A Comparative Evaluation for Question Answering over Greek Texts by using Machine Translation and BERT*

Michalis Mountantonakis[1,2*†], Loukas Mertzanis[1,2†], Michalis Bastakis[2†] and Yannis Tzitzikas[1,2]

[1]Institute of Computer Science, FORTH, Heraklion, Greece.
[2]Department of Computer Science, University of Crete, Heraklion, Greece.

*Corresponding author(s). E-mail(s): mountant@ics.forth.gr;
Contributing authors: loukas.mertzanis@gmail.com;
mbastakis@gmail.com; tzitzik@ics.forth.gr;
†These authors contributed equally to this work.

Abstract

Although there are numerous and effective BERT models for Question Answering (QA) over plain texts in English, it is not the same for other languages, such as Greek. Since it can be time-consuming to train a new BERT model for a given language, we present a generic methodology for multilingual QA by combining at runtime existing Machine Translation (MT) models and BERT QA models pretrained in English, and we perform a comparative evaluation for Greek language. Particularly, we propose a pipeline that a) exploits widely used MT libraries for translating a question and a context from a source language to the English language, b) extracts the answer from the translated English context through popular BERT models (pretrained in English corpus), c) translates the answer back to the source language, and d) evaluates the answer through semantic similarity metrics based on sentence embeddings, such as Bi-Encoder and BERTScore. For evaluating our system, we use 21 models, whereas we have created

*This is a preprint of the journal paper: Michalis Mountantonakis, Loukas Mertzanis, Michalis Bastakis and Yannis Tzitzikas, "A Comparative Evaluation for Question Answering over Greek Texts by using Machine Translation and BERT", Languages Resources and Evaluation Journal, Springer

a test set with 20 texts and 200 questions and we have manually labelled 4,200 answers. These resources can be reused for several tasks including QA and sentence similarity. Moreover, we use the existing multilingual test set XQuAD, with 240 texts and 1,190 questions in Greek language. We focus on both the effectiveness and efficiency, through manually and machine labelled results. The results of the evaluation show that the proposed approach can be an efficient and effective alternative option to multilingual BERT. In particular, although the multilingual BERT QA model provides the highest scores for both human and automatic evaluation, all the models combining MT and BERT QA models are faster and some of them achieve quite similar scores.

Keywords: Question Answering, Greek Language, Machine Translation, BERT, Short Sentence Similarity, Greek Annotated test set

1 Introduction

Recently, there is a high proliferation of approaches for the task of Question Answering (QA) over texts, which is usually called textual QA-based Machine Reading Comprehension (see a recent survey here [1]). A common way to answer questions for that task is by predicting the correct span of text. For achieving this target, there are several available pretrained BERT QA models in English, and they are quite effective [1, 2]. However, since most models have been pretrained by using English texts, it is not easy to adjust them to other languages, such as Greek, and this is a common problem for several research areas in the field of Natural Language Processing (NLP) [3].

However, there is a high need for providing multilingual solutions for the NLP field, given that 17% of people worldwide speak English[1]. Even in Greece, which is a popular tourist destination and millions of people work in tourism and speak English, a high percentage of people, i.e., 49%, cannot speak and read English[2]. Moreover, the percentages are even higher for other countries, which are also popular tourist destinations, such as Italy and Turkey, where 67% and 83% of people, respectively, cannot speak English.

For offering multilingual QA over texts through BERT for a language L, several methods can be applied, which can require Machine Translation (MT) mechanisms or/and pretrained BERT-based Question Answering (QA) models (e.g., see an example for Swedish [4]). Specifically, we can identify the following categories for multilingual QA over plain texts: C1) Machine-Translation based, i.e., to translate a question written in L language in English and possibly its context, i.e., when the context is not available in English language, to process it with English BERT, and finally to translate the answer back to L, C2) Multilingual BERT QA, i.e., to use existing multilingual pretrained models [5] that have been pretrained in several languages (including L), or C3)

[1] https://www.babbel.com/en/magazine/how-many-people-speak-english-and-where-is-it-spoken
[2] https://doublespeakdojo.com/how-common-is-spoken-english-in-greece/

pretrained BERT QA models for a language L, i.e., use an existing model or train a new BERT QA model by using a huge corpus from the desired language.

Since we desire our methodology to be generic, and given that there are not available monolingual QA approaches in Greek [6], we decided to evaluate the category C1 for the Greek language. Therefore, we present a pipeline that exploits MT and BERT-based pretrained models for QA in Greek language. On the contrary, multilingual models (C2) may not support any given language or all the words of a language and can be slower, while in some languages, there are no available pretrained BERT QA models (C3), e.g., in Greek [6], and it can be very expensive and time consuming to train a new BERT model. In this paper, we desire to answer the following Research Questions (RQ):

- **RQ1**: Can we answer questions effectively and efficiently in Greek, by combining at runtime MT tools and pretrained BERT QA models in English? For answering RQ1, we present a general configurable approach, which can be used for selecting any combination of MT and BERT QA model for answering a given question, and we evaluate the effectiveness and efficiency of the approach by using two test sets.
- **RQ2**: How does the machine translation affect the quality of the QA process? For answering this RQ, we first discuss the main problems that can arise due to the MT. Since Greek Language is quite complex and has a very rich vocabulary, i.e., as it has been written "the Greek vocabulary is the largest in the world and 3.5 times bigger than the English vocabulary."[3], it is challenging to detect such cases, e.g., the golden answer (i.e., the correct answer) may not be part of the translated context, however, the meaning may be the same.
- **RQ3**: Can we evaluate automatically in an effective way the answers, since due to the translation the words can change, but the meaning can remain the same? For tackling the mentioned problems, we exploit automatic similarity metrics that are based on token and sentence embeddings, such as BERTScore [7] and Bi-Encoder [8].

Concerning our contribution for the resources, we have created a test set, called *GreekTexts* containing 20 texts from a Greek text bank, and 10 questions for each such text (i.e., 200 questions in total), whereas we provide a manually annotated set with thousands of answers. These resources are available and can be exploited for QA and other tasks, e.g., sentence similarity. We use the mentioned resources for evaluating the proposed approach, by using 21 different models. Moreover we use the Greek version of an existing multilingual test set, called XQuAD [5], containing 240 texts and 1,190 questions.

Our target is to evaluate both the efficiency and the effectiveness for several different combinations of MT and pretrained BERT QA Models. We carry out a manual evaluation for *GreekTexts* test set, since a human evaluator can easily spot the correct translations that are different from the gold standard. Moreover, we provide an automatic evaluation for both test sets. We compare the mentioned combinations of models with a multilingual pretrained BERT QA model that supports Greek language. Concerning the results, all the models

[3] https://greekcitytimes.com/2021/03/28/english-words-greek-roots/

combining MT and BERT QA models are on average from 2-7 seconds faster comparing to the multilingual. Moreover, although the multilingual model provides the highest scores for both human and automatic evaluation, some models achieve quite similar scores; indicatively for the *GreekTexts* test set the best combination including a MT system and a BERT QA model achieves a human score 0.707 versus 0.715 of the multilingual BERT QA model (perfect score would be 1.0). To the best of our knowledge, this is the first work that evaluates the performance of textual QA in Greek language by combining MT and BERT QA models.

The rest of this paper is organized as follows: Section 2 describes the related work, Section 3 introduces the problem statement and Section 4 provides all the steps of the proposed approach. Section 5 describes the evaluation setup and Section 6 evaluates both the effectiveness and efficiency of MT and QA models over Greek texts. Finally, Section 7 concludes the paper and mentions directions for future work.

2 Related Work

This section analyzes approaches offering multilingual QA over texts, then it discusses BERT models over Greek languages and multilingual approaches over knowledge graphs, and it provides details about the novelty of this paper.

Multilingual QA over Texts. First, several approaches which are analyzed below exploit the Stanford Question Answering Dataset (SQuAD) dataset [9] (or variations of that dataset), which "is a collection of question-answer pairs derived from Wikipedia articles", e.g., the version SQuAD 1.1 contains 107,785 question-answer pairs on 536 Wikipedia articles.

