The AMARA Corpus: Building Parallel Language Resources for the Educational Domain

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, Stephan Vogel
Qatar Computing Research Institute
Doha, Qatar
{aabdelali,fguzman,hsajjad,svogel}@qf.org.qa

Abstract

This paper presents the AMARA corpus of on-line educational content: a new parallel corpus of educational video subtitles, multilingually aligned for 20 languages, i.e. 20 monolingual corpora and 190 parallel corpora. This corpus includes both resource-rich languages such as English and Arabic, and resource-poor languages such as Hindi and Thai. In this paper, we describe the gathering, validation, and preprocessing of a large collection of parallel, community-generated subtitles. Furthermore, we describe the methodology used to prepare the data for Machine Translation tasks. Additionally, we provide a document-level, jointly aligned development and test sets for 14 language pairs, designed for tuning and testing Machine Translation systems. We provide baseline results for these tasks, and highlight some of the challenges we face when building machine translation systems for educational content.

Keywords: Multilingual, parallel corpus, educational translation, lecture translation, crowd-sourcing.

1. Introduction

Lecture Translation has become an active field of research in the wider area of Speech Translation (Fügen et al., 2006; Fügen et al., 2007). This is demonstrated by large scale projects like the EU-funded translectures (Silvestre-Cerdà et al., 2012) and by evaluation campaigns like the one organized as part of the International Workshop on Spoken Language Translation (IWSLT) (Paul et al., 2010). 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 and delivered to thousands of students for free. 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, but also covering history, economy, psychology, medicine, and more.

Online education bridges 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. Most of this educational content is generated in English. This severely limits access for learners who are not able to 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 collection of TED Talks[4], for which so far more than $25,000$ volunteers have generated about $40,000$ translations into a total of 101 languages. However, it is clear 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 the language gap by automatically translating videos for which subtitles are not available. This can facilitate the task of volunteer translators, by providing an initial translation, which can be later post-edited (Green et al., 2013).

In this paper, we introduce the AMARA corpus, a new parallel corpus of subtitles of educational videos. While this corpus is designed specifically for SMT, it can be used for many other applications such as language recognition, bilingual dictionary generation, etc. Here, we describe in detail the gathering, validation and preprocessing of a collection of multilingual community-generated subtitles, which are publicly available through the Amara website[5]; and explore different approaches to align the subtitles, a prerequisite for the usage of this data in a machine translation task. Moreover, we provide parallel development and test sets, components that are required for building translation systems. Finally, we report and analyze the performance of baseline systems trained using the proposed corpus; and we identify the main challenges when translating educational content.

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) (Mostefa et al., 2007), 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[6] to its members.

More recently, the IWSLT10 (Paul et al., 2010) 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 (Cettolo et al., 2012).

[1] https://www.khanacademy.org
[2] https://www.coursera.org
[3] https://www.udacity.com
[4] http://www.ted.com
[5] http://www.amara.org
[6] http://www.elra.org

In the past, multilingual corpora creation from user-contributed movie subtitles has been addressed by Tiedemann (2008). Recently, a large collection of parallel movie subtitles from the OpenSubtitles[7] community along with tools for alignment of these has been made available through the Opus project (Tiedemann, 2012).

In this paper, we present the statistics 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 (Jansen et al., 2014) is a web-based platform for creating, editing and managing subtitles of on-line 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. The volunteers using Amara are typically organized into teams that carry specific translation and transcription efforts. The team hierarchy ensures data validation, and efficient task assignment to each team member according to his proficiency.

The Amara platform is used by many on-line educational organizations like KhanAcademy, TED, and Udacity. As a result, a large body of translations of educational content is available in multiple languages. This material usually consists of monologue video lectures, where a single instructor explains a variety of concepts. The genre of the lectures is informal speech, often with specific technical vocabulary, and with a large variety of topics. The transcriptions and translations of these videos are publicly accessible in the form of downloadable video subtitles.

