ANTS

— a complete system for automatic news programme annotation based on audiovisual content and text analysis

Giorgio Dimino, Alberto Messina and Roberto Borgotallo
RAI Centre for Research and Technology Innovation

This article describes an integrated system for the automatic annotation of television news programmes named ANTS (Automatic Newscast Transcription System). It consists of several analysis components, integrated within a unified architecture. Users have the possibility of accessing a large daily-growing database of news stories from the main national channels – all identified, categorised and published in a fully automatic way. The system identifies story boundaries, extracts texts from spoken content, classifies stories by subject and links external relevant information coming from the web.

The system's performance has been evaluated in a real-life scenario by a panel of professional users inside RAI. The strength of the approach behind ANTS is its ability to integrate several heterogeneous tools in a performant and ready-for-production environment. ANTS is capable of elaborating many hours of material per day, without significant service drops and with sufficiently good accuracy for industrial deployment in large broadcasting facilities.

Introduction and related work

Automatic programme segmentation is one of the most challenging and complex subjects of research.

Although being able to produce a correct segmentation is a key factor for improving the accessibility and precision of search and retrieval, we cannot count on an established approach at solving the problem in general. The common base of the approaches for news is constituted by the use of a combination of visual, audio and speech features.

The TRECVID initiative [1] had news segmentation among its tasks in 2003 and 2004. The works described in [2] and [3] illustrate several different approaches, identified and developed by the TRECVID participants in those two series. The best performing approaches presented at TRECVID 2004 included video and audio analysis, alone or supplemented with automatic speech-to-text transcripts, and showed an F-measure of between 0.6 and 0.7. The baseline features employed in several of the cases are (i) visual similarity between shots within a time window and (ii) the temporal distance between shots [4]. Other heuristics – such as the similarity of faces appearing in the shots and the detection of the repeated appearance of anchor person shots [5][6][7] – can add a supplemental layer of information to improve the accuracy.

The audio channel contribution can be employed to detect pauses, potential boundaries for topic changes [4][6][8], or to detect changes in audio classification patterns (e.g. from music to speech [6]), or to detect speaker changes [8].

As a third information source, text from transcripts or automated speech recognition is very often used, either by searching similar word appearances in different shots or by detecting text similarities between the shots [5][6].

The use of automatic speech-to-text transcripts introduces several issues in the news segmentation task, due to some typical errors such as missing words, word deletions and insertions, and wrongly transcribed words. The current system does not use Newsroom Computer System (NRCS) data, although recently we started investigating how to integrate this information as well, both for improving transcripts and for improving the performance of automatic segmentation.

Architecture

The system, the main components of which are a centralised workflow engine and a collection of generic *AntsClients*, has been designed to be highly distributed and scalable.

Each *AntsClient* is configured to carry out a specific task of the overall process, such as the speech-to-text activity and the news story segmentation. Running as daemons, the *AntsClients* communicate with the workflow engine via HTTP protocol to get jobs and to notify success/failure. The requested process is then executed by a specific command, launched locally by the *AntsClients*. Such an approach, described in *Fig. 1*, allows for deployment on multiple hosts within the network and/or multiple instances supplying the same service so that it is possible to scale the system as much as needed to reach the required throughput and to provide failover capabilities.

All the metadata produced are collected within a centralised repository until final delivery to the publication platform. The search & retrieve subsystem supports full text search, filtering by categories and/or by named entities, besides identification and publication information. Finally, as shown in *Fig. 2*, the page that is presented displays all the resulting components and allows them to be

