The AMARA Corpus: Building Resources for Translating the Web's Educational Content

Francisco Guzman, Hassan Sajjad, Stephan Vogel, Ahmed Abdelali

Qatar Computing Research Institute
Qatar Foundation
{fguzman,hsajjad,svogel,aabdelali}@qf.org.qa

Abstract

In this paper, we introduce a new parallel corpus of subtitles of educational videos: the AMARA corpus for online educational content. We crawl a multilingual collection community generated subtitles, and present the results of processing the Arabic–English portion of the data, which yields a parallel corpus of about 2.6M Arabic and 3.9M English words. We explore different approaches to align the segments, and extrinsically evaluate the resulting parallel corpus on the standard TED-talks tst-2010. We observe that the data can be successfully used for this task, and also observe an absolute improvement of 1.6 BLEU when it is used in combination with TED data. Finally, we analyze some of the specific challenges when translating the educational content.

1. Introduction

Lecture Translation has become an active field of research in the wider area of Speech Translation [1, 2]. This is demonstrated by large scale projects like the EU-funded translectures [3] and by evaluation campaigns like the one organized as part of the International Workshop on Spoken Language Translation (IWSLT), which introduced the challenge to translate TED talks [4] for the 2010 competition. However, the main limitation for the success of these projects continues to be the access to high quality training data.

With the emergence of Massive Online Open Courses (MOOCs), thousands of video lectures have already been generated. Sites like Khan Academy[1], Coursera[2], Udacity[3], etc., continuously increase their repertoire of lectures, which range from basic math and science topics, to more advanced topics like machine learning, also covering history, economy, psychology, medicine, and more.

Online education has bridged the geographical and financial gap, enabling students to access high quality content for free, irrespective of their location. However, the access to this content is still limited by language barriers. By far the most content available is in English. This severely limits access to this high-quality educational material for learners not being able to read and understand English. To overcome these language barriers, amazing efforts are undertaken by volunteers, to translate such lectures into many other languages. One example is the already mentioned TED Talks[4], for which so far more then $9,000$ volunteers have generated about $40,000$ translations into a total of 101 languages. While this and similar efforts at Khan Academy or MIT's Open Courseware[5] are highly commendable, the coverage is extremely skewed towards a small number of languages. It is therefore clear that manual translation trails behind, and that for many languages the small number of volunteers cannot keep up with the fast pace in which new content is appearing on these educational platforms.

Statistical machine translation (SMT) can bridge this gap by automatically translating videos for which subtitles are not available. It also can support volunteer translators, by providing an initial translation, which then can be post-edited [5]. Thus, SMT has the potential to increase the penetration of educational content, allowing it to reach a wider audience. To achieve this, an SMT system requires a large quantity of high-quality in-domain training data. Unfortunately, large data for machine translation has traditionally been constrained to domains such as legal documents, parliamentary proceedings and news. So far, the only openly accessible corpus for the lecture domain has been the TED talks [6].

In this paper, we introduce a new parallel corpus of subtitles of educational videos: the AMARA corpus for online educational content. We crawl a collection of multilingual community-generated subtitles[6]. Furthermore, we explore the steps necessary to build corpora suitable for Machine Translation by processing the Arabic-English part of the multilingual collection. This yields a parallel corpus of about 2.6M Arabic and 3.9M English words. We explore different approaches to align the subtitles, and verify the quality of the generated parallel corpus by building translation models, and extrinsically evaluating them on the standard TED-talks tst-2010 from IWSLT 2011, and on our proposed AMARA test set. We show that the AMARA corpus shares similar domain with TED-talks and leads to an increase of translation quality on the TED translation task.

[1] https://www.khanacademy.org/
[2] https://www.coursera.org/
[3] https://www.udacity.com/
[4] http://www.ted.com/
[5] http://ocw.mit.edu/index.htm
[6] Publicly available through the Amara website: http://www.amara.org

In the next section,we describe the related work and in Section 3 we present crawling, segmentation and statistics of the AMARA corpus. Section 4 shows the usability of AMARA alone and combined with IWSLT for machine translation. In Section 5, we present error analysis based on machine translation output. Section 6 presents our conclusions and future work.

