INFOLINK: ANALYSIS OF DUTCH BROADCAST NEWS AND CROSS-MEDIA BROWSING

Jeroen Morang, Roeland Ordelman, Franciska de Jong, Arjan van Hessen

Department of Electrical Engineering, Mathematics and Computer Science
University of Twente, The Netherlands

ABSTRACT

In this paper, a cross-media browsing demonstrator named InfoLink is described. InfoLink automatically links the content of Dutch broadcast news videos to related information sources in parallel collections containing text and/or video. Automatic segmentation, speech recognition and available meta-data are used to index and link items. The concept is visualised using SMIL-scripts for presenting the streaming broadcast news video and the information links.

1. INTRODUCTION

Recent years have shown large improvements in the performance of automatic speech recognition (ASR) systems and speech transcripts can now be generated at nearly the same quality as manual transcripts, at least for well-studied domains such as the broadcast news domain. Content-based searching in news archives on the basis of speech recognition transcripts can therefore successfully be applied, as demonstrated in scientific experiments [1, 2] as well as in commercial systems (see e.g., Virage's[1] *VS News Monitoring System*).

Searching audio or video material on the basis of automatically generated speech transcripts is often referred to as spoken document retrieval (SDR). SDR can be a convenient tool for identifying interesting parts in an unstructured audio or video document, without having to scroll manually through the document itself. Another advantage of SDR is that it enables cross-media retrieval, the functionality by which on the basis of a particular audio or video item, related content in other formats can be identified. It depends on the nature of available metadata and annotation what the added value of SDR can be.

Media collections can be enriched with metadata or related full-text content either via manual annotation, or automatically. Consider an archive with digital meeting recordings as an example. Documents of the first category are manually generated links to meetings agendas or policy documents pertaining to specific topics of a meeting. Such documents can be considered as part of the same archive. However, for a meeting archive links to external sources may also be useful, such as newspaper articles, an archive of broadcast news items and documentaries, and a 'Who-is-Who' of politicians. Given the potential size of such external sources, and the variety of perspectives that can be taken on them, automatically generated links could be very beneficial. (For work on cross-media disclosure for meeting recordings in the IST-project AMI, cf. [3].)

Cross-linking of both internal and external sources in various formats requires the generation and or conversion of annotations in compatible formats. Transcripts generated by ASR can contribute to the creation of textual representations for media files and thereby to SDR and the applicability of all kinds of text-based disclosure and browsing tools. ASR thus facilitates cross-media linking and enlarges the information potential of a collection substantially.

InfoLink is a cross-media browser for an archive of Dutch television broadcast news shows. The broadcast news archive is the internal content source. InfoLink links items from a broadcast news show to related items in a newspaper corpus and a historical video archive. A typical InfoLink case would be the automatic linking of a broadcast news item about the election of pope Benedict XVI to newspaper articles with background information or a historical video showing the inauguration of pope John Paul II in 1978. Links to background texts may add to the depth of information, and providing a historical video may have an educational and/or entertainment value. The scope of InfoLink could relatively easily be enlarged by including for example current affairs programs, electronic magazines, photo material or biographies of famous people in the collection set. Given a broadcast news collection with a large time-span, news topics from the recent past can be regarded as linkable sources themselves. Also outside the broadcast news domain, the concept of InfoLink can be usefully applied.

Broadcast archives, such as the video collection for which InfoLink is developed, can be seen as part of digital cultural heritage. InfoLink contributes to the disclosure of it and thereby to the aims of the Dutch research program CATCH (Continuous Access To Cultural Heritage; cf. `http://www.nwo.nl/catch/`.

InfoLink was developed in collaboration with the Institute for Sound & Vision (the broadcast archive for the public broadcast stations in the Netherlands) and deployed ASR tools available at the University of Twente. This paper describes the setup of the first InfoLink demonstrator. User experiments and performance figures for the retrieval and browse functionality are still to be collected. In section 2 some related work on cross-media search and ASR for the disclosure of cultural heritage will be described. In 3 the InfoLink setup will be described. The sections 4 and 5 specify the segmentation, indexing and linking part and in section 6 the presentation of the link concept is described. The performance of the demonstrator is addressed in section 7. First conclusions are presented in section 8.

2. RELATED WORK

For Dutch news content, a browser has been developed that facilitates the querying of a news archive with both video and newspaper content [4]. Outside the news domain, the concept of cross-media search has been studied in the context of the IST project MUMIS. A multilingual media archive with video recordings of

[1] `http://www.virage.com`

0-7803-9332-5/05/$20.00 ©2005 IEEE

soccer matches and ticker texts was disclosed on the basis of information extraction and an information merging component to support browsing in a consistent set of time-coded metadata, with redundancies filtered out and links to video fragments on a streaming server [5]. Also several spoken document retrieval initiatives can be mentioned. The IST project ECHO aimed at the disclosure of historical video archives partly on the basis of ASR for a number of EU languages. There is a growing interest in applying ASR to make oral history collections searchable at fragment level (Cf. [6] and [7]).

