Automatic Arabic Translation of English Educational Content Online using Neural Machine Translation:

the Case of Khan Academy

Dietrich College Senior Honors, Information Systems

Carnegie Mellon University - Qatar

Imane Bendou

Advised by: Prof. Houda Bouamor

Date: 05/04/2021

Abstract

Massive Open Online Courses (MOOCs) offer valuable and high quality learning opportunities and educational content in several disciplines to many students, to a large extent regardless of their background, location, and personal circumstances. However, language represents a major barrier for them, keeping non-native English speakers from benefiting from these online educational resources, since online content most available is in English. Given there are over 300 schools in Qatar covering all topics in Arabic, in order to make online educational resources more available to students in them, we designed and implemented an automatic machine translation solution based on deep learning techniques. It aims to make high-quality Arabic translations of subtitles available in English. We focused on the case of Khan Academy which provides a personalized learning experience that is mainly focused on videos. These videos have subtitles that are generally generated by volunteers for different languages. Our system covers several subjects ranging from Physics and Mathematics to Programming and Arts and Humanities, with a focus on high school level students. Our system was trained using a high-quality parallel corpus from the education domain developed by the Qatar Computing Research Institute (QCRI). Furthermore, the system underwent intrinsic evaluation by comparing its output to a high-quality reference translation, as well as extrinsic evaluation in a pilot study, where we aimed at testing the quality of the system's output in schools to evaluate its contribution to student understanding.

Acknowledgements

I would like to sincerely extend my gratitude to my professor and advisor Houda Bouamor, for guiding me and working closely and tirelessly with me on this thesis. I would like to thank her for encouraging me to drive this thesis to the end and for believing in me and this thesis work. I would also like to thank Dean Selma Limam Mansar for the guidance she provided with the educational aspect of this work and for believing in the potential of it as a whole. Furthermore, I would like to thank Mr. Fadhel Annan for facilitating our communication with government entities here in Qatar such as the Ministry of Education and Higher Education, as well as the schools. I also extend my thanks to Meg Rogers and Maha Kanso from the CMUQ research office for their help with obtaining an IRB approval.

I would also like to offer my gratitude to Dr. Ahmed Abdelali and Dr. Hassan Sajjad from QCRI for their comments, suggestions and making the corpus we used for our system available. I would also like to thank them for meeting with me in the beginning of this thesis work to help me adjust the scope of my thesis, as well as encouraging this thesis work.

Finally and importantly, I would like to extend a big thank you to my support system through this thesis: my parents, Amar and Salima, my fiance, Badis, and my friends and extended family. I would like to thank them for their patience with my indecisiveness throughout as well as the occasional pep-talks.

Table of Contents

Introduction **4**

Related Work **7**

2.1 OERs, MOOCs and the Language Barrier 7

2.2 Increasing Accessibility to MOOCs and OER 8

Background **10**

3.1 Machine Translation Overview 10

3.2 Arabic in Machine Translation 11

Data **11**

Methodology **12**

5.1 Our Machine Translation Models 13

5.1.1 Preprocessing 14

5.1.2 Training 16

5.1.3 Evaluation and Results 17

5.2 Experiment in Schools 20

5.2.1 Experimental Design 20

5.2.2 Our Results 22

Conclusion and Future Work **23**

1. Introduction

Technology has become an essential part of our daily life. This is apparent in education, for example, where Massive Open Online Courses (MOOCs) offer valuable and high quality learning opportunities and educational content in several disciplines to many students, to a large extent regardless of their background, location, and personal circumstances (Brown, 2014). Open Educational Resources (OERs) are also a great embodiment of the use of technology in education as they are teaching materials that are available for free for anyone with Internet connection (Kazakoff-Lane, 2014). MOOCs and OERs have become vital recently, given the COVID-19 outbreak that has forced schools to shut down and students to stay at home, giving these students much less frequent interactions with their teachers. This gave MOOCs, such as Khan Academy[1] and Coursera[2] the ability to emerge and become an incredibly useful source of information and knowledge to seekers. So, such resources would be super helpful in keeping the educational train running despite the pandemic.

However, the biggest barrier to learning and using these online platforms is language (Beaven et al. 2013). Online content is mostly available in English (Sanchez-Gordon & Luján-Mora, 2014), which means that students who follow curriculums based on other languages would find it difficult--and sometimes even impossible--to make use of the MOOCs, as opposed to students who go to English-speaking schools. There are 304 Independent schools in Qatar [13], and these schools rely on Arabic as the first language that is used for teaching all subjects including Science and Mathematics. With the shift to distance learning due to the COVID-19 outbreak, a lot of students would resort to online platforms for extra help but would most probably struggle because the content is not available in Arabic.

It is important to acknowledge the efforts made in making online educational content more accessible. For instance, Beaven et al. (2013) recognized the language gap and encouraged the use

[1] https://www.khanacademy.org/
[2] https://www.coursera.org/

of open translation, which makes use of crowdsourcing and open free software. Light & Pierson (2016) explored translating MOOCs to spanish for Chilean students manually, while Qatar Computing Research Institute (QCRI) worked on making some online educational content accessible in Arabic, relying on Statistical Machine Translation (SMT). Another method that can be considered and used to translate any content on the Internet would be Google Translate. This uses a Neural Machine Translation approach that utilizes predictive algorithms to guess ways to translate texts into other languages. Although Google Translate provides translations for a large number of languages, accuracy when it comes to Arabic is questionable. That is especially the case since it is significantly different from english, not only in terms of script and vocabulary, but also syntax, verb conjugation, sentence construction, etc. Google Translate may produce the correct meaning of a word (if we are lucky) but longer sentences are bound to go wrong at some point.