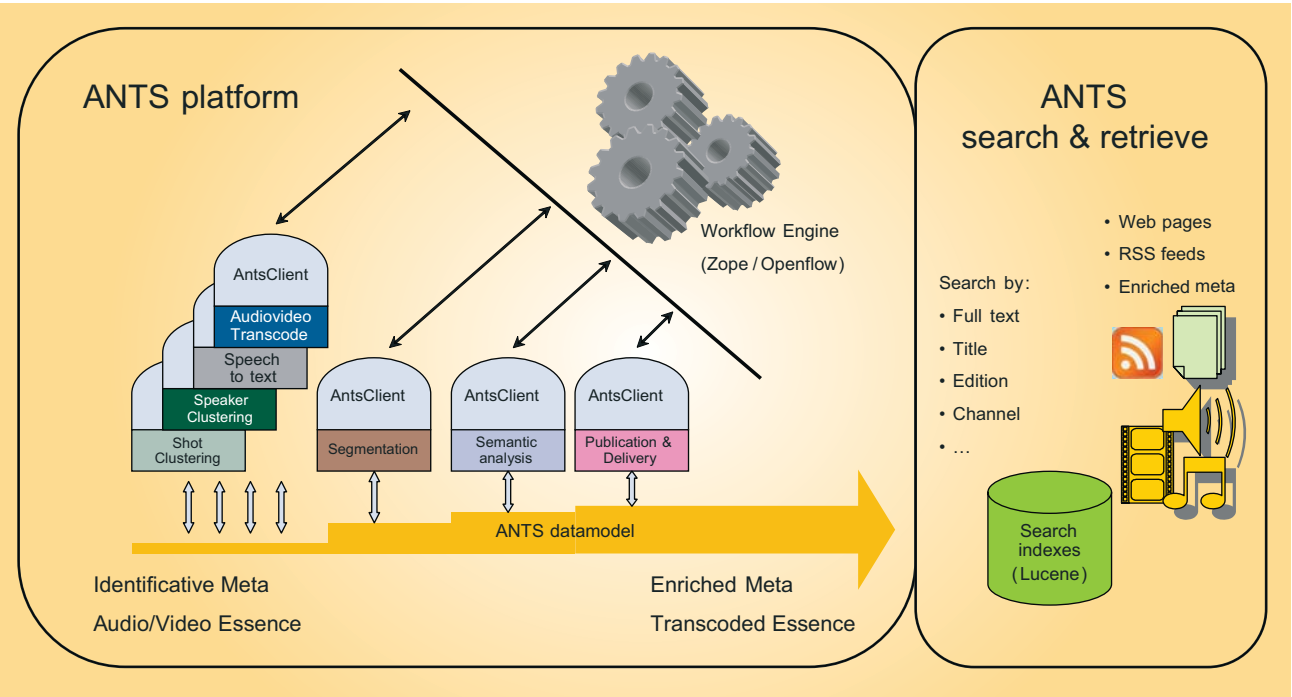

Figure 1
Ants architecture

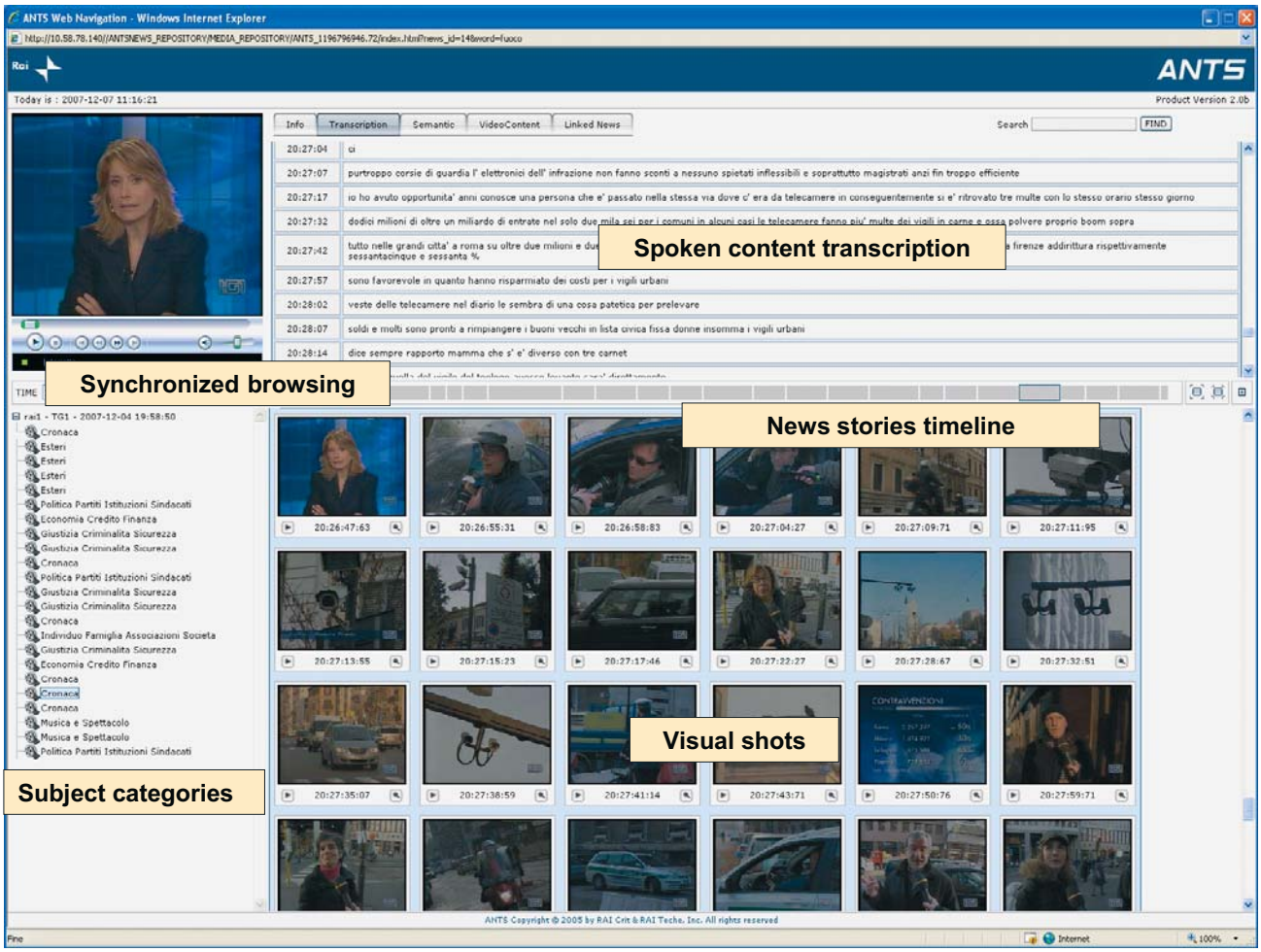

Figure 2
ANTS browsing interface

included synchronously along the common timeline. All the retrieval – as well as monitoring and administration – functionalities can be accessed over an IP network using a web browser.

Moreover, ANTS delivers the finished materials and the metadata, collected in XML format, to the RAI longterm archive catalogue system.