Concerning multilingual QA approaches, there are available either for a single language, or even multilingual approaches that cover multiple languages. Regarding a single language, [10] created a Korean Dataset based on SQuAD for QA through MT, which was tested by using an embeddings model based on the Korean version of Wikipedia. [4] evaluates two different approaches for QA in Swedish, i.e., by using a multilingual BERT model on the English SQuAD and by fine-tuning a Swedish BERT model. [11] describes how to perform multilingual QA by transforming both the question and the context in English. The pipeline is quite similar to the methodology that we follow, however, it has been tested for French and Japanese, and it follows different methods for evaluating the results. [12] proposed several back-translation based approaches, and managed to improve the results for the same datasets in French and Japanese. In [13], a new metric was used for evaluating QA approaches, that are based on MT, for many language pairs, whereas in [14], the Helsinki-NLP MT system [15] was used for translating German Questions and table contents to English, for performing QA over tabular sources.

Moreover, there are several multilingual BERT models that support several languages [5], and can be tested with the XQuAD dataset [5], which is a

Cross-lingual QA Dataset that is available for 11 languages (including Greek), and is based on the collection SQuAD v1.1 [9]. However, since XQuAD covers only a subset of the questions/answers of SQuAD, there have been proposed approaches for translating the whole SQuAD to another language through MT systems, e.g., for Arabic [16], for French [17] and for Spanish [18] language.

Greek BERT Models and QA. Several BERT models have been pretrained through a Greek corpus for numerous tasks, e.g., named entity recognition and linking [19, 20], part of speech tagging [19, 21], whereas in [22] the authors made available a rich set of resources for Greek language, which were evaluated for sentiment analysis. More approaches for Greek language are listed in a recent survey [6]. However, there is no available BERT model that has been pretrained exclusively in a Greek corpus and is suitable for the QA task, e.g., the GreekBERT [19] which is quite popular is used for the task of Masked Language Modelling. On the contrary, there are multilingual BERT models that can answer questions over Greek texts[4]. Finally, concerning QA over Greek, APANTISIS [23] is a modular QA system that can be plugged in relational databases. This approach transforms the question to an SQL query for retrieving the answer, thereby, it does not answer questions over Greek through BERT models.

QA over Knowledge Graphs using MT. The multilingualism in QA is a major challenge even in the area of knowledge graphs and semantic web [24], and there are available systems offering QA by using MT, i.e., over knowledge graphs [3], by MT tools [3] and BERT models in English language [25]. Moreover, multilingual evaluation collections over Knowledge graphs, e.g., QALD-9 [26] are available for evaluating such systems. Although we do not cover multilingual QA over knowledge graphs, this is one of our major future targets, by using the proposed pipeline.

2.1 Comparison with Related Work and Novelty

Although there are available approaches for NLP tasks for the Greek Language [19–21], there is a lack of monolingual BERT QA approaches in Greek [6] and of Greek QA test sets (and generally of Greek NLP collections). These are important problems that should be faced in the next years according to a recent report[5].

Concerning the novelty, to the best of our knowledge, it is the first work providing and evaluating a pipeline for supporting QA in Greek language, that i) combines in a configurable way a MT system and a BERT QA English pretrained model, and ii) compares the performance of several MT/BERT combinations for Greek QA through a comparative study (including also a

[4] https://huggingface.co/deepset/xlm-roberta-large-squad2

[5] https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_17__Language_Report_Greek_.pdf

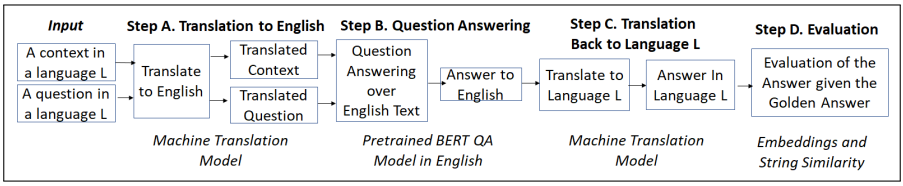

Figure 1 The steps of the proposed process for QA through Machine Translation to English

human evaluation). Thereby, this paper neither proposes a new monolingual BERT pretrained model in Greek nor a Multilingual BERT model.

Regarding the contribution for the resources, we introduce a) a new test set, i.e., *GreekTexts*, that can be used for evaluating any QA system over Greek in the future, and b) human annotated answers for that test set, which can be exploited for other NLP tasks, such as for sentence similarity.

Finally, comparing to our past work, in a recent workshop paper [27], we presented the research prototype *Tiresias*, where we followed the proposed pipeline, for offering bilingual QA in Greek and English over the abstracts of DBpedia [28]. However, comparing to [27], this paper presents all the details for each of the steps of the pipeline, by focusing on the difficulties of the Greek language and by proposing similarity metrics for an automated evaluation, whereas we provide more experiments over larger test sets.

3 Problem Statement

First, let L be a language, q_L a question in L and $cont_L$ a context in L containing the answer of q_L. Moreover, we define as M a specific MT system, and as B a specific BERT QA model.

A. Translation of the Input. The first step is to translate the question and the context in English. We define as $q_{L\to en}(M)$ the translation of the question in English, and as $cont_{L\to en}(M)$ the translation of the context in English, by using a MT system M.

B. Textual QA. Given the translated question and context in English, i.e., $q_{L\to en}(M)$ and $cont_{L\to en}(M)$, and a BERT QA model B the target is to provide a text span of $cont_{L\to en}(M)$ as an answer, i.e., $ans(q_{L\to en}(M), cont_{L\to en}(M), B) = Ans_{en}(q_L)$, where it holds that $cont_{L\to en}(M) = S_1 Ans_{en}(q_L) S_2$. In particular, $Ans_{en}(q_L)$ is a substring of the context, and S_1, S_2 are the string sequences before and after $Ans_{en}(q_L)$, respectively. Each of these sequences, i.e., S_1, S_2, can be possibly the empty string, i.e., if the answer is either the first or the final words of the context.

C. Translation of the Answer. The final step is to translate the answer back to language L by using the same MT system M in the opposite direction, i.e., to provide $Ans_{en\to L}(q_L, M)$. This is the case of using a multilingual MT system M. On the other hand, if the system M is bilingual, we should clarify that it is not the same MT system M, but the one trained in the opposite direction, i.e., say M'.

D. Answer Evaluation. For evaluating the answer, the language of the predicted and the golden answer, say $gAns(q_L)$, should be the same, i.e., the language L. This answer should be compared with the predicted answer $Ans_{en \rightarrow L}(q_L, M)$.

Objective and Challenges. The target is to propose a configurable approach for making it feasible to use any different pair M,B, i.e., MT and BERT QA model, for QA for any language L. Moreover, our objective is to evaluate the approach in Greek ($L =' gr'$), by exploiting different ways to evaluate the produced answer given a gold standard, for overcoming the problems arising from the double translation. Indeed, a common problem of the translation is that in some cases the produced answer of the translated context is not included in the initial context, although its meaning can be correct. For instance, the produced translated answer can be a synonym of the golden answer, whereas more extreme cases can be observed for languages with a complex morphology, such as Greek. Since in these cases, the most commonly used metrics, i.e., Exact Match (EM), and $F1_{score}$, which are based on the comparison of tokens between the golden and the predicted answer, are not so effective, the challenge is to exploit alternative ways for the evaluation, e.g., by using existing metrics that are based on embeddings.

4 How to perform QA using MT and BERT

Here, we introduce the exact steps of the proposed approach, which are depicted in Figure 1.

4.1 Input and Configuration

Rationale. The input is always a context $cont_L$ and a question q_L in a language L. Concerning the configuration, a combination of a MT tool M and a BERT QA model B should be selected. The major point is that we desire the proposed approach to be easily configurable, for supporting any new MT system or/and pretrained BERT QA model in the future. In this way, the efficiency and effectiveness of the introduced approach can be further improved in the future, e.g., by testing new novel combinations of such models.

