Amara: A Sustainable, Global Solution for Accessibility, Powered by Communities of Volunteers

Dean Jansen[1], Aleli Alcala[1], and Francisco Guzman[2]

[1] Amara.org, USA
{dean,aleli}@pculture.org
[2] Qatar Computing Research Institute, Qatar
fguzman@qf.org.qa

Abstract. In this paper, we present the main features of the Amara project, and its impact on the accessibility landscape with the use of innovative technology. We also show the effectiveness of volunteer communities in addressing large subtitling and translation tasks, that accompany the ever-growing amounts of online video content. Furthermore, we present two different applications for the platform. First, we examine the growing interest of organizations to build their own subtitling communities. Second, we present how the community-generated material can be used to advance the state-of-the-art of research in fields such as Statistical Machine Translation with focus on educational translation. We provide examples on how both tasks can be achieved successfully.

Keywords: Amara, online platform, user engagement, subtitles, translation, crowdsourcing, volunteer communities, lecture translation, statistical machine translation.

1 Introduction

In a globalized world, multilingual subtitles are absolutely essential to the future of web video: they bring more viewers, they make content searchable, and they allow access to content to billions of users who are left out because they are deaf, hard of hearing, or unable to access video because it's in a language different than their own.

Amara is a web service and software toolset for adding captions, subtitles, and translations to virtually any web video. Until now, captioning and translation services have been limited by closed, centralized, and expensive systems; and by tedious user experiences. The Amara system is an open, scalable, flexible, collaborative platform, which allows leveraging the power of crowdsourcing through volunteer engagement. It is the first large-scale, open platform with the potential of making large quantities of videos accessible through high quality captioning and translation.

In this paper, we describe the unique features of the Amara platform, which allow for faster transcription turnaround while maintaining high levels of user engagement.

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2014, Part IV, LNCS 8516, pp. 401–411, 2014.

Additionally, we showcase success stories where the platform has allowed volunteer translators to transcribe and translate hundreds of thousands of videos, in communities of varying sizes. Finally, we present a brief summary of cases where the community-generated translations can be used to advance research in the field of automatic translation focused on educational videos.

2 Features of the Amara Platform

To ensure the best user experience, the Amara platform is focused on five important areas: (i) ease of use, (ii) quality control, (iii) compatibility, (iv) ease of integration, and (v) engagement. Below, we explain each of these topics.

1. Ease of use

By separating the transcription step from the alignment step, the Amara interface for subtitle creation is extremely simple and enjoyable. Compared to existing desktop or web subtitling tools, it is easier for a new user to get started.

In addition, by separating subtitle services from video hosting, it lets users add subtitles with simplicity. There is no need to download videos and/or upload them to a new service.

2. Quality controls

The platform puts an emphasis on the production of high-quality subtitles and translations. This is accomplished through transparency, accountability, and a policy of open participation. To that end, Amara quality control works in the same way as Wikipedia's self-regulated community, where members collaboratively solve errors and problems with articles.

3. Compatibility

Amara supports the use of four of the most popular hosting sites (Youtube, Dailymotion, Kaltura, and Vimeo). Additionally, Amara lets users subtitle the vast majority of videos posted online, because it also supports video formats like .mp4 and .webm (the HTML5 standard for plugin-free video playback).

4. Ease of integration

Anyone can add Amara to their website by pasting a single line of JavaScript to their videos. Site owners and their visitors can immediately begin adding subtitles and translations to their videos. Any site using a compatible video host or player can become part of the Amara ecosystem in minutes.

5. Volunteer engagement

Volunteers that caption and subtitle videos are the core element of the Amara project. Therefore, Amara strives to adopt the best current practices and functionalities to

engage and motivate volunteers. Today, hundreds of thousands of volunteers participate on Amara and the numbers continue to grow.

2.1 Key Components

The following components are essential to Amara's success:

- A creation and editing interface that makes it easier to subtitle than any other system for captioning and translating video. This is an essential component needed to engage and involve millions of volunteers. A snapshot of the interface is shown in Figure 1.
- A collaborative editing process that facilitates the incremental improvement of subtitles and captions. This process includes features such as: revision history, rollbacks, email notifications when changes are made, etc.
- A platform that allows companies and organizations to: (a) collect and organize video into projects, (b) import / export and manage videos and subtitles through a powerful API, (c) manage subtitle creation processes involving volunteers, staff, contractors through flexible workflows, (d) manage the privacy of the content, (e) build and manage a volunteer team, including an application process and peer review for quality control, (f) synchronize completed subtitles to YouTube videos
- A comprehensive support for more video hosting platforms than any other system, enabling access to more videos.

