in the source sequence is calculated at each time step, highlighting the key information ("most", "appreciated", "employee", "firm") and downplaying the unnecessary information ("is", "the"). In the given example (Figure 9), the words "official" and "company" are produced as synonyms due to the strong effect of the words "employee" and "firm" that are already present in the original text.

3.5.2.2 Sentence Decoder (Generator)

Each stage of the decoding process makes use of the contextualized representation by combining it with the vector embedding of previously created words. A distribution of probabilities is obtained throughout the vocabulary, and the term with the highest

probability is generated. The decoder is composed of three layers: an embedding layer, a GRU layer, and an output layer:

Embedding layer. The decoder has an embedding layer that generates the corresponding embedding vector from the digital representation of each word (g_i) of the paraphrase.

The GRU layer. The encoder vector, the final concealed state that it produced, is the first thing the decoder receives. The latent state encapsulates the essential information included in every word of the input phrase. In order to produce a word at time step t, the decoder takes as input the hidden state, the output created at the previous time step, and the embedding of the previously formed word of the paraphrase.

The output layer. Following the output of the GRU layer, information is sent via the SoftMax function. This function serves as a classifier in the output layer of the decoder, predicting a multinomial probability distribution over integers that represent vocabulary items. To illustrate, in order to produce the word "official", the decoder takes as input the embedding of the previously formed word "friendliest," the context vector Ct, and the most recent GRU hidden state. When computing the distribution vocabulary using SoftMax, the Ct attention context vector will provide a substantial probability for the word "official," which is a synonym for "employee". The greatest probability value determines the output word. Thus, we shall produce a solitary paraphrase.

Although this strategy is frequently successful, it is suboptimal. We use the Beam Search Decoding Algorithm (Ashwin et al., 2018) to do an approximation search, which generates several paraphrases. The beam search method systematically explores all potential subsequent actions, retains the generated sequences, and regulates the quantity of parallel searches (known as beams) based on the sequence of probability. Multiple paraphrases are created based on the beam's size.

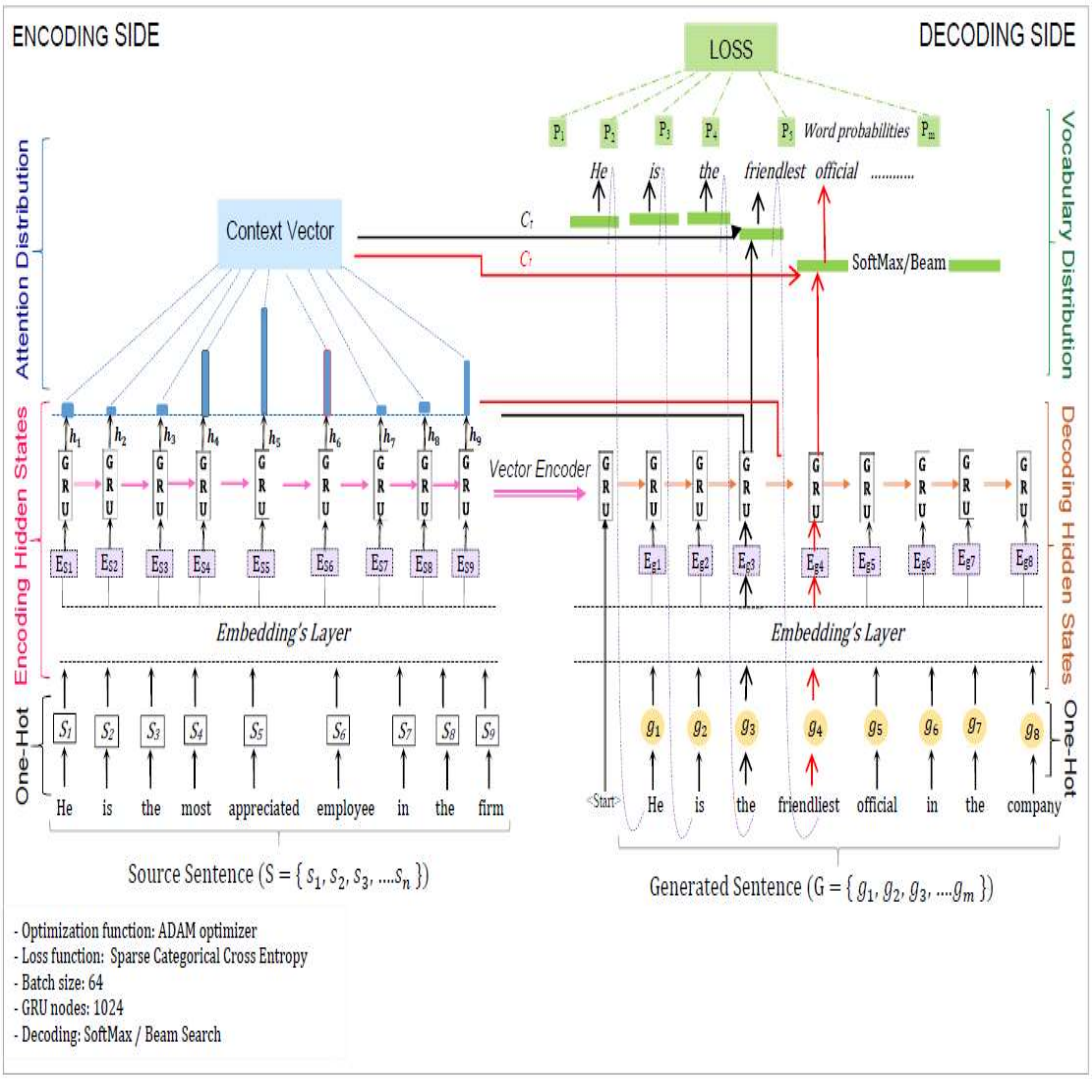

Figure 8 Proposed Alternative Sentence Generator Encoder-Decoder attentional model (Ouahrani & Bennouar, 2024).

3.6 TECHNICAL DESIGN

3.6.1 Integrating the paraphrase generator ARAG-ED into the grading System

As illustrated in Figure 10, the proposed grading model is implemented to support the ASAG tool. The ASAG tool is included in the online quiz system of our university's Learning Management System (LMS). The source code for the ASAG tool, including trained models for both English and Arabic, is shared here[18]. The ASAG tool incorporates the ARAG-ED generator (Ouahrani & Bennouar, 2024). The suggested grading mechanism is integrated into the Question Engine, which serves as the core

[18] https://github.com/leilaouahrani

module of the quiz system. The ARAG-ED algorithm produces many alternative reference responses based on a given reference answer. These alternative solutions are then combined with student answers and fed into the trained model. The model remains static and does not require retraining with multiple responses. Student response grading is conducted by comparing it to the reference answer. It returns the highest score obtained.

Scoring activities adhere to the process illustrated in Figure 7 previously presented. Various linguistic pre-processing techniques, such as sentence cleaning, stop-word removal, sentence tokenization, and sentence stemming, are employed to prepare answers for feature extraction. The Paraphrase Generator can assist teachers beyond the grading process by generating reformulated versions of their reference responses. This allows teachers to review and validate alternative model answers before use with the scoring system.

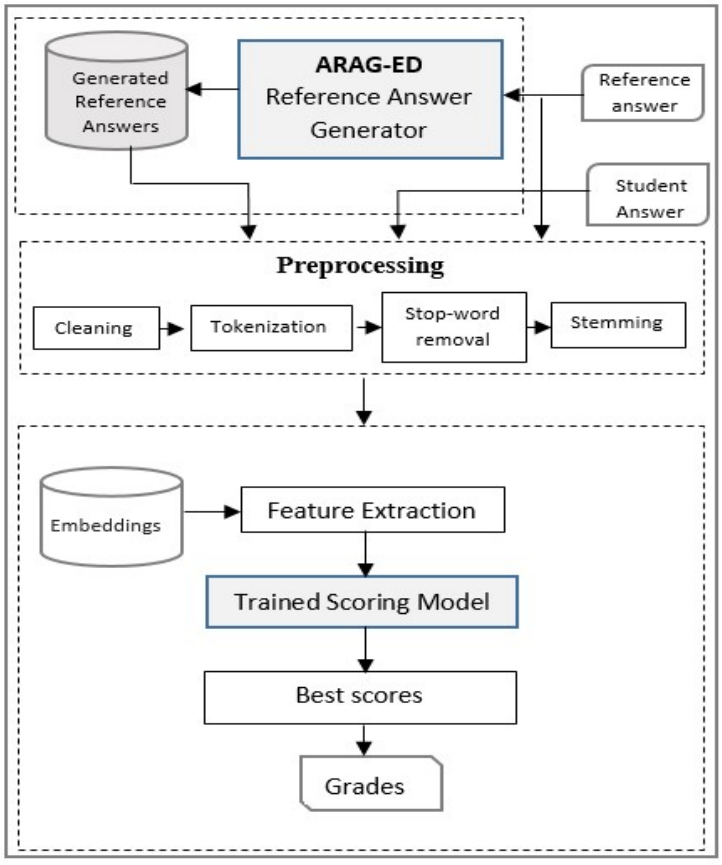

Figure 9 Integrating paraphrase generator ARAG-ED into the ASAG tool

3.6.2 Integrating the ASAG system into the Learning Management System

Although several LMS currently have automated grading systems, their functionality is generally limited to particular categories of objective questions. Integrating an ASAG system into an open-source LMS such as Moodle is a first step in tackling this problem. In line with most public universities in our country, our university has access to an e-learning environment that includes a Moodle Learning Management System. Figure 10 provides an architectural overview of the currently deployed ASAG system, highlighting two key conceptual modules: The Question Type ISAGe Plugin and the external grader.

The Question Type ISAGe Plugin. Guarantees transparency with the LMS and expands the quiz system's question engine to include the proposed ASAG question type. The quiz system's plugin manages communication between its many modules and takes over the LMS question bank, quiz reports, and question behavior.

The proposed ASAG Question Type Plugin has been added to the LMS Question Engine. It extends the quiz system's question engine by integrating the ASAG model; ensuring seamless interaction with the LMS. This integration demonstrates the feasibility of developing quizzes that incorporate ASAG-supported questions alongside existing question types, such as essays and multiple-choice questions, within the Learning Management System.

The plugin handles interactions with other quiz system modules and extends LMS functionality by inheriting from components such as question behavior, the question bank, and quiz reports. Both teachers and students utilize the system for assessments in a manner similar to other LMS activities.

Integrating the ASAG plugin into the quiz system offers the benefit of distinguishing between question types and question behaviors as separate concepts (as explained in Appendix A). Teachers can determine the question behavior, choosing options like "interactive mode" for instant feedback and multiple attempts, or "delayed mode" for a single attempt with feedback provided after submission. They can also impose penalties, assign weights to model answers with coefficients, and decide the sequence of model answers based on the assessment's objective (formative or summative). The plugin provides interfaces in Arabic, English, and French, adhering to Moodle LMS standards.

The external grader. It executes the scoring process. Deployed on the Cloud, it utilizes the trained grading model to predict grades. The Web gateway to the clients is provided by the API framework, which permits hosting the grader on the Cloud. The Control and View component is handled by HTTP requests from the client-side (the LMS), to the server (Hosting Cloud). The Integrated Development Environment is used as "PaaS" (Platform as a Service) where the grader ISAGe runs as "a service" separately and provides smooth integration to the LMS. The two modules communicate through a cURL interface; a command-line tool for getting or sending data using URL syntax. It (cURL) supports HTTPs to transmit students' answers and reference answers to the external grader; then it asynchronously waits for grades from the cloud and returns them to the LMS quiz system. When the evaluator operates in the cloud, it optimizes memory usage and execution time on the LMS side. The grades and feedback must be delivered within a limited time so that students do not become impatient. The current implementation of ISAGe provides comprehensive help to instructors and students throughout the whole assessment process for short free-text answers. This support is consistent across different question types, ensuring that instructors and students receive uniform assistance regardless of the question format. Both teachers and students use the system and carry out their respective role assessment assignments in a manner similar to how they access other activities. Additionally, the ASAG tool can also be deployed as a desktop application (e.g., in a classroom setting) without the need for an LMS platform or internet connectivity.

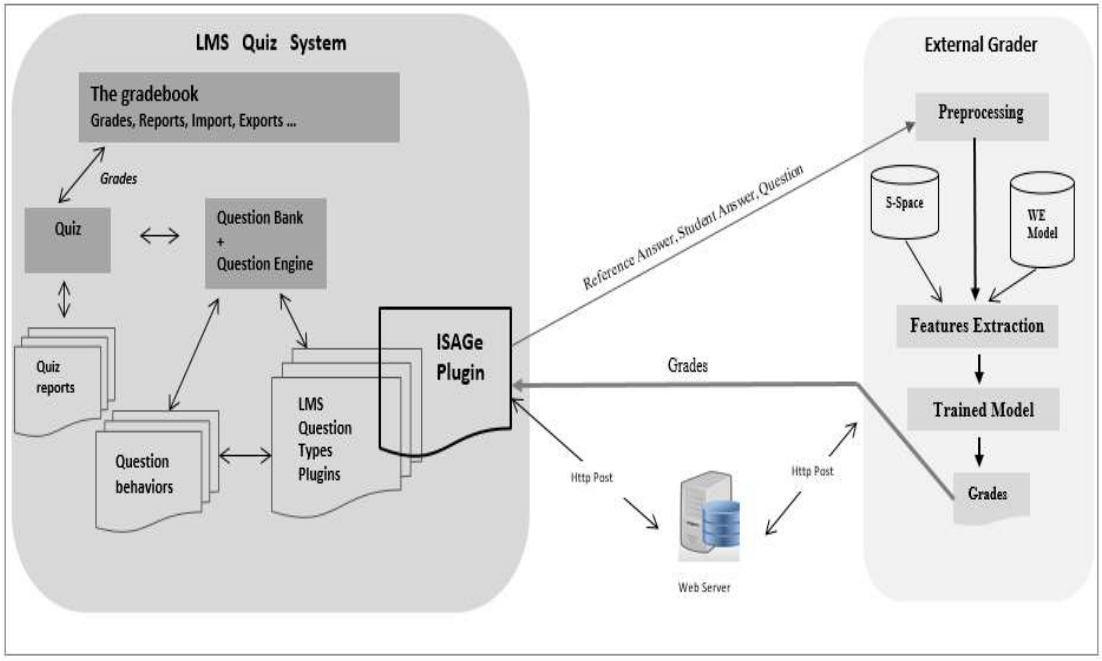

Figure 10 The ISAGe Architecture Overview.

Chapter 4: Results and Evaluation

In this chapter, we present the results and engage in discussions stemming from our empirical analysis, marking the culmination of our research efforts. The chapter is structured as follows: (1) Experimental Setup, (2) Dataset Evaluation and Baselines, (3) Intrinsic Evaluation of Paraphrased Reference Answers, (4) Evaluation of the Supervised Learning Model, (5) Analysis of Grading Errors and Limitations, and (6) Overall Discussions and Implications.

4.1 EXPERIMENT SETUP

The evaluation of the proposed approach within our incremental design framework involves assessing the effectiveness, performance, and suitability of design choices at different stages of the study. Our experimental design focuses on evaluating incremental goal alignment (addressing the research questions), grading performance testing, integration testing, and user feedback. By following a series of evaluation steps, we ensure that the system evolves iteratively, building upon the insights gained from previous iterations.

We conducted two types of experiments: quantitative and qualitative. Quantitative investigations assess scoring accuracy using datasets. The qualitative experiments analyze the effect of incorporating the ASAG on students' academic performance based on both formative and summative evaluations. The proposed methodology is founded upon a practical second case study carried out at the Bouira University. We used human expertise to conduct qualitative evaluation for the paraphrase generation task.

Baseline. Unsupervised grading model using COALS word distribution.

To determine answer-to-answer similarity, we integrate the vector summation model with syntactic similarity based on common words between the model answer (MA) and the student answer (SA). The vector summation model involves summing the context vectors of each word in both answers and then computing the cosine similarity between these vector sums. Additionally, we combine the Dice coefficient to emphasize cases with a significant number of common words between the answers, as it measures syntactic similarity based on shared terms. For the grading task, we utilize

the unsupervised K-means classifier, chosen for its effectiveness in cluster analysis. The value of K is set to 11 and remains constant, with each cluster representing one of the 11 possible scores (0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5) determined by the algorithm.

We utilize the corpus-based COALS word distribution method to derive word vector contexts. For evaluation, we create multiple semantic spaces, each varying in dimension and domain. In this section, the AR-ASAG dataset is employed to evaluate the grading model. The results are compared and analyzed from several perspectives: semantic space dimensions and domain specificity, word space distribution quality, term weighting and stemming effects. An unsupervised approach that maximizes the scores from the summation model is applied. Two versions of the combined model are proposed: one, referred to as the Baseline, includes term weighting as specified in the proposed features (Section 3.3.3 Word Weighting Features), and the other, referred to as U-Baseline (Unweighted-baseline), does not incorporate term weighting. The basic system refereed to the cosine similarity between the model answer (MA) and the student answer (SA).

Metrics. Pearson Correlation Coefficient (r) and Root Mean Square Error (RMSE)

Evaluation workflow. The evaluation workflow comprises the following steps:

1. **Assessment of Baselines: Assess and implement baselines for the Ar-dataset.**
   - Assess and implement baselines for the Ar-dataset.
   - Explore linguistic factors affecting evaluation quality and word distribution using the COALS approach, including stemming, term weighting, domain specificity, and word space dimensionality.
   - Unsupervised Kmeans scoring model to isolate the effects of the proposed features.

2. **Evaluation of the Paraphrase Generation Model:**
   - Conduct intrinsic quantitative and qualitative evaluation of the paraphrase generation model in relation to the NLP task.
   - Compare the results with those from existing literature in the field.

3. **Evaluation of the Supervised Regression Model:**
   - Evaluate the supervised regression model and the quality of the proposed features.
   - At this stage, the model utilizes only a single response.

4. **Assessing the Impact of the Paraphrase Generation Model:**
   - Perform extrinsic evaluation by analyzing the impact of the paraphrase generation model on the supervised ASAG model.
   - Assess the impact on the grading accuracy.
   - Compare the results with state-of-the-art in the ASAG field.
   - Analyze error grading and limitations.
5. **Evaluating the Approach's Integration into Real Educational Environments:**
   - Conduct qualitative evaluation of the approach's integration into a real-world environment. (Case study 2).
   - Assess the educational impact of the approach.

4.2 DATASET BASELINE AND COALS WORD DISTRIBUTION EVALUATION

4.2.1 Semantic space dimension and domain specificity evaluation

When utilizing corpus-based methodologies, it is crucial to take into account the impact of both the size and topic domain on the overall effectiveness of the system. The COALS algorithm should demonstrate flexibility by easily adjusting to changes in domain and semantic space dimensions due to its operational mechanism.

In-domain CYBER Corpus. We developed the Arabic Cyber Text Corpus *(Arabic Cyber Text Corpus, 2020)*[19] registered under the ISLRN 798-080-268-332-8, to address the absence of a dedicated corpus in Arabic for cybercrimes according to the topic of the dataset. We automated the acquisition of the domain-specific corpus by extracting texts from a set of URLs using predefined key phrases. The corpus was enriched by integrating various course notes related to the topics covered in the AR-ASAG dataset. To experiment domain and dimension effect, we used three publicly available corpora, specifically (BBC Arabic, CNN Arabic)[20], and Khaleej[21]. Semantic space are generated and stored in textual database of word vector context. The Vector-Context database consists of three columns: ID, Word, and Context Vector. The context vectors are stored as LONGTEXT variables and indexed based on the ID and Word columns to provide efficient time access and loading during similarity

[19] https://www.islrn.org/resources/798-080-268-332-8/
[20] https://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/
[21] https://sites.google.com/site/mouradabbas9/corpora

calculations. Table 10 provides an overview of the attributes of corpora utilized in constructing various semantic spaces for experimentation. In Table 11, we examine the corresponding semantic spaces generated from the four corpora, with dimensionalities ranging from 13733 to 28062. The basic model (cosine similarity) achieves optimal results by increasing dimensionality, with 17225 dimensions using light stemming and 23715 dimensions using root stemming. There has been a little alteration in performance. The performance gradually decreases when we decrease the vector dimension to 13000. This finding strongly supports the notion that the COALS method (Rohde et al., 2004) yields similar performance results when used with dimensionalities ranging from 14,000 to 100,000. Comparing the size of the Khaleej space to the CYBER space, we observe that the CYBER space is more effective for both stemming approaches. Upon comparing the data, it is evident that utilizing the in-domain CYBER space yields a correlation of r = 0.6550 and an RMSE of 1.10. This correlation is stronger than the correlation of r = 0.6379 and an RMSE of 1.14 obtained from a larger corpus focused on the broad domain. This implies that, in the context of the COALS algorithm, the significance of the texts lies more in their quality than their number.

