A system for automatic broadcast news summarisation, geolocation and translation

Peter Bell, Catherine Lai, Clare Llewellyn, Alexandra Birch, Mark Sinclair

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
{peter.bell, c.lai, a.birch, mark.sinclair}@ed.ac.uk, C.A.Llewellyn@sms.ed.ac.uk

Abstract

An increasing amount of news content is produced in audio-video form every day. To effectively analyse and monitoring this multilingual data stream, we require methods to extract and present audio content in accessible ways. In this paper, we describe an end-to-end system for processing and browsing audio news data. This fully automated system brings together our recent research on audio scene analysis, speech recognition, summarisation, named entity detection, geolocation, and machine translation. The graphical interface allows users to visualise the distribution of news content by entity names and story location. Browsing of news events is facilitated through extractive summaries and the ability to view transcripts in multiple languages.
Index Terms: multimedia archives, ASR, summarisation, named entity detection, geolocation, machine translation.

1. Introduction

The global media industry produces many thousands of hours of audio and video news content on a daily basis. A challenge for the industry is the need to balance the desire of news consumers for relevant localised content, against the objective of selecting global news items of importance to people located far from the source of the story. This task is highly labour-intensive. The former requires a high volume of news content to be collected and precisely targeted, a particular challenge where the consumers speak a minority language or dialect. The latter demands expensive multilingual media monitoring operations, which all but the largest media organisations struggle to afford.

This "Show & Tell" proposal presents a proof-of-concept automatic system for analysing news content, targeting it to potentially interested audiences on a geographic basis, and making it available in appropriate languages. Our hope is that a fully-developed version of this system could help news organisations operate more efficiently on a global scale. The system integrates our recent research outputs from a range of speech and natural language processing disciplines: audio scene analysis, automatic speech recognition, extractive summarisation, entity detection and spoken language translation. We describe these elements of the processing pipeline in more detail in the following sections.

In its current implementation, the system processes incoming broadcast media in an offline manner. Transcription, summarisation, entity detection and translation are performed for each story and imported into a web-based interactive user interface shown in Figure 1. We do not currently realign the translated content with the original media.

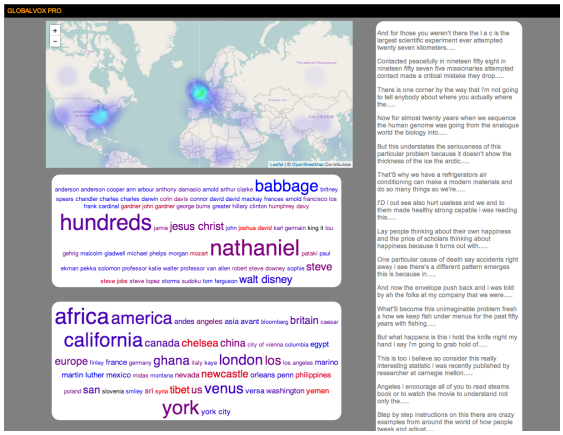

Figure 1: User interface

2. Speech recognition

Incoming video files are processed to extract the audio stream in wave format using FFmpeg[1]. After audio scene analysis using an unsupervised clustering technique, automatic speech recognition (ASR) is applied per item in an offline manner, allowing us to perform adaptation to each news story in a 2-pass configuration. The system uses sequence-trained deep neural networks in a hybrid configuration, following [1]. The models are trained on around 600 hours of multi-genre broadcast data from the British Broadcasting Corporation (BBC) taken from the training set defined for the 2015 MGB Challenge[2]. For ease of demonstration, the original news stories are all in English; we later translate from English to a range of target languages. In a final deployed system we would expect to include multiple source language options. The BBC news videos we use for the demonstration are pre-segmented by hand into discrete stories as part of the transmission process, so in this case we do not need to perform automatic topic segmentation, although clearly this could be required in a future version.

3. Summarisation

We use extractive summarisation techniques to select representative quotes from news stories. In addition to lexical features based on ASR output, the summariser uses prosodic features to rank utterances, estimating the probability of their appearing in

This work was supported by the European Union under the FP7 projects inEvent (grant agreement 287872) and EU-Bridge (grant agreement 287658), and by EPSRC Programme Grant grant EP/I031022/1, *Natural Speech Technology*.