2. Related Work

Several corpora have been developed to support the seminar and lecture translation efforts. One example is the corpus form Computers in the Human Interaction Loop (CHIL) [7], which consists of recordings and transcriptions of technical seminars and meetings in English. The content of the corpus includes a variety of topics: from audio and visual technologies to biology and finance. It is available through ELRA[7] to its members.

More recently, the IWSLT10 [4] evaluation campaign has turned its attention to the lecture and seminar domain by focusing on TED talks. To support this task, a collection of lecture translations has been automatically crawled from the TED website in a variety of languages and made publicly available through the WIT[3] project [8]. In this paper, we used such data as a point of comparison. We crawl parallel subtitles of educational videos and use several measures to show the quality of the crawled corpus in comparison with the closely related IWSLT data set.

In the past, multilingual corpora creation from user-contributed movie subtitles has been addressed by [9]. Recently, a large collection of parallel movie subtitles from the Opensrt[8] community along with tools for alignment of these has been made available through the Opus project [10].

Combination of corpora to improve the translation model has been explored with relative success in the past. For the NewsCommentary and OpenSrt corpora, [11] explore different ways to mix the phrase-table to adapt the Europarl corpus. For the Arabic-English IWSLT data, [12] achieve a relative improvement of 0.7 BLEU by mixing phrases from UN and IWSLT data using instance weighting with weights coming from the language model perplexity.

In this paper, we present the experimental results from data gathered from publicly available crowd-generated data, that has proved to be useful for the lecture domain, but that poses specific challenges, as it has a special focus on online education.

3. The AMARA Corpus

Amara is a web-based platform for editing and managing subtitles of online videos. It provides an easy-to-use interface, which allows users to collaboratively subtitle and translate those videos. The site uses a community-refereed approach to ensure the quality of the transcriptions and translations in the spirit of Wikipedia.

[7] www.elra.org

[8] www.opensrt.org

Amara works in collaboration with online educational organizations like KhanAcademy, TED, and Udacity. As a result, a large body of translations of educational content is available in multiple languages. For example, for Udacity, more than 25K subtitles for over 10K videos have been created by a team of 917 volunteers, since December 2012. These translations are publicly accessible through the Amara website in the form of downloadable video subtitles.

3.1. Languages

On the Amara website, the number of different languages into which a video has been subtitled varies from video to video. In Figure 1 we observe the overall distribution of the number of available languages per video by the total number of videos on the Amara website having translations available in that many languages. A few videos have subtitle translations in as many as 109 different languages. Furthermore, at least 1000 videos have translations available in 25 different languages, and 3000 have translations available in at least 6 different languages. However, the distribution quickly tails off, as many videos have been translated into only a few languages.

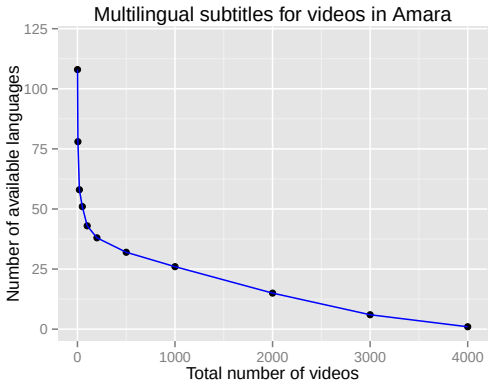

Figure 1: Distribution of the number of available languages per video by the total number of videos in the Amara website.

The most represented languages in the subtitles of this repository are: English with 90K subtitles, French with 20K subtitles, Spanish with 20K subtitles, Italian with 8.8K subtitles and Arabic with 5.9K subtitles. On the other hand, the original language of the videos is highly dominated by English with 135K videos, followed by Spanish with 8.7K videos, French with 6.1K videos, German with 5.0K videos and Russian with 4.3K videos.

