Automated Production of True-cased Punctuated Subtitles for Weather and News Broadcasts

Joris Driesen¹, Alexandra Birch¹, Simon Grimsey², Saeid Safarfashandi², Juliet Gauthier², Matt Simpson², Steve Renals¹

¹Centre for Speech Technology Research, University of Edinburgh, UK
²Red Bee Media, UK

{jdriesen, abmayne, srenals}@inf.ed.ac.uk,

Abstract

Providing subtitling for multimedia content is a highly costly process. Any system aimed at automating at least part of this process may therefore yield significant economic benefits for content providers. In this paper, we present an integrated automatic system capable of automatically subtitling weather forecasts and news broadcasts. In this system, a number of different modules are strung together, each performing a single processing step in the pipeline. An ASR (Automatic Speech Recognition) module first converts raw audio into an uninterrupted stream of written words. A decision tree classifier then marks sentence boundaries in the resulting word sequence. Finally, a SMT (Statistical Machine Translation) module ‘translates’ the resulting sentences into punctuated true-cased text. The system has been developed in close cooperation with Red Bee Media and will be deployed in their commercial production pipeline.

Index Terms: subtitles, multimedia, automatic speech recognition, machine translation

1. Introduction

Multimedia content providers endeavour to subtitle their programmes, with the goal of making them accessible for viewers who suffer from a hearing impairment. Subtitles are usually made by human subtitlers by means of re-speaking. This involves listening to a televised broadcast and repeating the spoken content in a noise-free environment to an ASR system that is highly tuned to the speech characteristics of the subtitler. The few transcription errors made in this process are fixed manually. An efficient and accurate method like this is necessary, since multimedia companies are often legally obliged to provide a very high coverage under very stringent quality demands. For example Red Bee Media, a UK-based content provider formerly part of BBC³, adheres to the Client Service Level Agreement, which demands readable punctuated subtitles for the majority of their produced content, with a Word Error Rate (WER) no higher than 2% for live subtitles (i.e. made on-the-fly) and 9% for prepared subtitles. Even though the process of subtitling has been optimized and streamlined, it is still a highly labour-intensive task which carries with it significant economic costs.

Furthermore, failure to meet the legal demands of coverage and accuracy is penalized monetarily, which may drive costs even higher. Because of these high costs and the sheer volume of produced media content, there are substantial commercial benefits to even modest improvements to the current subtitling process.

In this paper, we present a modular end-to-end system which takes a video as input and produces readable subtitles in a fully automated way. A schematic overview of this system is shown in figure 1. The video content that serves as input to this system comes from two different domains: ‘Weatherview’ and ‘Sky News’. Weatherview is the daily weather forecast for the UK on BBC. From an ASR perspective, this data is the easier of the two to handle. It consists of 3-minute segments, uttered by a single speaker in isolation, employing a restricted vocabulary. For more details, see [1]. The other input type consists of broadcast news, taken from the news channel ‘Sky News’. This data poses a bigger challenge, as it contains a multitude of different speakers, recording and noise conditions, a much larger vocabulary which tends to change over time as words and expressions rise and fade in popularity, etc. All inputs in this paper consists of data that was aired recently. As such, the proposed system is designed to cope with inputs of real-world scale and complexity.

2. Automatic Speech Recognition

In order to convert the input audio to text, we trained up two ASR systems, one for each of the input types of interest. For both ASR systems, 13 MFCC’s along with their first and second order differences were calculated within 25 ms frames which were shifted in steps of 10 ms. The resulting 39-dimensional spectral representations were then modified with Cepstral Mean

³British Broadcasting Corporation, the UK’s national radio and television corporation

Figure 1: A schematic overview of the proposed subtitling process.
Normalization (CMN). For Sky News data, where the speaker and recording conditions at any given time are unknown, we divided the input into segments of 40 seconds and assumed each of these segments contained a single speaker. Since most speakers in the data talk for longer than 40 seconds, this is a reasonable assumption to make for CMN. In a next step, the data is linearly transformed using MLLT and LDA [2].

4. Statistical Machine Translation for True-casing and Punctuation

SMT systems are designed to convert text from a source language into a target language. They achieve this by analyzing large parallel text corpora spanning both languages, and constructing a statistical mapping between the two. By training a SMT system on a text corpus, paired with a lower-cased depunctuated version of itself, the system can effectively learn to “translate” lower-cased word sequences into true-cased text with punctuation. The SMT system used in this paper is Moses [5]. This system regards text input as a sequence of “phrases”, i.e. word blocks of varying length that often occur together as a unit. Since the order of these phrases is not necessarily fixed between languages, the search space for a translation increases drastically with the length of the input, as all reorderings must be considered. This is the main reason for splitting up the ASR output, as discussed in section 3. The reason for splitting at sentence boundaries is that partial sentences tend to introduce ambiguities. For example, in a sentence like ‘‘I took a sip the coffee was hot’’, the erroneous phrase ‘‘sip the coffee’’ will be more likely to be selected if the first and last few words of the sentence are omitted. It is clear that when this happens, the system will fail to insert the full-stop which separates the two sentences.

5. Integration in Red Bee Media Workflow

As was stated above, multimedia content providers are under constant pressure to reduce the cost of subtitling, and increase both its efficiency and accuracy. The proposed system was designed from the ground up to be integrated smoothly into the production pipeline at RBM. The subtitles it produces are easily transcoded into the required formats for delivery, such as internet video, live playout, etc. Although the proposed system is very unlikely ever to replace human transcribers, it may simplify and speed up their work. Moreover, it may provide a valuable back-up system, ready to take over when the need arises. Lastly, by training up the SMT module with different language pairs, the system allows subtitling in different languages at almost negligible additional costs.

6. References


