Automatic Speech Recognition and Hybrid Machine Translation for High-Quality Closed-Captioning and Subtitling for Video Broadcast

Hassan Sawaf
Science Applications International Corporation (SAIC)
7990 Science Applications Ct.
Vienna, VA, USA
hassan.sawaf@saic.com

Abstract

We describe a system to rapidly generate high-quality closed captions and subtitles for live broadcasted TV shows, using automated components, namely Automatic Speech Recognition and Machine Translation. The human stays in the loop for quality assurance and optional post-editing. We also describe how the system feeds the human edits and corrections back into the different components for improvement of these components and with that of the overall system. We finally describe the operation of this system in a real life environment within a broadcast network, where we implemented the system to transcribe, process broadcast transmissions and generate high-quality closed captions in Arabic and translate these into English subtitles in short time.

1 Introduction

Automated closed captioning and subtitling for user-generated videos is not a new topic – for example Google offers this feature since November 2009 for videos uploaded to YouTube for select languages (Google (2009)). Many services by third-party providers also offer human post-editing to improve the quality of automated closed-captioning services, to improve the output quality of transcript for the closed caption text and subsequently also the translation for the subtitle text, for select languages, respectively.

The challenge is that given that the utilized automatic speech recognition (ASR) and machine translation (MT) technology are not tightly coupled, as well as that the offered solutions usually are not weaving the user feedback into the actual workflow and the respective core components, the results stays either below a certain an expected threshold of quality and additionally the service is very expensive to the end-user, as the work process complexity of the human in the loop is high.

This usually prevents TV broadcast networks to use these solutions for the content they want to provide to the consumer, and they have to pick-and-choose which content they will provide in multiple languages to their respective consumers, and that even often many hours or days post-factum.

We will describe a media processing, closed-captioning and subtitling system that serves as a tool to overcome the above-mentioned challenges in section 2, and we will describe the workflow and the implementation of the described system in real-life environments in section 3. We will outline then the achieved results, the experience gathered, concluding in section 4.

2 Baseline Subtitling System

The underlying system we describe here (SAIC's product "*Omnifluent Media*") consists of the following main components:

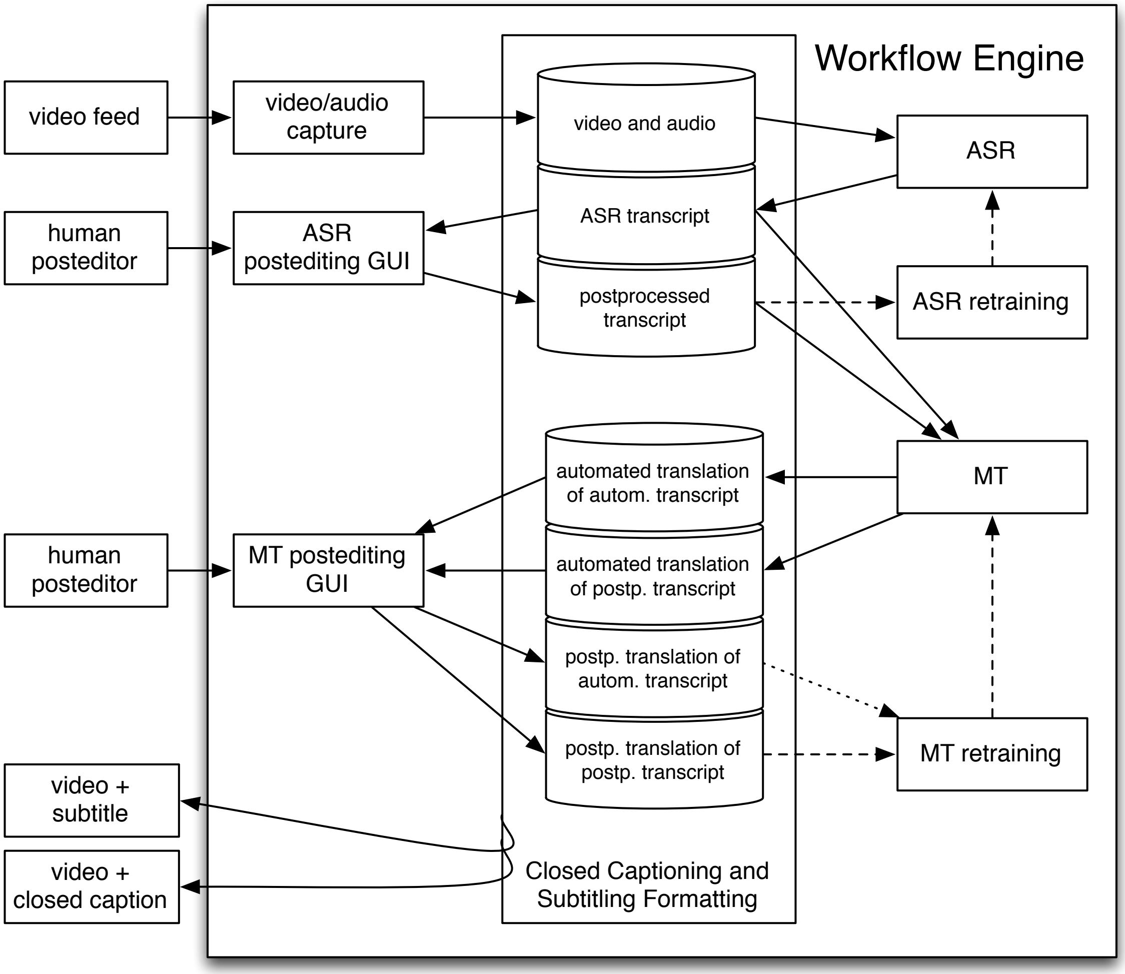

Figure 1. Simplified diagram of main components of the Omnifluent Media Subtitling System.

- **Video and Audio Captioning**. This component captions the audio stream from the live audio broadcast channel and streams it to the following components in the appropriate formats.
- **Segmentation of Audio**. To improve speech recognition and machine translation quality, the audio has to be pre-segmented. The segmentation has to take into account speaker change, channel change (different environments like studio vs. interview on the street), language change, prosodic clues (intonation on words and sentence), noise, music and jingle detection, etc.
- **Speech Recognition**. The actual text will be extracted from the audio stream using a high-accuracy large-vocabulary speech recognition subsystem. This system has special post-processing features that improve the overall readability, e.g. capitalization, punctuation and optional speaker identification and tagging.
- **Speech Recognition Post-editing**. To further improve the end-result of the transcription process, the system allows post-editing, so that a human can choose to listen to utterances captured from the broadcast source and compare it to the transcription from the automated process.
- **Machine Translation**. The machine translation subsystem converts the (optionally post-edited) transcript into the targeted language. It uses not only use the sequence of words from the transcription, but also use other meta-information like prosodic features and hesitations, etc.

- **Machine Translation Post-editing**. Also the machine translation result can be optionally post-edited to ensure that the results are high enough for publication. To do so, the post-editor can chose to read the original audio or the (raw or post-edited) transcript.
- **Closed-Caption and Subtitle Formatting.** Closed captions and subtitles can be encoded in different standards. Depending on the need of the user and consumer, these change, and might encode some additional information (e.g. certain "non-speech" events need to be exposed, or speaker identification tags need to be displayed).
- **Workflow Management.** This component allows the optimization of the overall broadcast process according to time constraints, resource availability, source data quality, and also the estimated quality of the results of the automatic components like ASR and MT.

Following, we will describe the two main components ASR and MT in more detail.

2.1 Automatic Speech Recognition (ASR)

The speech recognition system we use in the described setting is SAIC's product "*Omnifluent ASR*", which can be implemented SaaS-based or premised based, or in a hybrid way combining the two methods. For Arabic, the speech recognition engine is trained on more than 2,000 hours of manually transcribed data, in addition to more than 100,000 hours of automatically transcribed data that was used for unsupervised training. The system is a dialect adaptive system, i.e. there are sets of sub-models in the ASR system, that are dialect specific. Also in terms of channel and speaker and speaker-group, there are sub-models that focus on these.

The ASR system is capable of online learning from data corrected by the post-editors that then can be augmented to the background data. This can be done on the different levels:

1- For the *acoustic model* different adaptation techniques can keep the model "up-to-date" to the form and audio characteristics. The system is able to continuously adapt to new accents and dialects, the more the system sees from that type of input.

2- The *pronunciation dictionary* can be augmented with new words automatically, as they occur in the post-edited transcript. Additionally, the system can learn to adjust pronunciation variants according to the correction in the post-editing of the ASR output.

3- The *language model* can be adapted using the corrected and uncorrected speech recognition. The adaptation happens continuously, by using sub-sampling techniques against a big background corpus, by using category information ("politics" vs. "sports", etc.) and other adaptation techniques, to model style, dialect and domain.