3.1. Language Diversity

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 that many languages. A few videos have subtitles in as many as 109 different languages. Furthermore, at least 1000 videos have subtitles available in 25 different languages, and 3000 have subtitles available in at least 6 different languages. However, the distribution quickly tails off, as many videos have been subtitled into only a few languages.

The dominant languages of this repository are: English with 80K subtitles, French and Spanish with 19K subtitles for each language, Italian with 8.3K subtitles and Arabic with 5.6K subtitles. On the other hand, the original language of the videos is highly dominated by English with 120K videos, followed by Spanish with 8.3K videos, French with 5.7K videos, German with 4.8K videos and Russian with 4K videos. Under-resourced languages are also covered in the platform. For instance, a considerable amount of translations is available for rarer languages such as Thai or Hindi.

[7] http://www.opensubtitles.org

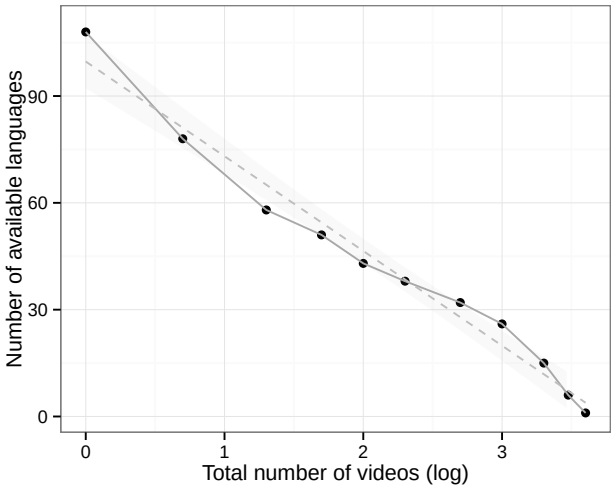

Figure 1: Distribution of the number of available languages per video by the total number of videos in the Amara website as of June 2013.

3.2. Crawling

The Amara site provides a list of videos and the number of languages the media has been subtitled into. Using an non-intrusive in-house crawler, and in cooperation with amara.org, we collected the list of videos available, and used it to collect corresponding transcripts and translations. The crawling yielded over 121K translated documents, corresponding to 43K videos in over 160 different languages. The initial collection was completed between Aug 10 and 20th, 2013.

3.3. Domain Filtering

In the Amara platform, videos come from very diverse sources with a wide range of domains and topics that include movies subtitles, music, advertisements, etc. However, we are interested only in educational videos. Thus, we only selected the resources that are related to education. To do so, we relied on (a) team information (as in the case of Khan Academy and Udacity), and (b) video metadata, (e.g. from youtube categories like Science and Education). Table 1 shows the various reference sources for the subtitled videos and their statistics.

Source	No. of Videos	No. of Subtitles
Khan Academy	2.7K	13.9K
Coursera	4.6K	6.6K
Udacity	3.7K	7K
Other-Education	9.4K	12.2K
Other-Science	2.6K	4.7K
Total	**23.1K**	**44.6K**

Table 1: Different educational sources of content and the number of videos and subtitles collected from them.

Language	Videos	Segments	Tokens*
English (en)	19357	2495K	25371K
Spanish (sp)	4506	479K	4898K
Portuguese (pt)	2341	291K	2806K
Chinese, Simp. (zhs)	1891	281K	368K†
Turkish (tr)	1618	205K	1199K
Polish (pl)	1430	197K	1430K
Arabic (ar)	1426	185K	1262K
Chinese, Yue (zht)	1354	231K	308K†
Russian (ru)	1274	146K	1162K
Italian (it)	1270	124K	1049K
French (fr)	1234	161K	1575K
Czech (cz)	1185	158K	1116K
Japanese (ja)	883	113K	169K†
Korean (kr)	786	109K	724K†
German (de)	760	99K	854K
Dutch (nl)	715	85K	753K
Thai (th)	687	95K	287K†
Bulgarian (bg)	668	100K	793K
Hindi (hi)	653	48K	509K
Danish (da)	582	58K	495K
Total	**44620**	**5.6M**	**47.1M**

Table 2: Distribution of the number of subtitles for the most popular video languages in the Amara platform. *Only approximate numbers as no language-specific tokenizer was used at this stage. † Numbers for certain languages are particularly low due to unsegmented text.