Running Example. Concerning our running example, the input contains a context and a question in Greek language (see Figure 2). The objective of this example is not only to show the whole process but also a possible problem that can occur due to the translation and concern the evaluation process.

4.2 Step A. Translation from a Language *L* to English

Rationale. It receives the input in natural language form (text) in a source language (e.g., Greek) and then translates it to English, e.g., step A of Figure 1, and the translations of the question and the context in English in Figure 2.

Why translate? Certainly, there are texts that have been already translated in many languages by humans, therefore, in such cases there is

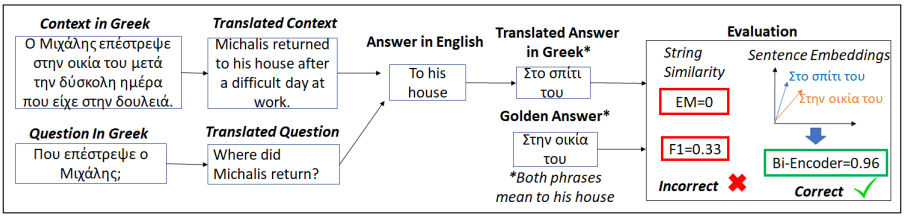

Figure 2 A running example for a context and question in Greek Language

no need for providing a translation for them. In these cases a solution could be to provide a translation only for the question and the answer. However, unfortunately it is not the case for all the texts, e.g., as it is stated here[6], many websites in Greece offering their content only in Greek language are high traffic sites. They mainly include websites having articles in Greek news pages, which are provided in Greek language and they are never translated by human experts to other languages. Moreover, even in multilingual webpages like Wikipedia, many pages are not offered in English language, e.g., see the page of the Greek village Ano Asites (`Wikipedia Page of Ano Asites`), whereas for many pages, the English version contains few (and maybe other) information comparing to other languages, e.g., see the English and Greek version for the page of ship MS Express Samina[7], which was sunk in 2000. Therefore, in these cases, the translation should be done for both the context and the question.

The Process. The first task is to choose which translation model M to use, for translating both the question and the context.

Output. The output is the translated question and context in English, i.e., $q_{L\rightarrow en}(M)$ and $cont_{L\rightarrow en}(M)$ which will be used in the next step.

4.3 Step B. QA through BERT models in English

Rationale. The objective is to perform QA over English language, i.e., by using the $q_{L\rightarrow en}(M)$ and $cont_{L\rightarrow en}(M)$, e.g., see Step B of Figures 1 and 2.

The process. We can use any existing BERT QA model B, which has been pretrained in English. Any of these models receives the translated question and context, and the answer is retrieved from the translated context.

Output. The output of this step is the answer in English, i.e., $Ans_{en}(q_L)$, e.g., in Step B of Figure 2 you can see the produced answer in English language.

4.4 Step C. Translate Back the Answer

Rationale. For providing the final output in language L, it is a prerequisite to translate back the answer to L, e.g., for offering a QA service in a language L.

The process and Output. It is done through the same MT system M as in Step A. The output is the answer in language L, i.e., $Ans_{en\rightarrow L}(q_L, M)$, e.g., in the running example, you can see the translated answer in the Greek

[6] https://w3techs.com/technologies/details/cl-el-
[7] https://en.wikipedia.org/wiki/MS_Express_Samina

language. A key observation in the running example is that the translated answer is not included in the initial Greek context. However, the translated answer and the golden answer have exactly the same meaning in Greek, i.e., both phrases 'Στο σπίτι του' and 'Στην οικία του' mean "to his house".

4.5 Step D. Evaluation of the Answer

Rationale. Here, we discuss ways to evaluate the answer given the golden answer. Since the translation can result to several complex cases (e.g., see Table 1), we provide both a) a manual evaluation by humans, and b) an automatic evaluation based on embeddings and not on exact string matching.

The process and output. We compare the golden answer $gAns(q_L)$ with the produced translated answer $Ans_{en \to L}(q_L, M)$, e.g., see step D of Figure 2, and the output is a similarity score, i.e., a continuous value. In this section we describe different ways for performing the comparison.

4.5.1 Manual Evaluation by Human Experts

For a predicted answer A (e.g., $Ans_{en \to L}(q_L, M)$) and a golden answer G (e.g., $gAns(q_L)$), we define a 3-scale annotation. For each annotation, we provide a continuous value, for enabling the comparison of the human annotation with automatic metrics. In particular, the 3-scales are:

1. $HumanL(A, G) = Correct$ (Score: 1.0): The answer A has exactly the same meaning as G, even it does not contain the same words/suffixes.
2. $HumanL(A, G) = Partially\ Correct$ (Score: 0.5): This includes the following cases: i) omissions, i.e., answer A covers a part of G (i.e., it misses certain details), and ii) substitutions/mistranslations, i.e., some parts of the answer A have been substituted/mistranslated and the meaning has changed.
3. $HumanL(A, G) = Wrong$ (Score: 0.0): The meaning of answer A is totally different compared to the golden answer G.

Although the human evaluation offers the most precise results, it requires a huge effort and can be very time-consuming for large test sets, especially for evaluating several combinations of MT systems and BERT QA models. For this reason, we describe below automatic metrics for evaluation.

4.5.2 Automatic Metrics for QA Evaluation

There are several available metrics for evaluating QA models, including a) general QA metrics, such as F1 score and Exact Match that are widely used for evaluating QA benchmarks [29, 30]. Moreover, there exists b) traditional string-matching based metrics based on n-grams such as [31, 32], BLEU [33], METEOR [34] ROUGE [35], and chrF [36] which is better for high-morphology languages, since it does not penalise very much differences related to endings, prefixes, etc. Finally, there are c) recent MT metrics that

ID	Category	Lang	Model Answer	Golden Answer	EM	F1	Bi-Encoder
1	**Paraphrasing**	en	A daily habit	A habit that i do every day	0	0.39	0.83
		gr	Μία καθημερινή συνήθεια	Μία συνήθεια που έχω κάθε μέρα	0	0.44	0.85
2	**Synonyms**	en	House	Home	0	0.00	0.93
		gr	σπίτι	οικία	0	0.00	0.86
3	**Addition of Articles**	en	the Dynasty	Dynasty	0	0.70	0.95
		gr	η Δυναστεία	Δυναστεία	0	0.70	0.96
4	**Changes in Tenses**	en	played	plays	0	0.00	0.96
		gr	έπαιξα	παίζω	0	0.00	0.87
5	**Grammatical Cases (Greek)**	gr	του εθνικού ύμνου	τον εθνικό ύμνο	0	0.00	0.98
6	**Addition of Special Characters**	en	multi-cultural	multicultural	0	0.00	0.92
		gr	πολυ-πολιτισμικός	πολυπολιτισμικός	0	0.00	0.96
7	**Abbreviations**	en	Mr. George	Mister George	0	0.70	0.91
		gr	κ. Γιώργο	κύριε Γιώργο	0	0.70	0.96
8	**String Representation of Numbers**	en	five to ten years	5 to 10 years	0	0.50	0.97
		gr	5 έως 10 χρόνια	πέντε έως 10 χρόνια	0	0.50	0.97
9	**Numbers Conversion**	-	55,1%	55.10%	0	0.00	0.97

Table 1 Problems of Evaluation due to the translation of the context and of the answer by using https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased [39], with examples of similar sentences in both English and greek language.

are based on embeddings, such as Sentence Mover's Similarity [37], BERTScore [7], POSSCORE [38], Bi-Encoder and Semantic Answer Similarity (SAS) [8].

Concerning the automatic evaluation, the common problems due to the translation are shown in Table 1, e.g., words/phrases can be replaced by synonyms (e.g., see also Figure 2), tenses/suffixes can be changed, etc. This becomes harder for more complex languages like Greek, which uses a very extensive inflection and the same word can be represented with many different suffixes, i.e., for denoting tenses, genders, singular and plural, and others. For example, grammatical cases like the example with ID 5 in Table 1 are quite common for the Greek Language; in that example the suffixes of all the words changed. Thereby, in such cases it is not effective to evaluate the results by using lexical similarity metrics of categories a) and b), such as F1 score, BLEU, i.e., since they do not take into account morphological variation, spelling variations, and others.