Table 10 Corpora Characteristics

	BBC Arabic	CNN Arabic	Khaleej	In-domain Cyber
Number of words	1 860 000	2 241 348	3 000 000	2009110
Number of documents	4763	5 070	5000	1273
contents	- Middle East News - News of the world - Economy &work - Sports - International press - Science & technology - Arts and cultures	- Middle East News - News of the world - Economy and works - Sports - Science & technology - Arts and cultures - Leisure	- International News - Local News - Sports - Economy	- Cyber crimes - Information Security culture - Cybercrimes classification - Cybercrime Algerian legislation

Table 11 Basic system results for different semantic spaces on AR-ASAG Dataset

	Root Stemming			Light Stemming		
Corpus	Vector Dimension	Pearson	RMSE	Vector Dimension	Pearson	RMSE
khaleej	13733	0,6306	1,13	18630	0,6087	1,21
cnn	16752	0,6317	1,14	21032	0,6090	1,20
bbc+cnn	24230	0,6379	1,14	28062	0,6115	1,22
Cyber	17225	**0,6550**	**1,10**	23715	**0,6340**	**1,14**

4.2.2 Word Space Distribution Quality

All the results presented in the subsequent sections are calculated using the CYBER semantic space. We compare the quality of word distribution against Zahran WE (Zahran et al., 2015) on the Ar-ASAG dataset and the Disco word space on the Arabic Cairo university dataset (Authors accepted to share the Cairo dataset).

COALS Word Space Distribution vs. Word Embedding. Zahran et al. (2015) employed the CBOW, SKIP-G, and GloVe models (Mikolov, Sutskever, et al., 2013) to construct a multidimensional word representation in vector space for Modern Standard Arabic. The Word Embedding model shared publicly has around 6.3 million entries, with a total word count of around 5.8 billion. In this thesis, we shall refer to them as Zahran-WE.

We assessed the distribution of words in the semantic space using Zahran-WE (Zahran et al., 2015). In our basic system, we replaced the words in the semantic space vector with Zahran-WE, derived from the Skip-Gram model. We then employed the AR-ASAG dataset to determine the correlation. The model mentioned here is refereed to Z-SkipGram Basic. The outcomes the comparison between the Basic system and Z-SkipGram Basic, are detailed in Table 12. The Basic system, utilizing the Cyber semantic space, surpasses Z-SkipGram Basic (root stemming: Pearson correlation coefficient increase of 0.03, RMSE increase of 0.2; light stemming: Pearson correlation coefficient decrease of 0.00008, RMSE increase of 0.09). This finding indicates the quality of the word distribution in semantic space, highlighting the effectiveness of the COALS model.

Table 12: Model performance using WE vs. COALS on the AR-ASAG Dataset

	Root Stemming		*Light Stemming*	
	Pearson	RMSE	Pearson	RMSE
Z-SkipGram WE	0.6281	1,30	**0.6348**	1,22
CYBER Space	**0.6550**	1.10	0.6340	**1.13**

COALS Word Space Distribution vs. Disco Word Space. DISCO (Kolb, 2008) is a similarity tool that quantifies and identifies terms with high distributional similarity to a particular word based on co-occurrences (the Disco space). It uses a pre-processed collection of Wikipedia data containing 267 million words and 220,000 distinct terms.

DISCO is compatible with nine languages, one of which is Arabic22. The tool utilizes statistical analysis of large text sets to evaluate the similarity between words. The Lin measure (Lin, 1998) is used to evaluate the similarity of words derived from vectors in the indexed data, specifically in the DISCO Word Space. DISCO utilizes two main similarity measurements: DISCO1 and DISCO2. The DISCO measure is widely used for evaluating ASAG and similarity tasks in Arabic (Gomaa and Fahmy, 2014b; Magooda et al., 2016; Zahran et al., 2015; Elghannam, 2016). It seems essential to assess our proposed COALS in comparison to the DISCO Word Space. Gomaa and Fahmy (2014b) presented their findings on the Cairo Arabic dataset, utilizing DISCO1 and DISCO2 similarity metrics on both the Arabic and translated English datasets.

According to the results presented in Table 13, our model's correlation is better than the highest achieved results by Gomaa and Fahmy (2014b) using DISCO1. Specifically, our system surpasses their findings by 8.07% on the translated dataset and by 13.07% on the Arabic dataset. Despite the RMSE of our basic system being -0.23, we interpret these findings as indicative of the good quality of the created COALS semantic vectors due to the significant improvement in correlation. It is crucial to note that Gomaa and Fahmy (2014b) obtained the findings shown in Table 13 using the same K-means clustering scaling technique that we did, ensuring the comparability of results based on a shared methodology.

Table 13 COALS vs. Disco on the Cairo University Dataset

		r	RMSE
IAA (Arabic and English Data set)		86.00	0.69
(Gomaa & Fahmy, 2014b) (English translated from Cairo)	Disco 1	68.00	0.84
	Disco 2	67.00	0.84
(Gomaa & Fahmy, 2014b) (Cairo Arabic)	Disco 1	63.00	0.88
	Disco 2	61.00	0.86
Our unsupervised Model (Cairo Arabic)		76.07	1.11

4.2.3 Unsupervised Grading model Assessment

Here, we analyze the quality of the suggested unsupervised grading model from two perspectives. Firstly, we consider the impact of word weighting and stemming on the correlation in the AR-ASAG dataset. Furthermore, on the outcomes of the SEMEVAL

[22] https://www.linguatools.de/disco/disco_en.html

2017 competition (specifically track 1: Arabic-Arabic) using the STS 250 dataset. As short Answer Grading and text similarity tasks are strongly related in our proposed approach, we consider SEMEval-2017 (Semantic Textual Similarity-Multilingual and Cross-lingual Focused Evaluation) (Agirre et al., 2017) Workshop which makes available STS 250 SEMEval 2017 Dataset[23]for track 1 Arabic-Arabic (Cer et al., 2017).The Dataset contains 250 pairs of sentences obtained by translation from English into Arabic. For each pair, a manual gold score that averages five human annotations is given.

Term Weighting Effect. The performance of the unsupervised model compared to the Inter-Annotator Agreement on the AR-ASAG dataset is presented in Table 14. Term weighting resulted in a significant enhancement in correlation. For root stemming, the Pearson coefficient increased by 0.028, and the RMSE increased by 0.04. For light stemming, the Pearson coefficient increased by 0.0478, and the RMSE increased by 0.07. The utilization of word weighting resulted in the system achieving its highest correlation of r = 0.7037. The root mean square error (RMSE) was significantly enhanced to 1.0240, an increase of 0.14. The retained baseline as a Pearson of 70.37% and a root mean square error (RMSE) of 1.0454.

Table 14: Baseline evaluation on the AR-ASAG Dataset

		Pearson	***RMSE***	***Av (RMSE)***
IAA (Manual scores)		0.8384	0.8381	*0,5629*
Basic System (Cosine)	*Root Stem.*	0.6550	1.10	
	Light Stem.	0.6340	1.14	
(Unweighted-baseline)	*Root Stem.*	**0.6830**	**1.06**	
	Light Stem.	0.6818	1.07	
Baseline (term weighting)	*Root Stem.*	0.7010	**1.0240**	***0.7841***
	Light Stem.	**0.7037**	1.0454	***0,8039***

Stemming Effect. . For the Arabic language, automating the identification of a word's root or stem is particularly challenging. In our study, we used both light stemming and root stemming techniques to analyze their effects on automatic short answer

[23]http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

grading.We used KHOJA' Stemmer [24](Khoja & Garside, 1999) for root stemming and (Zerrouki, 2010)[25] for light stemming.

Tables 11, 12 and 13 demonstrate that root stemming outperformed light stemming in the basic system. However, it is interesting that when term weighting and syntactic similarity were combined, light stemming yielded equivalent results (+0.0027) with a lower RMSE (-0.0214). Furthermore, we include the Root Mean Square Error (RMSE) for the dataset as well as the median error (RMSE) for each individual question. This provides an indication of the system's performance, enabling the observation of a single question in isolation. The average root mean square error (RMSE) for the root stemming (0.7841) is superior to the average RMSE for the light stemming (0.8039), but with a slight disparity. The average root mean square error (RMSE) indicates a small discrepancy for root stemming, but the Pearson correlation coefficient suggests the contrary. The distribution of human and automatic grades for both stemming approaches is shown in Figure 11. The two corresponding curves exhibit similar trends and are nearly indistinguishable. The basic system correlation was very responsive to the steaming procedure. The results were consistent when using term weighting and combination similarity.

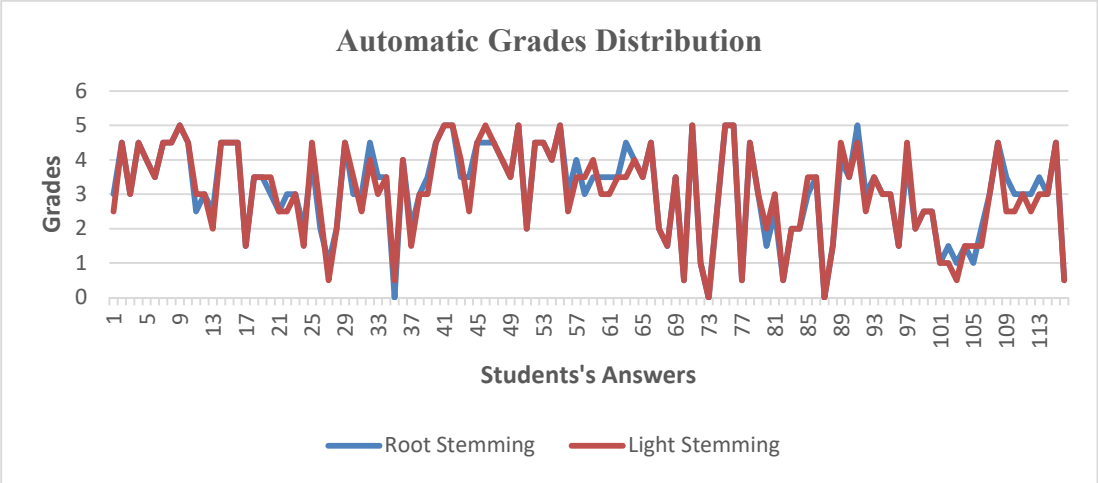

Figure 11 Automatic Grades Distribution : Root Stemming vs. Light Stemming

Comparing to STS 250 SEMEval Dataset. The results obtained on the general STS 250 SEMEval 2017 dataset are reported in Table 14. The unsupervised system outperformed the SEMEVAL baseline by 11.75% but lagged behind LIM-LIG

[24] http://zeus.cs.pacificu.edu/shereen/research.htm#stemming
[25] https://pypi.org/project/Tashaphyne/

(Nagoudi et al., 2017) by -2.43%. LIM-LIG achieved the second-highest score in Track 1. LIM-LIG used a vectorized word embedding-based approach similar to ours. Additionally, our system performed 3.23% lower than the winner of Track 1 (Huang and Su, 2017), who used a topological approach. The results from the STS 250 SEMEval dataset, which we deem satisfactory, reveal two key insights. Firstly, the proposed method demonstrates good generalization capabilities. Secondly, the AR-ASAG Dataset's quality is highlighted, as the grading model performs similarly across both datasets. Specifically, the system achieved a Pearson of 0.7037 and an RMSE of 1.0240 on the AR-ASAG Dataset, and a Pearson of 0.7220 and an RMSE of 1.03 on the SEMEval Dataset, both with comparable RMSE scores.

Finally, we approached the Automated Short Answer Grading (ASAG) task using an unsupervised learning method, specifically the K-means algorithm, to cluster answers based on the COALS word distribution. Although the initial results established a baseline and validated our choice of word distribution, they also highlighted the need for further refinement to enhance clustering accuracy. A more detailed discussion and our conclusions can be found in the "4.6. Overall Discussion and Implications" section. More detailed evaluation are available in (Ouahrani and Bennouar 2018; Ouahrani and Bennouar 2020).

Table 15 Baseline Evaluation on STS 250 SEMEval 2017 Dataset (Task 1)

	Pearson	RMSE
SEMEVAL 2017 track 1 Baseline	0.6045	-
SEMEval 2017 Winner Track 1 BIT System (Huang and Su, 2017)	0.7543	-
SEMEval 2017 2nd score Track 1 LIM-LIG (Nagoudi et al., 2017)	0.7463	-
Unsupervised model (combined)	**0.7220**	**1.03**

4.3 INTRINSEC EVALUATION OF GENERATED ALTERNATIVE REFERENCE ANSWERS

This section presents an intrinsic evaluation of the paraphrase generator for the task of paraphrase generation. Experiments are conducted in both the Arabic and English languages. English is chosen as a reference language due to its widespread use and abundance of resources and references, facilitating comprehensive comparisons with existing research. This comparison allows us to directly align our results with prior

studies, affirming the robustness of our model across multilingual contexts. In this section, we transition to an overview of the datasets to train and evaluate our model used and the review findings.

DATA. Two datsets are used to train the model in Arabic and English:

Al-Raisi Arabic Dataset[26] (Al-Raisi, Lin, et al., 2018). This is the first parallel monolingual corpus consisting of complete sentences in Arabic. The dataset consists of 100,000 pairs of sentences, where each pair includes an original phrase and its corresponding paraphrased reference sentence. This dataset is the most extensive parallel Arabic corpus that is accessible to the public. The dataset was created automatically by utilizing Google Translate APIs to translate a parallel bilingual corpus consisting of English and French texts into Arabic.

The Quora English Dataset[27]. The dataset comprises more than 400,000 pairs of phrases, consisting of an original text and its corresponding reference paraphrase in English. Each pair is labeled with a binary value, denoting whether the two statements are paraphrases of one another. Only sentence pairings that are paraphrases of each other are chosen to train the generator. To prioritize concise responses, we have excluded sentence pairings that exceed a length of 50 words from the datasets. As a result, 350,000 pairings were chosen from the Quora dataset, along with 77,371 pairs from the AL-Raisi dataset. 60% of each dataset was allocated for training, 20% for validation, and the remaining 20% for testing. These proportions are illustrated in Table 16.

Data pre-processing tasks include cleaning the data, removing lengthy phrases, normalizing the text, converting it to lowercase (for English), and tokenizing it. Preprocessing is done to enable one-hot encoding by extracting the vocabulary, which encompasses all unique words in the dataset. Sequential integers are assigned to each word in the retrieved vocabulary, establishing a numerical representation for the words. Subsequently, each input text is represented by a series of numbers that indicate the positions of words in the extracted vocabulary. The training was carried out utilizing the cloud-based service Google Collaboratory (Carneiro et al., 2018) with a Nvidia Tesla K80 GPU and 12GB of DDR5 RAM.

[26] http://www.cs.cmu.edu/~fraisi/arabic/arparallel/
[27] https://github.com/jakartaresearch/quora-question-pairs

Table 16 Dataset portions that are used to train and test the paraphrase generator

Dataset	Training (60%)	Validation (20%)	Test (20%)	Total (100%)
Quora dataset (English)	210,00	70,000	70,000	350,00
Al-Raisi Dataset (Arabic)	46,423	15,474	15,474	77,371

Evaluation Metrics. To assess the quality of the generated reference answers, we employed widely used metrics from the literature examined in our research study: BLEU (Papineni et al., 2002), GLEU (Napoles et al., 2015), and METEOR (Lavie & Agarwal, 2007).

Baseline. Currently, there is limited research available on the production of Arabic paraphrase, creating a gap in the existing literature. To establish a benchmark for comparison, the Bi-LSTM neural network is employed as a reference model. Bi-LSTM neural networks process sequential data using two LSTM sub-layers: one for forward input processing and another for backward input processing. The two concealed states of the two layers collectively retain information from both the preceding and subsequent time periods. Bi-LSTM is capable of acquiring long-term connections without the need to retain redundant contextual information. Each sub-layer is equipped with 256 LSTM nodes. The model takes into account all the words in the input phrase to predict an output word in both forward and backward directions. The model is composed of three layers: the embedding layer, responsible for converting input words into embeddings; the Bi-LSTM layer, which generates words; and the classification layer (SoftMax) that determines the output words. Word2Vec continuous Skip-gram pre-trained word vectors from the NLPL word embeddings repository [28] were utilized in the embedding layer, as detailed in Table 17. We utilized two pre-trained word embedding models, specifically the CoNLL17 English corpus and the CoNLL17 Arabic corpus. The embeddings have a dimension of 100 and encompass the mapping of 4,027,169 English words and 1,071,056 Arabic words. They are designed to represent the semantic and syntactic characteristics of the respective languages. We employ the ADAM optimizer with the Categorical Cross-Entropy loss function, utilizing a learning rate of 0.001.

[28] http://vectors.nlpl.eu/repository/

The model trains iteratively on fixed batches of 16 pairings. The nodes are discarded with a dropout probability of 0.3 (30%). This process applies regularization to the model in order to obtain the average predictions from all parameter values and combine them to produce the result. This guarantees that the model is generalized and, therefore, mitigates the issue of overfitting. As a result, in order to address this issue, the dropout rate first adjusts the weights, producing larger final weights than anticipated. Subsequently, the network possesses the capability to generate precise forecasts. Hence, when a unit is preserved during training using dropout, its outgoing weights are likewise scaled by the same dropout probability during inference. Inference during training multiplies a unit's outgoing weights by the dropout probability.

Table 17 Pre-trained used Word Embedding

Word Embedding	**Model**	**Dimension**	**Words**
English CoNLL17 corpus	Word2Vec Skip-gram	100	4,027,169
Arabic CoNLL17 corpus	Word2Vec Skip-gram	100	1,071,056

4.3.1 Automatic Evaluation of Generated Paraphrases

4.3.1.1 Results

An example of generated paraphrases in Arabic and English using ARAG-ED is presented in Table 18. Note that the training dataset is made up of pairs (source sentence, reference sentence). The output that the trained model produces is the generated sentence. We used the BLEU, GLEU, and METEOR metrics to automatically rate the three models: Bi-LSTM baseline, ARAG-ED without attention mechanism, and ARAG-ED with attention mechanism. These metrics always take a value between 0 and 1. Values closer to 1 indicate a higher similarity between the predicted text and the reference texts. Note that the Meteor metric is not calculated for the Arabic dataset since the calculation depends on the WordNet knowledge-based model. The limited richness of the Arabic version of WordNet affects the evaluation results by introducing limitations. The evaluation results are reported in Table 19.

For the Bi-LSTM model, the obtained results are very weak for all the metrics on the two datasets. The results were adversely affected by the poor quality of the pre-trained embeddings, leading to subpar performance. During the training, we noticed

that several words had no representation in word embeddings. In instances where words had no representation in the embeddings, they were substituted with zeros, affecting the overall output. We consider it a baseline to understand the advances in the suggested model.

A positive enhancement has been seen in the ARAG-ED model. It is important to mention that the model generates the embeddings within the embedding layer. The utilization of the attention mechanism has significantly improved the outcomes. The Arabic dataset achieved a BLEU score of 63 and a GLEU score of 59, whereas the English dataset achieved a BLEU score of 54, a GLEU score of 42, and a METEOR score of 42.

4.3.1.2 Generating multiple paraphrases with the beam search algorithm

In this analysis, we delve into the ARAG-ED model, highlighting its incorporation of an attention mechanism. Table 19 presents the results produced by employing the beam search-decoding technique to construct several paraphrases.

The model, trained on the Arabic dataset, generates outputs corresponding to beam sizes 1, 4, 7, and 10 for each source sentence. For the English dataset, the beam size is set to 1 and 10. The Average BLEU and the Best BLEU scores are computed. The terms "average" and "best" BLEU scores denote the quality levels of the multiple paraphrases generated. AVG-BLEU is the mean value obtained by averaging the BLEU scores determined for each pair of (generated sentence, reference paraphrase). Best-BLEU represents the top BLEU score attained by comparing produced sentences with reference sentences for every beam size. It is evident that the BLEU score declines proportionally with the rise in the number of produced paraphrases, ranging from 63% for beam = 1 to 53% for beam = 10. This phenomenon is expected because the average considers a spectrum of scenarios, including both optimal and less favorable outcomes. For Beam = 1, the SoftMax function prioritizes the probability assignment to the generated paraphrase.

Analysis of the BLEU scores (Lavie, 2010) indicates that the Arabic and English paraphrases produced are of exceptional quality, with BLEU scores of 53% and 49% for Beam = 10, respectively. According to the interpretation, BLEU scores exceeding 30 indicate understandable translations, while scores surpassing 50 signify good and fluent translations (Lavie, 2010). Therefore, a key aspect of our strategy involves integrating an attention mechanism with generated embeddings, leading to

enhanced model performance without the need for added computational complexity. This is crucial because the paraphrase generator is included in the online assessment system to enhance the accuracy of Automated Short Answer Grading (ASAG). In the context of implementing automated scoring in e-learning environments, it is crucial to consider aspects like server performance and practical efficiency to ensure effective system operation.

4.3.1.3 Comparison with previous work on the Quora dataset

English-related findings are sometimes the only accessible results in the literature review of paraphrase generation. The BLEU and METEOR scores obtained, with specific values, are displayed in Table 21. The scores obtained from our study are contrasted with those from similar research that employed other encoder-decoder types but were trained on the same dataset as ours. We conduct a comparative analysis with the VAE-SVG-eq (Gupta et al., 2018) and the GAP (Yang et al., 2020) results, focusing on their application to the Quora dataset, which has publicly accessible findings.

