Keyword Based Indexing of Multilingual News Videos

Sunil Kopparapu and Hiranmay Ghosh

Automated analysis of Indian language telecasts raises some unique challenges. Unlike most of the news channels in the western world, Indian television channels do not broadcast closed captioned text, which could be gainfully employed to index and analyze news broadcast in several languages. In this scenario, indexing of news video stream needs to rely completely on the strengths of audio and visual processing to extract keywords.

A list of *contemporary* words of interest are required to enable extraction of keywords in multilingual video broadcast. RSS feed(s) in English, made available and updated frequently by websites of the broadcasting channels can be exploited. Advantage of obtaining keywords from RSS feeds[1] is the currency of the keywords. The English keywords obtained from RSS feed form the list of keywords. These keywords need to be identified in multilingual news video streams.

Keywords in a news broadcast are typically proper nouns and common nouns. We can identify English keyword equivalents in other Indian languages by making use of a pronunciation lexicon for proper nouns and word dictionaries for common nouns. This can be used to build a dynamic multilingual keyword list which can be used for spotting keywords in audio (speech processing) and ticker text (visual processing).

The challenge is that both speech (for audio processing) and OCR technologies (used for ticker text analysis) for several Indian languages are not sufficiently mature; it is expected that the extraction of keywords from both audio and visual channels simultaneously, could significantly enhances the robustness of the indexing process.

The novelty of the approach is in constructing a multilingual keyword list from RSS feeds to enable keyword spotting simultaneously in both audio and visual streams for indexing news broadcast in different languages. A reasonable success in indexing multilingual news broadcast would have tremendous impact in terms of enabling video content search and retrieval in different and across languages.

Bits and pieces of what is proposed in available in terms of language word dictionaries, public domain speech engine which can be used for audio keyword spotting. However, this needs collaborative effort of researchers in several areas. Particularly, (a) Natural language processing (deriving keyword list from RSS feed; construction of a multilingual keyword list; extending keyword list in semantic sense), (b) speech processing (identifying keywords in audio; acoustic models for different languages; use of common dictionaries), (c) image or visual scene processing (ticker text extraction and Video OCR) and (d) storing keywords obtained from audio and visual processing for indexing video to enable search.

A major challenge would be to collect reasonable amount of data synchronously (RSS feed, news broadcast in different languages).

[1] Recently, RSS feeds in some Indian languages have become available