For this reason (and although we provide results for metrics of category a) in Section 6), we will mainly focus on metrics of category c), since they can tackle such problematic cases [40]. We selected to use BERTScore [7] and Bi-encoder [8] (e.g., see some real results for Bi-Encoder metric in the last column of Table 1), since these metrics support the Greek language, compared to other embedding-based metrics. Therefore, for a predicted Answer A (e.g., $Ans_{en \rightarrow L}(q_L, M)$) and a golden Answer G (e.g., $gAns(q_L)$), in Section 6 we provide results for the following metrics:

- Exact Match (EM) [9]: The "Exact Match" is 1 if the answer A is identical to the golden answer G, otherwise its value is 0.
- $F1score$ [9]: It is the fraction of shared tokens between the predicted answer A and the golden answer G among the tokens in the predicted or in the golden answer, respectively, with range [0,1].
- $Bi-Encoder$ [8]: it produces for both A and G a sentence embedding (a vector) by using a BERT model that is applied on A and on G separately to obtain the embeddings for each sentence. Their similarity score is computed as the cosine similarity of their vectors. This score is in the range $[-1, 1]$, where -1 means that A and G have totally different meanings and 1 indicates that they have the same meaning. In this paper, we will use the Bi-encoder metric with the BERT model "Distiluse-Base-Multilingual-Cased model"[8] [39], which supports Greek.
- $BERTScore$ [7]: It computes the similarity of two sentences, A and G as a sum of cosine similarities between their tokens' embeddings (a similarity score in range $[-1, 1]$). Since the evaluation concerns sentences in Greek, the BERT Model "bert-base-multilingual-cased"[9] is used.
- $BERTScore_{idf}$ [7]: It is a variation of the BERTScore metric that exploits inverse document frequency (idf) scores.

Finally, for evaluating these QA Metrics by comparing their scores with human scores, we will provide in Section 6 correlation metrics by using Pearson r, Kendall τ and Spearman ρ [40].

5 Experimental Setup

The purpose is to evaluate the performance (both effectiveness and efficiency) of using combinations of MT systems and BERT QA models, in comparison with a multilingual QA model that supports Greek (RQ1), to analyze how the MT systems can affect the quality of the QA process (RQ2), and to evaluate if the automatic metrics can tackle problems based on the translation (RQ3). Below, we provide all the details about the experimental setup.

5.1 Test Sets

Here, we list the two different test sets that we use for the experiments, by providing statistics and a comparison in Table 2.

[8] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased
[9] https://huggingface.co/bert-base-multilingual-cased

	XQuAD Test set (Greek version)	*GreekTexts* Test set
Source	Translated texts from SQuaD	Authentic Greek Texts from a Greek TextBank
Texts' Number	200	20
Words per Text	132.6	398.5
Questions' Number	1,190	200
Words per Question	10.5	7.9
Words per Answer	3.1	5.6
Quality of Greek Text)	Good (but since it is the outcome of a translation process, there were translation errors)	Very High (written by Greek Native Speakers
Stylistic Variety	Mainly Wikipedia Articles	Multiple, including dialogs, articles, narrations, etc.
Context Variety	Topics from multiple domains	Topics from multiple domains
Understanding level	Texts with similar difficulty (mainly articles)	Texts from Novice (Simple Dialogues) to Expert Level (Greek Literature)
Lexical Richness	Informal and formal words	Informal and formal words
Answer Type	Short Answers, mainly Named Entities and numbers	Long Answers
Evaluation in this paper	Automatic Evaluation	Human Evaluation, Automatic Evaluation

Table 2 The test sets for QA over Greek language. Comparison and Statistics

5.1.1 *XQuAD test set* (Greek version)

We use the cross-lingual test set *XQuAD*[10], which includes 1190 questions from 240 texts and is presented in the second column of Table 2. This test set is based on SQuAD v1.1, and it has been translated in 11 languages, by professional human translators (however we have detected some translation errors). Moreover, it has more texts and questions, however, compared to *GreekTexts* the average size of each text is quite smaller, i.e., 132.6 words per texts. Finally, in most cases it provides question with smaller answers, i.e., several answers are either numbers, dates, or a Named Entity (e.g., a person, a city).

5.1.2 *GreekTexts* Test set

Since *XQuAD* mainly contains Wikipedia articles (and not different text genres), we decided that we need a dataset to also cover as many different genres as possible (narrations, dialogues, etc.). For this reason we created the *GreekTexts* test set (it was created before evaluating any model), which is presented in the last column of Table 2. The texts were taken from a Greek text bank[11] and they are free to use for teaching and evaluation. Our target was to cover as many different cases as possible, that can be used for both manual and automatic evaluation. For this reason, the following types of variety were taken into account: a) stylistic variety, i.e., we selected different types of texts including a narration, an article, a historical text, an essay, a presentation,

[10] https://github.com/deepmind/xquad
[11] https://www.greek-language.gr/certification/dbs/teachers/index.html

ID	Machine Translation Model	URL
M1	Bing	https://pypi.org/project/translators/
M2	Helsinki-NLP	https://huggingface.co/Helsinki-NLP

Table 3 Machine Translation Systems that are used in the experimental evaluation

ID	BERT QA Model (pretrained in English)	Abbreviation	URL
B1	Deepset-Roberta-base-squad2	Deepset/RoBERTa	Webpage
B2	DistilBERT base cased distilled SQuAD	DistilBERT-cased	Webpage
B3	BERT large model (uncased) whole word masking finetuned on SQuAD	BERT-Mask-uncased	Webpage
B4	Deepset-Bert Large Uncased Whole Word Masking Squad 2	Deepset/Mask-uncased	Webpage
B5	DistilBERT base uncased distilled SQuAD	DistilBERT-uncased	Webpage
B6	rsvp-ai-Bertserini BERT base SQuAD	rsvp-ai/Bertserini	Webpage
B7	deepset-MiniLM-L12-H384-uncased	deepset/MiniLM	Webpage
B8	dmis-lab-BioBERT Large Cased v1.1 SQuAD	dmis-lab/BioBERT	Webpage
B9	deepset-Bert-Base-Cased-Squad2	deepset/BERT-cased	Webpage
B10	BERT large model (cased) whole word masking finetuned on SQuAD	BERT-Mask-Cased	Webpage

Table 4 BERT-based QA Models that are used in the experimental evaluation

an interview, a text aimed to children and others, b) context variety, i.e., we selected texts about various topics like history, animals, research and others, c) lexical richness, i.e., texts with a variety of different words, both formal and informal, and d) multiple understanding levels, i.e., texts from the A1 (novice) level to the C2 (expert) level.

5.2 Choosing MT systems

For selecting the translation models, we searched for the most popular MT systems that offer a Python API. We evaluated 5 models by using their readily available Python API; the two that are presented in 3, i.e., Bing and Helsinki-NLP, that were eventually selected for the experiments, and three others; the GoogleTrans[12], Goslate[13], and Textblob[14].

Preprocessing for MT. We had to overcome some issues in order to get all the MT systems to work with the texts of our test sets. Sometimes, texts needed to be split in order not to exceed the max text length limit of a MT system. Moreover, all the MT systems seem to have a difficulty to transfer the context from previous sentences to the next ones, i.e., every sentence was translated almost as if it was standalone and not part of a larger context. Finally, Helsinki-NLP does not accept an input >1000 characters. For this reason, we had to split automatically all the texts containing more than 1000 characters to smaller texts, to translate them separately and to concatenate the translation results.