Analysis tools and automatic annotation services

Fig. 3 illustrates the functional block diagram of ANTS, which includes semantic analysis of spoken text and automatic editorial segmentation.

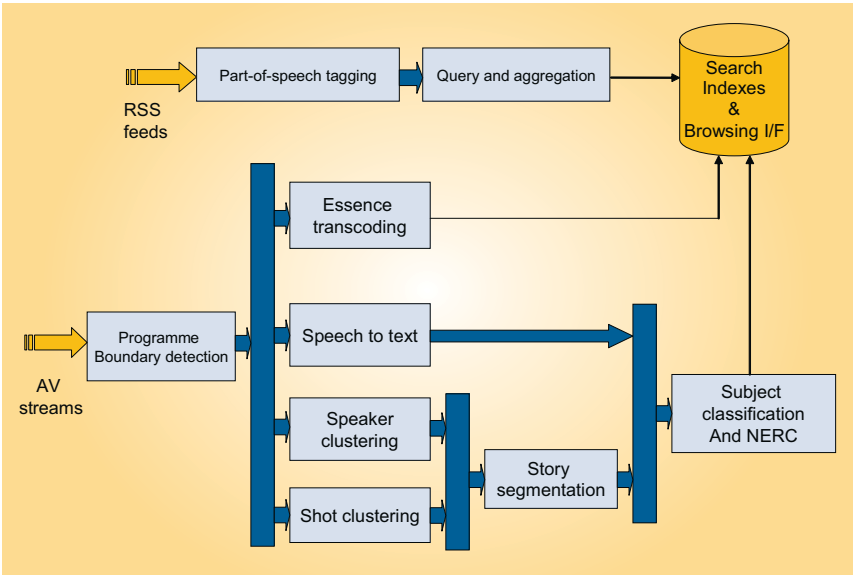

Figure 3
Ants functional block diagramm

Videoclip matching

To achieve automatic segmentation of live streams into pro-

grammes which can be further analysed in the subsequent chain, ANTS makes use of a videoclip-matching technique.