In Table 1 we present the distribution of videos from different languages that have been translated into English and Arabic. We observe that English is by far the most subtitled language, which should not be a surprise given the large number of available videos in the platform. Still, only 39% of all the English videos are subtitled into English. However, Arabic videos have an unusually high number of translations into English. In fact, for Arabic videos, there are more English subtitles than Arabic subtitles, which means that many

Videos		Subtitled into	
Language	Total	Arabic	English
English	135K	4463	54023
Arabic	3.8K	494	1286
Spanish	8.7K	33	1167
French	6.1K	38	1160
German	5.0K	11	1006

Table 1: Distribution of the number of translations into Arabic and English from the most popular video languages in the Amara platform. As of December 1st, 2013

videos are translated directly into English, without taking the route through generating Arabic subtitles first. At this point, about 33% of all Arabic videos are subtitled into English, which is a larger proportion when compared to Spanish (13%), French (19%) and German (20%). Note that this data could possibly mislabeled and contain wrong language information. Noisy data often results in poor word alignments and weak translation models.

To shed light on how valuable this data can be for machine translation, we examine the impact of the Arabic-English collection of subtitles, that we codename the AMARA Corpus, in a machine translation environment. These represent only a small fraction of the data available on the Amara website. In future, we plan to extend our work to other language pairs.

3.2. Crawling

The Amara site provides a list of videos and the number of languages the media has been subtitled into. Additionally, it allows filtering by languages. This resulted in 4338 videos that have subtitles in both English and Arabic[9]. In most cases, the original language of these videos is English. Using a non-intrusive in-house crawler, and in cooperation with amara.org, we collected the subtitle files for both Arabic and English. In the current version of the data, we did not perform any additional validation to verify that the documents are in the language they claim to be. Instead, we perform an indirect measurement of the quality by using the parallel data for a standard Machine Translation task.

The subtitle files are in Sub-Rip Text file format (.srt). It consists of segments that are formed by three components:

Segment ID: A number, in sequence, identifying the segment.
Time interval: The start and end times of the subtitle, which represent the timeframe the particular subtitle appears on the screen.
Content: The text for the subtitle segment, with one or more lines.

[9]This quantity includes videos originated in any language pair, not only Arabic and English. The date of collection was July 1st, 2013.

3.3. Data Filtering

From the crawled data for the Arabic-English language pair, we obtained subtitles for a total of 4338 videos, which originated from different organizations. These subtitle files also included transcriptions for the TED talks. To assess the usefulness of this data for translating a standard set for lecture translation such as the IWSLT-11 dataset, we decided to exclude all possible overlap with the IWSLT talk data to avoid contamination and thereby overly optimistic results. Unfortunately, the AMARA data does not have extensive meta-data that can be used for document-level filtering. Furthermore, the difference in sentence alignments, tokenization between our data and the IWSLT-talk data also posed a challenge.

To handle tokenization differences, we detokenized AMARA documents and re-tokenized them using the identical scheme as used for IWSLT. Furthermore, we calculated the percentage of overlap between each of the AMARA documents, and the IWSLT data (train, tune and test); and filtered out the documents ones that presented an overlap of more than a certain threshold (in this case 1% of the sentences in the document). However, due to the conversational nature of the data, frequent phrases such as "applause", "thank you", etc., match almost every document. As a consequence, the relative overlap of smaller documents was artificially inflated and they were filtered out. We fixed this by applying a strong constraint that prevented duplicated counts. Therefore, once a sentence from a specific document was matched with the IWSLT data, it could not be matched to any other document. Our assumption here is that there are no redundant documents in the pool of AMARA documents, so removing previously matched sentences would not cause any trouble. We tested filtering both with and without deduplication. In practice, there were not major differences between the two generated corpora. Thus, we kept the one with the strong constraint, which generated 2400 bilingual documents.

3.4. Segment Alignment

The collected subtitles are for the most part, parallel at the segment level. About 75% percent of all collected segments have identical time stamps on both sides. However, there are two cases, which lead to non-parallel segments:

Incomplete data: When the data in one language (mostly Arabic) is not complete. This could be the case when the translation is still in progress.
Different timestamps: When the text of source and target segment correspond to each other, but the timestamps are not synchronized across languages. This happens when the subtitles in the second language are not generated by translating the subtitles in the original language, but done directly by listening to the original sound track, and translating on the fly.