Selection of MT Systems. All the authors manually analyzed at least 10 pairs of questions and answers for each test set in both translation

[12] https://pypi.org/project/googletrans/
[13] https://pypi.org/project/goslate
[14] https://pypi.org/project/textblob/0.9.0

directions (from Greek to English and viceversa). We decided to use Bing and Helsinki-NLP due to the following reasons: a) mainly Bing and secondarily Helsinki-NLP provided a more human-readable translation comparing to the others, i.e., they grasped better the context, b) the Named Entities were usually interpreted by mistake as separate words for most tools, except for Bing (mainly) and Helsinki-NLP(secondarily), and c) Helsinki-NLP provided a more accurate translation regarding the tenses, i.e., it uses the appropriate one to translate the meaning of the original text. Also, d) Helsinki-NLP and Bing add some punctuations (mostly commas), to convey the meaning better, and e) Bing changed the order of the words in a sentence, to give a syntactically better translation.

5.3 The BERT QA Models for the Evaluation

Concerning the QA process, we use each of the 10 pretrained BERT QA models for the English Language[15] that are presented in Table 4, with Bing and Helsinki-NLP MT systems, therefore we use 20 different combinations. We decided to use all these models, since there was no way to evaluate them beforehand for the given task (i.e., Greek QA based on MT). Finally, we should note that we used the ready Python API for each model, which are offered through Hugging Face.

Multilingual BERT QA Model. We compare the performance of these models with a multilingual pretrained BERT model, that supports Greek, called "Deepset/XLM-RoBERTa" model[16]. The strong advantage of the multilingual model is that there is no need to translate the context and the question, thereby, the problems arising from translation are not faced. As for its limitations, it can be possibly slower since it has been pretrained on a larger corpus (training set) including data from many languages. Another limitation is that, although trained on many languages, it might not support the particular desired language. On the contrary, an advantage of the introduced process (that includes the translation) is that it can be easily configured for supporting any pair of MT system and BERT QA model, e.g., new models, improved versions of current models, etc., thereby the results can be further improved in the future, by following the same process.

5.4 Evaluation Process & Metrics

We desire to measure how effective and efficient is each model. For the GreekTexts test set we carried out both a manual and automatic evaluation. On the contrary, for XQuAD we carried out only an automatic evaluation, due to its huge size (i.e., 24,990 answers for all the models).

Process of Manual Evaluation (only for GreekTexts test set). For each question/answer and model (200 questions * 21 models), the authors (all the authors are Greek native speakers) annotated manually if the predicted

[15] The 10 most popular BERT models from https://huggingface.co for the QA task, at the time that we started the experiments (July 2022)

[16] https://huggingface.co/deepset/xlm-roberta-large-squad2

answer was correct, partially correct or wrong, with respect to the golden answer, i.e., in total 4,200 answers. Concerning the process, we divided the 200 questions in 4 different groups (i.e., 50 questions per author). Each author annotated manually all the answers of their questions' group, for each of the 21 models (i.e., 1,050 answers per author). Afterwards, for double-checking, each author rechecked the annotations of at least one other questions' group[17]. Finally, since each question/answer pair (of this test set) has a human and an automatic score, it enables the comparison with the correlation coefficients.

Process of Automatic Evaluation (Both test sets). We measure the average values of EM, $F1score$, $Bi-Encoder$ (by using the Distiluse-Base-Multilingual-Cased model), $BERTScore$ and $BERTScore_{idf}$. Moreover, for the *GreekText* test set we also evaluate their correlation with the human results (which can be considered as the golden value).

Resources and Code. Concerning the code for performing the automatic evaluation, it has been written in python and it is configurable for evaluating any combination of MT/BERT QA models. The code is available in GitHub[18], along with the evaluation datasets (in json format), and all the results.

Efficiency. We provide measurements for the average execution time for answering a question, for each model, including the multilingual.

6 Experimental Results

Here, we provide the results for the experimental setup of Section 5. All the experiments were performed in a machine with 32GB RAM, 4 CPU cores, and 500 GB disk space (SSD).

6.1 Results over *GreekTexts* test set

We discuss the results of both manual and automatic evaluation, and we offer correlation results between the automatic metrics and the human scores.

6.1.1 Manual Evaluation over *GreekTexts* test set

The results are described in Table 5 for all the models, in descending order according to score of the human evaluation.

Multilingual BERT QA Model vs MT+BERT QA. The multilingual BERT QA model (where translation is not needed) is first with respect to the score of the human annotation, however, by combining Bing and the "B8. Dmis-lab/BioBERT" model, we obtained similar results (i.e., 0.715 versus 0.707 for the human score), although we needed to translate both the context, the question and the final answer. Moreover, several other BERT QA models were quite effective, too, mainly by using Bing, such as "B10. BERT-Mask-Cased", "B4. BERT-Mask-uncased", "B1. Deepset/RoBERTa" and "B3. BERT-Mask-uncased".

[17] Approximately, each author spent on average 5-6 hours for the annotation process.
[18] https://github.com/Rantaplanb/question-answering-evaluation

Rank	MT System	BERT Model	Human Score	Bi-Encoder	BERT Score	$BERT\ Score_{idf}$	EM	F1 score
1	-	Multilingual Deepset/ XLM-RoBERTa	0.715	0.726	0.842	0.841	0.220	0.533
2	Bing	B8. Dmis-lab/BioBERT	0.707	0.686	0.820	0.818	0.130	0.436
3	Bing	B10. BERT-Mask-Cased	0.695	0.690	0.816	0.815	0.105	0.417
4	Bing	B4. Deepset/ Mask-uncased	0.692	0.665	0.809	0.808	0.105	0.406
5	Hel	B8. Dmis-lab/BioBERT	0.690	0.678	0.800	0.800	0.100	0.413
6	Bing	B1. Deepset/RoBERTa	0.680	0.679	0.813	0.812	0.125	0.420
7	Hel	B10. BERT-Mask-Cased	0.680	0.675	0.800	0.800	0.095	0.395
8	Bing	B3. BERT-Mask-uncased	0.677	0.675	0.813	0.813	0.130	0.434
9	Hel	B4. Deepset/ Mask-uncased	0.677	0.673	0.803	0.802	0.085	0.410
10	Hel	B3. BERT-Mask-uncased	0.662	0.654	0.794	0.793	0.075	0.389
11	Bing	B7. Deepset/MiniLM	0.652	0.653	0.803	0.802	0.100	0.393
12	Hel	B1. Deepset/RoBERTa	0.647	0.653	0.791	0.790	0.080	0.390
13	Bing	B6. Rsvp-ai/Bertserini	0.637	0.645	0.803	0.801	0.100	0.393
14	Hel	B7. Deepset/MiniLM	0.632	0.641	0.791	0.789	0.080	0.380
15	Bing	B5. DistilBERT-uncased	0.602	0.614	0.795	0.793	0.100	0.365
16	Hel	B6. Rsvp-ai/Bertserini	0.597	0.612	0.781	0.780	0.065	0.346
17	Bing	B9. Deepset/BERT-cased	0.590	0.611	0.788	0.788	0.105	0.364
18	Bing	B2. DistilBERT-cased	0.590	0.606	0.788	0.788	0.090	0.349
19	Hel	B5. DistilBERT-uncased	0.577	0.618	0.786	0.784	0.070	0.354
20	Hel	B9. Deepset/BERT-cased	0.575	0.593	0.779	0.778	0.065	0.348
21	Hel	B2. DistilBERT-cased	0.562	0.599	0.780	0.779	0.060	0.349

Table 5 Results for the GreekTexts test set for all the 21 models in descending order according to the scores of humans