Starting and ending programme jingles are used as references to be found in acquired streams, through the use of a shot-clustering algorithm.

As an offline process, feature signatures are extracted from each shot of the reference clips, which includes HSV colour space[1] and Luminance histograms, texture-signature histograms (contrast and directionality), and temporal-activity histograms.

All histograms have 65 bins, to form a 65 x 7 x N feature vector for each shot, where N is the number of pictures in the shot. During the online phase, clipshot signatures are used as multidimensional fixed centroids in a clustering algorithm.

At the end of the clustering, incoming shots aggregated with sufficient strength (i.e. closeness to the centroid) are classified as instances of the clipshot associated with the centroid. The average detection accuracy of the process is 0.86, while recall is about 0.87. (Precision is defined as the ratio between the correctly-identified items and the total number of detected items, while recall is defined as the ratio between the correctly-identified items and the effective number of items.)

Shot clustering

Shot clustering is done following an optimized bottom-up clustering method which makes use of the intersection of feature histograms as the core distance measurement.

The whole process is divided in the following steps:

- feature histograms extraction;
- shot detection;
- partial clustering on fixed-length segments;
- selection of relevant clusters found in the partial clustering phase, and;
- re-clustering on selected clusters.

Thus, the process has a sequence of images as the input and a set of cluster labels associated with each of the input images as the output. This means that the process performs shot detection as a side effect of the clustering process.

Audio clustering

Audio clustering is performed using the mClust tool [9], a free software under the GNU General Public Licence, developed at the LIUM laboratories of the University of Maine – Le Mans, France.

Elaboration results consist of a set of labelled clusters, each of which points to a different person talking in the analysed audioclip. The single cluster is a set of time intervals identified with relative boundaries from the beginning of the clip.

Segmentation of news stories

Segmentation of news programmes into news stories is done by exploiting both aural and visual cues with the help of a three-layered heuristic framework, deduced by the observation of editorial styles of a statistically significant set of programmes, spanning approximately 40 hours (~80 programmes).

1. For an explanation of HSV, see Wikipedia article: **http://en.wikipedia.org/wiki/HSV_color_space**

Abbreviations

ANTS	(RAI) Automatic Newscast Transcription System	**NRCS**	NewsRoom Computer System
HTTP	HyperText Transfer Protocol	**RSS**	Really Simple Syndication
IP	Internet Protocol	**XML**	eXtensible Markup Language

The basic heuristics, widely adopted in literature – e.g. by [10] – is that being able to detect boundaries of shots containing the anchorman is equivalent to detecting news story boundaries.

To detect anchorman shots we use another heuristics, namely that the most frequent speaker is the anchorman and that he/she speaks for periods of time spreading right along the programme timeline. This allows us to select the most probable candidate speaker among the ones identified by a speaker-clustering process.

This approach doesn't permit the system to discern situations in which the anchorman introduces several brief stories in sequence without external contributions (e.g. reportages). To overcome this limitation we use the third heuristic, consisting of the knowledge that, in the great majority of observed cases, the introduction of a new brief story is accompanied by a camera shot change (e.g. from a close-up shot to a wider one). To optimize the accuracy in selecting the camera shot changes, we perform a videoshot-clustering process based on the same features explained in *Fig. 3*. This allows us to detect and classify shot clusters as pertaining to studio shots containing the anchorman, following the same frequency/extension heuristic used for detecting the candidate speaker. This double clustering process (both on audio and on video) enables a very simple and effective recursive algorithm which selects alternatively video and audio clusters on the basis of their mutual coverage percentage.

Finally, story boundaries are identified as those in which either an audio or a video cluster boundary occurs among the clusters selected by the recursive algorithm, with an adaptive threshold to avoid oversegmentation. An outline of the experimental evaluation of the segmentation algorithm is presented on the next page.

Subject classification and link to external sources

In ANTS, spoken-content extraction is performed using a speech-to-text engine based on [11], which is capable of transcribing both Italian and English. Subject classification of segmented stories is done using a naive Bayesian classification model trained on a corpus made up of items of extracted text and annotated with a standard subject taxonomy of 28 classes. The corpus counts 25'000 items, 4/5 of which are used for training and the remaining 1/5 for testing. The overall observed subject classification accuracy on the running system is 0.82, while programme-level accuracy, i.e. the average classification accuracy calculated for the set of items belonging to the same programme, is 0.88.