In order to overcome the language barrier imposed on student's accessibility to online educational content, we propose to use the power of Machine Translation (MT), more specifically Neural Machine Translation (NMT), to build a model that translates English educational content into Modern Standard Arabic (MSA). Our model takes as input sentences extracted from the video subtitles available in the English version of Khan Academy videos and automatically generates their corresponding translation in MSA. In this work, we design and implement ten different translation models in which we explore different Arabic tokenization schemes. To train our models, we use the publicly available QCRI Educational Domain (QED) corpus, an open multilingual collection of subtitles for educational videos and lectures collaboratively transcribed and translated from English to several languages including MSA[3]. In this research work, we focus on subtitles available in MOOCs and take Khan Academy as a case study. Khan Academy is a platform that offers educational videos and respective exercises in different subjects and content areas and "has become a worldwide education phenomenon in just a few years" (Light & Pierson, 2016, p. 103). The platform is used by about 100 million people worldwide and "64 percent of first-generation students

[3] https://alt.qcri.org/resources/qedcorpus/

at top universities in the US say Khan Academy was a meaningful part of their education" (Skoll, 2013). However, despite how big and popular this platform is, there is little research done on it (Light & Pierson, 2016), and most of what exists is from Khan Academy itself (Khan, 2012; Maxwell, 2012; Schmitz & Perels, 2011). Considering this, and the fact that Khan Academy provides transcriptions for all its videos, we decided to use it as a case study after developing our NMT system.

To evaluate the usefulness of the Arabic subtitles generated by our models, we subject them to two types of evaluations: *intrinsic*, where we use BLEU, the de facto standard automatic evaluation metric, and extrinsic: where we test the usefulness of the subtitles generated by this system the learning of students in Qatari independent schools.

In this thesis, we aim at answering the following research questions:

1. **How can we use the power of Machine Translation to create more Arabic content of MOOCs (i.e., Arabic subtitles in Khan Academy)?**

2. **What kind of data do we need to build a system to translate educational content from English To Arabic?**

3. **How beneficial is this system to students in independent high schools in Qatar, where Arabic is the official language of courses?**

The remainder of this report is organized as follows: in Section 2, we provide an overview of the research work related to ours. In Section 3, we give a background on MT, its types, as well as a look at how Arabic behaves with MT. Section 4 introduces the data used to train our models, and section 5 dives into our methodology and results. We finally conclude with a section on future work.

2. Related Work

2.1 OERs, MOOCs and the Language Barrier

Education has been evolving rapidly, and with technology in constant growth, university courses and educational content in general was able to move online. Open Educational Resources (OERs) are an example of this, and they are "teaching, learning and research materials . . . that reside in the public domain or have been released under an open license that permits no-cost access, use and adaptation and redistribution by others" (Unesco, 2012, p. 3). The goal of OERs is "facilitating free education for the disadvantaged, addressing the need of developing countries for more seats in institutions of higher education, and providing people with affordable 24/7 continuing education" (Kazakoff-Lane, 2014). Along the same line, in 2008, the term MOOC was coined, referring to Massive Open Online Courses, by Stephen Downes, an online learning technology specialist and George Siemens, a psychology professor at the university of Abderdeen. The term refers to an "online course with the option of free and open registration, a publicly shared curriculum, and open-ended outcomes" (Vasiu & Andone, 2014). The intention behind this initiative was to utilize online tools to provide students with a richer learning experience. Peter Norvig, Director of Research at Google.Inc and Sebastian Thurn, chairman and co-founder of Udacity, an online learning platform, were the first to offer a MOOC for free in 2011 (McGill, 2015). Around the same time, other MOOC platforms emerged, among which was Khan Academy, founded by Salman Khan (Khan, 2020). Other MOOC providers have also appeared such as Coursera, Udemy and edX.

The importance of Massive Open Online Courses and Open Educational Resources (OER) in education is undeniable since it strives to make education accessible to students from overseas by making it "cost-effective" and developing "a collective sense of shared endeavour for participants" (Wolfenden et al., 2017). Moreover, OpenEdOz discusses the benefits of open education such as OER and MOOCs, among which are (1.) the opportunity to raise the quality of learning at decreased time and financial cost; (2.) provision of learning materials that are richer, and more

appropriate to the contexts and styles of learning of an increasingly diverse student community; (3.) opportunity to provide learning to disadvantaged communities globally; and (4.) greater levels of transparency into the teaching process (OpenEdOz, 2016, p. 43).

However, it is argued that even though open education platforms help bridge geographical and financial gaps, they are creating another type of gap: a language gap. For instance, Beaven et al. (2013) argue that "Language is one of the main barriers to the reuse of OER." They also argue that because most educational resources are available in English, this language gap is usually underestimated by the English-speaking world, and sometimes even disregarded. For instance, a quick review of Coursera shows that 5,044 out of a total of 6,957 courses, approximately 73%, are offered in English, while the other 27% is distributed across 16 languages (ClassCentral, 2021). Commenting on this, Sanchez-Gordon & Luján-Mora (2014) argue that "non-native speakers read at a slower speed than native speakers; the speed difference leads to information overload and cognitive issues," which means that language poses a learning barrier for non-native English speakers.

2.2 Increasing Accessibility to MOOCs and OER

There have been commendable efforts in enhancing accessibility to educational content. For instance, Beaven et al. (2013) stressed the importance of translation in making educational content more accessible. They praised Open Translation, which makes use of free/open software and open collaboration to engage a distributed volunteer workforce in the translation of resources that have been published openly on the web, and they believed it should be used as a solution for this problem. This method of translation can surely be used to translate MOOCs, but it is manual, making the process of translation tedious and not time-efficient. Moreover, since crowdsourcing is used, poor-quality entries are not always avoidable, which hinders the accuracy of the translations, and negatively affects the integrity of the translated educational content (Zuccon et al., 2011).