In order to deal with these issues, we used several algorithms to align the subtitle files. Below, we briefly summarize them:

Strict synchronization constraint (Baseline)
We only extracted the segments from the parallel files if they have identical segment IDs and timestamps. This is a strong constraint, yet gives a good notion of how much data is truly parallel at the segment level.

Automatic sentence alignment
This approach extends the assumption that translations tend to be similar in length [13] by using information from a bilingual dictionary to improve the alignment between parallel files. We used the implementation provided by Hunalign [14]. It aligns the parallel text in two passes.

First, sentence length and lexicon (if provided) information is combined to perform an initial alignment. A new, corpus specific lexicon is then generated from the resulting word alignment. A second pass is performed to align the text with the newly generated dictionary. Note that this approach allows merging of multiple consecutive segments into one longer segment.

Subtitle synchronization
This approach, as implemented in the Uplug subtitle alignment tool [10], exploits the timing information available in the subtitles to perform the alignment. It assumes that sentences that appear in close time-frames should be closer to each other. It can be enhanced by providing anchor-points from which timing offsets and speed ratios can be resolved [9].

The alignment can be enhanced by a bilingual dictionary or by exploiting cognates (LCSR) to establish better anchor points. To synchronize segments across different time-frames, this approach can merge several input segments into one output sentence.

Cascaded synchronization
This approach is a combination of the first two approaches. We started by enforcing a strict synchronization constraint on different subtitles. Then we performed word alignment on the concatenation of all of the strictly aligned data, and extracted a lexicon from the resulting alignment. This lexicon was then used to run the automatic sentence aligner on the unsynchronized portions of the subtitles. Finally, we concatenated both the strictly synchronized with the automatically aligned portions of the subtitles.

3.5. Synchronization Results

Table 2 presents the corpus statistics for the different parallel corpora resulting from the different alignment approaches. The strict synchronization loses a significant portion of the overall data, as shown by the lower total number of words. The segments are short, with only 9.4 words per segment.

	Corpus Statistics		
Algorithm	pairs	tokens	types
Strict Sync	306K	2.9M	55.2K
Hunalign	223K	3.9M	58.2K
Uplug+Cog	221K	3.9M	58.2K
Uplug+Dict	221K	3.9M	58.2K
Uplug+Cog+Dict	221K	3.9M	58.2K
Cascaded	382K	3.6M	58.2K
IWSLT11	93K	1.8M	43.1K

Table 2: Corpus statistics and translation results for different sentence alignment algorithms: strict synchronization (Strict Sync), automatic sentence alignment (Hunalign), subtitle synchronization (Uplug), and cascaded sentence alignment. IWSLT11 shows the statistics of the IWSLT 2011 data.

The sentence aligner (Hunalign) and all the variants of synchronization algorithm (Uplug) yield very similar results in terms of number of words and vocabulary size. However, the segments are now much longer, about 17 words per segment, showing that indeed, Uplug and Hunalign collapse different segments into one sentence pair.

The cascaded alignment preserves the original segment length (9.4 words), while diminishing the loss of tokens. Shorter sentence pairs typically yield better word alignment, which should help to improve the translation quality. On the other side, segmenting sentences into shorter segments means that longer phrases cannot be extracted, which would be extracted from concatenated segments. Segmentation for speech translation has been studied in the past, with somewhat conflicting results [15, 16] and needs to be revisited.

Despite observing a similar performance between all the synchronization variants, for the remainder of this paper we will use the corpus resulting from the cascaded synchronization alignment.

4. Experimental Results

In this section, we extrinsically evaluate the usefulness of the AMARA corpus by training models the data, and observing its performance on a IWSLT lecture translation task (2011). We explore different adaptation methods to better utilize the AMARA data for the IWSLT talk translation task.