Links to external information sources are implemented through linguistic analysis of the RSS feeds of a pool of six major newspapers, which are polled periodically. On each individual RSS item title, a part-of-speech tagging is performed in order to extract the most significant word cues, which are in turn used to perform a full text query on the text automatically extracted from the television news items. Query results are arranged in a browse list associated with the item title, and each individual news story is linked to the RSS item for which search scores are above a certain threshold. Aggregated news stories under a certain title can themselves be seen as RSS services provided by ANTS.

The result is that users can have a multimedia integration of RSS feeds coming from the major newspapers, made with relevant news stories collected from the major television newscasts. On the other hand, television news items can be automatically annotated using the RSS titles.

The mean average precision of the news-item aggregation process is 0.97, calculated as the ratio between the relevant news items and the total ones, with respect to the RSS feed title in a certain aggregation.

Experimental evaluation of the news stories segmentation algorithm

We tested our news story segmentation algorithm against a set of test programmes running to about 40 hours of material. The test set had been manually tagged, i.e. all true story boundaries had been identified. To assess the system performance we used an alignment measurement taking into account starting boundaries and ending boundaries with different weights, as well as considering missing material as having more impact than excessive material on the measurement.

In a first phase, we randomly selected a subset of the test material and then optimized empirically the parameters of evaluation in order to achieve a good match between the users' assessment of the segmentation quality and the objective measurement. We thus obtained a user-validated quality measurement. In the second phase, once the measurement has been verified according to the described procedure, we adjusted the segmentation model parameters in order to optimize the user-validated measurement output.

Table 1 shows the precision obtained for four subclasses of programmes.

Table 1 – Precision figures of automatic segmentation

Class	Tg1	Tg2	Tg3	TgR
Precision	0.81	0.69	0.80	0.73

The role of open source components

The system described in this article would not be viable without the availability of open source tools. Firstly, because almost all computers in service within this architecture run on the Linux operating system, which was also the development and testing platform. From components written in C programming language, compiled with the GNU Compiler Collection (GCC), to other components written in much higher-level programming languages, such as python, perl or ruby, or simple scripts in any variant of the Unix Shell ... the various ingredients were prepared and integrated in the open source domain.

To manipulate audio and video material, the *MJPEG Tools* [12], were successfully adopted; the workflow management relies on the couple *Openflow* over *Zope* [13][14]; the publication service is built as a web application which makes use of *Postgresql* [15] together with the *Apache* web servers http and *Tomcat* and the search engine *Lucene* [16].

The tool used for speaker segmentation, within the editorial segmentation process, is *mClust* [9], while for the subject categories, *Categorizer* [17] was adopted.

Beyond the concept of an open source toolkit, it is the open source environment, including the experience of the people involved, which made possible the achievement of such a complex system with continuous adjustments and requests for new features.

Conclusions

In this article we have given an overview of an automatic news programme annotation system named ANTS, developed at the RAI Research Centre in Turin. The strength of the underlying approach of ANTS consists in offering to the users a good global performance by integrating several analysis tools into a single fully-automated product.

Documentation of TV news items must provide usable results in a considerably shorter delay than for other television programmes. While with a manual annotation process – although assisted by automatic acquisition and shot detection – one item took a couple of working days before being

Giorgio Dimino received a degree in electrical engineering from the Polytechnic of Turin in 1987. In 1988 he joined the Research Centre of RAI - Radiotelevisione Italiana in Turin, working in the fields of digital audio and video processing, and archiving. His interests include the design of automated digital archives and the application of information technology in television production.

Mr Dimino has been leader of the work area on Metadata, Access and Delivery of the IST 6th Framework project PrestoSpace. He is also an active member of the EBU PMC and FIAT/IFTA.

Alberto Messina works for the Research Centre of RAI - Radiotelevisione Italiana in Turin: *Centro Ricerche e Innovazione Tecnologica* (CRIT). He is involved in several internal and international research projects in the field of digital archiving, with particular emphasis on automated documentation, and automated production. His current interests range from file formats and metadata standards to the domain of content analysis and information extraction algorithms, which is now his main focus. He is also author of various technical and scientific publications in this subject area. He collaborates extensively with the local University of Turin - Computer Science Department, which involves common research projects and students' tutorship.