3. INFOLINK SETUP

In this section the content collections incorporated in InfoLink demonstrator and its components will be described.

3.1. Content collection

The digital video collection for which InfoLink generates links to external sources consists of six hours of high quality resolution video and audio taken from the broadcast news archive ("*NOS Acht Uur Journaal*"), year 2001. The available metadata (short content descriptions) was manually created. The level of granularity of the description was insufficient and additional metadata generation (ASR) was needed in order to apply InfoLink.

The external information sources selected for linking are

- an archive of teletext subtitle files that were broadcasted along with the broadcast news shows, captured using a teletext capturing board.
- a collection of Dutch newspapers ("*de Volkskrant*") from 2001, available via the Twente News Corpus [2].
- a historical video archive ("*Polygoon Journaal*"); only the metadata descriptions were used. (The digital video can be accessed via http://www.beeldengeluid.nl/.) Because of the complexity of applying ASR to historical video (e.g., low quality audio, outdated word usage) and the ongoing research in this area, no efforts were spent on the generation of speech transcripts for this collection.

All newspaper articles and historic video descriptions were normalised and stop-words were removed on the basis of a list of common stop-words.

3.2. Infolink components

Three steps can be distinguished in the InfoLink process:

1. Preprocessing of selected collections: segmentation of digital video, extraction of a textual content descriptions (teletext; automatically generated speech transcripts), and indexing.
2. Linking: feeding content descriptions of a segment as queries into an retrieval engine in order to identify relevant documents from several textual databases. Between the original segment and the identified relevant external texts, links are generated
3. Presentation of 'InfoLinked' digital video.

In the following sections the three steps will be discussed in more detail.

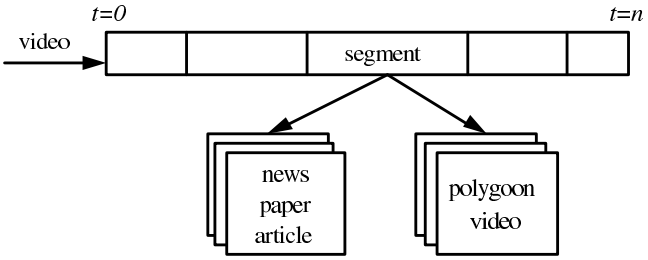

Fig. 1. A video-stream is segmented and every segment has several links to newspaper articles and historic videos.

4. PREPROCESSES: SEGMENTATION, ASR, INDEXING

Indexing is the process in which a descriptive (textual) representation of an object is created to make it available for retrieval. Though searching for similar images is feasible, automatic content analysis and indexing of video at a level that allows conceptual querying is known to be hard. Indexing of video can be done via manually generated content descriptions, low-level image features and speech transcripts generated via ASR. In InfoLink, time-coded indexing of the core video set was done mainly via ASR. A crucial step in ASR is segmentation. This section will describe each of the techniques applied.

4.1. ASR

Automatic speech recognition was applied to generate transcripts of the speech in the audio track of the news videos. For this purpose the UT-BN2002 broadcast news speech recognition system was deployed. The UT-BN2002 system is a hybrid RNN/HMM system with a $65K$ vocabulary and a statistical trigram language model trained on a newspaper corpus that has a word-error-rate of about 30% on broadcast news shows [2].

4.2. Segmentation

A crucial part of the InfoLink system is the segmentation part (Figure 1). When a video is not properly segmented and single segments cover multiple topics, the search for relevant documents will loose precision. Four segmentation approaches have been considered:

1. Static chunks

   As the simplest solution one could consider to segment the video into static chunks of for example 60 seconds. This would not necessarily be a disadvantage for InfoLink. Long items might be divided into sub-items with links to more precise information. In 2001, items about the war in Afghanistan sometimes could last the entire show but at the same time covered many aspects, such as terror in the United States, reactions in Europe and refugees at the Pakistan border.