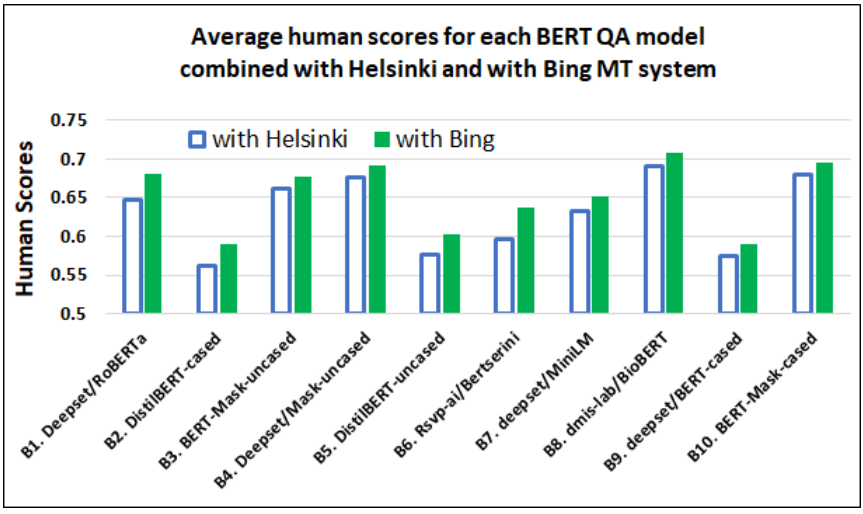

Figure 3 Average human scores for each BERT QA model combined with Helsinki and with Bing MT system

Bing vs Helsinki-NLP MT systems. Figure 3 shows that for all the BERT QA models the best results obtained by using Bing. The highest difference, i.e., 0.04 for the human score, obtained for the "B6. Rsvp-ai/Bertserini" model.

Statistics over the Human labels of Questions/Answers of *GreekTexts*. Table 6 shows that when using Bing, 30.5% of questions were

Category	with Bing	with Helsinki-NLP
Questions Answered Correctly by all 10 QA models	30.5%	27.0%
Questions Answered Correctly by 2-9 QA models	44.0%	42.0%
Questions Answered Correctly by only 1 QA model	3.5%	7.0%
Questions that were not answered correctly by any QA model (they answered either partially correct or Wrong)	22.0%	24.0%

Table 6 Statistics for the Questions of *GreekTexts* by using the human labels

answered correctly by all the 10 BERT models, while for Helsinki there were 27.0% such questions. On the contrary, 22% of questions (44 out of 200) were not answered from any BERT QA model when using Bing, while for Helsinki there were 24% of such questions. From this analysis, we can conclude that it would be an interesting direction to combine the results from different models for selecting the most correct answer.

Brief Analysis of Wrong Answers. By checking manually the results, the most wrong answers occurred due to the three following reasons, a) partially wrong translations, b) totally wrong translations and c) BERT QA models performance. Concerning a), there were in many cases substitutions/mistranslations that partially changed the meaning of the context or/and correct paraphrases (e.g., several worlds replaced by synonyms), which were different than the expected context. Regarding b), we observed wrong translations that totally changed the meaning for the question, context or even the final answer. This was a common problem especially for the Helsinki-NLP system, whereas we detected some few cases where the answer was correct in English, but it translated false back to Greek. Finally, for the last case c), several wrong answers occurred due to the performance of each BERT QA model, e.g., see in Table 5 the difference in scores between the "B2. DistilBERT-cased" and the "B8. Dmis-lab/BioBERT" QA models.

6.1.2 Results of Automatic Metrics over *GreekTexts*

Concerning the automatic metrics that are presented in the last five columns of Table 5, we can clearly see that the values of EM and F1 score are much lower than for embedding-based metrics, even for the multilingual model. This confirms that embedding metrics are more suitable for this type of tasks. Below we analyze the correlation of the automatic metrics with the human scores.

Correlation between Human Judgements & Automatic Metrics. Table 7 provides correlation coefficients between the scores of human evaluation and each of the automatic metrics, by using Pearson r, Kendall τ and Spearman ρ [40]. For computing the correlation coefficients, we used the scores of each question/answer pair separately, i.e., 4,200 question/answer pairs. As we can see, the metrics that are based on embeddings, and predominantly Bi-Encoder, align closest with human scores. On the contrary, the correlation between the lexical-based metrics and especially Exact Match (EM) is low.

Metric	Pearson r	Kendall τ	Spearman ρ
$Bi-Encoder$	0.86	0.69	0.83
$BERTScore$	0.80	0.65	0.80
$BERTScore_{idf}$	0.80	0.65	0.80
$F1Score$	0.69	0.59	0.69
EM	0.28	0.28	0.27

Table 7 Correlation coefficients between the human scores and each of the used automatic metrics (in descending order)

Rank	MT System	BERT Model	Bi-Encoder	BERT Score	$BERT\ Score_{idf}$	EM	F1 score
1	-	Multilingual Deepset/ XLM-RoBERTa	0.856	0.899	0.900	0.548	0.725
2	Bing	B4. Deepset/Mask-uncased	0.804	0.857	0.856	0.289	0.487
3	Bing	B3. BERT-Mask-uncased	0.800	0.857	0.856	0.292	0.492
4	Bing	B10. BERT-Mask-Cased	0.798	0.857	0.856	0.294	0.491
5	Bing	B8. Dmis-lab/BioBERT	0.797	0.856	0.855	0.294	0.491
6	Hel	B3. BERT-Mask-uncased	0.794	0.851	0.851	0.289	0.468
7	Bing	B1. Deepset/RoBERTa	0.792	0.852	0.852	0.288	0.482
8	Hel	B4. Deepset/Mask-uncased	0.790	0.847	0.847	0.290	0.469
9	Bing	B6. Rsvp-ai/Bertserini	0.790	0.851	0.850	0.284	0.475
10	Hel	B10. BERT-Mask-Cased	0.790	0.851	0.851	0.295	0.472
11	Bing	B7. Deepset/MiniLM	0.789	0.851	0.851	0.281	0.477
12	Hel	B1. Deepset/RoBERTa	0.788	0.848	0.848	0.290	0.464
13	Hel	B8. Dmis-lab/BioBERT	0.788	0.851	0.851	0.292	0.467
14	Bing	B9. Deepset/BERT-cased	0.782	0.847	0.847	0.269	0.470
15	Bing	B2. DistilBERT-cased	0.781	0.848	0.848	0.277	0.467
16	Bing	B5. DistilBERT-uncased	0.773	0.846	0.846	0.276	0.462
17	Hel	B7. Deepset/MiniLM	0.769	0.842	0.842	0.284	0.454
18	Hel	B5. DistilBERT-uncased	0.765	0.841	0.841	0.279	0.449
19	Hel	B6. Rsvp-ai/Bertserini	0.762	0.839	0.839	0.281	0.451
20	Hel	B2. DistilBERT-cased	0.761	0.839	0.840	0.278	0.443
21	Hel	B9. Deepset/BERT-cased	0.748	0.831	0.832	0.262	0.432

Table 8 Evaluation by using only Automatic Metrics for the XQuAD Dataset. The performance of MT/BERT QA pretrained models versus Multilingual Model (Ranking in descending order according to Bi-Encoder metric)

6.2 Results over XQuAD test set

The results are described in Table 8 in descending order according to the score of Bi-Encoder metric and are analyzed below.

Multilingual BERT QA Model vs MT+BERT QA. Below, we discuss the results and the tendencies and differences compared to *GreekTexts* test set. Concerning the same tendencies, for this test set Bing was also better than Helsinki, the multilingual model outperformed again all the combinations of MT/BERT QA models, the BERT models B3, B4, B8, B10 provided high scores, whereas the DistilBERT models provided lower scores. Regarding the differences, for this test set we observe a slightly different ranking for the MT/BERT QA models, e.g., the best MT/BERT QA model was the "B4. BERT-Mask-uncased" and not the "B8. Dmis-lab/BioBERT". However, the latter also achieved a high Bi-Encode score, which is quite close to the "B4. BERT-Mask-uncased" model.