Early adopters of the platform include TED Talks, Mozilla, PBS Newshour, and Udacity. This network of communities is growing exponentially and making tens of thousands of videos accessible to wider audiences.

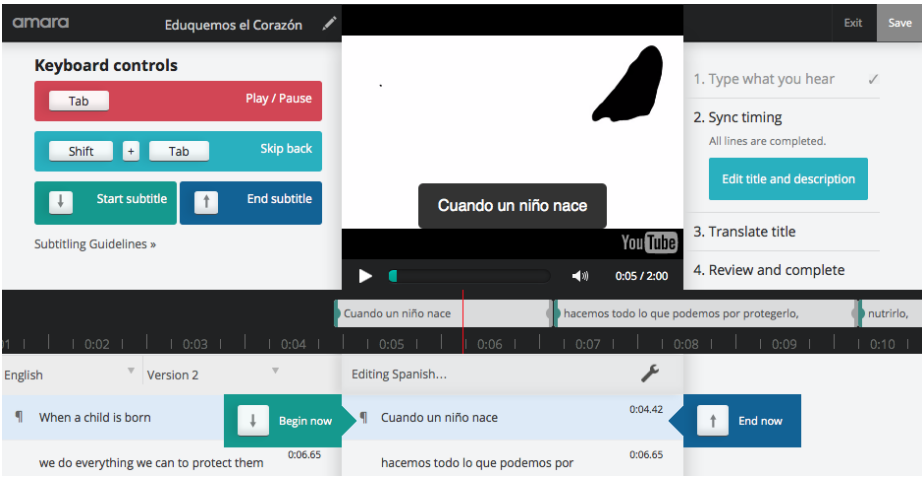

Fig. 1. A snapshot of the Amara subtitle editor

3 Using the Amara Platform: Case Studies

Below, we present different case studies that provide insight into how some organizations are successfully engaging their volunteers, fans, viewers, and students, through the Amara platform, to make their content accessible to more people, in more languages, and around the globe.

3.1 TED Talks

Around 2009, TED Talks[1] started a volunteer translation community using a homegrown system to create translations for their then 1100 TED Talks. During this time, TED built a community of about 7,000 members that focused solely on translating TED's 1100 or so TED Talks.

In 2012, TED migrated to the Amara platform given its ease of use. Additionally, this move was done to address their growing volume of TED, TEDx and TED Ed talks. They needed a subtitle creation tool that: (a) required very little training of their volunteers, (b) was scalable to support large amounts of content, and volunteers.

Since they switched to Amara in May of 2012, TED's volunteer base has grown to nearly 25,000 members. Their video content library is near 29,000 videos and growing steadily: TED Talks has 1600 videos; TEDx Talks, 27,000 videos; and TED Ed around 400 videos; all of which are now being translated by their team of volunteers. The rate at which the team completes subtitles continues to grow, as TED expands their use of the Amara platform.

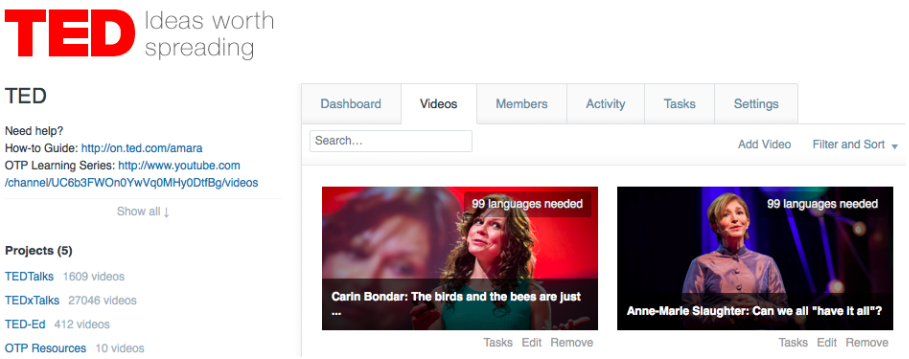

Fig. 2. The TED Translation Team on Amara

[1] http://www.amara.org/en/teams/ted/

3.2 Udacity

Udacity[2] is one of the largest Massive Open Online Courses (MOOC), education platforms. Udacity began using Amara in late 2012. Their courses focus on advanced math, science, and technology content. When they joined, their greatest concerns given the complexity of their material were: the quality of subtitles, and the volunteer engagement.