3.4. Data Validation

The collected data can be incomplete or contain wrong language information. Noisy data often results in poor word alignments and weak translation models; therefore, we had to carefully assess the content of the subtitles according to the following criteria: (a) *completeness:* we discarded around 30K subtitles that were empty or incomplete, and (b) *correct language:* we discarded subtitles, which were not in the language they claimed to be. To that end, we used the Cybozu Open Source Language Detector [8].
Filtering the collected data according to the above criteria resulted in 34K subtitles corresponding to 12.2K videos. However, for many languages, the quantity of documents is too small to be useful for Machine Translation purposes. Thus, we restrict the current release of the corpus to the 20 languages with the most resources available. Table 2. shows the monolingual statistics for each of those languages.

4. Parallel Segment Alignment

The gathered subtitles consist of segments that are formed by three components: (a) *segment id:* a number, in sequence, identifying the segment, (b) *time interval:* the start and end times of the subtitle, which represent the timeframe the particular subtitle appears on the screen, and (c) *content:* the text for the subtitle segment, with one or more lines.

[8]The library supports the detection of over 53 languages. In the case of Chinese, we trained our own classifier, given that Cybozu was unable to handle it correctly. The library is available at: `http://code.google.com/p/language-detection/`

Sentence Alignment	Spanish	Arabic	Russian
Baseline	241K	128K	56.3K
Cascade Sync	691K	318K	157K
Improvements	**287%**	**249%**	**279%**

Table 3: Comparison of the resulting number of segments using two different synchronization approaches.

In order to build parallel resources, we need to align the subtitle files at a segment level. About 75% percent of all collected segments from Amara have identical time stamps on both sides. However, there are two cases, which lead to non-parallel segments: (a) when the data in one language is not complete, and (b) when the text of source and target segment correspond to each other, but the timestamps are not synchronized across languages. This can happen if the subtitles are generated independently of the original language. To address the issues mentioned above, we used two algorithms to align the subtitle files:

- **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.
- **Cascaded synchronization**
  This approach is an extension to the previous synchronization approach. However, it tries to align the discarded segments by using length statistics and information from a bilingual dictionary in the spirit of Gale and Church (1993). We started by enforcing a strict synchronization constraint the subtitles. Then we performed word alignment on the concatenation of all of the strictly aligned data, and extracted a bilingual lexicon from the resulting alignment. For this, we used implementation provided by Hunalign (Varga et al., 2005). This lexicon was then used to run the automatic sentence aligner on the unsynchronized portions of the subtitles[9]. For development, and test sets, we required to have high-confidence alignments. Thus, to build those specific sets, filtered the aligned segments according to their final alignment score[10]. Finally, we concatenated both the strictly synchronized with the automatically aligned portions of the subtitles to generate the corpora.

In Table 3 we present the comparison of the two segment alignment strategies for the Spanish-English, Arabic-English and Russian-English language pairs. Guzman et al. (2013) provide a detailed comparison of different subtitle alignment methods. We observe that after synchronization, we are able to retrieve between 2 and 3 times more training data, that otherwise would have been lost. Table 4 presents the statistics for the different parallel corpora after applying the cascaded synchronization.