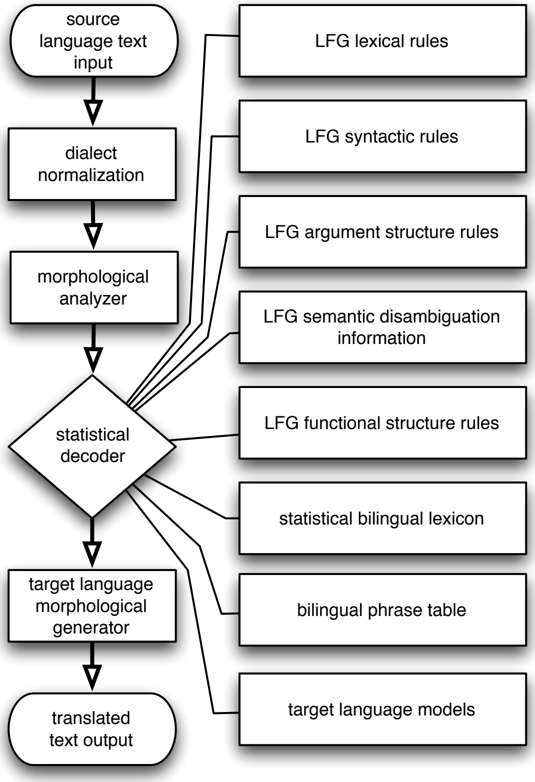
source language text input
dialect normalization
morphological analyzer
statistical decoder
target language morphological generator
translated text output
LFG lexical rules
LFG syntactic rules
LFG argument structure rules
LFG semantic disambiguation information
LFG functional structure rules
statistical bilingual lexicon
bilingual phrase table
target language models

Figure 2. Flow diagram of the machine translation process.

2.2 Hybrid Machine Translation (HMT)

For Machine Translation, we use SAIC's product "*Omnifluent HMT*" which is a state-of-the-art hybrid machine translation system. The core of the system is a statistical search that employs a combination of multiple probabilistic translation models, including phrase-based and word-based lexicons, as well as reordering models and target n-gram language models and rule-based parsers.

Our prime motivation for utilizing a hybrid machine translation system in this application is to take advantage of the possibility to cover certain systematic phenomena that can be described with an abstract rule, rather than trying to collect all possible samples for training a statistical approach, especially if the rule to cover deals with classes of words, or special word dependencies that can be distributed over the sentence, while not being continuous.

In our approach to HMT, the statistical search process has access to the information database available in the rule-based engine, as outlined in Figure 2. The components in the figure are described in more detail in Sawaf (2010) and Matusov (2012).

In rough strokes, Statistical Machine Translation is traditionally represented in the literature as choosing the target (English) sentence $\boldsymbol{e} = e_1 \ldots e_I$ with the highest probability given a source (French) sentence $\boldsymbol{f} = f_1 \ldots f_J$:

$$\hat{\boldsymbol{e}} = argmax_{\boldsymbol{e}} \{Pr(\boldsymbol{e}|\boldsymbol{f})\} . \qquad (1)$$

The rich syntactic/semantic information is derived from the rule-based engine parser that produces syntactic trees annotated with rich semantic and syntactic annotations.

The hybridization is then accomplished by treating all the pieces of information as feature functions in a log-linear framework:

$$Pr(\boldsymbol{e}|\boldsymbol{f}) = p_{l1..M}(\boldsymbol{e}|\boldsymbol{f}) = \frac{exp[\sum_{m=1..M} l_m \, h_m(\boldsymbol{e},\boldsymbol{f})]}{\sum_{\boldsymbol{e}'} exp[\sum_{m=1..M} l_m \, h_m(\boldsymbol{e}',\boldsymbol{f})]} ; \qquad (2)$$

we obtain the following decision rule out of (1):

$$\hat{\boldsymbol{e}} = argmax_{\boldsymbol{e}}\{Pr(\boldsymbol{e}|\boldsymbol{f})\} = argmax_{\boldsymbol{e}}\{\sum_{m=1..M} l_m \, h_m(\boldsymbol{e},\boldsymbol{f})\} . \qquad (3)$$

Incorporation of these different knowledge sources (rule-based and statistical) is then achieved by adding feature functions to the criterion, and allowing a training algorithm to train the weights of the feature in context to the other features in respect to the final translation quality measured by an error criterion, e.g. as described in Och and Ney (2002).

Thus, while the system can learn from example sentences and therefor corrections from the human post-editor, while we can learn from more abstract knowledge coded in hand-crafted rules by human linguists.

In addition to this learning process, the different models (phrase tables, language models) are optimized using various adaptation techniques, e.g. sub-sampling, multiple parallel processed domain-specialized phrase tables and language models, and category and class-based models. Similar to speech recognition for new dialects, the system is capable to learn new genres and dialects on the go, by learning new phrase tables with the post-processed translations, in combination with the high-confidence translations, which are generated automatically.

3 Workflow and Implementation

Figure 3 is a simplified outline of the workflow that is implemented to increase the processing speed from capturing the audio to generation of the closed captions and subtitles. As the system progresses and improves quality, the post-editing for both ASR and MT can be focused on utterances and sentences that have a lower confidence value.

In the implementation in real live environment in a broadcasting environment, the content is not all novel, i.e. some content is known prior to the actual broadcast, e.g. in scripts that the news anchor reads from the teleprompter. This content has to be processed by the workflow "out-of-order", i.e. independently from the live broadcast.