Today, Udacity has a strong membership of over 1500 volunteers. Although a smaller team than TED's, this team still manages to translate Udacity's videos into many languages. More importantly, this group of volunteers is passionate about making Udacity's content, education from some of the world's prestigious universities, available to people who do not speak English. Thus, allowing many around the world to benefit from this valuable content.

To ensure a high number of translations, Udacity also uses Amara's "On Demand" service to gain closed captions for all of their courses, making it easier and faster to gain translations from their volunteer community.

3.3 Github

Github[3] is one of the largest open source technology companies in the world. They joined Amara in October 2013 with just a handful of videos, and zero volunteers. Using Amara's standard outreach plan, they quickly grew their community.

To date, they are satisfied with the efforts of their volunteer community, and continue to add videos on a regular basis to keep them engaged.

Interesting Volunteer Behavior Developing across Communities. One of the most exciting things happening with the Github community is the passion of some of their members. So passionate with their participation in fact that they are blogging about it, tweeting and posting on Facebook to invite their friends and family to watch these videos and see the subtitles they created.

This volunteer behavior is a new phenomenon at Amara, and is also happening with TED and Epic Rap Battles volunteers.

This social behavior brings additional benefit to the translation efforts of the volunteer community. Team member communication gives members the opportunity to share their passion for the organization. At the same time, it also attracts more volunteers and viewers.

[2] http://www.amara.org/en/teams/udacity/
[3] http://www.amara.org/en/teams/github/videos/

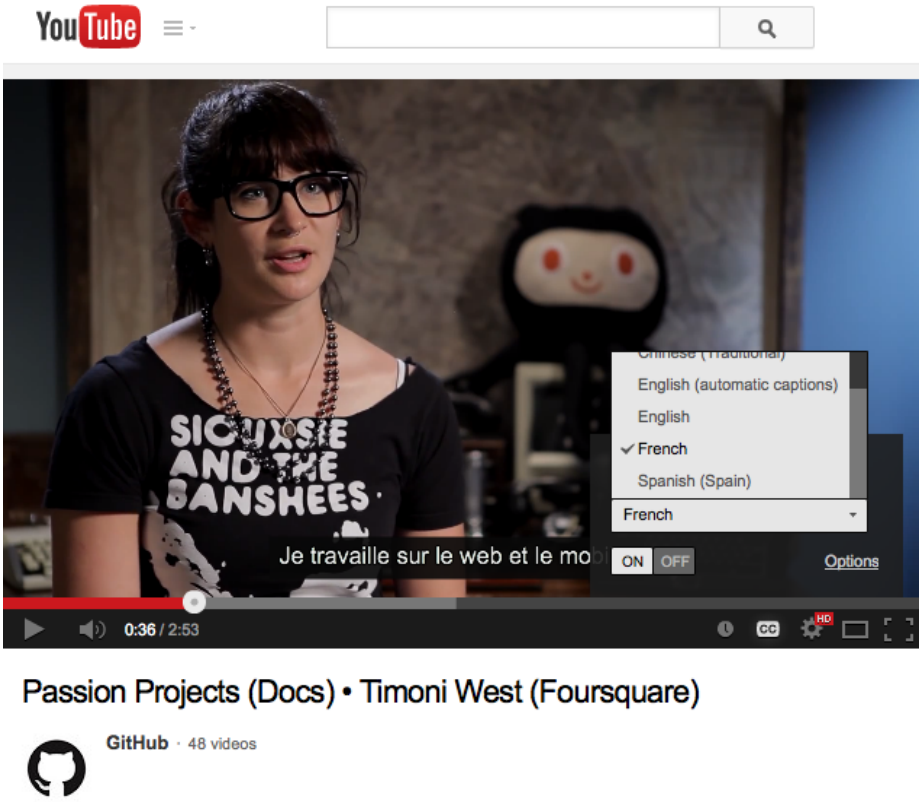

Fig. 3. Github's Youtube Videos, translated by community, using Amara

3.4 Maker Studios, Epic Rap Battles Series

Maker Studios is a media company focused primarily on the millennial generation, who are living a mobile, social, and on-demand life. Maker is the number one producer and distributor of online video to this diverse, tech-savvy group of people, attracting over 4.5 billion monthly views, and 340 million+ subscribers.