Bing vs Helsinki-NLP. By taking into account the Bi-Encoder metric, we observed better results for all the BERT QA models by using Bing instead of Helsinki-NLP, i.e., see Figure 4. The highest difference observed for the BERT

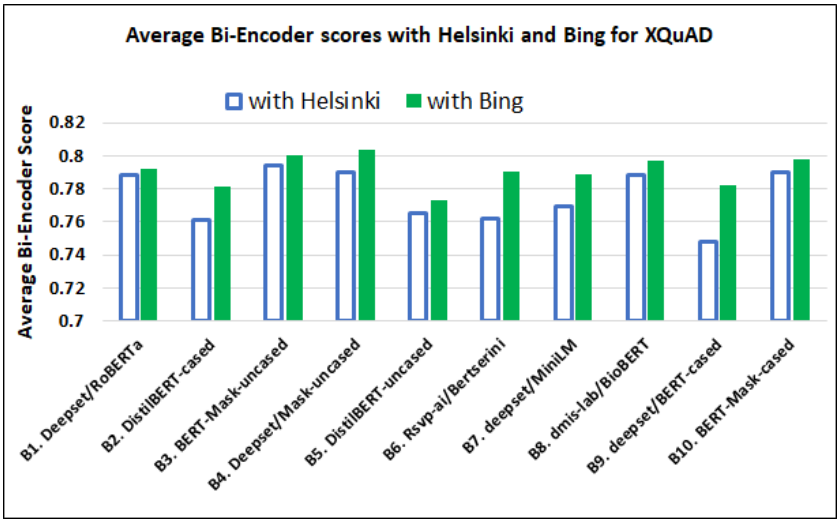

Figure 4 Bi-Encoder scores for BERT QA Models with Helsinki-NLP and Bing for XQuAD

QA model "B9. Deepset/Bert-Cased", i.e., +0.034 by using Bing instead of Helsinki-NLP MT system.

6.3 Efficiency

An important factor for a QA process is how fast it can produce the results. Here, we provide some indicative times for the average time for each text/question for each pair of MT/BERT models . We selected to use the test set having on average the most texts and questions/answer pairs, i.e., the XQuAD test set. Although the same context is used for answering multiple questions in XQuAD, we measured the execution time, by performing for each question the process from scratch. We measured the average time for each combination of MT/QA models and for the multilingual model. As we can see in Figure 5, the multilingual model is on average slower comparing to the combinations of MT/QA models, e.g., 1.16× slower comparing to the second one, i.e., B8. dmis-lab/BioBERT. Indeed, 14 seconds are needed on average for answering a question for the multilingual model. On the contrary, for the MT/BERT models we needed on average from 7 to 12 seconds for answering a question. These execution times include the time needed for translating the context, question and answer[19]. However, the translation steps were quite fast, i.e., on average less than 1.5s for each context/question. Additionally, the models using the Helsinki-NLP MT system were always faster for all the different combinations.

[19]Since the experiments performed in a machine having only CPUs, we should clarify that the time difference may be not significant in a production environment where a GPU will be used.

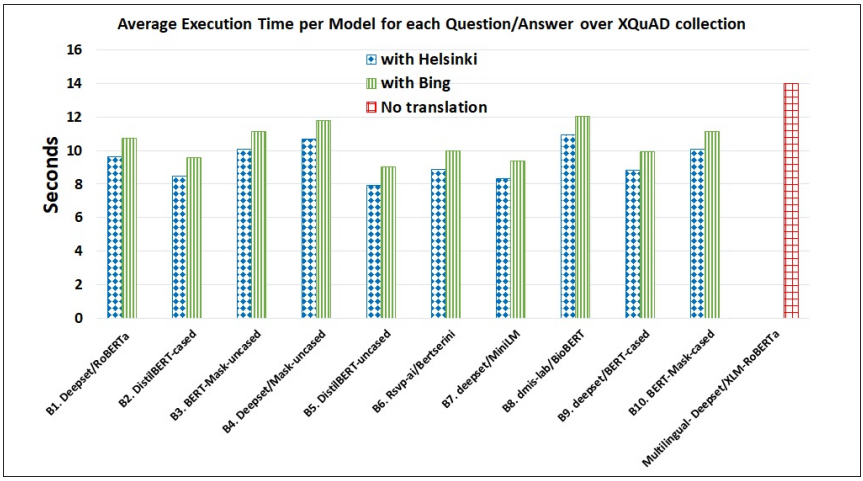

Figure 5 The efficiency of each model over XQuAD test set for the whole process

Category	**Pretrained QA Models using Bing**	**Pretrained QA Models using Helsinki-NLP**	**Multilingual Model (Deepset/ XLM-RoBERTa)**
Best Human Score for *GreekTexts*	0.707 (BERT Model B8. Dmis-lab/BioBERT)	0.69 (BERT Model B8. Dmis-lab/BioBERT)	0.715
Best Score by using Bi-Encoder for *XQuaD*	0.804 (BERT model B4. Deepset/Mask-uncased)	76.5% (BERT model B3. BERT-Mask-uncased)	0.856
BERT Models with high effectiveness	B3, B4, B8, B10	B3, B4, B8, B10	-
Key Advantages	1. Good Translation wrt Syntax and Named Entities. 2. Fast average execution time. 3. Can be evaluated with similarity metrics based on embeddings	1. Good Translation of Tenses. 2. Fastest average execution time 3. Can be evaluated with similarity metrics based on embeddings	1. Effective in small answers including Numbers and Named Entities. 2. Suitable for any automatic metric. 3. No translation errors
Key Drawbacks	1. Wrong answers due to bad translations, mainly of wrong translation of tenses. 2. Low scores for EM and F1 score	1. Wrong answers due to bad translations, mainly of named entities, of syntax, and of translations from English to Greek 2. Low scores for EM and F1 score	1. Wrong answers for questions having long answers or complex questions including formal Greek Words 2. Higher average execution time (due to larger training size)

Table 9 Some key observations from the experiments for the different models

6.4 Discussion

Concerning the key observations from our experimental evaluation with respect to the research questions (RQ1-RQ3), they are analyzed below. Also, Table 9 provides a comparison for the different models that we evaluated.

First, for the *RQ1* we observed that combining MT and English BERT QA models is a very good alternative option to a multilingual model that has been

pretrained with Greek texts for both test sets, since a) they can be faster, i.e., at least 2 seconds faster on average, and b) some models can produce similar results to the multilingual model.

Regarding *RQ2*, we have seen that the translation model can affect the quality of the QA process, i.e., Bing MT system seems to be the most effective one for the Greek language, since it provided more accurate translations comparing to Helsinki-NLP, (e.g., see the advantages and drawbacks in Table 9). Indeed, compared to Helsinki-NLP MT system, we observed better results for all the BERT QA models for both test sets.

Concerning *RQ3*, we have discussed the difficulties of evaluating the performance of MT/BERT models, since the translation can replace a word with a synonym, can change the suffixes of several words, etc. For this reason, the values for the metrics EM and F1 score were quite low for the combinations of MT/BERT QA models. However, by using metrics based on sentence embeddings, we observed a high correlation with the human judgements for the *GreekTexts* test set, mainly by using the Bi-Encoder metric [8] with the "Distiluse-Base-Multilingual-Cased" pretrained model that supports Greek. By using all the 4,200 human labelled answers, we observed the following values concerning the correlation of that metric with the human scores: 0.86 for Pearson r, 0.69 for Kendall τ and 0.83 for Spearman ρ.

Finally, since our pipeline (and code) is easily configurable, we expect that improvements in translation, BERT QA model, and sentence similarity models will have a positive impact for the proposed process.

7 Conclusion

We presented a multilingual approach for QA over Greek texts, by combining pretrained machine translation and BERT QA models in English at runtime. We presented all the steps and the difficulties for the studied problem, by predominantly focusing on the Greek language and its complexity. Regarding the evaluation, we performed experiments on two test sets by comparing 20 combinations of MT/BERT QA models and a multilingual BERT QA model (that supports Greek), by exploiting short sentence similarity embeddings.

Concerning the *GreekTexts* test set, which contains 20 texts and 200 questions, we performed an evaluation by using both human and automatic scores and the combination of the translation model Bing and the BERT model "dmis-lab/BioBERT" achieved similar results to the multilingual model, i.e., human score 0.707 versus 0.715. Moreover, we evaluated automatic metrics and we found out that the embedding-based metrics, especially Bi-Encoder, have higher correlation scores with the human judgments than the string-matching based metrics.