2. Using teletext subtitles

   Teletext subtitles for the Dutch television broadcast news are created and segmented manually by a velotypist. Between topics the segmentation symbols '$* \ * \ *$' are typed to generate a signal to the hearing-impaired that the current topic ends and a new one begins. This is similar to the

'>>>' symbols on closed caption systems on American television [8]. There are several problems when using teletext subtitles for segmentation. Teletext for news tends to summarise or compact the discourse. For live broadcasts, there is in addition a time lagging problem: the subtitles are out of sync with the spoken words. For archived content this problem can be solved by 'forced-alignment' of teletext by using time-coded speech recognition [9]. When the subtitles diverge too much from the actual speech, forced-alignment will be difficult, as is the case with subtitles generated before 2002. Only since 2002 news shows have been subtitled more accurately (in the sense of: similar to the actual speech). Moreover, as the segmentation symbols are often missing and sometimes appear in the middle of an item, a segmentation strategy that is purely based on the teletext subtitles would not be sufficient.

3. Using silence

   Analysis revealed that the shows commence with highlights followed by the introduction of the news-anchor and the main topics. The broadcast ends with the weather forecasts. The items are divided by a substantial silence (deliberately put there to mark the topic change by the news-anchor), but as these pauses also appear in interviews with for example correspondents, plain silence detection will not suffice. Instead a combination of silence detection and topic detection was proposed.

   Silence detection is used to segment the video into candidate segments. Transcripts of the segments are generated with ASR. Topic detection is used to decide on the final segmentation. This result is used to segment the teletext subtitles.

4. Using topic information

   Topic detection calculates the similarity between de transcripts of neighbouring candidate segments. When the similarity exceeds a certain threshold the segments are joined together. Low similarity scores suggest a topic boundary. To calculate the similarity between segments the cosine measure is used in combination with $tf.idf$ for calculation of the weights. All transcriptions of segments are regarded as the complete document collection. This technique is similar to Heart's *TextTiling*[10]. The used threshold is empirically established. To avoid segmentation errors in interviews the minimum segment length is set to 30 seconds. Reynar gives an extensive overview of topic segmentation techniques [11].

4.3. Indexing

Speech recognition transcripts and forced-aligned teletext subtitles describe the discourse of a video-segment. On the basis of the transcripts of the spoken audio the contents of a segment can be inferred.

Teletext subtitles have the tendency to be brief and omit parts of speech while a speech recognition system can only recognise the words in the vocabulary. Therefore many new names and places cannot be recognised. Even when both speech recognition and teletext subtitles do not provide complete error-free transcriptions, adequate retrieval of documents is still possible as long as WERs are below 50% [12, 1]. Results from TRECVID show that for retrieval of video the application of ASR transcripts contribute considerably to retrieval performance figures [13].

Fig. 2. Main window of the demonstrator. The video is displayed left and the hyperlinks are shown on the right. Using navigation-bar in the top-right corner the user can jump directly to other segments.

5. LINKING

To link video-segments to relevant documents from the databases described in section 3.1 the top five documents similar to a query were retrieved, using the full text search function of the MySQL database which is based on SMART retrieval system [14].

The search script is used by the 'linker'. The latter is a program that reads the transcription files and searches the MySQL databases for every segment. As described in section 4.3, for each segment two types of transcripts are available. The linker uses an entire segment transcription as query. A query to the newspaper article database results in two sets of five documents all ranked by relevance. The linker combines these results and selects the top five documents. Teletext subtitles are a more reliable source and they show a slightly better match to the actual discourse of a corresponding segment. This is primarily due to the correct representation of names of people, organisations, etc., which are often not recognised by the speech recogniser. Therefore, when combining the results, documents found with teletext subtitles are given a small bonus.

The retrieved documents are stored by rank in an XML-file. This XML-file serves as starting point for the generation of platform specific presentations. In this demonstrator a presentation was generated in SMIL.

6. PRESENTATION

InfoLink allows the user to view a Dutch broadcast news item via Internet as shown in Figure 2. Using normal navigation, the user is able to stop the streaming or to jump to another part with a slider. In a window next to the video-window, several hyperlinks will appear for the duration of the segment. The user can follow the hyperlinks to open documents with relevant data, for example a newspaper article. An extra navigation-bar enables users to jump directly to another segment. For the demonstrator we used SMIL 2.0, a W3C standard markup language for synchronised presentation of dynamic multimedia content and RealNetworks' RealText [15, 16].

7. DEMONSTRATOR PERFORMANCE

The demonstrator runs on a compilation video in SMIL format with several episodes of the eight o'clock news. The content has been manually segmented and ASR has been applied to the result-

ing segments. The transcripts have been stored together with the teletext subtitles. Automatic segmentation and forced-alignment of the video-stream is currently under development.

During a first user evaluation, a proof-of-concept for cross-media search was delivered and relevant documents turned out to be contained in the retrieval results. Sometimes documents were retrieved that seem relevant to the query but not pertaining to what users are likely to expect. For example, for a video item about 'KLM' and 'arbeidstijdverkorting' (English: shortening of working hours), documents describing other companies and 'arbeidstijdverkorting' seemed more relevant than documents just describing 'KLM'. It still needs to be sorted out which retrieval weighting schemes would fit the typical user need in a particular use scenario.