One of Maker's most popular YouTube channels is Epic Rap Battles (ERB) with about 8 million subscribers. Their viewer demographic for series, are 12 to 14 year-old boys. Therefore, they were cautious about engaging this young ERB community in volunteer subtitling. However, the idea of Amara and building a volunteer community was too compelling. So, Maker decided to do a small, highly controlled test.

Maker's Proof of Concept and Amazing Success Story. Maker chose four videos from their Epic Rap Battle[4] series and put them in an unlisted YouTube channel. They then sent just one tweet sharing this unlisted URL, inviting volunteers to join and help subtitle these four videos. Doing it this way, they figured only a very small subset of their subscribers would actually see the tweet, and an even smaller subset would probably join the effort.

What happened next was amazing: in less than 18 hours, they garnered 210 volunteers who subtitled each of the four videos into over 20 languages per video. Furthermore, after comparing the volunteer translations created in these languages to a set of refereed translations, their reviewers found the volunteer translations to be of higher quality, because they captured humor and tone of the content.

Great takeaway: volunteers who are passionate about the organization and the content have a greater understanding of the content, and are better positioned to create higher quality captions and translations.

After nearly two months, the Epic Rap Battle's community continues to grow both in membership, and the number of subtitles created per video, with no vandalism occurring.

An Epic Social Event and Maker's Upcoming New Amara Communities. Recently, Maker decided to try another experiment. They did a targeted Facebook promotion to different geographical regions around the world. Within hours, their volunteer membership went from around 400 members to over 3,000 members, shutting Amara down for a few minutes in the process. In fact, the Amara platform was not prepared for such levels of traffic jam, but quickly recovered. With this success and the earlier successes shared, Maker is now creating translation communities for three more of their YouTube channels.

Fig. 4. Maker Studios "Epic Rap Battles of History" on Amara

[4] http://amara.org/en/teams/erb/videos/

3.5 Summary

In Table 1 below, we present a summary of the engagement statistics for the different case studies introduced previously. Notice how the Amara platform has been successfully used in projects from different sizes. In all four cases, we have observed a widespread adoption by the community of volunteers.

Table 1. Engagement statistics from different organizations using Amara

Org	Videos	Volunteers	Languages (avg)
TED talks	29,000+ videos, 4 times the starting number	25,000 volunteers, 4 times the starting number	40+ languages per TED Talk, 10+ languages per TEDx and TED Ed Talks
Udacity	11,000 videos	1,500+ volunteers	10+ languages
Github	34 videos	450+ volunteers	Between 8 and 9 languages, some with up to 17 languages
Epic Rap	18 videos	3,295 volunteers	40+ languages, some with over 60 languages

These are but a few of the growing number of organizations using the Amara Enterprise platform. As an example, in the next week or so, Scientific American, the United States National Archives, World Vision, and Qatar Computer Research Institute (QCRI) are some of the top organizations launching their volunteer communities on Amara.

In addition, there are ongoing conversations happening with actors from diverse vertical markets, including higher education, film and TV, media companies, online education portals, and major YouTube channels. Amara is experiencing a strong growth not only in interest also in the rate adoption from many different vertical markets.

One market for which there is growing interest is the educational sector. The number of Amara partners, specialized in both higher education and online education, is increasing at a rapid pace. This is a particularly gratifying use of the platform, because it permits volunteers to generate translations for educational videos, allowing such videos to reach wider audiences. Furthermore, these translations are being used to explore how the translations created by volunteers can be used to generate automatic translations for videos that have not been translated by volunteers. This can further help to reduce the language barriers. Below, we briefly summarize such efforts.

4 Using Volunteer Translations for Educational Translation

In this section, we summarize the research presented by [5,6], which use translations generated in the Amara platform to train machine translation systems and improve the state-of-the-art in Lecture Translation in the educational domain.

The automatic translation of educational material has become an active field of research in the wider area of Speech Translation [1,2]. In this area of research, techniques from Speech Recognition and Machine Translation are applied to automatically translate technical lectures from one language (source) to another language (target). To advance the state-of-the-art in this field, researchers have proposed large-scale projects like the EU-funded translectures [3] and evaluation campaigns like the one organized as part of the International Workshop on Spoken Language Translation (IWSLT). However, the main limitation for the success of these is the access to high quality training data.