When the beam value is set to 1, the GAP model does not provide (report) any assessment. When evaluating the BLEU measure, it is apparent that our model outperforms both VAE-SVG-eq and GAP. Our model obtains higher scores, with an average BLEU improvement of 5.9% and the greatest BLEU improvement of 11% compared to VAE-SVG-eq. additionally, our model also surpasses GAP with a best BLEU improvement of 1.4%. This indicates that the decoder (generator) accurately replicates the n-gram alignments of the reference sentence in the created sentence. Furthermore, in terms of syntax, the sentences created are structurally correct and of high quality.

Utilizing METEOR, our model attains comparable scores to the best model but exhibits worse performance (Avgas-METEOR: -8%; Best-METEOR: -4.9%) in comparison to VAE-SVG-eq and (Best-METEOR: -2.94%) in comparison to GAP. The lack seen may be attributed to the caliber of the embeddings produced during the first phase. Increasing the number of training epochs may lead to higher-quality embeddings and enhance the METEOR score. The acquired scores undergo human manual evaluation, which will be explained in the next section.

Table 18 Examples of generated paraphrases using ARAG-ED

Arabic paraphrases	
Original sentence	ونحن لا نرغب في القيام بذلك وهكذا سنناقش مره اخرى هذه المسالة We do not wish to do so and so we will once again discuss this issue
Reference sentence	سنناقش مره اخرى هذه المسألة لأننا لا نرغب في القيام بذلك We will discuss this issue again because we do not wish to do so
Sentence generated by ARAG-ED (Arabic)	نحن لا تنوي القيام بذلك ونحن مره اخرى سنتحدث عن ذلك We do not intend to do that and we will talk about it again.
English paraphrases	
Original sentence	why do some people ask questions on Quora that could be asked directly to a search engine
Reference sentence	why do few people post questions on Quora check Google first
Sentence generated by ARAG-ED (English)	why do people ask questions on Quora that could simply be googled

Table 19 Proposed model evaluation for paraphrase generation task

	El-Raisi Arabic Dataset		Quora English Dataset		
	BLEU	GLEU	BLEU	GLEU	METEOR
Bi-LSTM (Baseline)	12	4	17	8	8
ARAG-ED (Without Attention Mechanism)	15	8	19	11	18
ARAG-ED (Beam=1)	**63**	**59**	**54**	**42**	**42**

Table 20 Proposed ARAG-ED-beam search decoding evaluation

	El Raisi Arabic Dataset		Quora Dataset			
BEAM	Avg-BLEU	Best- BLEU	Avg-BLEU	Best- BLEU	Avg-METEOR	Best-METEOR
1	63	63	54	54	42	42
4	59	-	-	-	-	-
7	55	-	-	-	-	-
10	53	54	43	49	24	28

Table 21 ARAG-ED vs. State-of-the-art on the Quora dataset - Comparative Analysis

	Beam =1		Beam =10			
			Average		Best	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
VAE-SVG-eq Gupta et al., (2018)	26,2	25,7	37,1	**32**	38	**32,9**
GAP Yang et al., (2020)	N/A	N/A	N/A	N/A	4 7,6	30,94
ARAG-ED	**54**	**42**	**43**	24	**49**	28

4.3.2 Manual Evaluation of Generated Paraphrases

Recent studies (Chaganty et al., 2018; Lai et al., 2022) have shown that automated measures like BLEU, GLEU, and METEOR are biased, which means that correlations between different systems are not always the same. Automated assessment measures prioritize n-gram similarities rather than semantic understanding. Consequently, artificial measures are prone to favoring certain systems over others, regardless of their true human assessment scores.

Arabic poses an additional challenge for natural language processing due to its intricate and extensive morphological characteristics. Arabic exhibits several word forms and word orderings, enabling the expression of each statement in multiple ways. Navigating Arabic morphology poses challenges, particularly in handling language tokens. Therefore, the token necessitates understanding the specific rules governing the combination of prefixes and suffixes in Arabic words.

Despite its widespread use, the BLEU scoring system does not take into account the concatenation limits specific to Arabic. BLEU calculates its score based on the matched n-grams in the texts being compared. It fails to examine the correct application of grammar and appears to provide precise evaluations for longer phrases. In our example, there is a sizable number of paraphrases that, in the judgment of humans, are correct but receive low scores from these metrics, indicating that the metrics have a low recall.

To enhance the accuracy of Automated Short Answer Grading (ASAG) with different reference answers, it is essential to conduct a qualitative assessment by human experts, as the effectiveness of paraphrases depends on specific grading. Human review is inherently more expensive due to the costly nature of the process, as opposed to automatic evaluation. However, it provides a more comprehensive assessment of the quality of the generated paraphrases across several aspects, such as relevance and readability (Babych, 2014). Relevance refers to the degree to which the paraphrase created aligns with the reference sentence. The objective is to assess the degree to which the produced sentence maintains the same meaning as the reference sentence. Readability refers to the level of understanding of the paraphrases produced, specifically about their structure and language. Thus, assessments that rely on human judgments serve as a valuable addition to automated evaluations that utilize metrics.

With a relevance score of 4.82 out of 5 and a readability score of 4.94 out of 5, Gupta et al.'s (2018) human evaluation revealed that the Quora dataset's paraphrases

were not entirely accurate. No manual qualitative assessment information is provided for the Al-Raisi Arabic dataset. In light of this, we performed a manual assessment of 100 sets of paraphrases that were chosen at random from the Al-Raisi Arabic dataset.

The assessment was conducted utilizing ARAG-ED, which was trained on the identical dataset known as Sample-Al-Raisi-Dataset-ARAG-ED in Table 22. We conducted a manual evaluation on a sample of 100 randomly chosen paraphrases generated by ARAG-ED, a model trained on the Quora English dataset. The paraphrases in question are denoted clearly as Sample-Quora-Dataset-ARAG-ED in Table 22. Three human experts were asked to evaluate the relevance and readability aspects of the sample of pairs (source, paraphrase) using a rating scale ranging from 1 to 5. We have retained these two components as they are found across the whole Quora dataset and VAE-SVG-eq (Gupta et al., 2018) on a randomly chosen subset. We conduct equivalent manual assessments for paraphrases in both English and Arabic. Indications to the experts for manual evaluation are presented in appendices B & C.

The human experts encountered difficulties when conducting manual assessments. The experts noted that the difficulty did not reside in determining if the rephrased language was clear and pertinent, but rather in precisely evaluating the degree of readability and relevance as compared to the original response. To assess the level of agreement among human experts, evaluations involve calculating Pearson's correlation coefficient, where a higher value indicates stronger consensus among the annotators. We calculated the Pearson correlation coefficient for every combination of annotators for all paraphrases that underwent manual review. The Pearson correlation coefficient among annotators suggests a robust link in terms of subjectivity. The experts who had the highest correlation had a consensus rate of over 68%. Experts with the weakest correlation earned a consensus rate of 63%. This is a logical conclusion, as subjectivity is inherent in any act of evaluation (Brown et al., 1999). To account for the potential margin of error resulting from subjectivity, we took into account the average of the three human manual scores shown in Table 22.

The English paraphrases achieved a relevance score of 3.58 out of 5, indicating that they retained 71.6% of the original meaning. This result aligns closely with the score obtained by VAE-SVG-eq on the Quora dataset, which scored 3.57 out of 5. Overall, the semantics of paraphrases are maintained effectively. Nevertheless, the sentences produced by our model exhibit more syntactic accuracy, boasting an average

textual readability score of (4.51/5) compared to (4.08/5) for the VAE-SVG-eq model. This validates the findings observed earlier using automated measurements.

While there are no existing findings from previous studies to compare with the Arabic dataset, the created sample demonstrated a high level of relevance and readability overall (relevance = 3.52/5; readability = 3.88/5).

Our intrinsic experiments focused on generating paraphrases demonstrate that the ARAG-ED model we propose produces highly accurate and precise paraphrases. The system produces diverse and precise rephrases based on the original text. The latter portion of the research examines the influence of implementing the suggested ARAG-ED on enhancing the precision of the ASAG system. The intrinsic experiments we carried out specifically for the task of paraphrase generation demonstrate that the ARAG-ED model we propose generates accurate paraphrases. The method generates a wide range of accurate and specific paraphrases derived from the original text.

The second part of the study aims to investigate how implementing the recommended ARAG-ED model enhances the accuracy of the ASAG system.

Table 22 Average of three human evaluations of generated reference texts (Ouahrani & Bennouar, 2024)

	Relevance /5	**Readability /5**
Quora dataset	4,82	4,94
Al Raisi Dataset	N/A	N/A
Sample-Quora-dataset-VAE-SVG-eq Gupta et al., (2018)	3,57	4,08
Sample-Al Raisi Dataset-ARAG-ED (ours)	**3,52**	**3,88**
Sample-Quora-Dataset-ARAG-ED (ours)	**3,58**	**4,51**

4.4 THE SUPERVISED SCORING MODEL EVALUATION

We performed two types of experiments: quantitative and qualitative. The AR-ASAG Arabic dataset and Mohler et al.'s (2011) English Short Answer Dataset are used in quantitative evaluation to measure the level of scoring accuracy. The datasets are divided into three parts: a training set (70%) for model training, an evaluation set (10%) for performance assessment, and a test set (20%) for final validation. The division is done randomly and stratified, which means that the question types in the Arabic dataset are represented proportionally in each group. Features are derived from the dataset. The dataset's features and grades are imported into a data frame to develop

scoring models and conduct regression analysis. Multiple model iterations are conducted to optimize parameters and enhance accuracy for achieving the highest precision.

Evaluation metrics. Pearson correlation (r: the higher the better↑) is the most frequently used metric for research in this area. We reported it for all our experiments. We report the Root Mean Squared Error (RMSE, the lower the better↓) and the Pearson coefficient.

Baselines. For Arabic, we utilize the dataset's baseline as documented in (Ouahrani & Bennouar, 2020) and presented in the section the 4.2 Section. The baseline is established using unsupervised K-means clustering, resulting in a Pearson correlation coefficient of **70.37%** and a root mean square error (RMSE) of **1.0454.** When focusing on English, the assessment of the baseline dataset is conducted to contrast it with previous studies that utilized the Mohler dataset. The evaluation's guiding principles are grounded in the reliability of human ratings and the tangible insights gained from the assessment of the ISAGe outcomes. The findings are presented and analyzed from various perspectives:

- Scoring Model Accuracy on the AR-ASAG Dataset,
- The impact of combining specific and general domain knowledge,
- The impact of using multiple reference answers on the ASAG accuracy,
- Comparison with previous works on the Mohler Dataset,
- The student achievement using the tool in formative and summative assessment,
- The analysis of the grading errors and limitations, and
- The computational complexity of the proposed ASAG System.

Two variations, ASAG-0 (with one reference solution) and ASAG-M (with teacher-provided and additional reference answers), are deployed for evaluation. ASAG-0 refers to ASAG with just one reference solution from the teacher, whereas ASAG-M refers to ASAG with the teacher-provided reference answer improved by M additional reference answers created via Beam search (where M is the number of generated responses). We selected ASAG-10, which had a beam size of 10, as it showed the highest accuracy during evaluation on the test set of the ASAG datasets.

4.4.1 Supervised Scoring Model Accuracy using one reference answer:

On the Arabic Dataset. In order to train the model specifically for Arabic, we created an Arabic word space that encompasses the domain of cybercrime using the COALS approach using the Cyber In-domain corpus presented in the section 4.2.1.

In order to acquire a broad understanding of Arabic, we utilized the pre-existing Skip-gram Word Embeddings for Modern Standard Arabic (Zahran et al., 2015). The model has around 6.3 million word vector entries that have been trained using a substantial quantity of unprocessed Arabic texts from several sources, including Arabic Wikipedia, Arabic Giga-word Corpus, Arabic Wiktionary, BBC and CNN Arabic news corpus, Microsoft crawled Arabic corpus, and Arabic books.

Here is Table 23 that shows how well the proposed model improved the baselines of the AR-ASAG dataset and the findings of the sample test, including the performance metrics and comparisons with baselines, are presented. Both the Support Vector Regression (SVR) and Mlpregressor Artificial Neural Network (ANN) models showed an increase in correlation compared to the baselines. The Pearson correlation for SVR increased by 5,33% and for Mlpregressor by 5,71%. Additionally, the root mean square error (RMSE) increased by 0,125. The Ridge Linear Regression (Ridge-LR) model had the highest correlation, with a Pearson coefficient of 77,95%, representing an increase of 7,58%. The root mean square error (RMSE) has been substantially reduced to 0,8967 (an improvement of 0,1487 compared to the baselines), approaching the RMSE achieved by humans, which is 0,8381.

An in-depth examination of the association between automatic grades and human grades, as depicted in Figure 12, enables a more accurate assessment of performance. We examine the disparity between the manual and automated grades for the Arabic sample test.

Significantly, the deviation falls within the range of 0 to 1 in 73,06% of cases, which is highly noteworthy on a 5-point scale. Out of a total of 91,56 cases (73,06 + 18,5), have a difference that is equal to or less than 1.5. Additionally, 96,47% of responses have a maximum difference of 2 on a 5-point scale. Only 3,52% of instances have a difference that exceeds 2. In contrast, there are no differences that are greater than or equal to 3. In conclusion, the results offer a reliable assessment of the suggested model, considering the subjective nature of the evaluation method.

Table 23 Proposed approach evaluation on the Arabic Dataset

System	Train Score	Test score	Pearson↑	RMSE↓
AR-ASAG Dataset Baseline (Ouahrani & Bennouar, 2020)	-	-	70,37	1,0454
SVR	53,88	57,25	75,70	0,9200
Mlpregressor-ANN	54,45	56,97	76,08	0,9200
Ridge-LR	59,25	60,18	**77,95**	**0,8967**

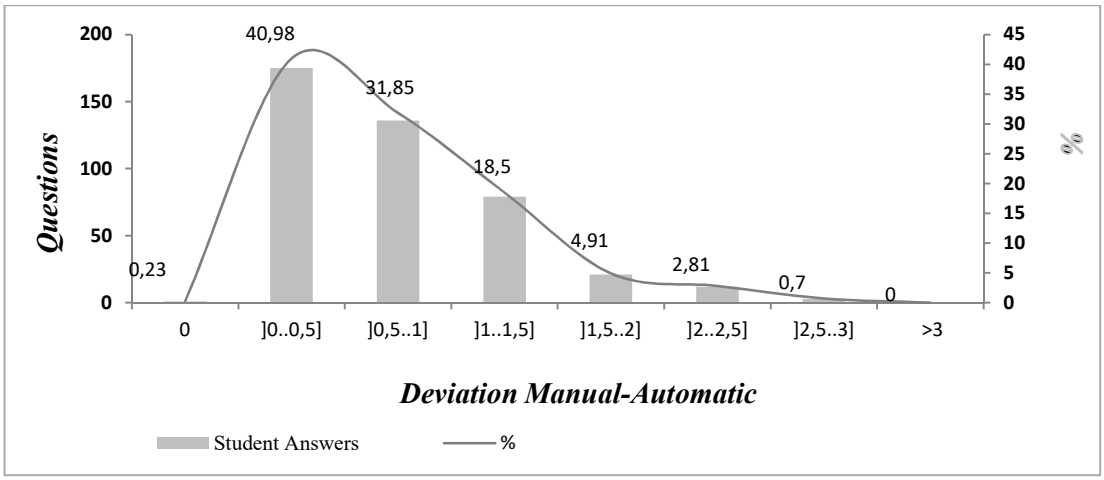

Figure 12 Automatic and Manual Scores Deviation on the AR-ASAG Dataset (Sample Test).

On the English Dataset. We assessed our approach using a different dataset for grading short answers, specifically the Mohler et al. (2011) dataset. The objective is to demonstrate that our approach is not limited to a certain category of short answers and can be effectively used with different datasets for constructing scoring models in various languages.

In order to train our English model, we constructed a domain-specific corpus focused on "computer science and programming" using the material from the English Mohler dataset. Specific domain knowledge is acquired by generating the relevant semantic space. In order to acquire information about the English language on a broad range of topics, we utilized a pre-trained Fasttext Skip-gram model with 300 dimensions[29]. The model includes a set of 2 million word vectors that have been

[29] https://fasttext.cc/docs/en/english-vectors.html

trained using Common Crawl data, which contributes to capturing the nuances and meanings of words in the English language context. These word vectors are designed to capture the nuances and meanings of words in the broader context of the English language.

Based on the results in Table 24, the correlation coefficient of our model across all test data samples is **66,89%,** indicating a strong relationship between the model's predictions and the actual scores. The root mean square error (RMSE) is **0,8206.**

These results imply that implementing the proposed paradigm results in improved performance, which could have significant implications for advancing scoring models in educational assessments. When we compare our Ridge-LR model's correlation to Mohler et al.'s (2011) dataset, the difference is substantial: the Pearson coefficient goes up by 15.09% and the RMSE goes down by 0.1574.

Figure 13 provides a presentation of the scoring model's performance on the sample set from the Mohler dataset, offering a close examination of the difference between the automatic score and the manual score. In 80,34% of the score pairings, the discrepancy is less than 1. The difference is smaller than 1.5 in 90,97% of cases, which is both interesting in practical terms and provides further confirmation of the results observed for the Arabic language.

Based on the findings from the Mohler dataset, we can conclude that the suggested method is not limited to a certain form of short response and may effectively apply to other languages as well.

Table 24 Proposed system evaluation results on the Mohler Dataset (one reference answer)

System	Pearson↑	RMSE↓
(Mohler et al. 2011) System (Mohler et al., 2011)	51,80	0,978
ISAGe Ridge-LR (Ours)	66,89	0,8206

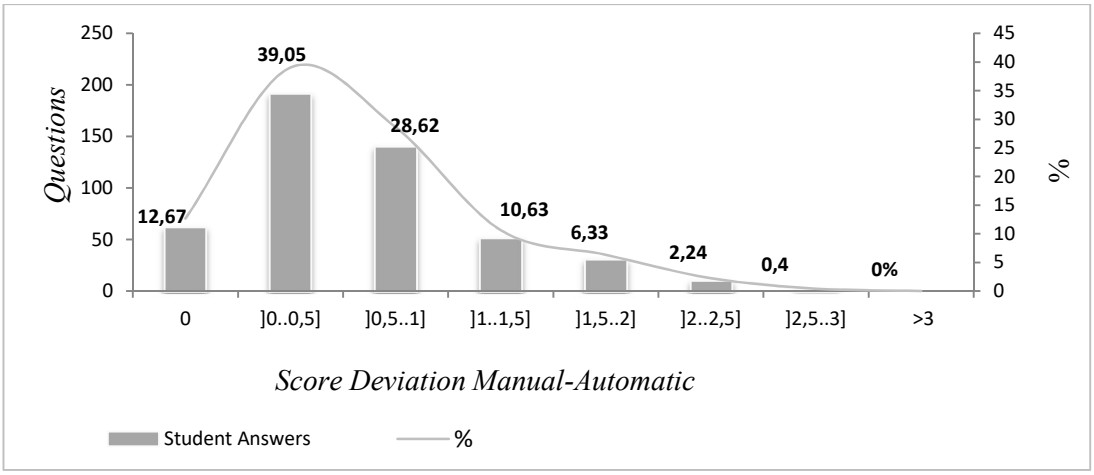

Figure 13 Automatic -Manual Grades Deviation on the Mohler Dataset (Sample Test).

4.4.2 Specific and general domain features impact

When presenting the findings for the Arabic dataset, we additionally demonstrate experiments using different versions of the suggested features to illustrate their effect on accuracy. We conduct experiments that encompass all aspects, excluding both domain-specific and domain-general features.

Table 25 shows a correlation drop when both the specialized and broad domains are excluded. Combining domain-specific and domain-general knowledge resulted in the highest correlation, highlighting the significance of integrating specific and general domain information.

Using distributional semantics facilitated the acquisition of additional syntactic and semantic domain knowledge, leading to enhanced accuracy. The training process utilizes characteristics derived from embeddings, eliminating the need for individual training for each question and the availability of several student replies for each question.

Table 25 Domain Features Ablation Study

Features	Train Score	Test score	Pearson↑	RMSE↓
All Features	59,25	60,18	**77,95**	**0,8967**
Without Specific-Domain Features	57,44	58,29	75,79	0,9100
Without General-Domain Features	55,87	57,28	75,95	0,9200

4.4.3 The impact of multiple reference answers on the ASAG task

The findings are presented only for the Ridge regression model in this section as it already presented the highest score. The evaluation results of the suggested supervised ASAG model on the Arabic and English datasets (sample test) are shown in Table 26.

The proposed ASAG models (ASAG-1 and ASAG-M) show significantly improved scoring performance over the baseline dataset. Incorporating paraphrase generation into the grader resulted in superior outcomes. Pearson's correlation increased significantly from 66,89% (without the use of paraphrases) to 73.50% (with the use of paraphrases) for the English dataset. The correlation of the Arabic dataset increased from 77.95% (without paraphrases) to 88.92% (with paraphrases). The correlation observed exceeds the agreement level among annotators for both datasets. The difference between manual and automated scores has significantly widened. The RMSE for the English dataset was reduced from 0,8206 (without paraphrases) to 0,7790 (with paraphrases). The Arabic dataset had a reduction from 0,8968 (without paraphrases) to 0,6955 (with paraphrases).