[1] http://www.ffmpeg.org
[2] http://www.mgb-challenge.org

an extractive summary via logistic regression. Previous work has suggested that the use of non-lexical features, such as word and utterance level prosody, can help ameliorate problems with ASR. The models were trained on manually annotated AMI meeting data. We found that using prosodically augmented lexical features provided the best performance on held out meeting data [2]. Even though the summariser was designed for multiparty dialogue, further experiments have shown it extends well to other spoken genres. In a pairwise preference test, we found that quotes ranked higher by the summariser were also selected as more representative by human subjects a significant majority of the time.

4. Named entity detection and geolocation

Our system employs methods to allow for searchable aggregation of summarised speech. The summariser provides the top 10 ranked utterances for further processing. However, text extracted from the spoken news reports lacks punctuation and capitalisation. To allow the use of richer punctuation and capitalisation features, we add a machine translation based punctuation module to the pipeline, described below. In addition, common names and places are capitalised after lookup in various people/place lexicons.

To identify and visualise items of interest, we use the Edinburgh Informatics information extraction tools [3, 4]. These are well established tools that process text to identify entity names and provide geographic coordinates for locations. The named entity recognition tool identifies word sequences as people or place name entities via a rule-based method that takes into account information about part-of-speech, capitalisation, local context and lexicon look-up. Places are then georesolved – the names are looked up in a geographic gazetteer and possible interpretations are returned. These are ranked in order to assign a specific latitude and longitude value. Entities are represented in the interface using wordles with the size of each entity reflecting the frequency; places are represented as a density map (Open Street Map and Leaflet are used[3]).

In order to gauge the general opinion towards each entity, sentiment analysis was performed on the sentences containing those entities using the rule based sentiment analyser Vader [5]. This tool is adapted for use with social media data and is therefore ideal for use with speech segments which can be presented using less formal language. This gave positive negative and neutral scores for each entity which were then represented using colour. Once processed the data was stored in a Mongo noSQL database[4]. The flexible schema of the database allows us to assign the entities counts and sentiment scores to each entity, enabling the recalculation of scores for various document sets.

5. Spoken language translation

Statistical machine translation of transcripts of BBC News reports requires special handling. The ASR output contains errors both in the words recognised and in sentence segmentations. It also lacks punctuation and capitalisation. We therefore used phrase-based machine translation which is more robust than structured syntax-based translation models. We trained two translation models: one which translates from unpunctuated ASR English output to punctuated English output, and one which translates from standard English text to the target language text. Casing was handled by a re-caser, which applies case to words according to their most common case in the training corpus. The training corpora used were Europarl [6], News Commentary, TED [7], and Commoncrawl [8]. We used the Moses SMT toolkit [9] with standard settings, including the use of 5-gram language models.

[3]http://www.leafletjs.com

[4]http://mongodb.org

6. Conclusion and Future Work

Feedback from initial demonstrations of our system to journalists was extremely positive, indicating that this system would be valuable in analysing large volumes of multilingual news content. While our system demonstrates the potential of existing speech and language technologies, it also highlights areas that need attention when building speech based end-to-end systems. For example, we found that segmentation of the speech stream can have a substantial effect on the usability and readability of transcribed speech through the pipeline. Improving segmentation can also improve the quality of extracted summaries and automatic translation. Thus, optimising initial audio segmentation is vital for overall system robustness. Downstream language processing trained on written texts often assume more information than is available from raw ASR output, e.g. punctuation. The problem is exacerbated by the frequency of non-sentential utterances in speech. Besides improving the links between our current modules, we intend to extend our system by including more higher level analysis such as topic and dialogue act detection. We also hope to make more use of audio event detection techniques for determining structure in longer broadcasts, for example detecting topic change indicators such as music.

7. Acknowledgements

This work was developed as part of a BBC News Labs *newsHACK* event. We are grateful to the BBC for their support.

8. References

[1] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc Interspeech*, 2013.

[2] C. Lai and S. Renals, "Incorporating lexical and prosodic information at different levels for meeting summarization," in *Proc. Interspeech 2014*, 2014.

[3] C. Grover, S. Givon, R. Tobin, and J. Ball, "Named entity recognition for digitised historical texts," in *Proc. LREC*, 2008.

[4] C. Grover and R. Tobin, "Rule-based chunking and reusability," in *Proc. LREC*, 2006.

[5] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. ICWSM*, 2014.

[6] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT Summit*, vol. 5, 2005.

[7] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proc EAMT*, 2012.

[8] J. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez, "Dirt cheap web-scale parallel text from the common crawl," in *Proc ACL*, 2013.

[9] 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. ACL*, 2007, demo session.