With the emergence of Massive Online Open Courses (MOOCs), thousands of educational video lectures have already been generated. Organizations like Khan Academy[5], and Udacity[6], etc., continuously increase their offerings of lectures, which range from basic math and science topics, to more advanced topics like machine learning.

However, language barriers limit the access to this content, given that most of the material is in English. This severely limits access to this high-quality educational material for learners who do not understand English.

To overcome these language barriers, volunteers continuously transcribe and translate such lectures into many other languages using the Amara platform. One example is the already mentioned TED Talks[7], for which, so far, more than 25,000 volunteers have generated more than 40,000 translations into a total of 101 languages. However, for many languages the small number of volunteers is insufficient to keep up with the rate 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. Thus, it 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, this type of data is very rare and expensive to create by hand. So far, the only openly accessible corpus for the lecture domain has been the TED talks [4], which is also based on volunteering efforts.

To address these issues [5], [6] have proposed to use the transcriptions and translations generated by volunteers in the Amara platform, to create corpora suitable for research in the Machine Translation field. They focus on generating corpora that targets educational content. They observe that the data generated by volunteers can be successfully used for such task. They analyze the output of the machine translation systems, and identify specific challenges that arise when translating highly technical

[5] https://www.khanacademy.org
[6] https://www.udacity.com
[7] http://www.ted.com

data, such as mathematical formulae. Furthermore, they observe that the gathered data can also be used to translate other lecture material such as the TED talks, thus showing that the data obtained from the Amara platform can be successfully used for improving lecture translation.

5 Conclusion

Today, video is the fastest growing form of content on the web, with videos being added daily on an exponential scale. Amara's mission of making all video content online accessible, is only possible by providing an open Wikipedia-like solution, where any user, is enabled to participate in addressing this challenge.

Amara is a unique volunteer driven platform that has been tested in real-world scenarios and has helped translate tens of thousands of videos. In the translation communities, volunteers are passionate about the content they translate, and generate high quality translations. Furthermore, contributing volunteers are proud of their achievements, and quickly spread-the-word to their networks, creating a social "contagion" effect. This allows translation communities to grow rapidly. Moreover, the corpora of translations and transcriptions generated by volunteers are being used to improve statistical machine translation systems for educational content.

6 About AMARA

6.1 Amara and PCF

Amara is a project of the Participatory Culture Foundation (PCF), a not for profit 501 (c)(3). Amara's mission is to ensure that all online video content is accessible to everyone regardless of hearing ability or language barriers.

6.2 Amara and Prosperity4All (P4ALL)

The Prosperity4All consortium is comprised of 25 partners from 13 countries who are developing a robust cross-platform spectrum of mainstream and assistive technology-based access solutions required for a sustainable Global Public Inclusive Infrastructure. The Amara platform will be the technology for creating captions and subtitles within this ecosystem.

References

1. Fügen, C., Kolss, M., Bernreuther, D., Paulik, M., Stücker, S., Vogel, S., Waibel, A.: Open domain speech recognition & translation: Lectures and speeches. In: Acoustics, Speech and Signal Processing, ICASSP 2006 (2006)
2. Fügen, C., Waibel, A., Kolss, M.: Simultaneous translation of lectures and speeches. Machine Translation 21(4), 209–252 (2007)

3. 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., Juan, A.: TransLectures. In: Online Proceedings of Advances in Speech and Language Technologies for Iberian Languages, IBERSPEECH 2012, Madrid, Spain (2012)
4. Paul, M., Federico, M., Stüker, S.: Overview of the IWSLT 2010 evaluation campaign. In: Proceedings of the International Workshop on Spoken Language Translation, IWSLT 2010, Paris, France (2010)
5. Guzman, F., Sajjad, H., Abdelali, A., Vogel, S.: The AMARA Corpus: Building Resources for Translating the Web's Educational Content. In: Proceedings of the International Workshop on Spoken Language Translation, IWSLT 2013, Heidelberg, Germany (2013)
6. Abdelali, A., Guzman, F., Sajjad, H., Vogel, S.: The AMARA Corpus: Building parallel language resources for the educational domain. To appear in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland (2014)