QAT^2 – The QCRI Advanced Transcription and Translation System

Ahmed Abdelali, Ahmed Ali, Francisco Guzmán,
*Felix Stahlberg, Stephan Vogel, Yifan Zhang**

Qatar Computing Research Institute, HBKU, Doha, Qatar
{aabdelali,amali,fguzman,fstahlberg,svogel,yzhang}@qf.org.qa

Abstract

QAT^2 is a multimedia content translation web service developed by QCRI to help content provider to reach audiences and viewers speaking different languages. It is built with establishing open source technologies such as KALDI, Moses and MaryTTS, to provide a complete translation experience for web users. It translates text content in its original format, and produce translated videos with speech-to-speech translation. The result is a complete native language experience for end users on foreign language websites. The system currently supports translation from Arabic to English.
Index Terms: speech recognition, speech-to-speech translation, machine translation, text-to-speech synthesis

1. Introduction

Every day, new online multimedia content is generated across the globe. Unfortunately, language barriers pose a limitation for the dissemination of such content. In recent years, language technologies have made important progress, and have become viable tools that enable users to consume this information. Online translators (e.g. Google Translate, Bing Translate) are good examples of how translation technologies can help reduce the language gap. However, most of the current technology is only focused on text, and it is not prepared to deal with the growing trend of multimedia content. For instance, many news articles and blog posts come accompanied by complementary videos. QCRI has developed the QAT^2 system to deal with the translation of web-based multimedia content from Arabic to English. More specifically, our system not only translates the textual information in a web-page, but it also translates and performs automatic voice dubbing of the embedded video content. The system is built on open source technologies to enable faster advancement by exploiting larger community contributions.

2. System description

The QAT^2 system works as a web service to translate multimedia Arabic content into English[1]. Our system uses the KALDI [1] toolkit for speech recognition, Moses [2] for machine translation and MaryTTS [3] for text to speech.

The system works as following: When a user submits a request for a webpage, the web-page will be downloaded together with its multimedia content. The text content is extracted without HTML markups and translated by our Arabic-to-English MT system. The translated text is then injected back into its original format preserving the style and arrangement of the original page. Video contents are identified and downloaded. Speaker diarization, gender detection, as well as speaker linking are performed on the extracted audio from the original video content. We use speaker boundaries to segment long speech utterances so that the recognition can run in parallel on our multicore servers for better turnaround time. Speaker linking and gender information will help TTS to choose one correct voice for each speaker in original speech so it will appear natural for end users. Speech recognition produces both the word-level and phoneme-level output with timestamps. The word level output is split into dialog-act segments based on pauses, speaker information and phrase-chunk boundaries. These segments are then pre-processed and fed into the MT system. The TTS will work with all information provided by previous steps; speaker information as well as word and phone level transcription to generate small audio clips. These audio clips get voiced over the original video. At the beginning of every speaker turn, the system lowers the original sound and introduces the TTS voice gradually until it reaches the volume level of the original voice.The system can also generate subtitles in both the source and target languages with coloring for each individual speaker.

2.1. Speech recognition

The ASR component in QAT^2 is a grapheme system developed to process Al Jazeera daily news feed on aljazeera.net. The system is built using our KALDI GALE Arabic recipe[4] with 270 hours audio data from GALE and transcribed Al Jazeera news programs. The system uses typical KALDI SGMM model with fMLLR-based speaker adaptation for first pass recognition and a sequence-discriminative trained Deep Neural Network for second pass recognition. Our language model has been trained with aljazeera.net news articles from the past 5 years together with the Arabic Gigaword corpus. The lexicon has more than 1.2M words with an Out of Vocabulary (OOV) rate less than 2.5% on our own test data. The evaluation of our system on broadcast news reports has a word error rate of 16% and the evaluation on broadcast conversational data has a word error rate of 27.25%, which result a combined WER of 24.1%. Detailed explanation of our system can be found in [5, 4, 6].

2.2. Machine translation

The translation was carried using a MSA to English phrase-based SMT system based on Moses [2], trained on the news portion of the NIST2012 data, as described in [7]. We used the MADA ATB segmentation for Arabic [8] and truecasing for English, a 5-gram language model, trained on GigaWord v.5 as described in [9]. For tuning, we used PRO-fix [7, 10] on the MT06 set. Before decoding, the Arabic transcription is subject to a *two-pass* dialog-act segmentation. First, we produce an initial segmentation based on speaker diarization, time span and number of words per segment. This is important to ensure that the

* Authors in alphabetical order

[1] Future support for new language pairs is planned.

subtitles generated fit in the screen comfortably. Then, we use the Arabic chunker from AMIRA [11] to recalculate the segment boundaries in order to match syntactic phrase-boundaries. Afterwards each dialog-act segment is pre-processed and translated independently, and a n-best list of 100 candidate translations is generated for each segment. The list will be used to select the best candidate translation that fits the time span of the original segmentation.