Another contribution to bridge this linguistic educational gap was brought about in Chile, where content from Khan Academy that was previously only available in English was translated to Spanish to make it more accessible to students who did not speak English. According to Light and Pierson (2016) “Intel and their Chilean education partner, the Centro Costadigital, selected and oversaw the translation of 650 math and biology videos into Spanish” (p. 106). With this initiative, the Chilean schools that were given access to Spanish subtitles for Khan Academy videos saw an improvement in students’ math skills (Light & Pearson, 2016) This forms another initiative that uses manual translation to bridge the language barrier formed in the context of MOOCs. Even though this initiative allowed for an enhanced access to online educational content, it still used manual translation, which carries the disadvantages aforementioned. It also solely focuses on Spanish-speaking students specifically.

Another effort made to translate online content from English to Arabic was done by Qatar Computing Research Institute (QCRI). They introduced the QCRI Educational Domain (QED) Corpus, which is an open multilingual collection of subtitles for educational videos and lectures collaboratively transcribed and translated over the AMARA web-based platform to over 20 languages, Arabic being the starting point (Abdelali, Guzman, Sajjad, & Vogel, 2014). This initiative made use of Statistical Machine Translation (SMT) as well as manual translation to help build this corpus. This effort is one step towards bridging this language gap and increasing educational content accessibility. However, given the QED courses use SMT, though a powerful MT model, it might not be the best in terms of accuracy (Kinoshita, Oshio, & Mitsuhashi, 2017) in comparison to other MT models.

In addition to these initiatives, “TraMOOCs” which stands for Translated MOOCs came in to solve this language accessibility issue. These TraMOOCs are essentially MOOCs that are machine-translated with an output quality that “relies on a multimodal evaluation schema that involves crowdsourcing, error type markup, an error taxonomy for translation model comparison, and implicit evaluation via text mining, and sentiment analysis on the students' forum posts”

(Kordoni et al., 2016). This initiative strives to make MOOCs more accessible to non-native English speakers by using SMT, using training data that is both in-domain--meaning educational content--and out-of-domain. It translates educational content from English to 11 BRIC languages, i.e. Bulgarian, Chinese, Croatian, Czech, Dutch, German, Greek, Italian, Polish, Portuguese, and Russian, but not Modern Standard Arabic.

Finally, an existing alternative to bridge the language gap in the world of education is Google Translate. It is widely resorted to on a daily basis by the general public--that is, not limited to students or educators. This method first started out using rule-based machine translation (Elliot, 2016) and then followed an SMT model. It eventually moved from SMT to NMT in 2017, which goes to show NMT is more accurate. However, as a result of a fine-grained evaluation of Google Neural Machine Translation (GNMT) output for English-Arabic translation, Alkhawaja et al. (2020) were able to identify and classify 19 different types of translation errors such as mistranslation, omission, lexical and subject-verb agreement errors. This touches upon the point that GNMT does not account for the morphological richness of Arabic, making using it for MOOCs and OERs incautious.

3. Background

3.1 Machine Translation Overview

For this research work, Machine Translation (MT) is used. Ping (2009) argues that MT is a discipline of translation that employs computer software to translate submitted texts. In other words, it is a system that takes as input source text (ST) in a certain language and produces a raw output referred to as target text (TT). There are mainly three types of MT: rule-based MT (RBMT), statistical MT (SMT) and neural MT (NMT). SMT works by measuring the statistical probability that a given sentence from the TT corresponds to the sentence from the ST, and after repeating this process for many TT sentences, the sentence with the highest probability is chosen (Koehn, 2009). RBMT is a “machine translation paradigm where linguistic knowledge is encoded by an expert in

the form of rules that translate from source to target language" (Torregrosa et al. 2019, p. 125). NMT systems "use artificial neural networks to predict the probability of an array of words, typically modeling entire sentences in a single integrated model," while using deep learning and representation learning. According to Castilho, Gaspari, Moorkens and Way (2017), NMT produces much better translations compared to SMT systems since NMT relies on artificial intelligence in sentence representation. NMT is also a better option than RBMT because the latter's "cost of formalising the needed linguistic knowledge is much higher than training a corpus-based system," such as NMT systems. This is why this research work uses NMT.

3.2 Arabic in Machine Translation

Arabic is a Central Sematic Language that is spoken by 300 million people and accepted as an official language in 27 countries. It is written from right to left and has 28 letters, without any form of capitalization, and it has a complex, not always consistent orthography. Arabic is a morphologically complex language which has rich inflectional and derivational morphology. It exhibits a high degree of morphological ambiguity due to the absence of the diacritics and inconsistent spelling of letters, such as Alif and Ya. Arabic words have numerous forms resulting from a rich inflectional system that includes features for gender, number, person, aspect, mood, case, and a number of attachable clitics. As a result, it is not uncommon to find single Arabic words that translate into five-word English sentences: وسنقولها wa+sa+naqulu+ 'and we will say it.' This challenge leads to a higher number of unique vocabulary types compared to English, which is challenging for machine learning models. This Morphological richness poses a challenge for MT. In this work, we need to really cater to the nature of the Arabic language and its morphological complexity by applying some sort of segmentation to the data we use to train our MT models.

4. Data

The data used for this work is from the QCRI QED parallel corpus--formerly known as the AMARA corpus. It is a multilingual corpus consisting of in-domain and out-of-domain data

collected from educational videos and lectures as well as ted talks. A parallel corpus is one that contains a collection of texts (sentences) in the ST and their translations in the TT (refer to Figure 1 for an example). Two main versions of this corpus were released, and both then have been used in order to train the NMT system. The first version of the ar-en QED corpus contains 158k EN-AR parallel sentences for training, 1,581 for validation, and 2,528 for testing "from monologue video lectures, where an instructor explains a concept. The genre of the lectures is informal speech, often with specific technical vocabulary, and with a large variety of topics" (Abdelali et al., 2014, p. 1857). The second version was released through Opus[4], and contains just over 492,000 EN-AR parallel sentences for training. It is important to note the QED corpus is made available in a raw format (where no tokenization or any pre-processing is applied). This corpus was chosen because it specifically contains subtitles from educational videos, which makes this data very relevant to this research work.