Regarding the *XQuAD* test set, which includes 240 texts and 1190 questions, we performed an evaluation by using only automatic metrics, and the multilingual BERT QA model achieved the highest Bi-Encoder score, i.e., 0.856, whereas the combination of Bing and "Deepset/Mask-uncased"

model was second with 0.804 (the same model was positioned fourth for the *GreekTexts* test set). On the other hand the MT/BERT QA model for the *GreekTexts* test set, i.e., "dmis-lab/BioBERT" by using Bing, was positioned fifth for the *XQuAD* test set. Concerning the efficiency, the MT/BERT QA models were faster comparing to the multilingual model, i.e., on average 7-12 seconds were needed for answering a question, whereas for the multilingual model we needed 14 seconds. From the experiments, we can conclude that the combination of machine translation and BERT QA English models is an effective and efficient alternative option for QA over Greek texts, especially by using sentence similarity metrics for the automatic evaluation of the answer.

As a future work, we desire to extend our approach for a) using more similarity metrics for evaluating the correctness of an answer, b) evaluating more MT systems such as DeepL[20] and including translations by Large Languange models (LLMs) such as ChatGPT [41], c) investigate answer type prediction mechanisms [30], e.g., for avoiding the translations belonging to categories like dates and numbers, d) supporting more languages, and e) studying methods for answering more complex questions [42], e.g., confirmation and list questions.

Data Availability Statement: All the datasets of this study, the annotated test set, all the experimental results, and the code for running the experiments are available in https://github.com/Rantaplanb/question-answering-evaluation.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Authors Contribution: All the authors wrote the main manuscript text, i.e., Michalis Mountantonakis (M.M.), Loukas Mertzanis (L.M.), Michalis Bastakis (M.B.), and Yannis Tzitzikas (Y.T.). M.M. proposed the methodology (with the help of Y.T., L.M and M.B). L.M. and M.B. created the software and performed the experimental analysis. M.M. prepared the figures and tables (with the help of L.M. and M.B). Y.T. and M.M. were the administrators and supervisors of this work. Finally all the authors reviewed and revised the manuscript.

References

[1] Zeng, C., *et al.*: A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. Applied Sciences **10**(21), 7640 (2020)

[2] Liu, Y., et al.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[3] Perevalov, A., *et al.*: Can machine translation be a reasonable alternative for multilingual question answering systems over knowledge graphs? In: Proceedings of the ACM WebConf, pp. 977–986 (2022)

[20] https://www.deepl.com/en/translator

[4] von Essen, H., Hesslow, D.: Building a swedish question-answering model. In: Proceedings of PaM 2020, pp. 117–127 (2020)

[5] Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations (2019) arXiv:1910.11856

[6] Papantoniou, K., Tzitzikas, Y.: Nlp for the greek language: A brief survey. In: 11th Hellenic Conference on Artificial Intelligence, pp. 101–109 (2020)

[7] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)

[8] Risch, J., Möller, T., Gutsch, J., Pietsch, M.: Semantic answer similarity for evaluating question answering models. arXiv preprint arXiv:2108.06130 (2021)

[9] Rajpurkar, P., et al.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)

[10] Lee, K., Yoon, K., Park, S., Hwang, S.-w.: Semi-supervised training data generation for multilingual question answering. In: Proceedings of the International Conference on Language Resources and Evaluation (2018)

[11] Asai, A., Eriguchi, A., Hashimoto, K., Tsuruoka, Y.: Multilingual extractive reading comprehension by runtime machine translation. arXiv preprint arXiv:1809.03275 (2018)

[12] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Cross-lingual machine reading comprehension. arXiv preprint arXiv:1909.00361 (2019)

[13] Krubiński, M., Ghadery, E., Moens, M.F., Pecina, P.: Just ask! evaluating machine translation by asking and answering questions. In: Proceedings of the Sixth Conference on Machine Translation, pp. 495–506 (2021)

[14] Schäfter, S., Zylowski, T.: German question answering in crm systems. INFORMATIK 2021 (2021)

[15] Tiedemann, J., Thottingal, S.: Opus-mt–building open translation services for the world. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (2020)

[16] Mozannar, H., Hajal, K.E., Maamary, E., Hajj, H.: Neural arabic question answering. arXiv preprint arXiv:1906.05394 (2019)

[17] d'Hoffschmidt, M., et al.: Fquad: French question answering dataset. arXiv preprint arXiv:2002.06071 (2020)

[18] Carrino, C.P., Costa-jussà, M.R., Fonollosa, J.A.: Automatic spanish translation of the squad dataset for multilingual question answering. arXiv preprint arXiv:1912.05200 (2019)

[19] Koutsikakis, J., *et al.*: Greek-bert: The greeks visiting sesame street. In: 11th Hellenic Conference on Artificial Intelligence, pp. 110–117 (2020)

[20] Papantoniou, K., Efthymiou, V., Flouris, G.: El-nel: Entity linking for greek news articles. In: ISWC (Posters/Demos/Industry) (2021)

[21] Partalidou, E., *et al.*: Design and implementation of an open source greek pos tagger and entity recognizer using spacy. In: 2019 International Conference on Web Intelligence (WI), pp. 337–341 (2019). IEEE

[22] Building and evaluating resources for sentiment analysis in the greek language. Language resources and evaluation **52**(4), 1021–1044 (2018)

[23] Marakakis, E., Kondylakis, H., Papakonstantinou, A.: Apantisis: A greek question-answering system for knowledge-base exploration. In: Strategic Innovative Marketing, pp. 501–510 (2017)

[24] Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., Ngonga Ngomo, A.-C.: Survey on challenges of question answering in the semantic web. Semantic Web **8**(6), 895–920 (2017)

[25] Abbasiantaeb, Z., Momtazi, S.: Text-based question answering from information retrieval and deep neural network perspectives: A survey. WIREs: Data Mining and Knowledge Discovery **11**(6), 1412 (2021)

[26] Perevalov, A., *et al.*: Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. In: 2022 IEEE 16th ICSC, pp. 229–234 (2022). IEEE

[27] Mountantonakis, M., Bastakis, M., Mertzanis, L., Tzitzikas, Y.: Tiresias: Bilingual question answering over dbpedia abstracts through machine translation and bert. In: DL4KG2022 Workshop (2022)

[28] Lehmann, J., *et al.*: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web **6**(2), 167–195 (2015)

[29] Adlakha, V., Dhuliawala, S., Suleman, K., de Vries, H., Reddy, S.: Topiocqa: Open-domain conversational question answering with topic switching. Transactions of the Association for Computational Linguistics **10**, 468–483 (2022)

[30] Dimitrakis, E., Sgontzos, K., Tzitzikas, Y.: A survey on question answering systems over linked data and documents. Journal of intelligent

information systems **55**, 233–259 (2020)

[31] Freitag, M., *et al.*: Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics **9**, 1460–1474 (2021)

[32] Rivera-Trigueros, I.: Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation, 1–27 (2021)

[33] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[34] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72 (2005)

[35] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[36] Popović, M.: chrf: character n-gram f-score for automatic mt evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 392–395 (2015)

[37] Clark, E., Celikyilmaz, A., Smith, N.A.: Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2748–2760 (2019)

[38] Vakulenko, S., Kiesel, J., Fröbe, M.: Scai-qrecc shared task on conversational question answering. arXiv preprint arXiv:2201.11094 (2022)

[39] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of Conference on EMNLP (2019)

[40] Chen, A., Stanovsky, G., Singh, S., Gardner, M.: Evaluating question answering evaluation. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pp. 119–124 (2019)

[41] Jiao, W., Wang, W., Huang, J.-t., Wang, X., Tu, Z.: Is chatgpt a good translator? a preliminary study. arXiv preprint arXiv:2301.08745 (2023)

[42] Etezadi, R., Shamsfard, M.: The state of the art in open domain complex question answering: a survey. Applied Intelligence, 1–21 (2022)