[9]This strategy might not be accurate for languages without word boundaries such as Chinese, Japanese, Korean and Thai

[10]We used a threshold of 0.1, and rejected segments with lower scores

	en	sp	pt	zhs	zht	tr	pl	ar	fr	cz	ru	it	ja	kr	bg	de	th	nl	da	hi
en	**2495K**																			
sp	335K	**479K**																		
pt	231K	117K	**291K**																	
zhs	139K	52K	46K	**281K**																
zht	117K	49K	48K	191K	**231K**															
tr	169K	72K	67K	47K	51K	**205K**														
pl	151K	88K	72K	47K	52K	65K	**197K**													
ar	158K	83K	73K	58K	59K	90K	69K	**185K**												
fr	125K	63K	58K	29K	29K	36K	59K	48K	**161K**											
cz	132K	61K	65K	55K	56K	66K	69K	56K	40K	**158K**										
ru	77K	39K	36K	22K	20K	18K	30K	29K	29K	25K	**146K**									
it	97K	52K	49K	29K	29K	38K	43K	44K	39K	38K	23K	**124K**								
ja	98K	44K	44K	46K	42K	47K	39K	48K	31K	43K	23K	29K	**113K**							
kr	83K	36K	36K	30K	32K	28K	32K	37K	25K	30K	14K	28K	26K	**109K**						
bg	79K	39K	44K	28K	33K	45K	44K	39K	33K	42K	15K	27K	21K	19K	**100K**					
de	77K	49K	45K	22K	25K	30K	45K	36K	42K	35K	22K	30K	24K	23K	28K	**99K**				
th	85K	50K	40K	31K	29K	56K	38K	45K	21K	29K	14K	22K	27K	15K	20K	20K	**95K**			
nl	73K	43K	42K	25K	29K	33K	41K	40K	33K	41K	19K	31K	25K	22K	25K	30K	19K	**85K**		
da	48K	27K	34K	18K	21K	29K	25K	30K	23K	32K	12K	21K	18K	16K	25K	16K	10K	21K	**58K**	
hi	43K	26K	22K	14K	14K	17K	24K	25K	16K	17K	8K	26K	16K	14K	13K	13K	13K	17K	15K	**48K**

Table 4: Total number of parallel segments resulting from the cascaded synchronization.

4.1. Test sets

For a specific subset of 14 languages, we also provide development and test sets required to build Statistical Machine Translation systems. The set of languages and documents was obtained by jointly maximizing the number parallel documents, subject to the following constraints: (a) each document had to be translated into each of the languages in the set, and (b) the total amount of sentences for each set should be at least 1000 parallel sentences (with the exception of Chinese, Japanese and Korean). This resulted in 13 (6+7) documents for test, and 8 for development. Table 5 provides further details of the distribution of documents according to their sources. Note how a large proportion of the data which is highly parallel, corresponds to Khan Academy.

The languages covered by these sets are: Arabic, Chinese Simplified, Czech, Danish, Dutch, English, French, German, Japanese, Korean, Polish, Portuguese, Russian, and Spanish. For the remaining languages, test sets are to be included in future versions of the corpus. In Table 6, we show the statistics of the provided test sets for translation between English and the other 13 languages.

4.2. Availability

The AMARA corpus is publicly available through the AMARA corpus website[11].

	Khan Academy	Other-Science
tst2014a	85%	15%
tst2014b	76%	24%
dev2014	92%	8%

Table 5: The distribution of data categories in the development and test sets.

[11] http://amaracorpus.qcri.org

Language	train	dev2014	tst2014a	tst2014b
Spanish (sp)	329K	1252	1126	1411
Portuguese (pt)	227K	1138	1066	1372
Arabic (ar)	152K	1160	1141	1425
Polish (pl)	146K	1159	1138	1455
Czech (cz)	127K	1101	1158	1344
French (fr)	120K	1220	1132	1410
Chinese S. (zhs)	119K	808	718	1083
Japanese (ja)	90K	1172	936	1455
Korean (kr)	75K	989	1028	1299
German (de)	73K	1117	1147	1412
Russian (ru)	73K	1164	1166	1381
Dutch (nl)	68K	1309	1137	1455
Danish (da)	44K	1138	1136	1455

Table 6: Statistics of development and test sets for translation between English and other 13 languages. For reference, we also show the data available for training.