The findings suggest that using paraphrases improves the performance of the regression proposed model.

Table 26 Proposed system evaluation on the Arabic and English Datasets (Test Set): multiple reference answers

	AR-ASAG Dataset		Mohler English Dataset	
	Pearson↑	RMSE↓	Pearson↑	RMSE↓
Inter-Annotator Agreement (Manual)	83,84	0,8381	64,43	-
Dataset baseline	70,37	1,0454	51,80	0,9780
ASAG-0 (Ours)	77,95	0,8968	66,89	0,8206
ASAG-M (M=10) (Ours)	**88,92**	**0,6955**	**73,50**	**0,7790**

Ablation study on the multiplicity of reference answers. When presenting the comparison findings for the Arabic dataset, we also demonstrate trials using different numbers of created reference responses to assess how the presence of multiple reference answers affects the accuracy of grading. We conducted studies without any paraphrases, referred to as ASAG-0. Subsequently, we incorporated other paraphrases answers (ASAG-1, ASAG-4, ASAG-7, and ASAG-10) that corresponded to beams 1, 4, 7, and 10.

Figure 14 demonstrates that failing to consider paraphrases leads to a fall in correlation. The correlation significantly increases with each subsequent rise in the

number of paraphrases. A Pearson correlation coefficient of +1.19% and a root mean square error (RMSE) of -0.0066 show that the improvement in paraphrasing between 7 and 10 paraphrases is minimal. This is because the best-paraphrased reference responses are achieved generally with just seven paraphrases. Configuring the creation of paraphrases to 7 could be beneficial during system implementation, ensuring optimal performance. A linear trend is observed, showing a consistent increase in Pearson correlation and a simultaneous decrease in RMSE.

According to the results, using the produced alternative reference responses along with the supervised grading model makes a big difference compared to using just one reference answer. This indicates that the suggested Encoder-Decoder produces believable rephrases. This validates the finding that the paraphrase generation significantly enhances Automated Short-Answer Grading systems.

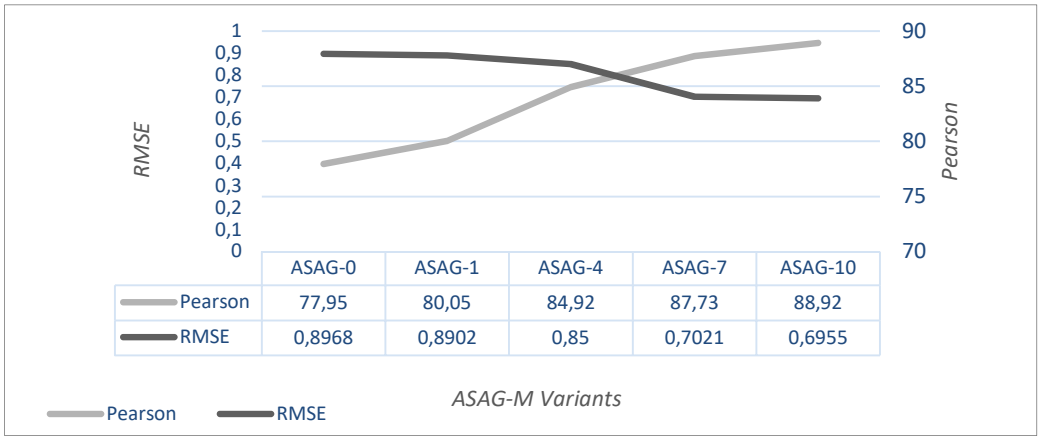

Figure 14 Ablation study on the multiplicity of reference answers on the Arabic Dataset.

4.4.4 Comparison with previous work on the Mohler Dataset (English SOTA)

According to the results shown in Table 26, our model achieves a correlation of 73,50% when employing multiple reference responses across test sample dataset. The root mean square error (RMSE) is 0,7790. Previous reference-based methods were used in (Ramachandran & Foltz 2015; Sultan et al. 2016; Kumar et al. 2017; Saha et al., 2018; Pribadi et al., 2018 ; Gomaa & Fahmy, 2020; Tulu et al., 2021; Agarwal et al., 2022) were discussed in chapter two. Their correlation and RMSE results are shown in Table 27. The proposed approach achieves near the state-of-the-art performance (Agarwal et al., 2022) with an RMSE of -0,017 (Pearson not reported).

Compared to works using *an unsupervised approach*, (Ramachandran and Foltz 2015; Pribadi et al. 2018) proposed the generation of alternative reference

answers using the Maximum Marginal Relevance (MMR) method (Pribadi et al., 2018) and summarization of the content of top-scoring student responses (Ramachandran & Foltz, 2015). Gomaa and Fahmy (2020) used a Skip-thought vector unsupervised approach to convert reference and student answers into embeddings to measure their similarity. Although they present the advantage of not requiring a training process, the impact of the paraphrase generator is more significant in the proposed approach: (Pearson: +26,7%, RMSE: +0,105) compared to the (Pribadi et al., 2018) system, (Pearson: +12,5%, RMSE: +0,081) compared to the (Ramachandran & Foltz, 2015) system, and (Pearson: +10,57%, RMSE: +0,131) compared to the (Gomaa & Fahmy, 2020) system. Achieving high-precision scoring remains challenging, particularly when ASAG evaluations carry significant stakes for students.

Compared to works using *a supervised approach* requiring a training process, the proposed system is similar to what Sultan et al. (2016) used. (Sultan et al., 2016) (That was the previous SOTA) trained a regression model involving semantic similarity, text alignment, question demoting, term weighting, and length ratio features. The use of multiple reference answers generated by paraphrase generation and the combination of specific and general domain knowledge features, enabled the proposed approach to outperform Sultan et al.'s (2016) system (Pearson +10,5% and RMSE -0,071). The proposed approach outperforms (Kumar et al., 2017) and (Saha et al., 2018) systems that used a deeper approach. Kumar et al. (2017) used a Siamese bidirectional LSTM applied to a reference and a student answer based on earth-mover distance across all hidden states from both LSTMs and a final regression layer to output grades. Saha et al. (2018) combined handcrafted features and sentence embedding features to train an end-to-end deep neural network to learn embeddings and a neural network to train a grading classifier. This task is challenging because training embeddings necessitates extensive amounts of data to be effective.

Attention-based and transformer models (Vaswani et al., 2017) have been utilized in new ASAG methods to enhance the representation of structural and semantic features, leading to more accurate assessments. (Agarwal et al., 2022) used the Multi-Relational Graph Transformer to represent short text matches and added relation-enriched structural information. They got the best results on the Mohler dataset (RMSE: 0.7620, Pearson not reported). Transformer-based language models, like BERT, have demonstrated exceptional performance on a range of natural language processing tasks, including question answering, sentiment analysis, text

summarization, and others. Although they possess significant power, the computing requirements, as highlighted by Huang et al. (2022), are excessively high for practical use. Two main obstacles are encountered in the ASAG assignment. One of the primary challenges in implementing these models for Automated Short Answer Grading is the limited size of the ASAG dataset, which restricts the availability of adequate training or fine-tuning data, hindering practical implementation. The study by Gaddipati et al. (2020) used transfer learning models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), GPT (Radford et al., 2018), and GPT-2 (Radford et al., 2019) to test how well the embeddings worked on the ASAG task using cosine similarity. Table 27 indicates that the performance of these models is subpar. When employed extensively, as is the situation in e-learning settings, they pose difficulties for professionals. The heightened memory and processing demands result in expensive computational costs during inference, limiting the utility of storing and loading model parameters in resource-constrained contexts. For instance, in educational settings with limited computing resources, the computational demands of these processes can hinder their practical use. The study of (Tulu et al., 2021) notably produced remarkable state-of-the-art outcomes by training on the Mohler dataset. The study significantly benefited from utilizing sense representations derived from synsets and their interconnections in the WordNet lexical-semantic network through the SemSpace method. Additionally, the improved accuracy in semantic analysis shows that the use of SemSpace sense vectors from the English WordNet significantly increased precision. Using individual training files for each question, the system achieved a Pearson's correlation of 0,95. However, when all questions, student answers, and reference answers are included in a single training file, the Pearson correlation drops significantly to 0,15 which is disastrous. Training each question individually necessitates a substantial number of student responses per question. Consequently, the system lacks the ability to generalize to new, unseen questions, presenting a significant scalability challenge within an e-learning environment. In the other hand, relying on an English lexical-semantic network like WordNet poses a significant challenge for languages with limited resources, such as Arabic, due to the scarcity of available linguistic resources. Lexical databases, as the one employed in (Tulu et al., 2021), require extensive linguistic resources and comprehensive datasets, which are readily available for English but lacking for languages that are not as extensively studied. The absence of comprehensive lexical databases and semantic networks in languages such as Arabic

renders it impractical to employ the same methodology. As a result, under these conditions, the effectiveness and practicality of this method are constrained, hindering its full potential in practice.

Finally, our proposed approach employs a simplified model augmented by paraphrase generation and achieves a near-SOTA (Agarwal et al., 2022) system (RMSE: -0.017) on the Mohler dataset. Our model's simplicity in construction, loading, and integration into the LMS question engine enhances efficiency and user-friendliness, for example, by reducing setup time and improving accessibility. Utilizing a small dataset not only simplifies data acquisition from the teaching archive or LMS question bank but also enhances model training efficiency and accuracy due to reduced complexity and noise in the data

Table 27 Comparison with previous work on the Mohler Dataset (Ouahrani & Bennouar, 2024)

ASAG System			Pearson↑	RMSE↓
Pribadi et al. (2018)			46,80	0,8840
Saha et al. (2018)			57,00	0,9000
Ramachandran and Foltz (2015)			61,00	0,8600
Gomaa and Fahmy (2020)			63,00	0,9100
Sultan et al. (2016)			63,00	0,8500
Kumar et al. (2017)			55,00	0,8300
Agarwal et al. (2022)			N/R	**0,7620**
Tulu et al. (2021)	Individual training files for each question		94,90	0,0400
	Single training file for all questions		**15,00**	
Ouahrani and Bennouar (2024)			**73,50**	**0,7790**
Gaddipati et al., (2020)	ELMo	WE	48,50	0,9780
	GPT	WE	24,80	1,0820
	BERT	WE	31,80	1,0570
	GPT-2	WE	31,10	1,0650

4.4.5 Formative and summative assessment using the ASAG. (Case study 2).

The second case study involved integrating our ASAG solution into an online evaluation system for both formative and summative assessments. This integration was tested at the Computer Science Department of Bouira University using the university's Moodle web platform. The ASAG system was seamlessly incorporated into the Question Engine for the "Cybercrimes" course, facilitating both formative and summative evaluations for students. Master's students specializing in Information Systems and Software Engineering (ISIL) took the "Cybercrimes" course, which was

a mandatory part of their curriculum. The course combined in-person sessions with online components; in-person classes provided condensed instruction, while homework assignments were delivered online. The university's web platform, which supports course delivery and student interaction, was the focus of our experiments. Students engaged in continuous online evaluations to monitor their progress and improve learning within a hybrid teaching strategy that blended online and in-person instruction. The Learning Management System (LMS) was accessible both on-campus and remotely, allowing students to participate from their homes. The ASAG plugin was integrated into the LMS Question Engine for the "Cybercrimes" course, which was taught in Arabic.

Experiments were conducted with a sample of 30 students to evaluate the system's impact, although the ASAG system was available to all students. During formative evaluations, students were given various assignments, including short-answer questions and objective questions such as multiple-choice and fill-in-the-blank items. Students could review their exam history, which included their responses, grades, and the reference answers.

To motivate students to fully engage with the formative tests, each test was graded. The average of these grades was used as a control grade, weighted together with the summative assessment grade. The summative assessment was conducted online within the institution using secure internal PCs. Feedback was gathered to assess the proposed approach and its impact.

We conducted an analysis of the results obtained by students in both formative and summative assessments. For the final test, the automatic grades were collected and evaluated on a per-question and per-assignment basis:

- *Per-Question Evaluation.* This involved examining the correlation of all short answer responses from all students. This method allowed us to assess the tool's performance in a real-world setting as compared to evaluations using datasets.
- *Per-Assignment Evaluation.* This focused on the overall summative grade of each student in the assignment. In addition to the automated grading, tutors manually graded the final test to assess the ASAG system's performance in a practical context (in term of correlation and RMSE).
- Correlation Analysis. We calculated the correlation between human scoring and automatic scoring specifically for short answers, as all other questions were objective and had a perfect correlation (correlation = 1).

- Configuration for Multiple Attempts. The ASAG tool was set up to allow students to attempt the same question multiple times without penalty across all tests conducted in formative assessment.

The results are presented and discussed in the following:

Per-question evaluation. The automated scoring provided a strong correlation (Pearson = 88,09% and RMSE = 0,6464 on a five-point scale) with the human grades, taking into account the students' answers to the summative assignment. In Figure 15, the scatter plot depicting the difference between human and automatic grades shows a significant concentration of points within the [0, 1] range, with a smaller concentration in the [1, 1.5] interval. Despite the inherent subjectivity in grading short answers, these results are considered acceptable, though not entirely reliable. This correlation is better than that observed for the Arabic dataset.

After the formative sessions, students improved their ability to respond effectively by focusing on the target themes, resulting in better performance of the tool. The study found that longer student responses were more prone to incorrect assessments, highlighting a negative impact of response length on grading accuracy. The grader struggled to evaluate lengthy answers due to difficulties in assessing similarities. To help students concentrate on synthesizing key elements and concepts in short answers during formative sessions, a maximum length constraint will need to be implemented.

Per assignment evaluation. We assess the overall evaluation of students' final grades by comparing manual and automated grading methods. Figure 16 depicts the correlation between these two grading approaches. The cumulative error in total scores is minimized by balancing it across multiple questions within the same assignment, resulting in a minimal difference between manual and automatic final grades. The correlation coefficient with human scoring is approximately 91.81%, which is highly encouraging. In the formative evaluation, the average score of the assignments for each student is calculated and compared to the final summative grade.

As shown in Figure 16, the average formative grade distribution curve is generally lower than the final summative grade distribution curve, indicating a trend of student improvement. Overall, the final grades are significantly enhanced due to the tool's capability to improve learning through formative assessments provided throughout the course.

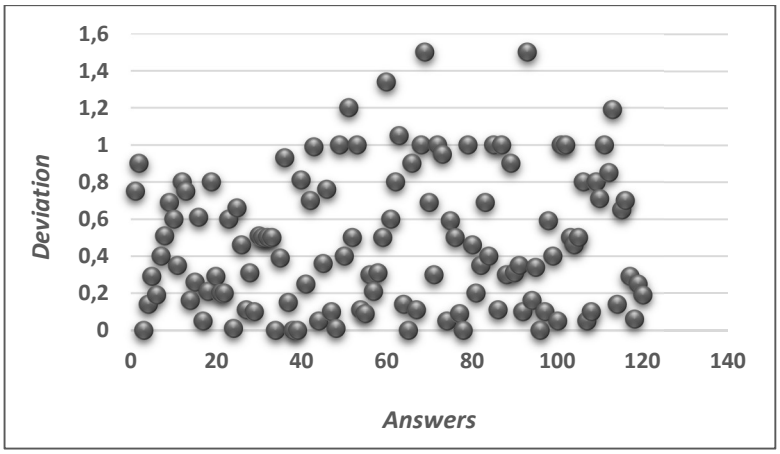

Figure 15 Manual-Automatic Grades Per-Question Distribution on Final Summative Assignment.

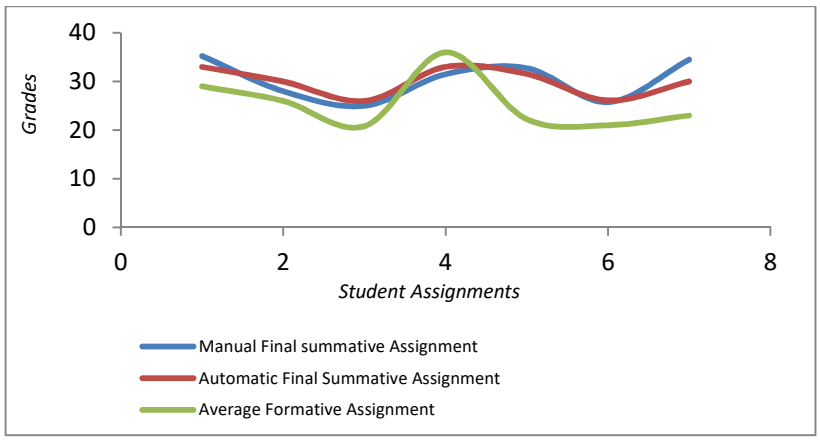

Figure 16 Per-Assignment Distribution Grades-Average Formative Assignment vs.

4.4.6 Computational Complexity

Despite advancements in software and technology for e-learning and online learning management systems (LMSs), there are still several significant limitations, primarily related to technical issues. These include slow server and browser response times, delays in resolving technical problems, insufficient availability of equipment for students, and the potential high cost involved in developing initial programs (Marczak et al., 2016).

When implementing autonomous scoring in e-learning settings, it is important to take into account issues beyond only accuracy, including server performance and practical efficiency. Low computational complexity is crucial for the scoring system due to its integration with online learning. Computational complexity was assessed by measuring CPU utilization and runtime while manipulating the number of combinations of (student answer, reference answer).

The experiments were conducted on a workstation featuring an Intel (R) Core(TM) i7 CPU running at 2.50GHz and 8.0 GB of RAM. The experiments were conducted on a workstation with a 64-bit Windows 10 operating system, and an internet speed of approximately 4 Mbps was utilized.

As seen in Figure 17, consumption usage has a linear tendency. The percentage increased from 6.2% to 20.9%. The runtime exhibited a range of 36 to 177 seconds, as the number of pairs of answers to assess ranged from 50 to 427. This is acceptably accurate in an online examination session. The findings indicate that the technique can be adjusted without major alterations to current LMS systems and without necessitating a high-performance computer. The application may be implemented on any Moodle platform, enabling a consistent evaluation method for all courses in the LMS.

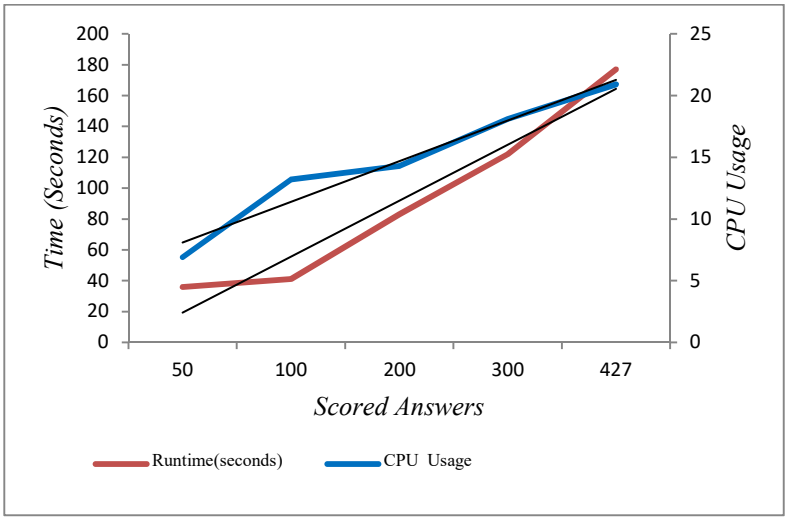

Figure 17 Runtime and CPU usage.

4.5 ANALYSIS OF THE GRADING ERRORS AND LIMITATIONS

Deploying ASAG systems in practical environments presents difficulties owing to the need for accurate representation and a thorough comprehension of the answer text. When the system fails to meet acceptable performance standards, it leads to concerns about trust. For example, errors in automated grading have been shown to significantly impact students' academic (Azad et al., 2020; Hsu et al., 2021; Schneider et al., 2023).

The barrier to confidence in the tool will persist until ASAGs can consistently match or exceed human performance standards. Given their statistical character, dependable ASAG systems play a crucial role in creating confidence and practical

usefulness, such as in predicting student performance trends accurately. Although errors are certain to occur, using good management practices can help minimize their consequences. For instance, it may be more desirable to grant a passing grade to a student who should have failed rather than give a failing grade to a student who should have passed (Schneider et al., 2023).

Comparing the grades assigned automatically with those assigned by humans, as depicted in Figures 18 and 19, allows for a more comprehensive evaluation of performance, enabling the identification of an acceptable margin of error. The objective of our inquiry is to analyze the disparities between manually assigned grades and those assigned automatically on a 5-point scale for the Arabic set test, comprising 428 question-answer pairs, to enhance the efficiency and accuracy of ASAG systems.

As shown in Figure 18, the correlation between the predicted score (by rounding off the scores) and the human score is perfect for 52 responses (12.15%). In 86.91% of the cases, the difference is between 0 and 1 on a scale of 5. This is reasonable since the correlation exceeds the overall agreement among human annotators (IAA = 83,84% for the entire Arabic dataset). The results provide a strong indication of the effectiveness of the proposed model, especially considering the subjectivity inherent in the evaluation process. In 96.26% of cases, the difference is less than or equal to 1.5. Figure 19 shows a clear clustering of data points between 0 and 1, with a slightly reduced concentration in the range of 1 to 1.5. The points inside the interval [1, 1.5] are more likely to be in closer proximity to 1 than 1.5. Furthermore, the discrepancy in scores poses a significant issue, especially in assessments where the outcomes have substantial implications. In 3.73% of the answers (16 out of 428), the difference surpasses 1.5%. In order to comprehend this scoring bias, a more thorough examination is carried out, focusing on different question types.