Mr Messina is an active member of several EBU projects including P/TVFILE, P/MAG and P/CP, and is chairman of the P/SCAIE project dealing with automatic metadata extraction techniques.

Roberto Borgotallo graduated in Telecommunication Engineering at *Politecnico di Torino* in 1999. Since 2001, he has been working for RAI - Radiotelevisione Italiana at the R&D department (*Centro Ricerche e Innovazione Tecnologica*) in Turin. Initially, he was involved in several projects gravitating around the RAI multimedia catalogue, called CMM, regarding metadata ingestion and transformation.

More recently, Mr Borgotallo has been working in a team that is developing an automatic metadata extraction platform which is actually used extensively in RAI for experimental purposes, and even for real in the production evironment. His major professional interests are metadata and essence transformation, system integration and workflow management.

available to users, after the introduction of ANTS a newscast becomes searchable at single-story level within two hours after publication.

The ANTS system is currently in use in RAI to index the main newscast editions of the three national generalist channels RAI1, RAI2 and RAI3. Users can directly query the system or subscribe to personalized RSS feeds on specific topics and be notified when new relevant content is available. The service will be extended progressively to cover also the regional news.

Another interesting application of ANTS – which is being provided to some Italian regional administrations – is the monitoring of local stations' transmissions. In this case, all the relevant channels' emissions are recorded and indexed. Administration officers can then browse the content for statistical analysis purposes or for verification of compliance to the Authority's regulations on TV emissions.

Future works will be directed towards extending editorial segmentation to other kinds of programmes and to the development and integration of emerging techniques in metadata extraction.

Acknowledgements

The authors wish to thank Laurent Boch and Daniele Airola, from RAI CRIT, for their outstanding contributions to the design and development of the ANTS system.

References

[1] Trec video retrieval evaluation. Internet site: **http://www-nlpir.nist.gov/projects/t01v/**

[2] T. Chua, S. Chang, L. Chaisorn and W. Hsu: **Story boundary detection in large broadcast news video archives techniques, experience and trends**
In *Proc. of ACM Multimedia 2004*.

[3] W. Kraaj, A. Smeaton and P. Over: **Trecvid 2004: An overview**
In *Proc. of TRECVID Workshop 2004*.

[4] D. Eichmann and D.-J. Park: **Boundary and feature extraction at the university of Iowa**
In *Proc. of TRECVID Workshop 2004*.

[5] M.J. Pickering, L. Wong, and S.M. Rueger: **Anses: Summarization of news video**
In *Proc. of International Conference on Image and Video Retrieval (CIVR)*, 2003.

[6] T. Volkmer, S.M.M. Tahahoghi and H.E. Williams: **Rmit university at trecvid 2004**
In *Proc. of TRECVID Workshop 2004*.

[7] Y. Zhai, X. Chao, Y. Zhang, O. Javed, A. Yilmaz, F. Rafi et al: **University of central Florida at trecvid 2004**
In *Proc. of TRECVID Workshop 2004*.

[8] G.M. Qu'enot, D. Mararu, S. Ayache, M. Charhad and L. Besacier: **Clips-lis-lsr-labri experiments at trecvid 2004**
In *Proc. of TRECVID Workshop 2004*.

[9] Internet site: **http://www-lium.univ-lemans.fr/tools**

[10] M. De Santo, G. Percannella, C. Sansone and M. Vento: **Unsupervised news video segmentation by combined audio-video analysis**
In *MRCS*, pages 273–281, 2006.

[11] F. Brugnara, M. Cettolo, M. Federico and D. Giuliani: **A system for the segmentation and transcription of Italian radio news**
In *Proc. of RIAO, Content-Based Multimedia Information Access*, 2000.

[12] Internet site: **http://mjpeg.sourceforge.net**

[13] Internet site: **http://openflow.sourceforge.net**

[14] Internet site: **http://www.zope.org**

[15] Internet site: **http://www.postgresql.org**

[16] Internet site: **http://www.apache.org**

[17] Internet site: **http://search.cpan.org**