4.1. Datasets

To evaluate the usefulness of the crawled data, we experimented with the Arabic-English datasets from the IWSLT 2011 Evaluation Campaign[6]. The IWSLT dataset contained train, dev-2010 and tst2010 sets which consist of 90.5K , 934, 1.6K parallel sentences respectively. In these experiments, we did not make use of the additional IWSLT monolingual data, i.e. the language models in most experiments use only the English side of the parallel corpora, but we also report results using a GigaWord LM.

We used the AMARA corpus resulting from the cascaded synchronization. We divided this corpus into several datasets by randomly sampling the available subtitles. This generated 370K, 5K, 3.6K and 4.4K sentences to be used for train, tune, test and a second test set[10], respectively.

We used IWSLT dev-2010 set for tuning and then tested on two datasets: the IWSLT tst-2010 and AMARA tst-2013, each with a single reference translation. This allowed us to benchmark the improvements obtained by using the AMARA corpus with a standard test set (the former), and to gain insights about translating online educational data (the latter).

In Table 3 we present the 5-gram, Kneser-Ney smoothed, open-vocabulary language-model perplexity for the target side of the test sets given the training corpora. Observe that while the IWSLT10 has similar perplexity w.r.t. the AMARA and IWSLT language models, the reverse relationship does not hold. The AMARA test data has a broader domain, which is not fully captured by the IWSLT language model, which is limited to TED lectures.

training LM	**testset**			
	AMARA13		IWSLT10	
	PPL	OOV	PPL	OOV
AMARA	107.5	1.3	116.7	1.6
IWSLT	204.5	2.6	107.7	1.5

Table 3: Target side per word perplexity (PPL) and out-of-vocabulary rate (OOV %) of the test sets with respect to the language model built on the training data

4.2. Experimental Setup

Preprocessing: We tokenized the English side of all bi-texts as well as the monolingual data (GigaWord) for language modeling using the standard tokenizer of the Moses toolkit [17]. We further truecased this data by changing the casing of each sentence-initial word to its most frequent casing in the training corpus. For the Arabic side, we segmented the corpus following the ATB segmentation scheme with the Stanford word segmenter [18].

Training: We built separate directed word alignments for English→Arabic and for Arabic→English using IBM model 4 [19], and symmetrized them using *grow-diag-final-and* heuristic [20]. We extracted phrase pairs of maximum length seven. We scored these phrase pairs using maximum likelihood with Kneser-Ney smoothing,as implemented in the moses toolkit, thus obtaining a phrase table where each phrase-pair has the standard five translation model features. We also built a lexicalized reordering model : *msd-bidirectional-fe*. For language modeling, we trained a separate 5-gram Kneser-Ney smoothed LM model on each available corpus (target side of a training bi-text or monolingual dataset) using KenLM [21]; we then interpolated these models minimizing the perplexity on the target side of the tuning dataset (IWSLT dev-2010). Finally, we built a large joint log-linear model, which used standard SMT feature functions: language model probability, word penalty, the parameters from the phrase table, and those from the reordering model.

We used the phrase-based SMT model as implemented in the Moses toolkit [17] for translation, and reported evaluation results over two datasets. We reported BLEU calculated with respect of the original reference using NIST v13a, after detokenization and recasing of the system's output.

Tuning: We tuned the weights in the log-linear model by optimizing BLEU [22] on the tuning dataset, using PRO [23] with the fixed BLEU prosposed by [24]. We allowed the optimizer to run for up to 10 iterations, and to extract 1000-best lists for each iteration.

Decoding: On tuning and testing, we used monotone-at-punctuation decoding (this had no impact on the translation length). On testing, we further used cube pruning.

4.3. Baseline B_1

For the baseline system, we trained the phrase and the reordering models on the IWSLT training dataset. The language model was trained on the English side of the IWSLT training data. We tuned the weights on IWSLT-dev2010. Below, we present the experimental results when using the AMARA data for the translation model, the language model and both.

4.4. AMARA Data and the Translation Model

We investigated several ways to maximize the impact of the AMARA corpus for translation by building variations of the translation and reordering models. The systems presented in this section used the same language model built on the English side of the IWSLT training data. As for the baseline, the weights are tuned on the IWSLT-dev2010. Following are different translation settings that we experimented with.