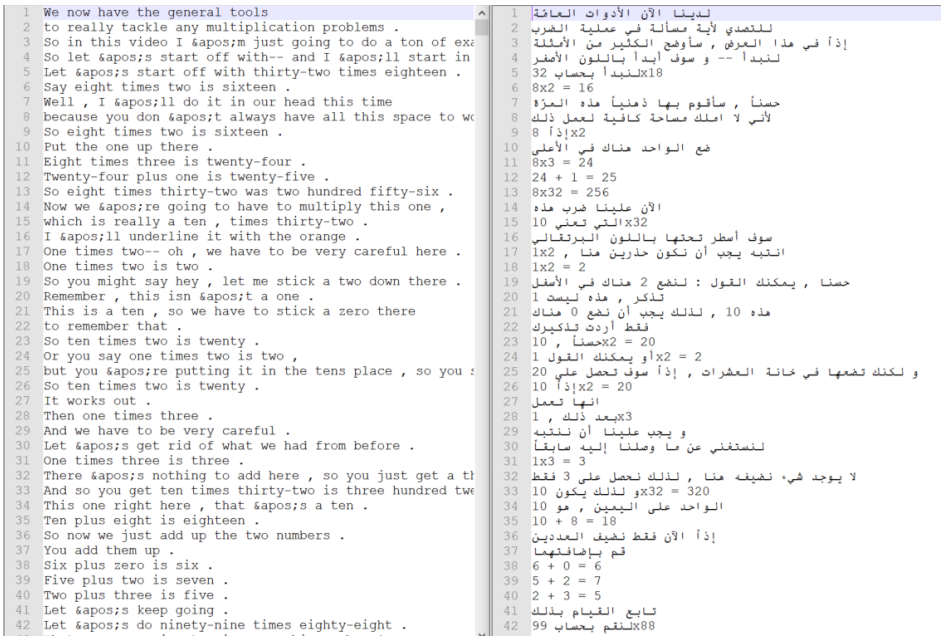

Figure 1: Example of EN-AR parallel sentences within the QED parallel corpus

[4] https://opus.nlpl.eu/

5. Methodology

Our methodology in this research work is divided into two main parts: (i) building the NMT system to automatically generate the Arabic subtitles; (ii) evaluating the usefulness of the subtitles added to a Khan Academy video for students in independent schools in Qatar, the government schools where Arabic is the main language of teaching. We built 10 different NMT models (under two sets of MT experimental setups) where 4 tokenization schemes were explored in addition to the raw version of the corpus. We then used our best performing system to translate subtitles obtained from a Khan Academy video. Then, in order to evaluate the usefulness of the subtitles output by our MT system, we designed an experiment to be conducted in 3 Qatari government schools.

5.1 Our Machine Translation Models

In this research work, we built several systems that take as input an English sentence (i.e., an English video subtitle) and generate its translation to Arabic as output. An example of a subtitle from a Khan Academy video that we put through this system can be seen in Figure 2.

timestamp	source - English	translation_output - Arabic
00:00	In this video, I want to talk a little bit	في هذا العرض، اريد ان اتحدث قليلاً

Figure 2: Example of a subtitle from a Khan Academy Video translated by our MT system

In order to build this system, we made use of NMT, since it is the state-of-the -art technology for MT (Castilho et al., 2017). NMT is likely to produce the most accurate translations, especially given the morphological richness of Arabic. We specifically made use of OpenNMT[5], an Open-Source toolkit for NMT developed by the Harvard NLP[6] group and SYSTRAN[7]. It relies on an attention-based 2-layer LSTM encoder-decoder architecture. This is where "an input sequence--say, an English sentence--is read in its entirety and encoded in a fixed-length internal

[5] https://opennmt.net/
[6] https://nlp.seas.harvard.edu/
[7] https://www.systransoft.com/

representation. A decoder network then uses this internal representation to output words until the end of the sequence token is reached" (Brownlee, 2017). This model is "attention-based" where attention is an interface that connects the encoder to the decoder and provides the latter with information on every encoder's internal representation (i.e., every encoder's word and its context). Figure 3 presents a simple diagram of the encoder-decoder architecture.