5. Experimental Results

In this section, we provide baseline machine translation results using the proposed datasets. To be able to better compare and analyze the results, we built several systems that translate into English. The languages considered in these experiments were: Spanish, Russian, Arabic, Portuguese, Polish, Danish, Dutch, French, Czech and German.[12]

5.1. Experimental Setup

Preprocessing: We tokenized the English side of all bitexts for language modeling using the standard tokenizer of the Moses toolkit (Koehn et al., 2007). 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 source languages except Arabic, we tokenized using the standard tokenizer of the Moses toolkit. For Arabic, we segmented the corpus following the ATB segmentation scheme with the Stanford word segmenter (Green and DeNero, 2012).

[12] In this paper, we do not provide experiments for Japanese, Korean, or Chinese as they require a language dependent word segmenter.

Training: We built word alignments using IBM model 4 (Brown et al., 1993), and symmetrized them using *grow-diag-final-and* heuristic (Koehn et al., 2003). We extracted phrase pairs up to a maximum length of seven words. We scored these phrase pairs using maximum likelihood with Kneser-Ney smoothing, 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 the target (i.e. English) side of the training bi-text using KenLM (Heafield, 2011). 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 (Koehn et al., 2007) for translation, and reported evaluation results over two AMARA test sets. We reported BLEU calculated with respect to 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 (Papineni et al., 2002) on the AMARA dev2014 dataset, using PRO (Hopkins and May, 2011) with the fixed BLEU+1 (Nakov et al., 2012; Nakov et al., 2013). We allowed the optimizer to run for up to 25 iterations, and to extract 1000-best lists for each iteration.

Decoding: During tuning and testing, we used monotone-at-punctuation decoding. On testing, we further used cube pruning, minimum Bayes risk decoding (Kumar and Byrne, 2004) and the operation sequence model (Durrani et al., 2011).

5.2. Translation Results

In Table 7 we present the results (BLEU) for each of the test-sets and each of the 10 systems. The BLEU scores are fairly high, particularly for Portuguese, Spanish, and Danish, given that the test sets have one reference translation only. For instance, the scores for Spanish are 48.2 and 41.4 for the tst2014a and tst2014b, respectively. As a comparison, the BLEU score for training on TED, and testing on TED-tst2010 is 36.6. For morphologically rich languages like Arabic, Czech, Russian, the performance is also higher than expected. E.g. for Arabic, BLEU scores are 38.0 and 34.4. As a comparison, TED BLEU scores are 23.6 for testing on TED-tst2010.

However, when we take a closer look at the characteristics of the data, these results are understandable: the utterances in the educational videos, when segmented into subtitles, are typically short. For example, the English side of dev2014a set is 7.54 words long on average. The corresponding Arabic subtitles average 7.5 words, Russian it is 7.7 words, to name a few. In addition, the segments exhibit mostly simple syntactic structures with very little reordering, making the decoding task mostly monotone. This suggests that translating educational subtitles might be an easier task.

	BLEU NIST v13		OOV	
Source Lang.	**tst2014a**	**tst2014b**	**tst2014a**	**tst2014b**
Spanish (sp)	48.2	41.4	0.6%	0.8%
Portuguese (pt)	52.1	46.6	0.7%	0.9%
Arabic (ar)	38.0	34.4	1.0%	1.2%
Polish (pl)	34.7	29.4	2.5%	2.4%
Czech (cz)	33.7	32.9	2.3%	2.6%
French (fr)	31.5	35.1	0.8%	1.1%
German (de)	35.2	34.0	2.3%	1.8%
Russian (ru)	34.3	38.6	1.8%	1.7%
Dutch (nl)	39.8	45.6	1.3%	1.4%
Danish (da)	40.5	35.3	2.5%	2.6%

Table 7: Results for systems trained, tuned and tested on the AMARA corpus, and translating into English.

A second observation can be made from the results in Table 6. For some languages, e.g. Spanish, Portuguese, and Danish, the first test set gets higher BLEU scores than the second, whereas for other languages, like French and Russian, it is the other way round. To understand this behavior we need to keep in mind that different language pairs have different quantities of data for training, which in turn leads to different degrees of coverage for the two test set. This can be seen in the differences of OOV rates for the different languages and test sets. Although OOVs only explains part of the issue, there is a moderate correlation (-0.578) between the BLEU score and the OOV rate i.e. more OOVs, lower BLEU score.