AMARA only (TM_1): Instead of using the IWSLT training data, we built the translation and reordering models using only the AMARA corpus.

Concatenation (TM_2): In this setting, we concatenated AMARA with IWSLT for training of the translation and reordering models. This generally improves word alignment, reduces OOV rate and improves translation quality if two corpora are from similar domain. However, if the added corpus is noisy or of out-of-domain, (e.g. UN data), we can observe a degradation in performance.

Phrase table combination (TM_3): We applied phrase table combination as described in [25]. We built two phrase tables and reordering models separately on the IWSLT and AMARA data. Then, we merged them by adding three additional indicator features to each entry to inform the decoder if the phrase was found in the first, second or both tables. This can be seen as a form of log-linear interpolation.

[10]We did not use the second test set for the experiments in this paper.

SYS	TM	IW10	OOV	AM13	OOV
B_1	IWSLT	22.97	1.9	23.26	3.9
TM_1	AMARA	22.40	2.4	23.66	1.7
TM_2	IW+AM	23.41	1.2	27.63	1.8
TM_3	PT(IW,AM)	**23.57**	1.2	**27.65**	1.8

Table 4: Results of the translation system tested on IWSLT-tst2010 and AMARA-tst2013. All systems use identical language model built on the IWSLT training data and use IWSLT-dev2010 for tuning.

4.4.1. Results

Table 4 shows the results of using the different translation models. Using only AMARA for translation model (TM_1) showed competitive results with our baseline B_1 that is built on IWSLT data. The comparable BLEU score on IWSLT10 shows the value of the AMARA corpus as a parallel corpus in the IWSLT10 translation task. Furthermore, the concatenation and merging of AMARA and IWSLT are able to further reduce the OOV rate. From these combinations, we observe a BLEU improvement up to 0.6 for IWSLT10 and 4.4 for AMARA[11].

4.5. AMARA Data and the Language Model

In this section, we explore the usability of the AMARA data for language modeling. For every system, the translation and reordering models were trained on the IWSLT data and tuned on IWSLT-dev2010. We experimented with different approaches to build the language models:

AMARA only (LM_1): used a LM trained exclusively on the target side of the AMARA corpus.

Concatenation (LM_2): used a concatenation of the English side of both the IWSLT and AMARA corpora.

Interpolation (LM_3): used an interpolated from B_1 and LM_1. The interpolation weights were set to minimize perplexity on the target side of IWSLT-dev2010.

Gigaword (LM_4): uses LM built on the English GigaWord (v5) corpus. This was only included as a reference.

4.5.1. Results

Table 5 summarizes the results of our experiments. Using only AMARA for language model slightly hurts the performance on IWSLT10 by 0.14 BLEU points. However, it has better results when tested on AMARA13. Both the concatenated and interpolated language models show improvements in the translation quality of both sets.

4.6. Best Combination

We combined the best translation model and language model settings from Table 4 and Table 5 respectively and summarize the results in Table 6. From these results we can observe that using AMARA data with IWSLT gives up to a 1.69 improvement in BLEU for the IWSLT-tst2010 and 8.84 BLEU for the AMARA-tst2013. While the results on the AMARA set might seem unrealistically high, we need to remember that the IWSLT baseline is out-of-domain for the AMARA test set, as explained by the high perplexity in table 3. Improving an out-of-domain baseline with in-domain data with translation model adaptation has been observed to give such high jumps in performance [11].

[11]The higher gain in BLEU for AMARA13 might be an artifact of using IWSLT target side for LM and IWSLT-dev for tuning.

SYS	LM	IW10	AM13
B_1	IWSLT	22.97	23.26
LM_1	AMARA	22.83	24.05
LM_2	IWSLT+AMARA	**23.69**	**25.90**
LM_3	INTERPOL	23.59	25.62
LM_4	GW	24.24	24.79

Table 5: Results of the translation system tested on IWSLT-tst2010 and AMARA-tst2013. All systems use identical translation model built on the IWSLT training data and use IWSLT-dev2010 for tuning.