2.3. Speech synthesis

The speech synthesis is based on the open-source platform MaryTTS. The TTS module gathers information emerging in different steps of the previous pipeline. The original Arabic audio track is used to mix the synthesized voices with the background sounds of the video to simulate a natural voice-over effect. The voice is selected based on the gender and speaker identification in the diarization step. We assign female voices to female speakers and male voices to male speakers, and reuse the same voice for all speech acts of the same speaker. ASR does not only provide the textual input for MT but also defines fine-granular timing constraints for TTS: if the temporal mismatch between the original Arabic speech segment and the corresponding synthesized English utterance is too large, the audio and video do not match up well any more. Finally, the MT system determines what needs to be said in English.

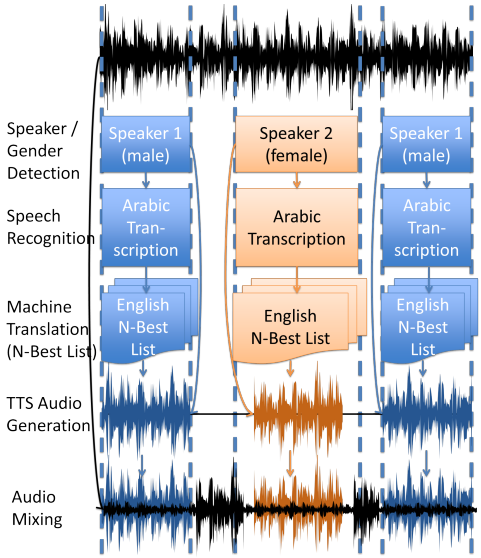

Figure 1: *Final audio generation.*

Fig. 1 illustrates how the English audio in the final video is generated. As discussed above, the diarization, ASR, and MT modules are sequentially applied first. For a tighter integration of MT and TTS, the MT module passes through n-best lists instead of the single best translation. We estimate the duration of each hypothesis using the `REALISED_DURATIONS` output type in MaryTTS and select the hypothesis which fits best in the specific time slot. If all hypotheses are too long, we speed up the speaking rate accordingly. In a final step, we mix the original audio with the synthesized voice-over. For better understanding, we reduce the volume of the original track to 10% during the time the TTS voice speaks.

3. Conclusion

This paper presents our translation system for online multimedia content to lower the language barriers and improve the reach of news content. Currently, the system provides a near native end-to-end experience to users. Our aim is to further expand this effort by improving the system in two directions: a) deploying an Automatic Dialect Identification (ADI) to distinguish between Modern Standard Arabic (MSA) and Dialectal Arabic (DA) to improve both ASR and MT performance; and b) adding more languages to our system. The demo has been presented in the BBCnewsHack and won the "Best in Show" award.

4. References

[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[2] 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 *Proc. of ACL'07*, 2007, pp. 177–180.

[3] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[4] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass, "A complete Kaldi recipe for building Arabic speech recognition systems," in *SLT*, 2014.

[5] P. Cardinal, A. Ali, N. Dehak, Y. Zhang, T. Al Hanai, Y. Zhang, J. Glass, and S. Vogel, "Recent advances in ASR applied to an Arabic transcription system for Al-Jazeera," in *Interspeech*, 2014.

[6] A. Ali, H. Mubarak, and S. Vogel, "Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr," in *Proc. of IWSLT'14*, 2014.

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

[8] R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin, "Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking," in *Proc. of ACL-HLT 2008*, Columbus, OH, USA, June 2008, pp. 117–120.

[9] H. Sajjad, F. Guzmán, P. Nakov, A. Abdelali, K. Murray, F. Al Obaidli, and S. Vogel, "QCRI at IWSLT 2013: Experiments in arabic-english and english-arabic spoken language translation," in *Proc. of IWSLT'13*, vol. 13, Heidelberg, Germany, December 2013.

[10] P. Nakov, F. Guzmán, and S. Vogel, "A tale about PRO and monsters," in *Proc. of ACL'13*, Sofia, Bulgaria, August 2013, pp. 12–17.

[11] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic tagging of arabic text: From raw text to base phrase chunks," in *Proc. of HLT-NAACL 2004: Short Papers*, ser. HLT-NAACL-Short '04, 2004, pp. 149–152.