Figure 20 illustrates an in-depth examination of the association between automatic grades and human grades for each type of question, using the Pearson correlation coefficient. The questions may be classified into five categories: defining the concept, explaining, exploring the consequences, justifying, and identifying differences. The question "What consequences?" demonstrated the highest correlation (Pearson = 87.15%), exceeding the manual correlation between the two human annotators for the entire Arabic dataset, highlighting its importance in the evaluation process. The "Explain" question received a score of 83.58%, while the "Define the Concept" question received a score of 82.34%. The question "What is the difference?"

yielded less favorable results, with percentages of 57.97% and 75.78%, in contrast to the question "Justify?" For instance, inside the test sample and when assessing student submissions, there are several responses that receive the lowest scores when graded automatically. These answers pertain to two specific questions:

- What is the difference between hacking and penetration testing?

ما الفرق بين القرصنة واختبار الاختراق؟ (In Arabic)

- Justify the truth of this statement: “A false sense of security is more dangerous than a true sense of insecurity”.
- علل صحة العبارة: الإحساس الخاطئ بالأمن أخطر من الإحساس الصحيح بعدم الأمن (In Arabic)

An analysis of the answers may reveal certain factors that contribute to this bias. Initially, teachers and students use different assessment methods, like grading rubrics and scoring criteria, which may highlight assessment discrepancies. Second, the topic of short answers to the "justify" question is so broad, like discussing abstract concepts, that it encourages veering off topic quite easily, leading to less focused responses. Third, the length of the student response has a negative effect on the grade. Usually, when they include irrelevant information, students exceed the word limit of the response.

In terms of similarity, the grader has poorly assessed a long answer. These findings support the statement that “the issues on which ASAG systems perform poorly are often also the ones on which humans do not agree" (Adams et al., 2016). Prioritizing the most effective method of formulating the question is crucial when generating automatic short response questions. Short responses to questions often tend to deviate rapidly from the main subject matter.

It is crucial to establish rules for students and provide precise definitions for short answer questions. Providing learners with clear guidance is crucial to ensure their precise comprehension of the expectations placed on them. It is advisable to prioritize individual ideas or concepts over a large topic.

A hybrid assessment strategy, combining manual and automatic methods, is significantly more successful for evaluating challenging topics than relying solely on automatic assessment. The LMS quiz system is set up to allow for human scoring as needed. These insights indicate challenges and problems associated with grading in practical situations.

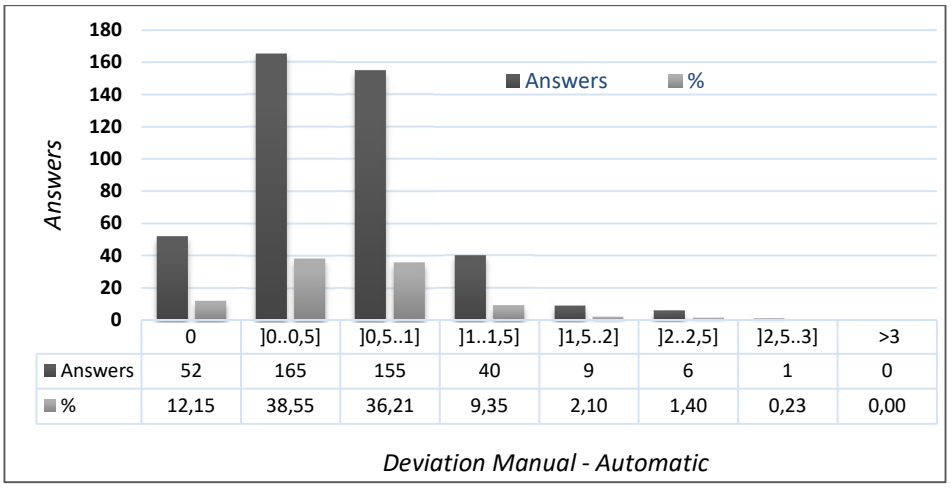

Figure 18 Automatic and Manual Scores Difference on the AR-ASAG Set Test).

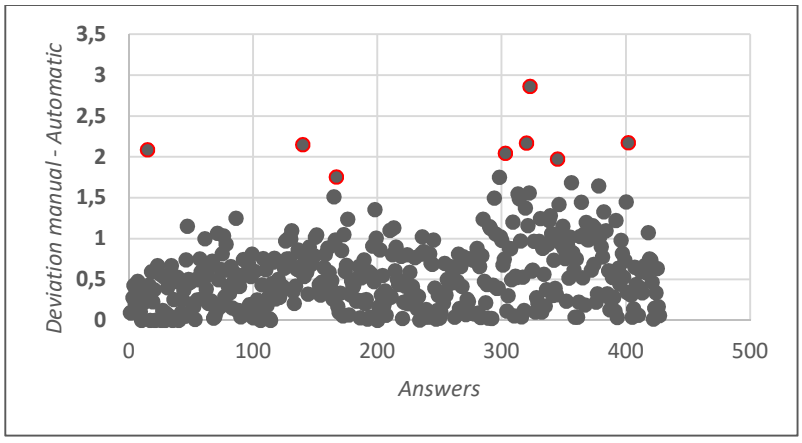

Figure 19 Automatic and Manual Scores Difference on the AR-ASAG Dataset (Set Test).

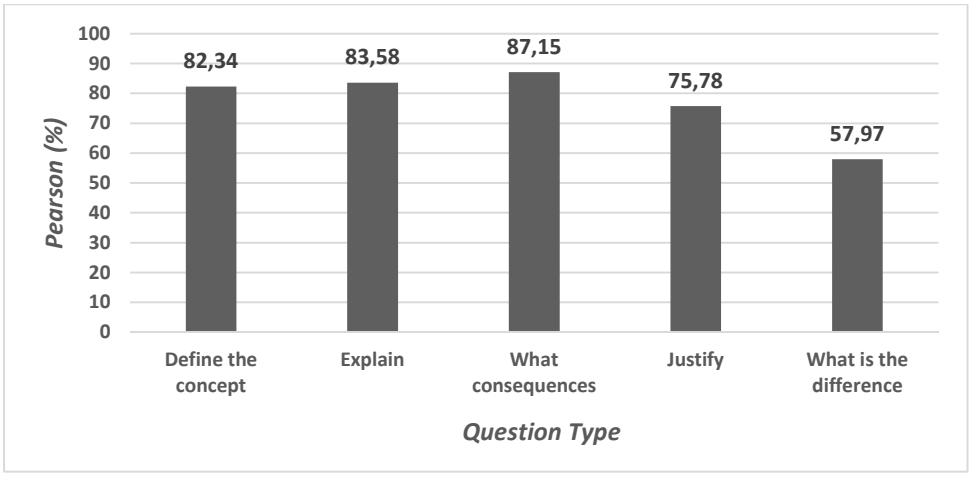

Figure 20 Distribution of Automatic-Manual grades per question type on the AR-ASAG Dataset Set test.

4.6 OVERALL DISCUSSION AND IMPLICATIONS

In practice, a few ASAG tools are implemented and made available in e-learning systems. ASAG tools still require significant manual oversight. Our thesis attempts to give input for improvement in this context. While our case study was carried out in our academic e-learning setting, the results are expected to have relevance for a broader educational community in theory and practice. The responses to our questions research may argue it:

Addressing the lack of resources challenge, we have attempted to address three main issues: the scarcity of datasets, linguistic complexities, and the lack of extensive lexical knowledge database: We have developed the AR-ASAG, an Arabic dataset designed for automatic short answer grading. As far as we know, AR-ASAG is the first Arabic dataset for automated short answer grading that is openly accessible for download in Arabic. Specifically, the AR-ASAG dataset offers a diverse range of examples and a comprehensive grading system evaluation, which can enhance the quality and efficiency of assessments.

Ultimately, this enhancement in grading practices can lead to fairer evaluations, better student feedback, and increased overall academic performance in institutions. The tests and explorations undertaken provide valuable insights into the effectiveness of the proposed grading method combining specific and general knowledge domain. The technique employed in this study yields encouraging outcomes for Arabic when assessed on the dataset.

(Question 1.1) Indeed, employing a medium-sized domain-specific COALS semantic space rather than a large corpus for latent semantic analysis can yield superior outcomes. These findings indicate that, in the context of the COALS approach, the significance of text quality outweighs that of text quantity. This outcome offers two significant advantages for the field: Firstly, it advocates for using domain-specific corpora, allowing for comparison of outcomes within this dimensionality range. This suggests that creating a specific collection of relevant texts for the evaluated course or field is appropriate. It also simplifies the process of creating or finding a suitable corpus, which does not need to be excessively large. Furthermore, it decreases the need for substantial computational resources in order to build ASAG system.

(Question 1.2) & (Question 1.3) our empirical evaluated evidence showed that a simple stemmer is as effective as a more complex root stemmer is. This is particularly

interesting for the Arabic language, as root stemmers are still far from achieving full development. The proposed approach is versatile and can be applied not only to the Arabic language but also to other languages facing challenges such as limited linguistic resources and complex syntax. Our focus is on improving response grading accuracy by addressing real-world challenges in grading assignments. The utilization of term weighting optimizes the system's correlation, while also maintaining simplicity and practicality. The combined impact of word weighting is more significant than the use of the stemming technique (still challenging for Arabic).

To mitigate the lack of extensive lexical linguistic resources, we employed two primary strategies to enhance feature engineering and effectively utilize both domain-specific and general knowledge. The first strategy, *feature enrichment*, involves augmenting features to better align with human evaluation criteria (combined and enriched similarities, length statistics, POS tagging, term weighting, difficulty and question gap). The second strategy, *combined training*, integrates the semantic strengths of the domain-specific COALS model with those of a general-domain skip-gram embedding model.

We drew from the fact that: what we can find through distributional semantics can be as rich, or in some cases richer, than what is found in traditional dictionaries and lexical databases such as WordNet.

A*ddressing the challenging diversity of student responses*,
(Question 2.1) The major finding of this study is the observation that the use of paraphrase generation significantly helps ASAG improve. The proposed model demonstrates improvement and comparable accuracy compared to the current best methods on both the Arabic and English datasets.
(Question 2.1) Generating higher paraphrases in NLP remains a difficult task. The lack of linguistic resources, such as lexicons, dictionaries, and corpora, significantly hinders the task of paraphrase generation in Arabic NLP. We tackled this gap in Arabic NLP, laying the groundwork for future research and introducing an innovative deep learning-based approach to the Arabic paraphrase generation task.

Addressing the integration and the scalability challenge in a real LMS environment, the proposed approach can offer significant scalability benefits on a large scale without requiring a large number of student responses to train each question. Our

proposed approach has been shown to be just as effective as approaches that rely on complex models, while also enhancing the system's scalability and usability.

(Question 3.1) Supervised methods like BERT, GPT, and T5 require a substantial dataset to function effectively. Conversely, other methods may effectively train the model with a smaller dataset. The suggested model requires a limited dataset for training in the section of the course. After undergoing training, the model may be imported and utilized in a different plugin.

The PMatch system (Jordan, 2013), from the Open University and regarded as the most advanced ASAG system in e-learning settings, operates by matching terms and their synonyms. Training each question in the model necessitates a substantial number of student replies, which is a challenge. In contrast to the Open Mark system, our proposed model utilizes a general-question approach that trains on the complete collection of questions in the dataset. Instead of training the model individually for each question, many questions are trained in the same model.

In the other hand, deploying ISAGe as a service on an external cloud offers two key advantages. Firstly, it enhances response times by offloading the scoring task from the platform, thereby separating scoring from learning activities within the e-learning environment. Achieving fast and accurate short answer scoring in the context of e-learning presents a significant challenge, especially when numerous quizzes are submitted simultaneously, necessitating prompt grading responses. Secondly, deploying ISAGe on an external cloud promotes a culture of innovation in e-assessment by integrating machine learning and fostering openness to learning from experience. Recognizing the inherent probabilistic nature of model predictions and the impossibility of achieving perfection, continuous retraining of the scoring model with fresh data becomes essential.

By shifting system training and scoring outside the Learning Management System, the approach facilitates adaptation to new educational requirements through retraining on new features or datasets, all without requiring a complete overhaul of existing LMS systems.

(Question 3.2) Moodle, as the leading open-source virtual learning environment, has been instrumental in developing our grader engine. Recently, the growing use of Moodle in our country's universities prompted us to combine expertise and resources to enhance the assessment of free-text short answer questions in Moodle.

Deploying the proposed ASAG as a plugin would be beneficial to both students and teachers: It would motivate teachers and students to engage actively with the LMS for assessments, offering them valuable benefits and opportunities for learning. LMSs are commonly utilized for delivering course materials and conducting objective assessments, offering educators and students streamlined processes and effective evaluation tools. Teachers would receive valuable insights into individual learners' progress, aiding in personalized instruction

Following the analysis of the feedback from qualitative experiments, the following consequences of incorporating the proposed ASAG into the educational e-learning environment have been identified: The suggested approach initially provides opportunities for practicing and assessing learning progress. Assignments are designed to empower students to progress at their own pace, fostering a student-centered learning experience. Students find a series of tests to be less daunting than a solitary online final examination. Consequently, there is an increase in the rate of participation and an enhancement in scores. Moreover, when utilized as a smoothing extension, the ASAG effectively leverages all the features and functionalities of the LMS for comprehensive utilization.

Detailed performance data for students on individual questions, including metrics on accuracy and time taken, is readily accessible. All attempts for each question are stored, even those created during the initial reflection. Analyzing the individual advancement of students as they progress through a question allows for personalized feedback and targeted support, enhancing the learning experience. Instructors can pinpoint challenging components within questions or poorly formulated items, enabling tailored interventions and content improvements to enhance student comprehension. Furthermore, he has the ability to modify the assessment design.

Discussing the viability of our approach , utilizing a well-trained model to predict outcomes on new data, unseen by the model previously, enables a comprehensive assessment of the methodology's viability and potential effectiveness. The model we have developed is trained specifically to evaluate short answer questions in the field of cybercrime. It is capable of assessing fresh, short answers in this particular subject area. This highlights the model's flexibility and efficiency, enabling instructors to integrate seamlessly new questions into the course without requiring

model retraining. He should just focus on preparing the question and the reference answer, as is common in exams. The trained model is uploaded into the Learning Management System (LMS), which is integrated into a plugin to guarantee accurate predictions. In order to implement a new ASAG (Automated Short Answer Grading) on a different course, it is necessary to train our model using a fresh dataset, as the ASAG task is highly dependent on the specific domain. The most challenging aspect is obtaining the dataset.

Discussing the limitations of the ASAG model, or rather its challenges, the risk of over scoring poses significant challenges in the grading process, potentially leading to inflated grades and undermining the accuracy of assessment outcomes. Given the lack of a performance grading system that matches human performance, it is generally more acceptable to give students a higher score than to give them a lower grade (Schneider et al., 2023). As indicated in the evaluation report, the ASAG model exhibited subpar performance in areas related to wide-ranging topical questions. The broad nature of short-answer topics often results in students veering off-topic and digressing. The length of the student's response significantly sways the grader's assessment. For instance, a lengthy response that includes irrelevant information may result in a lower rating. Furthermore, clear instructions and precise recommendations are essential when formulating short-answer questions. Learners must be provided clear guidance to ensure a thorough understanding of expectations. Finally, it will take time to transition to fully automate testing for short answer questions. Ensuring the reliability and validity of scoring is paramount in high-stakes tests, necessitating a rigorous evaluation of any automated system before it is employed to determine test takers' results.

Finally, currently, no ASAG system can eliminate human teachers. ASAG systems should be utilized to enhance and complement human scoring until they can achieve performance comparable to that of humans. Our study intends to give feedback to enhance performance in this regard. During low-stakes formative activities, when students are studying independently and responding to test preparation questions, the grader can evaluate their answers. When teachers have limited time to give feedback on students' answers, they can utilize the ASAG with an acceptable margin of error to assess their responses. Customizing the application to enable learners to make multiple attempts at the same question without negative consequences can enhance motivation.

In high-stakes tests, a hybrid assessment strategy that combines manual and automatic methods appears to be more successful in evaluating items that are challenging to assess automatically. Automated grading should be used in conjunction with human grading rather than being used as a replacement until automated short answer graders can achieve the same level of performance as humans.

Chapter 5: Conclusions

5.1 SUMMARY OF THE RESEARCH

In conclusion, the incorporation of ASAG systems in education has a substantial influence on the process of teaching and learning. ASAG technology has already revolutionized assessment practices in education by streamlining grading processes, enhancing feedback mechanisms, and supporting personalized learning experiences.

From a pedagogical perspective, ASAG technology offers several benefits for teaching and learning. ASAG technology frees up instructors' time and resources by automating the grading process. This allows them to focus on designing engaging learning experiences, providing personalized feedback, and fostering deeper levels of student engagement. Furthermore, ASAG systems have successfully supported innovative assessment practices such as formative assessments and self-assessments, promoting active learning and reflective thinking among students. However, addressing the technical challenges, such as algorithm biases, pedagogical challenges like ensuring feedback quality, and ethical challenges such as data privacy concerns associated with ASAG, requires ongoing research, collaboration, and innovation.

In this thesis, the goal has been to identify key factors to manage the ASAG project, in which the increase in accuracy and the methodological deployment in practice can improve the quality of learning and teaching. We intended to demonstrate that ASAG, rather than being a basic technology, could improve the performance of the educational process. We aim to increase knowledge in theory and practice of ways in which integrated short answer grading into e-learning environments can promote assessment in higher education.

To achieve this goal, we explored advances in compositional distributional semantics for semantic understanding, machine learning for predictive modeling, natural language processing techniques for text analysis, paraphrase generation for varied responses, and cloud sourcing for computational scalability, aiming to improve accuracy and enhance feasibility and scalability in practice.

Our approach to feature engineering is based on three key aspects. First, text similarities between the reference and the student answer using question information. Second, the extension to word weighting, POS tagging, answer length

statistics, and difficulty features. Finally, the integration of specific and general domain information using compositional distributional semantics. Given the variability in student responses and the potential for improved scoring accuracy, we proposed automatically generating various reference answers to handle effectively the diversity of student answers, enhancing the assessment process. We designed a sequence-to-sequence deep learning model with the aim of generating alternative reference answers from a provided reference answer. Furthermore, we offer a supervised grading model that refines specific attributes to improve accuracy by considering multiple reference responses. We assessed the performance of our integrated system in our academic e-learning environment through experiments using Arabic and English datasets, involving active participation from both students and teachers. There is a closer correlation between the grades given by humans and the grades given by the model. It demonstrates superior outcomes compared to the most advanced English dataset available, as demonstrated by a significant improvement in accuracy and efficiency.

These findings are anticipated to have a wide-reaching impact on the education community, offering valuable insights and practical applications for educators and learners alike. Along with quantitative performance reporting, we uniformly conducted qualitative evaluations to understand better strengths and weaknesses.

5.2 CONTRIBUTIONS

By advancing the state of the art in ASAG technology and its integration into teaching and learning practices, we contribute to the improvement of educational outcomes and the advancement of digital learning environments. This thesis presents several significant contributions:

First, we presented AR-ASAG, an Arabic dataset designed for automatically evaluating short answers. Subsequently, we investigated a supervised grading model. AR-ASAG is the first Arabic dataset for automated short answer grading that is openly accessible for download in Arabic.

Our evaluation makes a significant contribution, as more academics are likely to utilize the publicly accessible dataset instead of relying on data from limited internal sources. Public datasets are critical because they provide the controlled experimental circumstances required for an "apples-to-apples" meaningful comparison of different ASAG capabilities that is critical for determining true progress in the field. The

regulated experimental circumstances enable researchers to evaluate correctly the efficacy of various ASAG systems, especially as underscored.

By creating this dataset, we address the scarcity of public datasets for short answer grading and provide a valuable resource for developing and evaluating automated grading systems. This dataset not only facilitates the training of neural models but also enhances the ability to compare and improve different grading algorithms. Its design ensures that it is representative of real-world educational settings, making it an invaluable tool for advancing research in automated short answer grading.

Numerous studies have already utilized the dataset for training or evaluating models. Badry et al. (2023) focuses on developing an ASAG model utilizing semantic similarity techniques. They achieved an F1-score of 82.82% and an RMSE (Root-Mean-Square Error) of 0.798 on the Ar-ASAG dataset. Ljungman et al. (2021) conducted an extensive Classification Benchmark on Automated Grading of Exam Responses using the Ar-ASAG. A Comparative Study in Arabic Short Answer Grading benchmarking In-Context Meta-Learning vs. Semantic Score-Based Similarity is conducted using the Ar-ASAG dataset by Fateen and Mine (2023). A Comprehensive Review of Arabic Question Answering Datasets is presented in (Saoudi & Gammoudi, 2024).