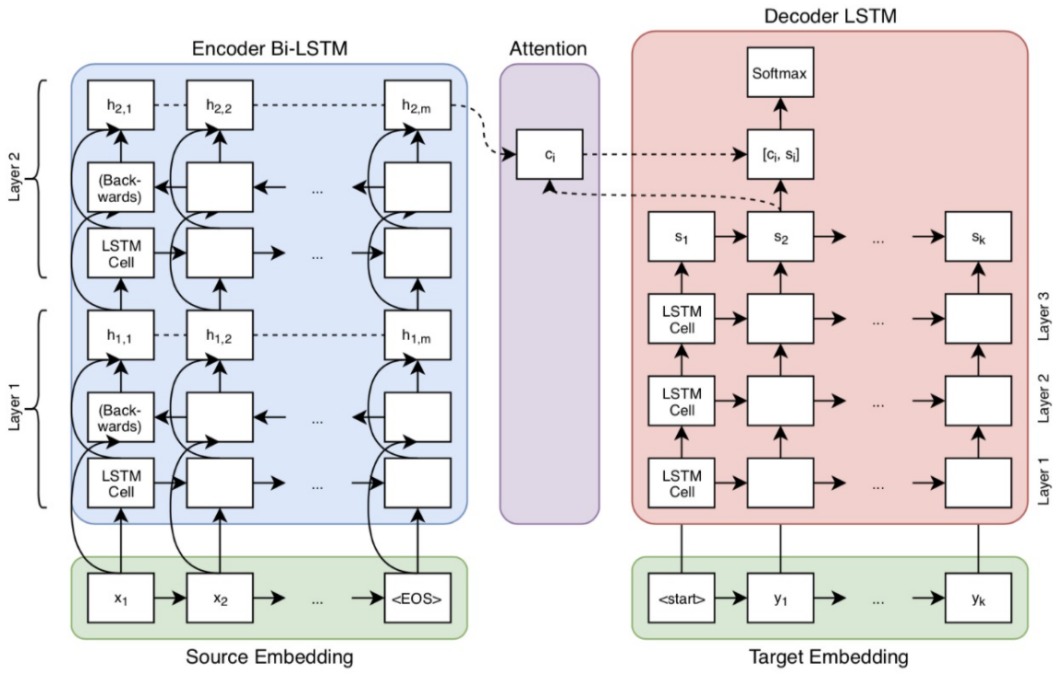

Figure 3: Diagram of an attention-based 2-layer LSTM encoder-decoder architecture

More details about OpenNMT and its underlying infrastructure can be found in Klein, Kim, Deng, Senellart, & Rush's (2017) overview paper. The experimental setup is as follows:

5.1.1 Preprocessing

After the data (from both corpora) collection, the data was stripped from empty lines and unnecessary characters in order to avoid problematic model training. We then subjected the data into four tokenization schemes: Moses tokenization, D34MT tokenization, Byte-Pair-Encoding (BPE) tokenization, and character-level tokenization.

1. *Moses Tokenization*: the first tokenization scheme that was used is the standard tokenizer of the Moses toolkit (Koehn et al., 2017) for both the Arabic and English languages. In order to do this, we used the perl Moses tokenizer which was within the OpenNMT package. This type of tokenization separates punctuation and normalizes characters (e.g., quotes Unicode variants). An example of what an English sentence tokenized through the Moses tokenizer would look like is:

   “And we will say it.” → “And” + “we” + “will” + “say” + “it” + “.”

   An example of this same sentence in Arabic using Moses tokenization is as follows:

   "وسنقولها." ← "وسنقولها" + "."

   We subjected both Arabic and English texts to Moses tokenization.

2. *D34MT Tokenization:* the second type of tokenization we used is D34MT which splits words into refix(es), stem, and suffix(es) and “improves matching of core units of meaning, namely stems” over the MADAMIRA platform (Pasha et al., 2013). Here is an example of what that would look like:

   "وسنقولها." ← "و" + "س" + "ن" + "قول" + "ها" + "."

   Since this type of tokenization is dedicated to Arabic morphology, we only applied it to the Arabic text.

3. *Byte-Pair-Encoding (BPE):* we also used BPE tokenization, which separates words into sub-words in an unsupervised way. It “splits words into symbols (a sequence of characters) and then iteratively replaces the most frequent symbols with their merged variants” (Sajjad et al., 2017). Below is an example that shows this tokenization in English.

   “And we will say it.” → “And” + “we” + “wi” + “ll” + “say”

An example of this same sentence in Arabic using BPE tokenization is as follows:

"وسنقولها." ← "و" + "سن" + "قول" + "ها" + "."

Both English and Arabic texts were subjected to BPE tokenization.

4. *Character-level tokenization:* character-based approach, where words are separated into characters. This is among the harshest types of tokenization since it works on obtaining the smallest possible token (character).

   An example of what an English sentence tokenized at the character level would look like is:

   And we will say it. → A n d # w e # w i l l # s a y # i t .

   In Arabic, this statement, tokenized at a character-level would look like this:

   وسنقولها ← و س ن ق و ل ه ا .

5.1.2 Training

In order to train the models we built, we first started by building the vocab(s), which is necessary for training. This was done on the basic configuration of OpenNMT. Then, we trained the different models on their respective datasets (i.e., we trained the first model on the raw version of QED, the d34mt model on the d34mt tokenized data, etc.). This training was run on a single GPU (world_size 1 & gpu_ranks [0]) (“Quickstart — OpenNMT-Py Documentation,” 2021).

The training data was obtained from the QED corpus, and we ran two main sets of MT experiments:

- **Experiment Set 1:** building and training 5 models (raw_model, moses_tokenized, d34mt_tokenized, bpe_tokenized, and character-level_tokenized) on the first version of QED, which contains 158K EN-AR parallel sentences for training.

- **Experiment Set 2:** building and training 5 other models (raw_model, moses_tokenized, d34mt_tokenized, bpe_tokenized, and character-level_tokenized) on a merged version of the first

and second versions of the QED corpus to form a total of 650K EN-AR parallel sentences for training.

5.1.3 Evaluation and Results

After running our 2 sets of experiments, we used the two testing sets from the QED corpus (tst1 contains 1131 parallel sentences, tst2 contains 1397 parallel sentences), unknown to the system (that is, not part of the training data). After obtaining the translations for each of the training sets, we ran a BLEU evaluation and obtained the results shown in Table 1. BLEU refers to the Bilingual Evaluation Understudy, which is the de facto metric for evaluating the output of machine translation by comparing it to reference translation produced by experts. So, in our case, we compared the output from our test set experiments to the test sets provided in the QED parallel corpus, using the OpenNMT Multi-BLEU[8].

Experiment #	Tokenization	tst1 evaluation	Tst2 evaluation
Set 1			
1	**raw**	28.84	18.56
2	**moses**	34.70	23.53
3	**d34mt**	41.39	30.54
4	**BPE**	35.65	28.51
5	**character_based**	50.62	45.72
Set 2			
6	**raw**	33.04	25.11
7	**moses**	38.62	30.75
8	**d34mt**	37.45	33.30
9	**BPE**	36.93	29.97
10	**character_based**	51.72	47.96

Table 1: BLEU Results for Experiment Sets 1 and 2

There are many observations that could be made when looking at the BLEU results in Table 1. As an initial observation, we can automatically see that the tst1 and tst2 evaluations differ in that tst1

[8] https://github.com/OpenNMT/OpenNMT/blob/master/benchmark/3rdParty/multi-bleu.perl

tended to evaluate higher in all the tokenization schemes, when trained on both training sets. For example, when looking at the character-based model trained on Set 1, the score for tst1, at 50.62, is much higher than that of tst2, at 45.72. This could be due to the fact that tst2 is a larger set to evaluate than tst1, meaning that there would be more variation to take into account, which penalized our system a little more than when evaluated on the smaller tst1 set.