In summary, we observed that the AMARA corpus is useful for training translation systems to translate new educational material. However, there are some challenges particular to this genre. In the following section, we provide some examples thereof.

5.3. Discussion

To shed light on the characteristics of the educational data, we analyzed the output of the Russian, Spanish and Arabic to English machine translation systems. Besides having short sentences, the datasets contain a good percentage of mathematical equations and formulas, which require mostly monotonic translations. Figure 2 shows a few output examples of the Russian, Spanish and Arabic to English machine translation systems. There, we split the examples into two categories: (a) the *best* translations, and (b) the *worst* translations; according to their BLEU score. From these examples, we can draw the following observations:

- Some of the translations require little to no reordering, thus are easier to translate. To illustrate this, observe the second example of the *best* translations for the Spanish to English system.
- A few of the *worst* translation examples exhibit a data sparseness problem, as shown by OOV words. This can be solved using an in-domain dictionary or with larger training corpora. However, this is not a serious issue, given that for most of the language pairs, we observed an OOV rate lower than 3%.

	Best Translations	Worst Translations
Spanish		
Src	Luego el brazo volverá a su posición original	Ej. -3 < +3
Trans	Then the arm will return to its original position	It. Minus 3 < +3
Ref	Then the arm will return to its original position	You know, minus 3 is less than positive 3.
Src	Entonces esto es igual a 4x al cuadrado menos 2x más 8.	La opción B es (x + 3) por (x - 4) por una
Trans	So this is equal to 4x squared minus 2x plus 8.	Choice B is (x + 3) times (x - four) by a
Ref	So this is equal to 4x squared minus 2x plus 8.	So choice B is x plus 3, times x minus 4, times
Russian		
Src	3 плюс минус 7 является минус 4.	Затяните каждый винт на шайбу до тех пор, пока шайба останавливается поворота
Trans	3 plus minus 7 is minus 4	Tighten every screw on шайбу until шайба stop rotating
Ref	3 plus minus 7 is minus 4	Hand-tighten each screw onto the puck until the puck stops turning
Src	Так минус раз х квадрат это минус х в квадрате.	Я просто умножу на -2 и раскрою скобки
Trans	So minus times x squared is minus x squared.	I just multiply it by 2, and let me distribute
Ref	So minus times x squared is minus x squared.	I'm just going to distribute the -2 times all of this.
Arabic		
Src	نطرح 3 من كلا الطرفين	4(2- س3 1) 3 6 =(س5)
Trans	Subtract 3 from both sides.	4 times 2 is 5x is equal to 6 minus 3 times 1 minus 3 is going
Ref	Subtract 3 from both sides.	4(2-5x)=6-3(1-3x)
Src	ما هو الرسم البيانى لـ-x^2 = y ؟	لذلك قالوا 2س+3س=5س
Trans	What is the graph of y is equal to minus x squared?	so they said, 2x plus 3, x is equal to 5 is
Ref	What is the graph of y is equal to minus x squared?	so they said 2x + 3x = 5x.

Figure 2: A few best and worst translation examples of the Spanish, Russian and Arabic to English machine translation systems.

- The evaluation of this mathematical content presents a serious challenge, given its diversity of representations. To illustrate this, observe the first example of the *worst* translation for the Spanish system. The translation system produces an adequate translation, but it gets penalized with a low BLEU score given that it outputs the symbolic representation of *less than* "<" instead of its name.

  A similar behavior can be seen in the second example where the reference contains a mix of representations, i.e "− four", while the system outputs the representation "minus 4". This highlights the inconsistency of human translators when translating digits and symbols across languages. The *worst* translation examples for Arabic and Russian show similar phenomena. These issues can be resolved by either introducing a preprocessing step to standardize representations, or by using an evaluation metric such as METEOR (Lavie and Denkowski, 2009), that can handle paraphrases.