Second, we have presented the first approach that uses paraphrase generation to improve automatic short answer scoring. Our model for generating alternative answers can generate several paraphrases for a given reference answer. The proposed model demonstrates improvement and comparable accuracy compared to the current best methods on both the Arabic and English datasets. The fact that humans have judged the generated paraphrases to be well formed, grammatically correct, and pertinent to the input sentences demonstrates this.

By customizing the paraphrase generation method for ASAG, we were able to offer a wide range of reference answers capable of accommodating the diverse nature of student responses. The paraphrase generator relieves the teacher of the tedious task of manually constructing multiple formulations of the reference answer.

Third, generating higher paraphrases in NLP remains a difficult task. The rich and intricate morphological structure of the Arabic language poses unique challenges for paraphrase generation. The lack of linguistic resources, such as lexicons, dictionaries,

and corpora, significantly hinders the task of paraphrase generation in Arabic NLP. Our thesis addresses this gap in Arabic NLP and provides the basis for further studies as well as a novel deep learning-based formulation of the Arabic paraphrase generation task.

Finally, automating the assessment of short answers for large student cohorts while delivering immediate feedback necessitates a comprehensive computer-based solution. We explored key concepts for the straightforward deployment, scalable real-world application, and incremental development of ASAG systems within higher education settings.

5.3 PRACTICAL IMPLICATIONS

The research conducted in this thesis on ASAG systems has shown significant improvements in educational settings, including enhanced efficiency, objectivity, and scalability. Specific real-world applications and case studies have illustrated the transformative impact of ASAG on teaching practices, learning outcomes, and educational equity. It shows how integrating short answer assessment into the e-learning environment can improve traditional higher education teaching and learning and facilitate an adaptive assessment design.

- We have developed ISAGe as a free open-source plugin for Moodle, the leading open-source VLE globally. The LMS has proven to be a suitable platform for developing our grader question engine due to its user-friendly interface and robust features. In recent years, the increasing adoption at universities has led us to combine expertise and resources to fully embed and enhance the assessment of free-text short answer questions available in Moodle through a supervised learning approach. Results from experiments and feedback indicate that both teachers and students find ISAGe user-friendly and comfortable to use.
- Students value the immediate feedback provided by ISAGe as it plays a crucial role in sustaining their motivation and engagement with the learning process. Assessment, course materials, and learning activities are all on the same platform. Because students know the environment, the stress of the exam is reduced greatly. Furthermore, ISAGe offers repeat practice opportunities for

formative assessment. Assignments are designed to allow students to complete them at their own pace.

- Teachers, especially in the larger classes, may combine automated ISAGe questions in their quizzes because they do not impose writing constraints on student responses. Traditionally, on Moodle, the short answer question grader is based on grammars, or pattern-matches, the formulation of the answer is constrained to respect several constraints. The students were penalized for additional space, a spelling error, etc. Conversely, teachers are required to anticipate the diverse formulations that students may use in their responses. This task is often laborious or sometimes unachievable in practice due to the wide range of possible student answers.
- Using the grader in a formative test that does not contribute to the grade of the course may allow students to use continuous assessment as extra practice. Having a sequence of assessment tasks (training, final test) is less intimidating to students than a single final test and results in a higher participation rate and an improvement in scores.
- In an attempt to align teaching and assessment, and as we cannot ensure that the tool works most appropriately all the time, ISAGe also ensures the possibility of intervention by the teacher, who can re-examine the quality of the submissions of students' answers.
- ISAGe is integrated into the LMS platform as a smoothing extension, allowing it to fully utilize all of the platform's features. Data on the performance of students on individual questions is collected and made available. All efforts for each question are kept, even those created during the initial reflections. An analysis can be conducted to evaluate the progress of individual students throughout an examination. The instructor can discern the elements of the course that may warrant more teaching endeavors in the future. They can also identify inquiries that are troublesome, potentially due to inadequate formulation by the teacher. The instructor can then adapt the assessment design by incorporating insights gained from student progress and feedback.
- By integrating the ASAG system into Moodle, teachers may unleash their creativity to create customized materials and exams tailored to each course's

needs. Furthermore, it offers timesaving features like randomly generated questions with numerous viable responses, along with constructive and motivating automatic feedback for both summative and formative assessments. Research studies have shown that integrating the ASAG system boosts student engagement and academic achievement while enhancing the adaptability of learning settings.

5.4 RECOMMENDATIONS

In light of the comprehensive examination and analysis conducted in this thesis on automatic short answer grading (ASAG) systems, it is evident that there are several recommendations for further exploration and enhancement. These recommendations derived from the research findings, observations, and cases studies aimed at guiding in the development, use, and implementation of ASAG systems aimed at guiding in the development, use, and implementation of ASAG systems and enhancing educational outcomes:

- Due to the subjective nature of the grading process one aspect where our grader may lack lies in the reference-based scoring approach in general. The grader predicts score according to the provided reference answer. Although we proposed to automatically, generate several paraphrases of the reference answer, it is important and strongly recommended that the question be asked appropriately to ensure that correct student responses are predictable (objective design of the question). When formulating the question, the teacher should assess if it corresponds to the desired learning objectives and if its scope is clearly stated. He should ensure that the scope is sufficiently defined for students to respond to within the designated period. In addition, the question may offer adequate instructions to guide the learner towards the anticipated answer.
- Short answer questions sometimes tend to deviate rapidly from the main topic. Establishing clear rules for students and being precise when formulating short response questions in asynchronous e-learning settings are crucial for maintaining academic integrity and ensuring fair assessment practices. Learners should be given detailed instructions and examples to ensure a precise understanding of the expectations placed on them. Focusing on specific ideas or concepts that learners

must contemplate is more effective than covering broad topics as it promotes deeper understanding and critical thinking skills.

- Along with quantitative performance reporting, researchers should uniformly describe failure modes and conduct qualitative error analyses to understand strengths and weaknesses. The field needs consensus around which metrics best capture different aspects of grading accuracy, reliability, consistency, and calibration from simple accuracy to inter-annotator agreement scores.
- Finally yet importantly, we would recommend developing a strategic planning framework in higher education in Algeria that considers the resistance to change against e-assessment in general and particularly short answers in formative and summative evaluations. The framework should offer the higher education organization a tool with which top management could follow a planning cycle to adopt e-assessment practices in general.

  The lack of such Strategic Information System Planning seems to be an obstacle to the effective use of our developed approach and tool and to any other approaches that promote e-assessment in general. For investments to have a significant impact on the quality of teaching and learning, we believe that this must go through the definition of the expected impacts and changes with a large-scale strategic adoption.

  This research opens the discussion on the need to take charge of the governance aspect in the field of Educational Information Systems. The implementation of the ASAG provokes and requires changes in the organization and supposes the commitment to new strategies of student's assessment and teachers behaviors.

  Designing a roadmap for deploying ASAG tools across various courses requires collaboration between researchers and subject matter experts in LMS governance for broad adoption. This collaboration facilitates user-friendly interfaces, making it easier for instructors from diverse technical backgrounds to use the tools without requiring extensive developer assistance. Moreover, the ASAG tool should be deployable as a web application or standalone tool to ensure accessibility across different platforms and environments, enhancing its usability and reach.

  This underscores the necessity for academic institutions to adopt innovative strategies to address the challenges of knowledge and skill assessment. Enhancing

efficiency and elevating the quality of education are key motivators for transitioning to automated knowledge assessment systems. Strategically, it is crucial to refine assessment delivery methods and establish the appropriate organizational structure and infrastructure to support the e-assessment process.

5.5 FURTHER RESEARCH

Several areas emerged during this research where more research would have been required:

- Despite the introduction of a new dataset, the scarcity of datasets remains a significant challenge for ASAG systems. Therefore, it is crucial to invest more effort into developing diverse and large-scale datasets for training and evaluating ASAG models. To address this issue, we investigate the potential of data augmentation and synthetic data generation in mitigating this challenge. Future research may focus on how various data augmentation techniques, such as paraphrasing, synonym replacement, and noise injection, can enhance the size and diversity of the proposed dataset. This, in turn, could improve the generalizability of ASAG models to new and unseen questions and answers. Additionally, exploring how models trained on augmented and synthetic data can become more robust to variations in student responses, including differences in writing styles, grammatical errors, and informal language usage, will be a key area of investigation.
- In relation to generating paraphrases, there is still a need for us to focus on enhancing the quality of the alternative reference responses in terms of their readability and relevance. The process necessitates the retraining of the generator model using extensive and contextually diverse datasets. The primary hindrance in this situation is the significant shortage of parallel corpora available in Arabic. This matter requires a serious and focused approach. Our objective for future work is to generate automatically Arabic datasets for the purpose of paraphrase creation. We approach this objective through two distinct methods. Firstly, by leveraging pre-existing parallel bilingual corpora in other languages. Secondly, by utilizing unstructured web information to automatically build various datasets pertaining to certain areas, A recently released Arabic version of the mt5 model language, known as AraT5 (Nagoudi et al., 2022), is now accessible (Xue et al.,

2021). An interesting direction would involve examining the enhancement of paraphrased reference answers by fine-tuning the AraT5 model using Arabic parallel corpora.

- An important aspect of ASAG task is the focus on understanding audience knowledge and tailoring responses accordingly. When formulating concise response questions, the teacher had to consider carefully the intended audience. Integrating the planned ASAG into the LMS poses challenges in acquiring information about the public's history prior to question development. Our objective is to investigate the efficacy of feedback in predicting student behavior, specifically focusing on comprehension levels, vocabulary, and time allocation for completing assessments. The ASAG's integration into the e-learning environment enables the availability of all this input to support Student's Performance Analyzer Tools. The student's assessment and performance analyzers may be crucial for evaluating learning behavior in an online learning environment. Moodle's integrated analytics and reporting tools offer instructors comprehensive insights into student performance and learning progress.
- Emphasizing the critical nature of prioritizing security and reliability in online assessments, as ASAG systems may be vulnerable to cheating if not managed effectively. Addressing common ethical issues such as fairness and plagiarism in the development of an ASAG system for integration within a standalone LMS also brings up concerns and challenges that may be considered. ASAG systems should include capabilities like plagiarism detection in order to sustain assessment integrity in e-learning contexts. Through the identification and prevention of dishonest actions, these systems maintain the integrity of academic norms and guarantee the reliability of online evaluations. The idea is to investigate the integration of a plagiarism detection plugin in order to provide a similarity report to identify instances of textual plagiarism in student responses. Our initial proposal is to smoothly incorporate a plagiarism detection extension into the Learning Management System (LMS) and establish a connection with the ASAG plugin in a manner that aligns with the university's academic integrity strategies.

Bibliography

Ab Aziz, M. J., Ahmad, F. D., Ghani, A. A. A., & Mahmod, R. (2009). Automated marking system for short answer examination (AMS-SAE). *2009 IEEE Symposium on Industrial Electronics and Applications, ISIEA 2009 - Proceedings*, *1*, 47–51. https://doi.org/10.1109/ISIEA.2009.5356500

Abbirah, A., Joorabchi, A., & Hayes, M. J. (2022). On Deep Learning Approaches to Automated Assessment: Strategies for Short Answer Grading. *International Conference on Computer Supported Education, CSEDU - Proceedings*, *2*, 85–94. https://doi.org/10.5220/0011082100003182

Abdeljaber, H. A. (2021). Automatic Arabic Short Answers Scoring Using Longest Common Subsequence and Arabic WordNet. *IEEE Access*, *9*, 76433–76445. https://doi.org/10.1109/ACCESS.2021.3082408

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2013). On the evaluation and improvement of Arabic WordNet coverage and usability. *Language Resources and Evaluation*, *47*(3), 891–917. https://doi.org/10.1007/s10579-013-9237-0

Adams, O., Roy, S., & Krishnapuram, R. (2016). Distributed Vector Representations for Unsupervised Automatic Short Answer Grading. *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, 20–29. https://aclanthology.org/W16-4904

Agarwal, D., Gupta, S., & Baghel, N. (2020). A Dataset for Automated Short Answer Grading of Children{'}s free-text Answers in Hindi and Marathi. *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 430–436. https://aclanthology.org/2020.icon-main.58

Agarwal, R., Khurana, V., Grover, K., Mohania, M., & Goyal, V. (2022). Multi-Relational Graph Transformer for Automatic Short Answer Grading. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2001–2012. https://doi.org/10.18653/v1/2022.naacl-main.146

Al-Raisi, F., Bourai, A., & Lin, W. (2018). NEURAL SYMBOLIC ARABIC PARAPHRASING WITH AUTOMATIC EVALUATION. *Computer Science & Information Technology*, 01–13. https://doi.org/10.5121/CSIT.2018.80601

Al-Raisi, F., Lin, W., & Bourai, A. (2018). A Monolingual Parallel Corpus of Arabic. *Procedia Computer Science*, *142*, 334–338. https://doi.org/10.1016/J.PROCS.2018.10.487

Al-Shalabi, E. F. (2016). An Automated System for Essay Scoring of Online Exams in Arabic based on Stemming Techniques and Levenshtein Edit Operations. *International Journal of Computer Science Issues*, *13*(5), 45–50. https://doi.org/10.20943/01201605.4550

Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, *49*(3), 721–751. https://doi.org/10.1007/s10579-014-9284-1

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, I. S. (2020).

Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, *1*(May), 1–7. https://github.com/codelucas/newspaper

Alkhatib, M., & Shaalan, K. (2018). Paraphrasing Arabic Metaphor with Neural Machine Translation. *Procedia Computer Science*, *142*, 308–314. https://doi.org/10.1016/j.procs.2018.10.493

Allison, L., & Dix, T. I. (1986). A bit-string longest-common-subsequence algorithm. *Information Processing Letters*, *23*(5), 305–310. https://doi.org/10.1016/0020-0190(86)90091-8

Altinpulluk, H., & Kesim, M. (2021). A SYSTEMATIC REVIEW OF THE TENDENCIES IN THE USE OF LEARNING MANAGEMENT SYSTEMS. *Turkish Online Journal of Distance Education*, *22*(3), 40–54. https://doi.org/10.17718/TOJDE.961812

Amur, Z. H., & Hooi, Y. K. (2022). State-of-the-Art: Assessing Semantic Similarity in Automated Short-Answer Grading Systems. *Information Sciences Letters*, *11*(5), 1851–1858. https://doi.org/10.18576/isl/110540

Appiah, M., & Van Tonder, F. (2019, April 1). Students' Perceptions of E-assessment at a Higher Education Institution. *5th International Conference on Computing Engineering and Design, ICCED 2019*. https://doi.org/10.1109/ICCED46541.2019.9161088

Ashton, H. S., Beevers, C. E., Milligan, C. D., Schofield, D. K., Thomas, R. C., & Youngson, M. A. (2005). Moving beyond objective testing in online assessment. In *Online Assessment and Measurement: Case Studies from Higher Education, K-12 and Corporate* (pp. 116–128). IGI Global. https://doi.org/10.4018/978-1-59140-497-2.ch008

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, D. B. (2018). Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 7371–7379.

Azad, S., Chen, B., Fowler, M., West, M., & Zilles, C. (2020). Strategies for deploying unreliable AI graders in high-transparency high-stakes exams. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12163 LNAI*, 16–28. https://doi.org/10.1007/978-3-030-52237-7_2

Babych, B. (2014). Automated MT evaluation metrics and their limitations. *Tradumàtica: Tecnologies de La Traducció*, *12*, 464. https://doi.org/10.5565/rev/tradumatica.70

Bachman, L. F., Carr, N., Kamei, G., Kim, M., Pan, M. J., Salvador, C., & and Sawaki, Y. (2002). A Reliable Approach to Automatic Assessment of Short Answer Free Responses. In and Y.-F. S.-C. Tseng, T.-E. Chen & E. Liu (Eds.), *Proceedings of the Nineteenth International Conference on Computational Linguistics, volume 2 of COLING '02* (pp. 1–4). Association for Computational Linguistics.

Badry, R. M., Ali, M., Rslan, E., & Kaseb, M. R. (2023). Automatic Arabic Grading

System for Short Answer Questions. *IEEE Access*, *11*, 39457–39465. https://doi.org/10.1109/ACCESS.2023.3267407

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015, September 1). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1409.0473v7

Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, 107–115. https://doi.org/10.5555/1631836.1631849

Bär, D., Biemann, C., Gurevych, I., & Zesch, T. (2012). UKP: Computing semantic textual similarity by combining multiple content similarity measures. **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, *2*, 435–440. https://aclanthology.org/S12-1059

Barrón-Cedeño, A., Rosso, P., Agirre, E., & Labaka, G. (2010). Plagiarism detection across distant language pairs. *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, *2*, 37–45. https://www.researchgate.net/publication/221102799_Plagiarism_Detection_across_Distant_Language_Pairs

Beckman, K., Apps, T., Bennett, S., Dalgarno, B., Kennedy, G., & Lockyer, L. (2019). Self-regulation in open-ended online assignment tasks: the importance of initial task interpretation and goal setting. *Studies in Higher Education*. https://doi.org/10.1080/03075079.2019.1654450

Benharzallah, M. (2020). E-learning at the Algerian University Reality and challenge. *The Journal of Distance Learning and Open Learning*, *November*. https://www.riemysore.ac.in/ict/index.html

Bennouar, D. (2013). Challenges of e-Test in a University Context. *Proceeding of 1st Elearning Spring School*, 27–30.

Bennouar, D. (2017). An Automatic Grading System Based on Dynamic Corpora. *The International Arab Journal of Information Technology*, *14*(4A). https://pdfs.semanticscholar.org/f188/a42968741ca733178701766d1eebb9f0a410.pdf?_ga=2.110053895.340036640.1535244474-1251382057.1535244474

Blokdyk, G. (2017). *Iterative and Incremental Development: Practical Design Techniques*. CreateSpace Independent Publishing Platform. https://books.google.dz/books?id=exg1swEACAAJ

Bloom, B. S. (1984). Taxonomy of Educational Objectives Book 1: Cognitive Domain. In *nancybroz.com*. http://nancybroz.com/nancybroz/Literacy_I_files/Bloom Intro.doc

Boitshwarelo, B., Reedy, A. K., & Billany, T. (2017). Envisioning the use of online tests in assessing twenty-first century learning: a literature review. *Research and Practice in Technology Enhanced Learning*, *12*(1). https://doi.org/10.1186/s41039-017-0055-7

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Brown, S., Glasner, E., & Angela, E. (1999). *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches.* https://eric.ed.gov/?id=ED434545

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. In *International Journal of Artificial Intelligence in Education* (Vol. 25, Issue 1, pp. 60–117). Springer New York LLC. https://doi.org/10.1007/s40593-014-0026-8

Burstein, J., Wolff, S., & Lu, C. (1999). *Using Lexical Semantic Techniques to Classify Free-Responses* (pp. 227–244). https://doi.org/10.1007/978-94-017-0952-1_11

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers and Education*, *55*(2), 489–499. https://doi.org/10.1016/j.compedu.2010.02.012

Byrnes, K. G., Kiely, P. A., Dunne, C. P., McDermott, K. W., & Coffey, J. C. (2021). Communication, collaboration and contagion: "Virtualisation" of anatomy during COVID-19. In *Clinical Anatomy* (Vol. 34, Issue 1, pp. 82–89). John Wiley & Sons, Ltd. https://doi.org/10.1002/ca.23649

Cahuantzi, R., Chen, X., & Güttel, S. (2021). *A comparison of LSTM and GRU networks for learning symbolic sequences*. http://eprints.maths.manchester.ac.uk/

Callear, D., Jerrams-Smith, J., & Soh, V. (2001). CAA of short non-MCQ answers. *Proceedings of the 5th CAA Conference, Loughborough: Loughborough University*. /articles/conference_contribution/CAA_of_Short_Non-MCQ_Answers/9488924/1

Carbonell, J., & Goldstein, J. (1998). Use of MMR, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 335–336. https://doi.org/10.1145/290941.291025

Carneiro, T., Da Nobrega, R. V. M., Nepomuceno, T., Bian, G. Bin, De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, *6*, 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. In Association for Computational Linguistics (Ed.), *Proceedings ofthe 11th International Workshop on Semantic Evaluations (SemEval-2017)* (pp. 1–14). https://doi.org/10.18653/v1/S17-2001

Chaganty, A. T., Mussmann, S., & Liang, P. (2018). The price of debiasing automatic metrics in natural language evaluation. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the*

Conference (Long Papers), *1*, 643–653. https://doi.org/10.48550/arxiv.1807.02202

Chen, M., Tang, Q., Wiseman, S., & Gimpel, K. (2020). Controllable paraphrase generation with a syntactic exemplar. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 5972–5984. https://doi.org/10.18653/v1/p19-1599

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1724–1734. https://doi.org/10.3115/v1/d14-1179

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. https://arxiv.org/abs/1412.3555v1

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *Alt-J*, *13*(1), 17–31. https://doi.org/10.1080/0968776042000339772

Cummins, R., Zhang, M., & Briscoe, T. (2016). Constrained multi-task learning for automated essay scoring. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *2*, 789–799. https://doi.org/10.18653/v1/p16-1075

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://github.com/tensorflow/tensor2tensor

Dias, S. B., Hadjileontiadou, S. J., Diniz, J., & Hadjileontiadis, L. J. (2020). DeepLMS: a deep learning predictive model for supporting online learning in the Covid-19 era. *Scientific Reports*, *10*(1), 1–17. https://doi.org/10.1038/s41598-020-76740-9