As said the quality of the linking is a function of the indexing. The quality of the indexing could also be assessed according to a formal system performance evaluation protocol, TREC-style or otherwise. Because a test collection for Dutch video retrieval is not available (yet), this remains a topic for future research.

8. CONCLUSION

During the implementation of InfoLink several problems were encountered, of which segmentation is the most prominent. The current segmentation method is rather general, which is fine in case InfoLink should be applied to a non-news archive. Nevertheless a segmentation method tuned to news broadcast programs could yield an improvement. For example, video-based segmentation could detect program specific features (shots, news-anchor, captions) which could make segmentation more precise. However the silence detection and text segmentation methods described in this paper can be very useful add-ons to manual indexing and segmentation.

An aspect that has received little attention is the query formulation and query expansion with use of synonyms, co-occurrence information, etc. By expanding a query or generating a 'smarter' query from a transcription retrieval performance could be improved.

For InfoLink, teletext subtitles and speech recognition transcripts seem to be useful additional metadata in a multimedia search environment. They contribute to a search facility that works across different modalities. The combination of subtitles and speech transcripts turned out powerful, but even if in isolation they contribute to the power of the cross-media search functionality.

9. ACKNOWLEDGEMENTS

The InfoLink demonstrator has been developed with the help of Netherlands Institute for Sound and Vision within the Dutch research program CATCH (`http://www.nwo.nl/catch/`). Part of the work on speech recognition has been funded by MultimediaN (`http://www.multimedian.nl`) and Waterland (`http://hmi.ewi.utwente.nl/project/Waterland`).

10. REFERENCES

[1] J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees, "The TREC SDR Track: A Success Story," in *Eighth Text Retrieval Conference*, Washington, 2000, pp. 107–129.

[2] Roeland Ordelman, *Dutch Speech Recognition in Multimedia Information Retrieval*, Ph.D. thesis, University of Twente, The Netherlands, October 2003.

[3] S. Renals, "Ami: Augmented multiparty interaction," in *Proc. NIST Meeting Transcription Workshop*, Montreal, 2004, AMI-10.

[4] NOVALIST, ," `http://dis.tpd.tno.nl/druid/folders/novalist.pdf`.

[5] J. Kuper, H. Saggion, H. Cunningham, T. Declerck, F.M.G. de Jong, D. Reidsma, Y. Wilks, and P. Wittenburg, "Intelligent multimedia indexing and retrieval through multi-source information extraction and merging," in *18th International Joint Conference of Artificial Intelligence (IJCAI)*, Acapulco, Mexico, 2003, pp. 409–414.

[6] W. Byrne, D. Doermann, and M. Franz, "Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives," *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, July 2004.

[7] Roeland Ordelman, Marijn Huijbregts, and Franciska de Jong, "Unravelling the Voice of Willem Frederik Hermans: a Oral-History Indexing Case Study," Technical report, University of Twente, CTIT, Enschede, 2005.

[8] Andrew Merlino, Daryl Morey, and Mark Maybury, "Broadcast news navigation using story segmentation," in *MULTIMEDIA '97: Proceedings of the fifth ACM international conference on Multimedia*, New York NY, USA, 1997, pp. 381–391, ACM Press.

[9] Shih-Fu Chang Gary Huang, Winston Hsu, "Automatic closed caption alignment based on speech recognition transcripts," *Columbia DVMM Technical Report 005*, 2003.

[10] Marti Hearst, "Multi-paragraph segmentation of expository text," in *32nd. Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, New Mexico, 1994, pp. 9–16.

[11] Jeffrey C. Reynar, *Topic segmentation: algorithms and applications*, Ph.D. thesis, 1998, Adviser-Mitchell P. Marcus.

[12] E. Zuurbier, "Afstudeerverslag," Tech. Rep., University of Twente, 2004.

[13] Tzvetanka Ianeva, Liudmila Boldareva, Thijs Westerveld, Roberto Cornacchia, Djoerd Hiemstra, and Arjen de Vries, "Probabilistic Approaches to Video Retrieval," in *Proceedings of the TRECVID workshop*, 2005.

[14] Gerard Salton, "The smart information retrieval system after 30 years - panel," in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum)*, Abraham Bookstein, Yves Chiaramella, Gerard Salton, and Vijay V. Raghavan, Eds. 1991, pp. 356–358, ACM.

[15] "Synchronized Multimedia Integration Language (SMILL)," `http://www.w3.org/AudioVideo/`.

[16] Inc. RealText Authoring Guide RealNetworks, "See http://service.real.com/help/library/guides/realtext/realtext.htm," .