As a second observation, this same idea of variation benefits our system when comparing the results of the two training sets. Since set 2 is much larger than set 1, the translations created in the experiments using set 2 produced higher evaluation results. In fact, when comparing Set 1 character-based model to its equivalent in Set 2, we can see that the scores improved from 50.62 and 45.72, to 51.72 and 47.96, respectively. In this case, due to its larger size, Set 2 helped train the system on many different instances of variation, which allowed it to perform better when evaluated using both tst1 and tst2 evaluation sets.

Finally, when analyzing the different tokenization scheme models, it is shown that the harshest tokenization scheme, character-based, produced the most accurate translations when evaluated against both test sets. The scores of the character-based model when trained on set 2, which are 51.72 and 47.96, are much higher than those of the raw data, which scored 33.04 and 25.11. This was also observed when looking at the translation output manually rather than just relying on BLEU (see Table 2).

English	Arabic_raw	Ar_char_based	Reference text
Which step should be Step 3 in the solution?	وهذه الخطوة التي يجب أن تكون الخطوة الثالثة في البحث ثلاث خطوة	اي خطوة يجب أن تكون الخطوة 3 في الحل؟	اي خطوة يجب ان تكون الثالثة في الحل؟
And this is another one I need to cut and paste.	وهذا واحد آخر يجب ان cut والصق	وهذا واحد آخر بحاجة لقص ولصق	وهذه ايضاً سأحتاج لقصها ولصقها
to get to the Sun, and just to (get) put that in perspective:	Sun, عن Sun, و حتى (get) في :perspective	وحتى نحصل على الشمس، وفقط للحصول على نظرية فيثاغورس:	وحتى نحصل على الشمس، وفقط لنضعها في منظور ما

Table 2: Examples of character-based model translation output compared to reference text

Table 2 also shows a comparison to the target text translated by experts (i.e., reference text). It shows that, despite having minimal errors that we corrected in post-editing later, the character-based model's output produced a translation that is much better than the raw model, and one that is much closer to the reference translation produced by an expert.

Moreover, the Moses-tokenized model performed better than the raw model across both experiment sets, and we believe this is due to the fact that tokenization allows meaning to be more easily interpreted. The D34MT model performed better than the Moses model in Set 1 (i.e., 7 BLEU points), and that is because D34MT takes into account the morphological complexity of Arabic, which leads to overall better translation. Contrary to what we expected--since we assumed the harsher the tokenization scheme the better the translation, both D34MT models performed better in terms of BLEU scores than BPE models. We attributed that to the fact that Arabic does not respond well to unsupervised tokenization schemes, and that it requires a scheme that caters to its morphological complexity.

We also compared our results to the results achieved by QCRI on their initiative with QED (Abdelali et al., 2014). Specifically, we compared our best-performing model's output results to QCRI's on the same test sets (Figure 4). For both test sets, our character-based system scored 13 BLEU points higher than QCRI's, which is very significant in the BLEU metric ("Evaluating Models," 2021). This is primarily because QCRI researchers used SMT which normally produces lower quality translations than NMT. It is also the case because QCRI used Moses tokenization which is less specific and less harsh than what we used, which character-level tokenization.

Ours

Experiment #	Tokenization	tst1 evaluation	tst2 evaluation
10	character_based	51.72	47.96

QCRI's

	tst1 evaluation	tst2 evaluation
English to Arabic (QCRI)	38.0	34.4

Figure 4: Comparison between the output of our system and QCRI's

5.2 Experiment in Schools

After building our MT model and evaluating it quantitatively, we aimed to evaluate the usefulness of the subtitles generated by this system to students in Qatari schools, since they are our primary end users.

In order to conduct this experiment, we required approval from the Ministry of Education and Higher Education (MOEHE) in Qatar to run this experiment. After designing the experiment and obtaining approval from the MOEHE, and securing the CITI[9] certificate, we started the process of applying for the Carnegie Mellon University in Qatar's Institutional Review Board (IRB) approval since we would be dealing with human subjects. The IRB-approved experimental design is in the following section.

5.2.1 Experimental Design

To assess the usefulness of the subtitles obtained by our MT system, we designed a pilot study with grade 10 students as participants. Students (a sample of 60 students) are split into 2 groups. The group assignment was meant to be done automatically using a random selection algorithm. Group 1 was the control group where students watched a short video (10 minutes long maximum) extracted from Khan Academy about a physics concept (Newton's First Law) in English with no Arabic subtitles provided. They were then given a set of four comprehension questions related to the content of the video (provided in Arabic and English) and were asked to answer them after watching the video. As for Group 2, students watched the same short Khan Academy video about

[9] https://about.citiprogram.org/en/homepage/

the same concept given to the ones in Group 1 but for this group, the video is supplied with Arabic subtitles generated automatically using the machine translation system we built. These students also answered the same four comprehension questions given to Group 1.

The experiment was conducted online, where students were provided with a link to the online google forms, one google form for Group 1[10] and one for Group 2[11], where a link to the respective videos was shared. The Google Forms include the questions to be answered by the students as well. These forms were sent to the students through two intermediaries, Mr. Fadhel Annan, the associate Dean of government and corporate affairs for CMU-Q, and the school principals to ensure abiding by the IRB guidelines.