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. https://doi.org/10.2307/1932409

Ditters, E. (2013). Issues in Arabic Computational Linguistics. In *The Oxford Handbook of Arabic Linguistics* (pp. 213–240). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199764136.013.010_update_001

Dzikovska, M. O., Moore, J. D., Steinhauser, N., Campbell, G., Farrow, E., & Callaway, C. B. (2010). BEETLE II: A system for tutoring and Computational Linguistics experimentation. *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 13–18. https://aclanthology.org/P10-4003

Dzikovska, M. O., Nielsen, R. D., Brew, C., Giampiccolo, D., Clark, P., Leacock, C., Bentivogli, L., Dagan, I., & Dang, H. T. (2013). SemEval-2013 Task 7: The joint student response analysis and 8th recognizing textual entailment challenge. **SEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics*, *2*, 263–274. https://aclanthology.org/S13-2045

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, *24*(3), 284–332. https://doi.org/10.1007/s40593-014-0017-9

Elghannam, F. (2016). Automatic Measurement of Semantic Similarity among Arabic Short Texts. *Communications on Applied Electronics*, *6*(2), 16–21. https://doi.org/10.5120/cae2016652430

Fateen, M., & Mine, T. (2023). In-Context Meta-Learning vs. Semantic Score-Based Similarity: A Comparative Study in Arabic Short Answer Grading. *ArabicNLP 2023 - 1st Arabic Natural Language Processing Conference, Proceedings*, 350–358. https://doi.org/10.18653/v1/2023.arabicnlp-1.28

Florjancic, V. (2016). Learning Technology for Education in Cloud – The Changing Face of Education. *Communications in Computer and Information Science*, *620*, 168–180. https://doi.org/10.1007/978-3-319-42147-6

Furst, E. J. (1981). Bloom's Taxonomy of Educational Objectives for the Cognitive Domain: Philosophical and Educational Issues. *Review of Educational Research*, *51*(4), 441. https://doi.org/10.2307/1170361

Gaddipati, S. K., Nair, D., & Plöger, P. G. (2020). *Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading*. https://arxiv.org/abs/2009.01303v1

Galhardi, L. B., & Brancher, J. D. (2018). Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11238 LNAI* (pp. 380–391). Springer, Cham. https://doi.org/10.1007/978-3-030-03928-8_31

Gamage, S. H. P. W., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International Journal of STEM Education*, *9*(1). https://doi.org/10.1186/s40594-021-00323-x

Ghouali, K., & Cecilia, R. R. (2021). Towards a moodle-based assessment of algerian efl students' writing performance. *Porta Linguarum*, *2021*(36), 231–248. https://doi.org/10.30827/PORTALIN.VI36.17866

Gomaa, W., & Fahmy, A. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, *68*(13), 13–18. https://doi.org/10.5120/11638-7118

Gomaa, W. H., & Fahmy, A. A. (2012). Short Answer Grading Using String Similarity And Corpus-Based Similarity. *International Journal of Advanced Computer Science and Applications*, *3*(11).

https://doi.org/http://dx.http://dx.doi.org/10.14569/IJACSA.2012.031119

Gomaa, W. H., & Fahmy, A. A. (2014a). Arabic Short Answer Scoring with Effective Feedback for Students. *International Journal of Computer Applications*, *86*(2), 35–41. https://doi.org/10.5120/14961-3177

Gomaa, W. H., & Fahmy, A. A. (2014b). Automatic scoring for answers to Arabic test questions. *Computer Speech and Language*, *28*(4), 833–857. https://doi.org/10.1016/j.csl.2013.10.005

Gomaa, W. H., & Fahmy, A. A. (2020). Ans2vec: A Scoring System for Short Answers. *Advances in Intelligent Systems and Computing*, *921*, 586–595. https://doi.org/10.1007/978-3-030-14118-9_59

Goyal, T., & Durrett, G. (2020). Neural Syntactic Preordering for Controlled Paraphrase Generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 238–252. https://doi.org/10.18653/v1/2020.acl-main.22

Gratta, R. del, Frontini, F., Khan, A. F., Mariani, J., & Soria, C. (2014). The LREMap for Under-Resourced Languages. *In Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, 78.

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2021). Arabic natural language processing: An overview. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 33, Issue 5, pp. 497–507). King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2019.02.006

GUEMIDE, B., & Maouche, S. (2020). Assessment of Online Learning in Algerian Universities during COVID-19. *The International Journal of E-Learning and Educational Technologies in the Digital Media*, *6*(3), 28–58. https://doi.org/10.17781/p002676

Gupta, A., Agarwal, A., Singh, P., & Rai, P. (2018). A deep generative framework for paraphrase generation. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 5149–5156. https://doi.org/10.5555/3504035.3504666

Gusfield, D. (1997). Algorithms on Strings, Trees and Sequences. In *Algorithms on Strings, Trees and Sequences*. Cambridge University Press. https://doi.org/10.1017/cbo9780511574931

Gütl, C. (2007). *e-Examiner: Towards a Fully-Automatic Knowledge Assessment Tool applicable in Adaptive E-Learning Systems* (pp. xxx–xxx). . https://graz.pure.elsevier.com/en/publications/e-examiner-towards-a-fully-automatic-knowledge-assessment-tool-ap

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18. https://doi.org/10.1145/1656274.1656278

Hall, P. A. V., & Dowling, G. R. (1980). Approximate String Matching. *ACM Computing Surveys (CSUR)*, *12*(4), 381–402. https://doi.org/10.1145/356827.356830

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). *Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers*. https://arxiv.org/abs/2204.03503v1

Harris, Z. (1968). *Mathematical Structures of Language*.

Hassan, S., Fahmy, A. A., & El-Ramly, M. (2018). Automatic short answer scoring based on paragraph embeddings. *International Journal of Advanced Computer Science and Applications*, *9*(10), 397–402. https://doi.org/10.14569/IJACSA.2018.091048

Hettiarachchi Enosha, Antonia, H., Mor Enric, Guerrero-Roldán, & Ana-Elena. (2015). Improving student performance in high cognitive level courses by using formative e-assessment. *International Journal of Technology Enhanced Learning*, *7*(2), 116–133. https://doi.org/10.1504/IJTEL.2015.072027

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., Flor, M., Madnani, N., Tetreault, J., Blanchard, D., Napolitano, D., Lee, C. M., & Blackmore, J. (2014). *Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring*. http://arxiv.org/abs/1403.0801

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Horn, R. A., & Johnson Frontmatter, C. R. (2012). *Matrix Analysis Second Edition*. www.cambridge.org

Hsu, S., Wentin, T., Zhang, Z., & Fowler, M. (2021, May 6). Atitudes surrounding an imperfect ai autograder. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445424

Huang, S., Wu, Y., Wei, F., & Luan, Z. (2019). Dictionary-guided editing networks for paraphrase generation. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 6546–6553. https://doi.org/10.1609/AAAI.V33I01.33016546

Huang, X., Bidart, R., Khetan, A., & Karnin, Z. (2022). Pyramid-BERT: Reducing Complexity via Successive Core-set based Token Selection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 8798–8817. https://doi.org/10.18653/v1/2022.acl-long.602

Islam, A., & Inkpen, D. (2008). Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data*, *2*(2), 1–25. https://doi.org/10.1145/1376815.1376819

Jayashankar, S., & Sridaran, R. (2017). Superlative model using word cloud for short answers evaluation in eLearning. *Education and Information Technologies*, *22*(5), 2383–2402. https://doi.org/10.1007/s10639-016-9547-0

Jordan, S. (2012). Short-answer e-assessment questions: five years on. In G. W. and L. G. D. Whitelock, W. Warburton (Ed.), *Proceedings of CAA 2012*

International Conference, Southampton. http://caaconference.co.uk/pastConferences/2012/caa2012_submission_3.pdf

Jordan, S. (2013). E-assessment: Past, present and future. *New Directions*, *9*(1), 87–106. https://doi.org/10.11120/ndir.2013.00009

Jordan, S., & Butcher, P. (2013, April). Does the Sun orbit the Earth? Challenges in using short free-text computer-marked questions. *HEA STEM Annual Learning and Teaching Conference 2013: Where Practice and Pedagogy Meet*. http://www.heacademy.ac.uk/events/detail/2012/17_18_Apr_HEA_STEM_2013_Conf_Bham

Jordan, S., & Mitchell, T. (2009). e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, *40*(2), 371–385. https://doi.org/10.1111/j.1467-8535.2008.00928.x

Kazemnejad, A., Salehi, M., & Soleymani Baghshah, M. (2020). Paraphrase Generation by Learning How to Edit from Samples. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6010–6021. https://doi.org/10.18653/v1/2020.acl-main.535

Kenter, T., & de Rijke, M. (2015). Short Text Similarity with Word Embeddings. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*. https://doi.org/10.1145/2806416.2806475

Khan, S., & Khan, R. A. (2019). Online assessments: Exploring perspectives of university students. *Education and Information Technologies*, *24*(1), 661–677. https://doi.org/10.1007/s10639-018-9797-0

Khoja, S., & Garside, R. (1999). Stemming arabic text. In *Lancaster, UK, Computing Department, Lancaster University*.

Kingma, D. P., & Ba, J. L. (2015, December 22). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1412.6980

Kingma, D. P., & Welling, M. (2014, December 20). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. https://arxiv.org/abs/1312.6114v10

Király, S., Nehéz, K., & Hornyák, O. (2017). Some aspects of grading java code submissions in MOOCs. *Research in Learning Technology*, *25*. https://doi.org/10.25304/rlt.v25.1945

Kolb, P. (2008). Disco: A multilingual database of distributionally similar words. *Proceedings of KONVENS-2008, Berlin*, *2003*, 37–44. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:DISCO+:+A+Multilingual+Database+of+Distributionally+Similar+Words#0

Koleva, N., Horbach, A., Palmer, A., Ostermann, S., & Pinkal, M. (2014). Paraphrase Detection for Short Answer Scoring. *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, *107*, 59–73.

https://www.aclweb.org/anthology/W14-3505

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., Van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., … Adeyemi, M. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Kumar, A., Ahuja, K., Vadapalli, R., & Talukdar, P. (2020). Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, *8*, 330–345. https://doi.org/10.1162/tacl_a_00318

Kumar, S., Chakrabarti, S., & Roy, S. (2017). Earth mover's distance pooling over siamese LSTMs for Automatic short answer grading. *IJCAI International Joint Conference on Artificial Intelligence*, *0*, 2046–2052. https://doi.org/10.24963/ijcai.2017/284

Kumar, V., & Sharma, D. (2016). Creating Collaborative and Convenient Learning Environment Using Cloud-Based Moodle LMS: An Instructor and Administrator Perspective. *International Journal of Web-Based Learning and Teaching Technologies*, *11*(1), 35–50. https://doi.org/10.4018/IJWLTT.2016010103

Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019). Get IT Scored Using AutoSAS — An Automated System for Scoring Short Answers. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9662–9669. https://doi.org/10.1609/AAAI.V33I01.33019662

Kumaran, V. S., & Sankar, A. (2015). Towards an automated system for short-answer assessment using ontology mapping. *International Arab Journal of E-Technology*, *4*(1), 17–24. https://dblp.org/db/journals/iajet/iajet4.html%0Ahttp://www.iajet.org/Pages/archive-vol-4.aspx%0Ahttp://www.iajet.org/documents/vol.4/no.1/3.pdf

Lai, H., Mao, J., Toral, A., & Nissim, M. (2022). Human Judgement as a Compass to Navigate Automatic Metrics for Formality Transfer. *HumEval 2022 - 2nd Workshop on Human Evaluation of NLP Systems, Proceedings of the Workshop*, 102–115. https://doi.org/10.18653/v1/2022.humeval-1.9

Lavie, A. (2010, November 19). Evaluating the output of machine translation systems. *AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas*. https://www.cs.cmu.edu/~alavie/Presentations/MT-Evaluation-MT-Summit-Tutorial-19Sep11.pdf

Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, *June*, 228–231. https://aclanthology.org/W07-0734/

Leacock, C., & Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In *WordNet: An Electronic Lexical Database*. MIT Press. https://doi.org/10.7551/mitpress/7287.003.0018

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405. https://doi.org/10.1023/A:1025779619903

Levy, O., Zesch, T., Dagan, I., & Gurevych, I. (2013). Recognizing partial textual entailment. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *2*, 451–455. https://aclanthology.org/P13-2080

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138–1150. https://doi.org/10.1109/TKDE.2006.130

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches out (WAS 2004)*, *1*, 25–26. https://aclanthology.org/W04-1013

Lin, D. (1998). Extracting Collocations from Text Corpus. *In Workshop on Computational Terminology, Montreal.* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=35e950f10818b578c685d7ca2b688b95d8022931

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12028

Ljungman, J., Lislevand, V., Pavlopoulos, J., Farazouli, A., Lee, Z., Papapetrou, P., & Fors, U. (2021). Automated Grading of Exam Responses: An Extensive Classification Benchmark. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12986 LNAI*, 3–18. https://doi.org/10.1007/978-3-030-88942-5_1

Lovins, J. (1963). *Development of a stemming algorithm.* M.I.T. Information Processing Group Electronic Systems Laboratory.

Lubis, F. F., Mutaqin, Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A. A., & Rosmansyah, Y. (2021). Automated Short-Answer Grading using Semantic Similarity based on Word Embedding. *International Journal of Technology*, *12*(3), 571–581. https://doi.org/10.14716/ijtech.v12i3.4651

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Luo, W., Liu, F., & Litman, D. (2016). An Improved Phrase-based Approach to Annotating and Summarizing Student Course Responses. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, 53–63. http://www.coursemirror.com/download/dataset2

Machado, M., & Tao, E. (2007). Blackboard vs. Moodle: Comparing user experience of learning management systems. *Proceedings - Frontiers in Education Conference, FIE*. https://doi.org/10.1109/FIE.2007.4417910

Madnani, N., Burstein, J., Sabatini, J., & O'reilly, T. (2013). Automated Scoring of a Summary Writing Task Designed to Measure Reading Comprehension. In editors J. Tetreault, J. Burstein, and C. Leacock (Ed.), *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–168). Association for Computational Linguistics.

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

Mahmoud El-Haj, Kruschwitz, U., & Fox, C. C. (2015). Creating Language Resources for Under-resourced Languages : Methodologies , and experiments with Arabic. *Language Resources and Evaluation*, *49*(3), 549–580. https://doi.org/10.1007/s10579-014-9274-3

Marczak, M., Krajka, J., & Malec, W. (2016). Web-based assessment and language teachers - From Moodle to WebClass. *International Journal of Continuing Engineering Education and Life-Long Learning*, *26*(1), 44–59. https://doi.org/10.1504/IJCEELL.2016.075048

Marvaniya, S., Foltz, P., Saha, S., Sindhgatta, R., Dhamecha, T. I., & Sengupta, B. (2018). Creating scoring rubric from representative student answers for improved short answer grading. *International Conference on Information and Knowledge Management, Proceedings*, 993–1002. https://doi.org/10.1145/3269206.3271755

Mezher, R., & Omar, N. (2016). A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring. *Journal of Applied Sciences*, *16*(5), 209–215. https://doi.org/10.3923/jas.2016.209.215

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 775–780. https://doi.org/10.1.1.65.3690

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://arxiv.org/abs/1301.3781

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2019). Advances in pre-training distributed word representations. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 52–55. https://arxiv.org/abs/1712.09405v1

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013, October 16). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*.

https://arxiv.org/abs/1310.4546v1

Mitchell, T., Russel, T., Broomhead, P., & N., A. (2002). Towards robust computerised marking of free-text responses. In M. Danson (Ed.) (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughboroug University, Loughborouh, UK*. https://www.researchgate.net/publication/28576280_Towards_robust_computerised_marking_of_free-text_responses

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 752–762.

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, 567–575. https://doi.org/10.3115/1609067.1609130

Moodle-Stats. (2024). *Home | stats.moodle.org*. https://stats.moodle.org/

Moodle. (2011). *Regular Expression Short-Answer question type*. https://docs.moodle.org/310/en/Regular_Expression_Short-Answer_question_type

Mustafa, M., Eldeen, A. S., Bani-Ahmad, S., & Elfaki, A. O. (2017). A Comparative Survey on Arabic Stemming: Approaches and Challenges. *Intelligent Information Management*. https://doi.org/10.4236/iim.2017.92003

Nababteh, M., & Deri, M. (2017). arabic. *IJCSNS International Journal of Computer Science and Network Security*, *17*(2), 131–140.

Nael, O., ELmanyalawy, Y., & Sharaf, N. (2022). AraScore: A deep learning-based system for Arabic short answer scoring. *Array*, *13*, 100109. https://doi.org/10.1016/j.array.2021.100109

Nagoudi, E. M. B., Ferrero, J., & Schwab, D. (2017). LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 134–138. https://doi.org/10.18653/v1/s17-2017

Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2015). Ground Truth for Grammaticality Correction Metrics. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 588–593. https://doi.org/10.3115/v1/p15-2097

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nielsen, R. D., Ward, W., Martin, J. H., & Palmer, M. (2008). Annotating students' understanding of science concepts. *Proceedings of the 6th International*

Conference on Language Resources and Evaluation, LREC 2008, 3441–3448. http://www.lrec-conf.org/proceedings/lrec2008/pdf/873_paper.pdf

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers and Education*, *56*(2), 337–345. https://doi.org/10.1016/J.COMPEDU.2010.07.013

Oguguo, B. C. E., Nannim, F. A., Agah, J. J., Ugwuanyi, C. S., Ene, C. U., & Nzeadibe, A. C. (2021). Effect of learning management system on Student's performance in educational measurement and evaluation. *Education and Information Technologies*, *26*(2), 1471–1483. https://doi.org/10.1007/S10639-020-10318-W/METRICS

Omran, A. M. Ben, & Ab Aziz, M. J. (2013). Automatic essay grading system for short answers in English language. *Journal of Computer Science*, *9*(10), 1369–1382. https://doi.org/10.3844/jcssp.2013.1369.1382

Ott, N., Ziai, R., & Meurers, D. (2012). *Creation and analysis of a reading comprehension exercise corpus* (pp. 47–69). John Benjamins Publishing Company. https://doi.org/10.1075/hsm.14.05ott

Ouahrani, L., & Bennouar, D. (2018). A Vector Space Based Approach for Short Answer Grading System. *ACIT 2018 - 19th International Arab Conference on Information Technology*. https://doi.org/10.1109/ACIT.2018.8672717

Ouahrani, L., & Bennouar, D. (2019). Key Challenges for Automatic Short Answer Grading for the Arabic Language. *The National Forum on Digitization Challenges in the Arabic Language*. https://www.researchgate.net/publication/381489195_Key_Challenges_for_Automatic_Short_Answer_Grading_for_the_Arabic_Language

Ouahrani, L., & Bennouar, D. (2020). AR-ASAG An ARabic Dataset for Automatic Short Answer Grading Evaluation. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2634–2643. https://aclanthology.org/2020.lrec-1.321

Ouahrani, L., & Bennouar, D. (2024). Paraphrase Generation and Supervised Learning for Improved Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*. https://doi.org/https://doi.org/10.1007/s40593-023-00391-w

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/v1/d14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies - Proceedings of the Conference, *1*, 2227–2237. https://doi.org/10.18653/v1/n18-1202

Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., & Farri, O. (2016). Neural Paraphrase Generation with Stacked Residual LSTM Networks - ACL Anthology. *Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2923--2934. https://aclanthology.org/C16-1275/

Pribadi, F. S., Permanasari, A. E., & Adji, T. B. (2018). Short answer scoring system using automatic reference answer generation and geometric average normalized-longest common subsequence (GAN-LCS). *Education and Information Technologies 2018 23:6*, *23*(6), 2855–2866. https://doi.org/10.1007/S10639-018-9745-Z

Qiu, R. G. (2019). A systemic approach to leveraging student engagement in collaborative learning to improve online engineering education | International Journal of Technology Enhanced Learning. *International Journal of Technology Enhanced Learning*, *11*(1), 1–19. https://dl.acm.org/doi/10.5555/3302810.3302811

Qu, X., Gu, Y., Xia, Q., Li, Z., Wang, Z., & Huai, B. (2024). A Survey on Arabic Named Entity Recognition: Past, Recent Advances, and Future Trends. *IEEE Transactions on Knowledge and Data Engineering*, *36*(3), 943–959. https://doi.org/10.1109/TKDE.2023.3303136

R. Abbas, A., & S.Al-qazaz, A. (2015). Automated Arabic Essay Scoring (AAES) Using Vectors Space Model (VSM) and Latent Semantics Indexing (LSI). *Engineering and Technology Journal*, *33*(3B), 410–426. https://doi.org/10.30684/etj.33.3b.4

Radford, A., Jeffrey, W., Rewon, C., David, L., Dario, A., & Ilya, S. (2019). Language Models are Unsupervised Multitask Learners | Enhanced Reader. *OpenAI Blog*, *1*(8), 9. https://github.com/codelucas/newspaper

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *Homology, Homotopy and Applications*, *9*(1), 399–438. https://www.bibsonomy.org/bibtex/273ced32c0d4588eb95b6986dc2c8147c/jonaskaiser

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 1–67. https://doi.org/10.48550/arxiv.1910.10683

Ramachandran, L., Cheng, J., & Foltz, P. (2015). Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 97–106. https://doi.org/10.3115/v1/W15-0612

Ramachandran, L., & Foltz, P. (2015). Generating reference texts for short answer scoring using graph-based summarization. *10th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2015 at the 2015 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015, 207–212. https://doi.org/10.3115/v1/w15-0624

Ras, E., & Brinke, D. J. Ten. (2015). Computer assisted assessment: Research into E-assessment. *18th International Conference, CAA 2015 Zeist*, *571*(July). https://doi.org/10.1007/978-3-319-27704-2

Raza, S. A., Qazi, W., Khan, K. A., & Salam, J. (2021). Social Isolation and Acceptance of the Learning Management System (LMS) in the time of COVID-19 Pandemic: An Expansion of the UTAUT Model. *Journal of Educational Computing Research*, *59*(2), 183–208. https://doi.org/10.1177/0735633120960421

Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. In *Systematic Biology* (Vol. 45, Issue 3, pp. 380–385). Taylor and Francis Inc. https://doi.org/10.1093/sysbio/45.3.380

Regragui, Y., Abouenour, L., Krieche, F., Bouzoubaa, K., & Rosso, P. (2016). Arabic WordNet: New content and new applications. *Proceedings of the 8th Global WordNet Conference, GWC 2016*, 330–338. https://aclanthology.org/2016.gwc-1.47

Resnik, P. (1995, November 29). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. https://arxiv.org/abs/cmp-lg/9511007v1

Rocchio, J. (1971). Relevance feedback in information retrieval. In editor Salton, G. (Ed.), *The Smart Re- trieval System - Experiments in Automatic Document Processing* (pp. 313–323). Prentice-Hall, Inc. https://www.bibsonomy.org/bibtex/1c18d843e34fe4f8bd1d2438227857225/bsmyth

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2004). An Improved Method for Deriving Word Meaning from Lexical. *Cognitive Psychology*, *7*, 573–605.