It showed also to be very beneficial for the quality of the ASR and HMT engines to have direct access and integration to information retrieval systems that have content of similar type and genre like the data that has to be processed from the broadcast input. This access allows online adaptation of language models for ASR and MT, and of the pronunciation dictionary for ASR.

This fact simplifies the work for the post-editors insofar, that they can process the content with less time pressure, but adds more overhead to the workflow management, as the post-edited data "has to wait" for the actual broadcast. It also adds complexity by the fact that the uttered speech might not be exactly what the anchor sees and reads on the teleprompter.

4 Results

Human (expedient and professional) transcribers usually have a speed of transcription of 60-90 words per minute for Arabic, i.e. one hour of broadcast needs between 2 and 3 transcriber plus quality control (editor) to achieve reasonable results that can be used for publication. In addition to that, the translation speed of a human translator is usually around 400-500 words per hour for expedient translators from Arabic to English, which means around 12 human translators (plus editors, usually each 5-8 translators one editor) to process one hour of broadcast data in realtime. The challenge to have this converted in very short order to be readily available in the closed captioning and/or subtitling format expected by the end-user is not even addressed at this point.

With the described system, we are able to process data from capturing the broadcast from either the satellite dish or the actual content source via video-over-IP to the delivery of correctly formatted closed captions and subtitling information in less than one hour, usually even far below 30 minutes. At the same time, the throughput with the same team of human transcribers and translators increases by a factor of 2 (in the initial phase) to 4 (as soon as the team reaches proficiency in using the workflow and post-editing portal).

In future, we will add more HLT processing, ASR and MT adaptation features to the described system to increase the efficiency even more. Also we will make use of more (statistical and linguistic) knowledge sources to improve MT, and lastly, we target an even tighter integration of ASR and MT.

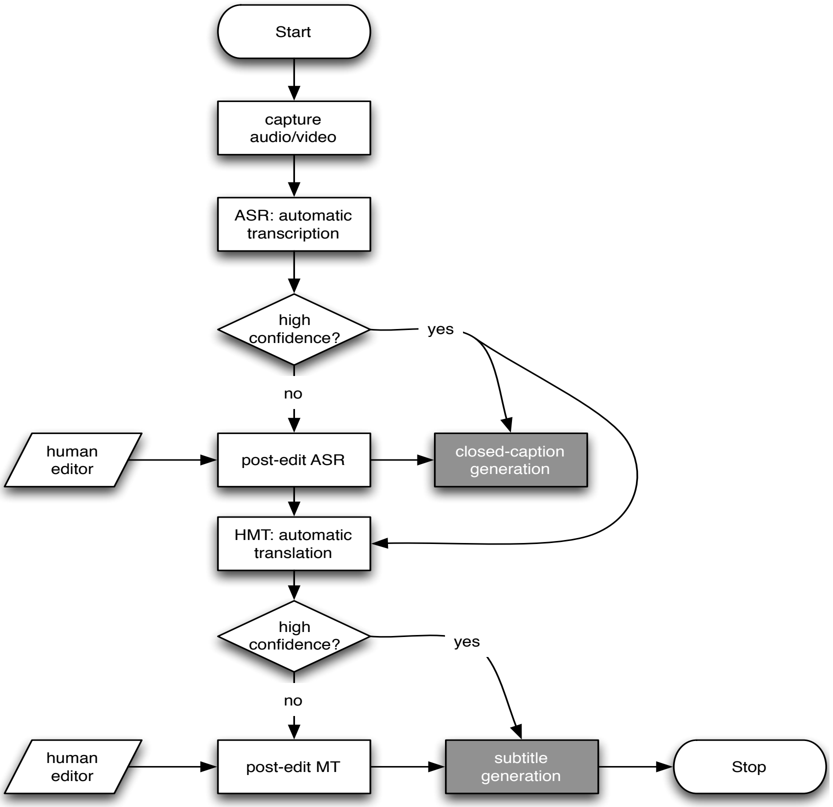

Figure 3. Simplified diagram of the workflow for closed-captioning and subtitling using the Omnifluent Media Subtitling System.

References

Evgeny Matusov, 2012. Incremental re-training of a hybrid English-French MT system with Customer Translation Memory data. To appear in *AMTA 2012: The Tenth Conference of the Association for Machine Translation in the Americas,* San Diego, CA.

Franz Josef Och, and Hermann Ney, 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the Ninth Machine Translation Summit*, pp. 295–302. New Orleans, LA.

Google, 2009. Automatic captions in YouTube . Google Official Blog. *http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html*.

Hassan Sawaf, 2010. Arabic Dialect Handling in Hybrid Machine Translation. In *AMTA 2010: The Tenth Conference of the Association for Machine Translation in the Americas,* Denver, CO.