Using our best-performing model, the character-based model, we translated a set of subtitles extracted from an English Khan Academy video[12] that explains a physics concept that the students are yet to study in their classes (i.e, Newton's First Law of Motion). Due to the time constraints of this study, we needed to post-edit those subtitles. We then edited those subtitles onto the Khan Academy video and used it for our experiment (Figure 4).

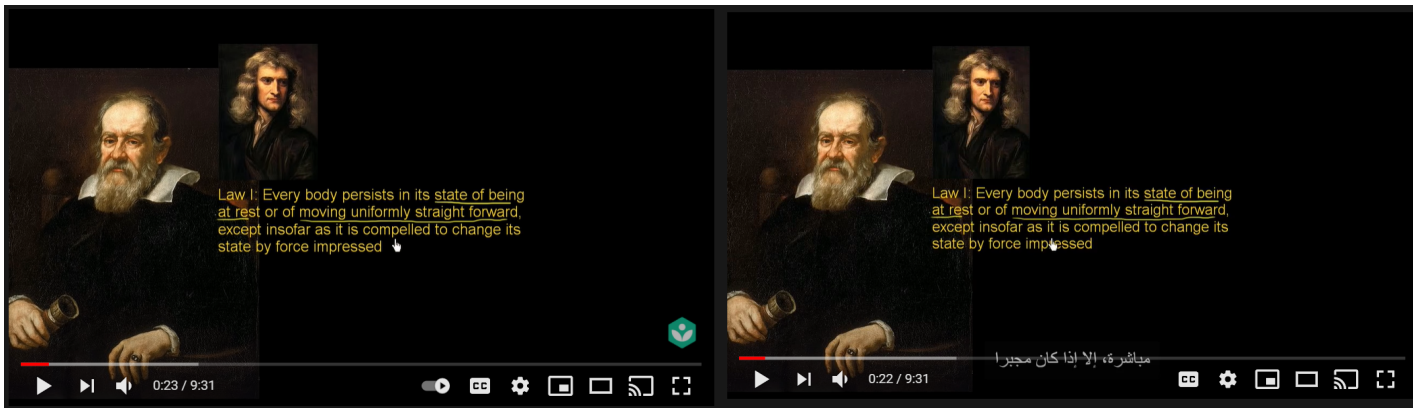

Figure 5: Khan Academy video used for the experiment, with and without the subtitles

The questions that the students answered can be found in Figure 5.

[10] https://forms.gle/N6KzDyzLLvaTTMDc8
[11] https://forms.gle/XknbhhKqpTGcp82QA
[12] https://youtu.be/5-ZFOhHQS68

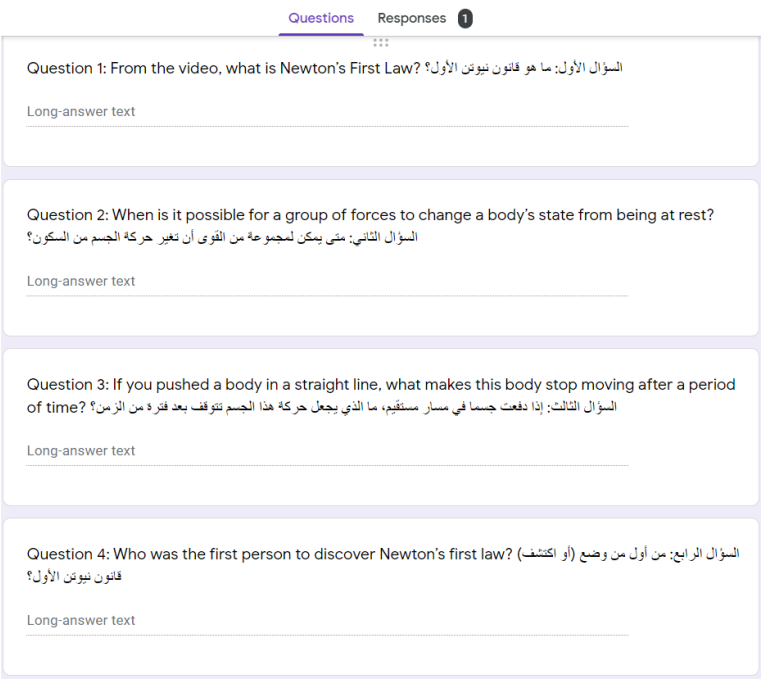

Figure 6: Questions to evaluate the students' learning from the Khan Academy Video

5.2.2 Our Results

Overall, the response rate was 70%, where we received entries from 52 students out of a 70 student sample. Unfortunately, due to the current COVID-19 pandemic restrictions that made this experiment an inherently online one, and given that we were not able to have direct communication with our participants, we were only able to obtain answers for Group 1 in the duration of this research work. In fact, the entire sample dedicated for this study (52 students) ended up being part of Group 1, who watched the Khan Academy video without the subtitles, due to the unavoidable intermediaries of this process. However, even though this meant that the experiment was incomplete, we were still able to obtain useful insights from Group 1's answers for larger-scale future experiments.

The first observation we made when looking at the entries was that almost half (48%) of the answers were in English. This could mean two things: either that the students knew English and were able to answer the questions without the help of subtitles, or that these students used basic

English to answer the questions word-for-word from the video. In the first case, future, bigger experiments should account for the students' English proficiency levels.

Secondly, some students mentioned in their "comments and suggestions" that some of their answers were from previous knowledge outside of that video, which is something to be taken into account with future studies. That could be done through a subject knowledge test that the students could take place before the experiment itself. Among the comments as well, there were 3 comments on the need for Arabic subtitles for better understanding of the contents of the videos.

In terms of correctness of answers, we translated English answers provided by the students in English to Arabic in order to have as fair of a comparison as possible. When comparing the answers to the answer key, the answers were about 70% accurate. This percentage was intended to be compared to the percentage obtained from Group 2.