SYS	TM	LM	IW10	AM13
B_1	IWSLT	IWSLT	22.97	23.26
S_1	TM_3	LM_3	**24.66**	31.62
S_2	TM_2	LM_2	24.33	**32.10**

Table 6: Results of the translation system tested on IWSLT-tst2010 and AMARA-tst2013. S_1 uses interpolated language model and merged phrase table to build translation model. S_2 uses concatenated training data for both translation model and language model.

In summary, we observed that both in isolation and in combination, the parallel and monolingual data from the volunteer-fueled AMARA corpus, is of sufficient quality to be used for a lecture translation task.

5. Error Analysis

For this section, we analyzed the errors performed during the translation of the AMARA13 testset. This was done to determine what are the specific challenges found when translating this set. We further provide a brief discussion of ways in which these problems can be fixed in the future. To do so, we classify the most important errors in two categories:

5.1. Mathematical quantifiers and numbers

One specific case of problem where recall is particularly low, refers to the translation of certain mathematical forms and numbers. This phenomenon is observed in instances where the numbers and operations were spelled out in the English side while in Arabic they are provided in their mathematical notation. For instance, the expression “is equal to” had a recall of 0 out of 41 times. The “the derivative of” was correctly translated only 6 out of 23 times. These problems

arise from the non-homogeneity with which mathematical texts are translated. For example:

Ar: ومرة اخرى هذا يساوي $+3+2$ ويساوي 5
En: Once again that's two plus plus three, so that equals five.

Ar: نحن بحاجة لتقييم نهاية اقتراب x من ما لا نهاية لـ $4x \hat{} 2 - 5x$ ، وكل ذلك مقسوم على $1 - 3x \hat{} 2$
En: We need to evaluate the limit, as x approaches infinity, of 4x squared minus $5x$, all of that over 1 minus $3x$ squared .

We observe that on the Arabic side, the mathematical symbols and digits are preferred, while in English, these are spelled out. A similar problem is the text-to-number conversion, which has been previously solved using rule-based approaches. In this case, a more refined set of rules can be devised to homogenize mathematical notation on both the source and target side of the corpus.

5.2. OOVs and transliteration

OOVs from languages with different scripts pose a challenge for readability. In an educational context, these need to be minimized and dealt correctly.

In the AMARA set, we observed that English terms are sometimes used in Arabic to denote English named entities. Examples of such cases are: Nevis, Yukon, Blanc, which are names of mountains used for math problems. These words can be left "untranslated" and the issue will be resolved.

A different problem, specific to Arabic-to-English translation, particularly for the technical domain, is the occurrence of OOVs related to neologisms. Fortunately many of these can be tackled by simple transliteration. For instance: جافاسكربت (javascript), ميدياغوبلن (media goblin), وركرافت (Warcraft), etc.

Together, these two problems account for 8 of the top 10 most frequent OOVs, this represents at least 12% of all the OOV words found in the testset.

6. Conclusion and Future Work

In this paper, we used data generated by a community of volunteers to advance the state-of-art of machine translation for educational content. This data, available through the AMARA platform, provides an opportunity to build a large, multilingual corpus, which can help to provide automatic translations in cases where no manual translation is available.

At this time, we explored the Arabic-English parallel portion of the data, and we evaluated its usefulness by translating the TED task of the IWSLT data. We presented different ways to process the data, especially to deal with problems in the original segment alignment. We showed that this data can be successfully used to translate lectures.

In addition, we used a new test set with AMARA specific data, geared towards educational translation. We observed that this data covers a broader domain than the IWSLT, and has specific challenges, some of which we analyzed. For instance, stylistic preferences when translating mathematical expressions, are prevalent and crucial for the content to be translated correctly.

In the future, we plan to extend the processing of the AMARA corpus to include at least 25 languages. Adding meta-data, like domain and topic, speaker, transcriber, and translator IDs, will allow using this corpus for speech translation research. For example, studying model adaptation or developing translation strategies to deal with the specific language and notation used in mathematics, biology, chemistry, etc. Finally, we plan to leverage the social graph of volunteers to be able to assign confidence to their translations depending on their characteristics (e.g. number of translations completed, domain of expertise, etc.). In summary, this data presents many possible lines of research. We are currently evaluating the different alternatives to make this corpus publicly available, while respecting copyright.