6. Conclusion and Future Work

In this paper we described the version 1.4 of the AMARA corpus: a new multilingual corpus, covering a wide range of educational lectures. The corpus was derived from the AMARA platform, where volunteers contribute subtitles and translations for on-line video material. These volunteers are highly motivated to make the educational material available in their own language. As a result, translations are of high quality, particularly in cases where an explicit quality control mechanism has been established. This makes these resources very valuable to build machine translation systems for educational material. The spoken and educational nature of the data leads to new challenges for both translation and evaluation. Moreover, the multilingual nature of this corpus can have many other applications such as language recognition, bilingual dictionary generation, etc.

In this paper, we established a first set of training, development and test corpora, which we used to build and evaluate translation systems for many of the languages covered in the AMARA corpus. We demonstrated the usefulness of the data for the purpose of translating educational lectures.

In the future, we will continue to update the corpus as new transcripts and translations become available. In next releases of the corpus, we expect to include more languages, as the amount of volunteer translations increases. We are particularly interested in extending the corpus to cover low-resource language pairs as this is where translation technology for the educational domain will have the biggest impact.

Besides growing the AMARA corpus, we will address some of the specific challenges when translating educational videos containing mathematical formulas, chemical notation, and specialized terminology. Furthermore, we plan to explore the usage of this data in translation applications. For instance, one task will be to performmanual assessments of the usefulness of automatic translation of educational videos, e.g. how much information is lost in automatic translations. A second task will be to evaluate whether machine translation output can facilitate the task of the volunteer translators, i.e. by performing post-editing on the machine translation output instead of starting from scratch, or by presenting the user with translation alternatives, lexicon, etc.

7. Acknowledgements

We would like to thank the Amara.org staff, especially Nicholas Reville, Aleli Alcala and Dean Jansen for their support in the development of this corpus. We would also like to thank the reviewers for their useful recommendations.

8. References

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, EAMT '12, Trento, Italy.

Durrani, N., Schmid, H., and Fraser, A. (2011). A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, USA.

Fügen, C., Kolss, M., Bernreuther, D., Paulik, M., Stücker, S., Vogel, S., and Waibel, A. (2006). Open domain speech recognition & translation: Lectures and speeches. In *Acoustics, Speech and Signal Processing*, ICASSP '06, Toulouse, France.

Fügen, C., Waibel, A., and Kolss, M. (2007). Simultaneous translation of lectures and speeches. *Machine Translation*, 21(4):209–252.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

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

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *ACM Human Factors in Computing Systems*, CHI '13, Paris, France.

Guzman, F., Sajjad, H., Abdelali, A., and Vogel, S. (2013). The AMARA corpus: Building resources for translating the web's educational content. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT '13, Heidelberg, Germany.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, Edinburgh, UK.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Edinburgh, Scotland, United Kingdom.

Jansen, D., Alcala, A., and Guzman, F. (2014). Amara: A sustainable, global solution for accessibility, powered by communities of volunteers. In *Proceedings of the 16th International Conference on Human-Computer Interaction*, HCII '14, Heraklion, Greece.

Koehn, P., Och, F. J., and Marcu, D. (2003). 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*, HLT-NAACL '03, Edmonton, Canada.

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.

Kumar, S. and Byrne, W. (2004). Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Main Proceedings*, Boston, Massachusetts, USA.

Lavie, A. and Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., et al. (2007). The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3-4):389–407.

Nakov, P., Guzman, F., and Vogel, S. (2012). Optimizing for sentence-level BLEU+1 yields short translations. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING '12, Mumbai, India.

Nakov, P., Guzman, F., and Vogel, S. (2013). A tale about PRO and monsters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, Philadelphia, USA.

Paul, M., Federico, M., and Stüker, S. (2010). Overview of the IWSLT 2010 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT '10, Paris, France.

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

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, Marrakech, Morroco.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC '12, Istanbul, Turkey.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, RANLP '05, Borovets, Bulgaria.