Roy, S., Bhatt, H. S., & Narahari, Y. (2016). *An Iterative Transfer Learning Based Ensemble Technique for Automatic Short Answer Grading*. https://www.cs.york.ac.uk/semeval-2013/task7/

Roy, S., Narahari, Y., & Deshmukh, O. D. (2015). A Perspective on Computer Assisted Assessment Techniques for Short Free-text Answers. *Communications in Computer and Information Science*, *571*(June 2015). https://doi.org/10.1007/978-3-319-27704-2

Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., & Sengupta, B. (2018). Sentence level or token level features for automatic short answer grading?: use both. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10947 LNAI*, 503–517. https://doi.org/10.1007/978-3-319-93843-1_37

Sahlgren, M. (2005). An introduction to random indexing. *IN METHODS AND APPLICATIONS OF SEMANTIC INDEXING WORKSHOP AT THE 7TH INTERNATIONAL CONFERENCE ON TERMINOLOGY AND KNOWLEDGE*

ENGINEERING, TKE 2005. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.96.2230

Said, M. M. T., Aravind, V. R., James, D. F., & Umachandran, K. (2019). Dissecting assessment: A paradigm shift towards technology-enhanced assessments. *World Journal on Educational Technology: Current Issues*, *11*(2), 162–170. https://doi.org/10.18844/wjet.v11i2.4256

Sakaguchi, K., Heilman, M., & Madnani, N. (2015). Effective feature integration for automated short answer scoring. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*.

Salam, M. A., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. *PLoS ONE*, *17*(8 August), e0272269. https://doi.org/10.1371/journal.pone.0272269

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Saoudi, Y., & Gammoudi, M. M. (2024). A Comprehensive Review of Arabic Question Answering Datasets. *Communications in Computer and Information Science*, *1961 CCIS*, 278–289. https://doi.org/10.1007/978-981-99-8126-7_22

Schneider, J., Richner, R., & Riser, M. (2023). Towards Trustworthy AutoGrading of Short, Multi-lingual, Multi-type Answers. *International Journal of Artificial Intelligence in Education*, *33*(1), 88–118. https://doi.org/10.1007/s40593-022-00289-z

Scikit-learn. (2019). *scikit-learn: machine learning in Python — scikit-learn 0.21.0*. https://scikit-learn.org/stable/

Setiadi, P. M., Alia, D., Sumardi, S., Respati, R., & Nur, L. (2021). Synchronous or asynchronous? Various online learning platforms studied in Indonesia 2015-2020. *Journal of Physics: Conference Series*, *1987*(1). https://doi.org/10.1088/1742-6596/1987/1/012016

Shehab, A., Faroun, M., & Rashad, M. (2018). An automatic Arabic essay grading system based on text similarity algorithms. *International Journal of Advanced Computer Science and Applications*, *9*(3), 263–268. https://doi.org/10.14569/IJACSA.2018.090337

Shen, L., Liu, L., Jiang, H., & Shi, S. (2022). On the Evaluation Metrics for Paraphrase Generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 3178–3190. https://doi.org/10.18653/v1/2022.emnlp-main.208

Shermis, M. D. (2015). Contrasting State-of-the-Art in the Machine Scoring of Short-Form Constructed Responses. *Educational Assessment*, *20*(1), 46–65. https://doi.org/10.1080/10627197.2015.997617

Shore, J., & Warden, S. (2008). The art of agile development. In *Theory in practice*. https://www.jamesshore.com/v2/books/aoad1/incremental_design

Siddiqi, R., Harrison, C. J., & Siddiqi, R. (2010). Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies*, *3*(3), 237–249. https://doi.org/10.1109/TLT.2010.4

Sima, D., Schmuck, B., Szöll, S., & and Miklós, A. (2009). Intelligent Short Text Assessment in eMax. *Studies in Computational Intelligence*, *243*, 435–445.

Sukkarieh, J. Z., & Blackmore, J. (2009). c-rater: Automatic Content Scoring for Short Constructed Responses. *Proceedings of the 22nd International FLAIRS Conference. Association for the Advancement of Artificial Intelligence*, 290–295. https://www.ets.org/research/policy_research_reports/publications/chapter/2009/imsb

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003). Auto-marking: using computational linguistics to score short, free text responses. *The Annual Conference of the International Association for Educational Assessment (IAEA), Manchester, UK*, 1–15. http://www.ucles.org.uk

Sultan, M. A., Salazar, C., & Sumner, T. (2016). Fast and easy short answer grading with high accuracy. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 1070–1075. https://doi.org/10.18653/v1/n16-1123

Sun, J., Ma, X., & Peng, N. (2021). AESOP: Paraphrase Generation with Adaptive Syntactic Control. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5176–5189. https://doi.org/10.18653/v1/2021.emnlp-main.420

Sychev, O., Anikin, A., & Prokudin, A. (2020). Automatic grading and hinting in open-ended text questions. *Cognitive Systems Research*, *59*, 264–272. https://doi.org/10.1016/j.cogsys.2019.09.025

Takano, S., & Ichikawa, O. (2022). Automatic scoring of short answers using justification cues estimated by BERT. *BEA 2022 - 17th Workshop on Innovative Use of NLP for Building Educational Applications, Proceedings*, 8–13. https://doi.org/10.18653/v1/2022.bea-1.2

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003*, 252–259. https://doi.org/10.3115/1073445.1073478

Tulu, C. N., Ozkaya, O., & Orhan, U. (2021). Automatic Short Answer Grading with SemSpace Sense Vectors and MaLSTM. *IEEE Access*, *9*, 19270–19280. https://doi.org/10.1109/ACCESS.2021.3054346

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser,

Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*, 5999–6009.

Wali, W., Gargouri, B., & Ben Hamadou, A. (2015). Supervised learning to measure the semantic similarity between Arabic sentences. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9329*, 158–167. https://doi.org/10.1007/978-3-319-24069-5_15

Wang, H. C., Chang, C. Y., & Li, T. Y. (2008). Assessing creative problem-solving with automated text grading. *Computers and Education*, *51*(4), 1450–1466. https://doi.org/10.1016/j.compedu.2008.01.006

Whitelock, D., & Bektik, D. (2018). *Progress and Challenges for Automated Scoring and Feedback Systems for Large-Scale Assessments* (pp. 1–18). https://doi.org/10.1007/978-3-319-53803-7_39-1

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research, American Statistical Association*, 354–359. https://www.researchgate.net/publication/243772975_String_Comparator_Metrics_and_Enhanced_Decision_Rules_in_the_Fellegi-Sunter_Model_of_Record_Linkage

Winter, R. J. (2014). Agile Software Development: Principles, Patterns, and Practices. *Performance Improvement*, *53*(4), 43–46. https://doi.org/10.1002/pfi.21408

Wu, H., Huang, H., Jian, P., Guo, Y., & Su, C. (2017). BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity. In Association for Computational Linguistics (Ed.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017): Vol. L* (pp. 77–84). http://www.aclweb.org/anthology/S17-2007

Wu, W.-S. (2008). The Application of Moodle on an EFL Collegiate Writing Environment. *Journal of Education and Foreign Languages and Literature*, *7*, 45–56. https://www.researchgate.net/publication/255622991_The_application_of_Moodle_on_an_EFL_collegiate_writing_environment

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1994-June*, 133–138. https://doi.org/10.3115/981732.981751

Wubben, S., van den Bosch, A., & Krahmer, E. (2010). Paraphrase generation as monolingual translation: Data and evaluation. *Belgian/Netherlands Artificial Intelligence Conference*. http://ilk.uvt.nl/

Xu, P., Kumar, D., Yang, W., Zi, W., Tang, K., Huang, C., Cheung, J. C. K., Prince, S. J. D., & Cao, Y. (2021). Optimizing deeper transformers on small datasets. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2089–2102. https://doi.org/10.18653/v1/2021.acl-long.163

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). *mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer*. 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

Yang, Q., Huo, Z., Shen, D., Cheng, Y., Wang, W., Wang, G., & Carin, L. (2020). An end-to-end generative architecture for paraphrase generation. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3132–3142. https://doi.org/10.18653/v1/d19-1309

Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H., Rashwan, M., & Atyia, A. (2015). Word Representations in Vector Space and their Applications for Arabic. In A. Gelbukh (Ed.) (Ed.), *16th international conference, CICLing 2015 Cairo, Egypt, april 14* (Vol. 9041, Issue April, pp. 430–443). Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-18111-0_32

Zeng, D., Zhang, H., Xiang, L., Wang, J., & Ji, G. (2019). User-Oriented Paraphrase Generation with Keywords Controlled Network. *IEEE Access*, *7*, 80542–80551. https://doi.org/10.1109/ACCESS.2019.2923057

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2019). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2019.1648300

Zhao, J., Zhu, T., & Lan, M. (2014). *ECNU: One Stone Two Birds: Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment*. 271–277. https://doi.org/10.3115/v1/s14-2044

Ziai, R., Ott, N., & Meurers, D. (2012). Short Answer Assessment : Establishing Links Between Research Strands. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics*, *2*(2005), 190–200.

Appendices

Appendix A

Question types and behaviors in the Moodle quiz system.

1) Question types [30]**:** The most common question types in Moodle are:

a) *Calculated.* Calculated questions offer a way to create individual numerical questions by the use of wildcards that are substituted with individual values when the quiz is taken.

b) *Calculated multi-choice.* Calculated multi-choice questions are like multi-choice questions with the additional property that the elements to select can include formula results from numeric values that are selected randomly from a set when the quiz is taken.

c) *Calculated simple.* Simple calculated questions offer a way to create individual numerical questions whose response is the result of a numerical formula which contain variable numerical values by the use of wildcards (i.e. {x} , {y}) that are substituted with random values when the quiz is taken.

d) *Drag and drop.* Students select missing words or phrases and add them to text by dragging boxes to the correct location. Items may be grouped and used more than once.

e) *Essay.* This allows students to write at length on a particular subject and must be manually graded.

f) *Matching.* A list of sub-questions is provided, along with a list of answers. The respondent must "match" the correct answers with each question.

g) Embedded Answers (Cloze Test / Gap Fill). These very flexible questions consist of a passage of text (in Moodle format) that has various answers embedded within it, including multiple choice, short answers and numerical answers.

h) Multiple choice. With the Multiple Choice question type you can create single-answer and multiple-answer questions, include pictures, sound or other media in

[30] https://docs.moodle.org/4x/sv/Question_types

the question and/or answer options (by inserting HTML) and weight individual answers.

i) Short Answer. In response to a question (that may include an image), the respondent types a word or phrase. There may several possible correct answers, with different grades. Answers may or may not be sensitive to case.

j) *Numerical.* From the student perspective, a numerical question looks just like a short-answer question. The difference is that numerical answers are allowed to have an accepted error. This allows a continuous range of answers to be set.

k) Select missing words. Students select a missing word or phrase from a dropdown menu. Items may be grouped and used more than once.

l) *True/False.* In response to a question (that may include an image), the respondent selects from two options: True or False.

2) Questions behaviors[31]**:** The following question behaviors are available when creating a quiz:

a) Deferred feedback. Students must enter an answer to each question and then submit the entire quiz before anything is graded or they get any feedback.

b) Adaptive mode (no penalties). Allows students to have multiple attempts at the question before moving on to the next question. This behavior requires that the "Whether correct" box is ticked under "During the attempt" in the "Review options" section, at a minimum.

c) *Manual grading.* Used for essay questions (irrespective of what the quiz is set to) but you can now choose to have every question in the quiz manually graded if you wish.

d) Interactive with multiple tries. Used for allowing multiple attempts on the same question (perhaps with a grade penalty). Students answer the question and click the 'Check' button. If the answer is wrong, the student can click the 'Try again' button to try a new response. Importantly, the question definition must contain hints that will be shown after each incorrect attempt, though the hint text can be as minimal as an HTML non-breaking space. Once the student has got the

[31] https://docs.moodle.org/4x/sv/Question_types

question right, they can no longer change their response. Once the student has the question wrong too many times, they are just graded wrong (or partially correct) and get shown the feedback.

e) *Immediate feedback.* Similar to interactive mode in that the student can submit their response immediately during the quiz attempt, and get it graded. However, they can only submit one response, they cannot change it later.

f) *Deferred feedback or Immediate feedback with Certainty-based marking (CBM).* The student does not only answer the question, but they also indicate how sure they are they got the question right. The grading is adjusted by the choice of certainty so that students have to reflect honestly on their own level of knowledge in order to get the best mark.

g) *Conditional questions.* If using the Interactive with multiple tries or Immediate Feedback behavior and with the navigation method set to 'Free', it is possible to make the display of a question dependent on a previous question being answered first.

Appendix B

Manual evaluation of generated paraphrases - Indications to Arabic experts

السياق والهدف

نحن مهتمون بعملية " إعادة الصياغة " التي تسمح، من جملة معينة، بتوليد عدة جمل ذات معنى مكافئ) إعادة صياغة الجملة الأصلية .(نستخدم نموذج التعلم العميق لعملية البناء .بمجرد تدريب النموذج، قمنا باختباره على مجموعة من جمل الاختبار(الجملة الأصلية، الجملة المولدة) باستخدام المقاييس المحسوبة اليا. نظرًا لأن التقييم الدقيق لإعادة الصياغة يمثل مشكلة مفتوحة نعتقد أن التقييم التلقائي لا يكفي لتقييم إعادة الصياغة من منظور دقيق، من حيث جانبين:

- **الملاءمة :** يعبر عن أهمية إعادة الصياغة المتولدة مع الجملة المدخلة .هنا يتعلق الأمر بمسألة ملاحظة إلى أي مدى تحافظ الجملة المولدة على نفس المعنى مثل الجملة الأصلية.
- **سهولة القراءة** : سهولة قراءة الجملة من حيث الشكل والقواعد دون النظر إلى معنى الجملة المولدة.

لتحديد الجوانب التي لم يتم تناولها بواسطة مقاييس التقييم التلقائية، يصبح التقييم البشري ضروريًا لمشكلتنا . لذلك نقوم بجمع الأحكام البشرية على عينة من 100 زوج من (الجملة الأصلية، الجملة المولدة) يتم أخذ هذه الأزواج عشوائيًا من مجموعة الاختبار التي تم تقييمها تلقائيًا. يتم التحقق من حيث الملاءمة وقابلية القراءة لكل زوجين في العينة. يقوم الخبير البشري بتعيين درجة على مقياس مستمر من 1 إلى 5 لكل جانب لكل إعادة صياغة تم إنشاؤها، حيث 1 هوالأسوأ و 5 هو الأفضل. نقوم بإجراء العديد من التقييمات البشرية لنفس الأزواج من قبل العديد من الخبراء البشريين المتطوعين، وسنحتفظ بمتوسط العلامات التي حصلنا عليها لكل زوجين .عن طريق الارتباط بين الخبراء لأن التقييم البشري أيضًا يظل غير موضوع . يمكننا بالتالي إكمال تقييم نموذجنا على هذين الجانبين المتعلقين بالملاءمة وسهولة القراءة.

العمل المطلوب من الخبير البشري:

1. اقرأ الجملة الأصلية
2. اقرأ الجملة التي تم إنشاؤها.
3. أعط درجة بين 1 و 5 : 1هو الأسوأ و 5 هو الأفضل لجانب" الملاءمة " للجملة التي تم إنشاؤها مقارنةً بالجمل الأصلية.
4. أعط درجة بين 1 و5 (1 هي الأسوأ و 5 هي الأفضل) لجانب" القراءة "من الجملة التي تم إنشاؤها.

Appendix C

Manual evaluation of generated paraphrases - Indications to English experts

We are interested in the process of paraphrasing, which consists of starting from an input sentence to generate one or more sentences equivalent in meaning (a reformulation of the original sentence). We use a deep learning model for the generation process. Once the model was trained, we tested it on a test set of sentences (original and generated) using automatically calculated metrics. The precise evaluation of paraphrases is an open problem. We believe that automatic evaluation is not enough to evaluate paraphrases from a fine perspective, in terms of two aspects.

– **Relevance:** expresses the relevance of the paraphrase generated with the input sentence. Here, it is a question of noting to what extent the generated sentence preserves the same meaning as the original sentence.

– **Readability** (readability in form): the readability of the generated paraphrase in terms of form and grammar without considering the meaning of the generated sentence.

For our issue of producing sentences with the same meaning as an original sentence, human evaluation becomes necessary to quantify the aspects that automatic evaluation metrics do not address. Therefore, we collect human judgments on a sample of 100 pairs of (original phrase, generated phrase). These pairs are taken randomly from the test set already evaluated automatically. The two aspects of relevance and readability are verified in the human evaluation of each pair in the sample. The human expert scores on a continuous scale of 1 to 5 for each aspect by paraphrase generated, where 1 is the worst and 5 is the best. We do multiple human evaluations for the same couples by several volunteer human experts, mastering the language, and we will keep the average of the scores obtained for each couple. Through the correlation between experts (because human evaluation remains subjective), we can thus complete the evaluation of our model on these two aspects relating to relevance and readability.

Work requested from the human expert:

1. Read the original sentence 2.
2. Read the generated sentence
3. Score between 1 & 5 for the "Relevance" aspect of the generated sentence versus the original sentence (1 is the worst and 5 is the best)
4. Score between 1 & 5 for the "Readability" aspect of the generated sentence (1 is the worst and 5 is the best).

Appendix D

Program Teaching "Cybercrimes" Course: Case study 1
Master Title: Software Engineering, Computer Systems & Networks, and Information Systems Security.
Semester: three
Teaching Unit: Discovery Unit
Course: Introduction to Cybercrimes
Learning outcomes:
- Gain a comprehensive understanding of the various categories of cybercrimes.
- Learn the proactive measures to prevent these crimes.
- Importance of integrating IT security considerations from the initial stages of solution development, addressing both organizational and technological aspects (Security by Design).
- Motivation of the human factor against cybercrimes.

Recommended prior knowledge:
- Mandatory: None
- Desirable: Computer Security, Systems and Networks.

Content of the material:
- Concepts of cybercrime (Definitions and examples, Hackers, Crackers, Dark Net, etc.)
- Typology of cybercrimes:
  - Offenses against the confidentiality, integrity and availability of data and computer systems;
  - Offenses relating to content;
  - Offenses linked to infringements of intellectual property,
  - Hostile sites (Religious, Political, Terrorism, etc.)
  - Espionage and invasion of privacy crimes, industrial espionage, etc.
  - Financial crimes (Theft of credit card numbers, Data fraud, Cyber-money laundering, etc.)
  - Organized crimes,
  - Some common infringement techniques: "Defacing" of sites, "Spoofing", "Key loggers", "cracking" of passwords, Computer viruses, "Proxy Bypass", VPN, TOR, etc.
- The challenges of the fight against cybercrime
  - IT security culture (Identification, Authentication, Authorization, Responsibility, Non-repudiation, Privacy),
  - Means of protection: material, technical, and administrative.
  - Social engineering,
  - Biometrics as a means of protection (Definition, Advantages, some techniques (Fingerprint Scanning, Hand Geometry, Iris/Retina/Facial/DNA Scanning, Voice/Signature/Verification, etc.)
- Legal aspects linked to Cybercrime in Algeria
  - Issues,
  - Algerian legislation and cybercrimes

Evaluation method:
  - Final summative exam
  - Online formative activity.