6. Conclusion and Future Work

During this research work, we built a neural machine translation system that translates educational video subtitles from English to Arabic, focusing on the case of Khan Academy. In order to select our best-performing system, we built 10 models based on different tokenization schemes and the training sets we had available. We used 4 tokenization schemes: Moses tokenization, D34MT tokenization, BPE tokenization and character-level tokenization. We deduced that our best-performing model was the character-based one, which outperformed the baseline model, that is QCRI's model.

In order to evaluate our system extrinsically, we designed a pilot study that was conducted in qatari schools where Arabic is the official language of instruction. Even though this study did not go as planned, we were able to learn more about what can be done in the future. One thing to be done if this experiment were to be replicated in the future would be to do it offline in a way that allows full control over group assignment and entry collection. Another area future work could address is the

students' previous knowledge of the subject at hand, the English language, and scientific English (where students are not proficient in English, but are able to answer questions because they have scientific terms).

As for the MT aspect of this work, future work can implement multi-model translations where the system takes some input text and decides what type of tokenization model to use with that ST to produce the best translations. Another area future work can address would be Out-Of-Vocabulary (OOVs), which are terms that cannot be found in the training sets, in which case the system cannot translate them. This can be potentially addressed by training the model on larger, more inclusive datasets in the educational domain.

References

Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A fast and Furious Segmenter for Arabic. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 1-6. doi:10.18653/v1/n16-3003

Abdelali, A., Guzman, F., Sajjad, H., & Vogel, S. (2014). The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). doi:https://www.aclweb.org/anthology/L14-1675/

Beaven, T., Comas-Quinn, A., Hauck, M., Arcos, B. de los, & Lewis, T. (2013, August 2). Journal of Interactive Media in Education. Retrieved from https://jime.open.ac.uk/articles/10.5334/2013-18/

Brown, S. (2014). MOOCs: Opportunities, Impacts, and Challenges. Massive Open Online Courses in Colleges and Universities by Michael Nanfito. American Journal of Distance Education, 28(2), 139–141. doi: 10.1080/08923647.2014.896558

Brownlee, J. (2017, December 31). Encoder-Decoder recurrent neural network models for neural machine translation. Retrieved April 28, 2021, from Machine Learning Mastery website: https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/

Castilho, S., Gaspari, F., Moorkens, J., & Way, A. (2017). INTEGRATING machine translation into moocs. Proceedings of EDULEARN17 Conference, 9360-9365. doi:10.21125/edulearn.2017.0765

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art?. The Prague Bulletin of Mathematical Linguistics, (108).

El-Kahlout, I., Bektas¸, E., Erdem, N., & Kaya, H. (2018). Translating Between Morphologically Rich Languages: An Arabic-to-Turkish Machine Translation System. Proceedings of the Fourth Arabic Natural Language Processing Workshop, 158–166. https://doi.org/https://www.aclweb.org/anthology/W19-4617.pdf

Habash, N. Y. (2010). Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1), 1-187.

Kazakoff-Lane, C. (2014, March 4). Environmental scan and assessment Of OERs, MOOCs and libraries. Retrieved April 9, 2021, from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Environmental%20Scan%20and%20Assessment.pdf

Khan, S. (2012, February 21). The rise of the tech-powered teacher (opinion). Retrieved April 11, 2021, from https://www.edweek.org/teaching-learning/opinion-the-rise-of-the-tech-powered-teacher/2012/10

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017). Opennmt: Open-source toolkit for neural machine translation. Proceedings of ACL 2017, System Demonstrations, 67-72. doi:10.18653/v1/p17-4012

Koehn, P. (2009). Probability Theory. In Statistical Machine Translation (pp. 63-78). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511815829.004

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (Demonstration session)., ACL '07, Prague, Czech Republic.

Light, D., & Pierson, E. (2016). Increasing Student Engagement In Math: The Study Of Khan Academy Program In Chile. ICERI2016 Proceedings. doi: 10.21125/iceri.2016.0209

Maxwell, L. (2012, December 01). Q&A: Khan Academy creator talks ABOUT K-12 INNOVATION. Retrieved April 11, 2021, from https://www.edweek.org/ew/articles/2012/03/07/23biz-qanda-khan.h31.html

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., … Roth, R., M. (2013). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 1094–1101. https://doi.org/http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf

Ping, K. (2009). Machine Translation. In Baker, M. & Saldanha, G. (Ed.). Encyclopedia of Translation Studies. (2nd ed.). New York: Routledge.

Quickstart — OpenNMT-py documentation. (n.d.). Retrieved April 28, 2021, from OpenNMT website: https://opennmt.net/OpenNMT-py/quickstart.html

Quickstart¶. (n.d.). Retrieved April 11, 2021, from https://opennmt.net/OpenNMT-py/quickstart.html#step-1-prepare-the-data

Sajjad, H., Dalvi, F., Durrani, N., Abdelali, A., Belinkov, Y., & Vogel, S. (2017). Challenging Language-Dependent Segmentation for Arabic: An Application to Machine Translation and Part-of-Speech Tagging. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2. https://doi.org/10.18653/v1/P17-2095

Sanchez-Gordon, S., & Luján-Mora, S. (2014). MOOCs gone wild. In Proceedings of the 8th International Technology, Education and Development Conference (INTED 2014) (pp. 1449-1458).

Schmitz, B., & Perels, F. (2011). Self-monitoring of self-regulation during math homework behaviour using standardized diaries. Metacognition and Learning, 6(3), 255-273.

Skoll - Khan Academy. (2013). Retrieved April 11, 2021, from https://skoll.org/organization/khan-academy/#:~:text=Khan%20Academy%20has%20been%20translated,users%20globally%20subscribe%20to%20KA.

Torregrosa, D., Pasricha, N., Chakravarthi, B. R., Mesoud, M., & Arcan, M. (2019). Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models. Proceedings of MT Summit XVII, 2, 129-133. doi:https://www.aclweb.org/anthology/W19-6725.pdf