7. Acknowledgments

We would like to thank Nicholas Reville and the Amara staff for their support.

8. References

[1] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stücker, S. Vogel, and A. Waibel, "Open domain speech recognition & translation: Lectures and speeches," in *Acoustics, Speech and Signal Processing*, ser. ICASSP '06, 2006.

[2] C. Fügen, A. Waibel, and M. Kolss, "Simultaneous translation of lectures and speeches," *Machine Translation*, vol. 21, no. 4, pp. 209–252, 2007.

[3] J. A. Silvestre-Cerdà, M. A. del Agua, G. Garcés, G. Gascó, A. Giménez, A. Martínez, A. Pérez, I. Sánchez, N. Serrano, R. Spencer, J. D. Valor, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan, "TransLectures," in *Online Proceedings of Advances in Speech and Language Technologies for Iberian Languages*, ser. IBERSPEECH '12, Madrid, Spain, 2012.

[4] M. Paul, M. Federico, and S. Stüker, "Overview of the IWSLT 2010 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, ser. IWSLT '10, 2010.

[5] S. Green, J. Heer, and C. D. Manning, "The efficacy of human post-editing for language translation," in *ACM Human Factors in Computing Systems*, ser. CHI '13, 2013.

[6] M. Federico, S. Stüker, L. Bentivogli, M. Paul, M. Cettolo, T. Herrmann, J. Niehues, and G. Moretti, "The IWSLT 2011 evaluation campaign on automatic talk

translation," in *Proceedings of the Eight International Conference on Language Resources and Evaluation*, ser. LREC '12, Istanbul, Turkey, 2012.

[7] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, *et al.*, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 389–407, 2007.

[8] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation*, ser. EAMT '12, Trento, Italy, 2012.

[9] J. Tiedemann, "Synchronizing translated movie subtitles," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, ser. LREC '08, 2008.

[10] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eight International Conference on Language Resources and Evaluation*, ser. LREC '12, 2012.

[11] B. Haddow and P. Koehn, "Analysing the effect of out-of-domain data on SMT systems," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*, ser. WMT '12, Montreal, Canada, June 2012.

[12] S. Mansour and H. Ney, "A simple and effective weighted phrase extraction for machine translation adaptation," in *Proceedings of the International Workshop on Spoken Language Translation*, ser. IWSLT '12, 2012.

[13] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," *Computational linguistics*, vol. 19, no. 1, pp. 75–102, 1993.

[14] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, "Parallel corpora for medium density languages," in *Proceedings of the Recent Advances in Natural Language Processing*, ser. RANLP '05, 2005.

[15] S. Rao, I. Lane, and T. Schultz, "Optimizing sentence segmentation for spoken language translation," in *Proceedings of International Speech Communication Association*, ser. INTERSPEECH '07, Antwerp, Belgium, 2007.

[16] M. Paulik, S. Rao, I. Lane, S. Vogel, and T. Schultz, "Sentence segmentation and punctuation recovery for spoken language translation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, ser. ICASSP '08, Las Vegas, Nevada, USA, 2008.

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

[18] S. Green and J. DeNero, "A class-based agreement model for generating accurately inflected translations," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '12, Jeju Island, Korea, 2012.

[19] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[20] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, ser. HLT-NAACL '03, Edmonton, Canada, 2003.

[21] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, ser. WMT '11, Edinburgh, UK, 2011.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '02, Philadelphia, PA, USA, 2002.

[23] M. Hopkins and J. May, "Tuning as ranking," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11, Edinburgh, Scotland, United Kingdom, 2011.

[24] P. Nakov, F. Guzmán, and S. Vogel, "Optimizing for sentence-level BLEU+1 yields short translations," in *Proceedings of the 24th International Conference on Computational Linguistics*, ser. COLING '12, Mumbai, India, 2012.

[25] P. Nakov and H. T. Ng, "Improved statistical machine translation for resource-poor languages using related resource-rich languages," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '09, Singapore, 2009.