Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment

Martha Larson[1], Eamonn Newman[2], and Gareth J. F. Jones[2]

[1]Multimedia Information Retrieval Lab, Delft University of Technology, 2628 CD Delft, Netherlands
[2]Centre for Digital Video Processing, Dublin City University, Dublin 9, Ireland
m.a.larson@tudelft.nl,{enewman,gjones}@computing.dcu.ie

Abstract. VideoCLEF 2009 offered three tasks related to enriching video content for improved multimedia access in a multilingual environment. For each task, video data (Dutch-language television, predominantly documentaries) accompanied by speech recognition transcripts were provided. The *Subject Classification Task* involved automatic tagging of videos with subject theme labels. The best performance was achieved by approaching subject tagging as an information retrieval task and using both speech recognition transcripts and archival metadata. Alternatively, classifiers were trained using either the training data provided or data collected from Wikipedia or via general Web search. The *Affect Task* involved detecting narrative peaks, defined as points where viewers perceive heightened dramatic tension. The task was carried out on the "Beeldenstorm" collection containing 45 short-form documentaries on the visual arts. The best runs exploited affective vocabulary and audience directed speech. Other approaches included using topic changes, elevated speaking pitch, increased speaking intensity and radical visual changes. The *Linking Task*, also called "Finding Related Resources Across Languages," involved linking video to material on the same subject in a different language. Participants were provided with a list of multimedia anchors (short video segments) in the Dutch-language "Beeldenstorm" collection and were expected to return target pages drawn from English-language Wikipedia. The best performing methods used the transcript of the speech spoken during the multimedia anchor to build a query to search an index of the Dutch-language Wikipedia. The Dutch Wikipedia pages returned were used to identify related English pages. Participants also experimented with pseudo-relevance feedback, query translation and methods that targeted proper names.

1 Introduction

VideoCLEF 2009[1] was a track of the CLEF[2] benchmark campaign devoted to tasks aimed at improving access to video content in multilingual environments. The overall goal of the VideoCLEF benchmarking initiative, now referred as "MediaEval"[3], is to develop new, forward-looking multimedia retrieval tasks and data sets with which to evaluate these tasks. During VideoCLEF 2009, three tasks were carried out. The *Subject Classification Task* required participants to automatically tag videos with subject theme labels (e.g., 'factories,' 'physics,' 'poverty', 'cultural identity' and 'zoos'). The *Affect Task*, also called "Narrative peak detection," involved automatically detecting dramatic tension in short-form documentaries. Finally, "Finding Related Resources Across Languages," referred to as the *Linking Task*, required participants to automatically link video to Web content that is in a different language, but on the same subject. The data sets for these tasks contained Dutch-language television content supplied by the Netherlands Institute of Sound and Vision[4] (called in Dutch *Beeld & Geluid*), which is one of the largest audio/video archives in Europe. Each participating site had access to video data, speech recognition transcripts, shot boundaries, shot-level keyframes and archival metadata supplied by VideoCLEF. Sites developed their own approaches to the tasks and were allowed to chose the method and features that they found most appropriate. Seven groups made submissions of task results for evaluation.

In 2009, the VideoCLEF track ran for the first time as a full track within the Cross-Language Evaluation Forum (CLEF) evaluation campaign. The track was piloted last year as VideoCLEF 2008 [11]. The VideoCLEF track is successor to the Cross-Language Speech Retrieval (CL-SR) track, which ran at CLEF from 2005 to 2007 [12]. VideoCLEF seeks to extend the results of CL-SR to the broader challenge of video retrieval. VideoCLEF is intended to complement the TRECVid benchmark [15] by running tasks related to the topic or subject matter treated by video and emphasizing the importance of speech and language (e.g., via speech recognition transcripts). TRECVid has traditionally focused on objects, entities and scenes that are depicted in the visual channel. In contrast, VideoCLEF concentrates on what is described in a video, in other words, what a video is about.

This paper describes the data sets and the tasks of VideoCLEF 2009 and summarizes the results achieved by the participating sites. We finish with a conclusion and an outlook for MediaEval 2010. For additional information concerning individual approaches used in 2009, please refer to the papers of the individual sites in this volume.

[1] http://www.multimediaeval.org/videoclef09/videoclef09.html
[2] http://www.clef-campaign.org
[3] http://www.multimediaeval.org
[4] http://www.beeldengeluid.nl

1.1 Data

VideoCLEF 2009 used two data sets both containing Dutch-language television programs. Note that these programs are predominantly documentaries with the addition of some talk shows. This means that the data contains a great deal of conversational speech, including opinionated and subjective speech and speech that has been only loosely planned. In this way, the VideoCLEF data is different and more challenging than broadcast news data, which largely involves scripted speech.

The VideoCLEF 2009 Subject Classification Task ran on TRECVid 2007 and 2008 data from *Beeld & Geluid*. The Affect Task and Linking Task both ran on a data set containing material from the short-form documentary *Beeldenstorm*, also supplied by *Beeld & Geluid*. For both data sets, Dutch-language speech recognition transcripts were supplied by the University of Twente [5]. The shot segmentation and the shot-level keyframe data were provided by Dublin City University [1]. Further details are given in the following.

TRECVid 2007/2008 data set In 2009, VideoCLEF attempted to encourage cross-over from the TRECVid community by recycling the TRECVid data set[5] for the Subject Classification Task. Notice that the Subject Classification Task is a fundamentally different task than what ran at TRECVid in 2007 and 2008. Subject Classification involves automatically assigning subject labels to videos at the episode level. The subject matter of the entire video is important, not just the concepts visible in the visual channel and not just the shot-level topic.

Classifying video, i.e., taking a video and assigning it a topic class subject label, is exactly what the archive staff does at *Beeld & Geluid* when they annotate video material that is to be stored in the archive. The class labels used for the VideoCLEF 2009 Subject Classification Task are a subset of labels that are used by archive staff. As a result, we have gold standard topic class labels with which to evaluate classification. Additionally, we can be relatively certain that if these labels are already used for retrieval of material from the archive then they are relevant for video search in an archive setting, and we assume, beyond. Original Dutch-language examples of subject labels can be examined in the search engine.[6]

In the VideoCLEF 2009 Subject Classification Task, archivist-assigned subject labels were used as ground truth.[7] The training set is a large subset of TRECVid 2007 and contains 212 videos. The test set is a large subset of TRECVid

[5] http://www-nlpir.nist.gov/projects/tv2007/tv2007.html#3

[6] Visit the search engine at http://zoeken.beeldengeluid.nl. A keyword search will return a results list with a column labeled *Trefwoorden* or keywords. These are the topic class subject labels that are used in the archive.

[7] In total 46 labels were used: aanslagen (attacks), armoede (poverty), burgeroorlogen (civil wars), criminaliteit (crime), culturele identiteit (cultural identity), dagelijks leven (daily life), dieren (animals), dierentuinen (zoos), economie (economy), etnische minderheden (ethnic minorities), fabrieken (factories), families (families), gehandicapten (disabled), geneeskunde (medicine), geneesmiddelen (pharmaceutical drug),

2008 and contains 206 videos. The videos are most drawn from a subset of the overall *Beeld & Geluid* collection called *Academia*,[8] which is a collection that was created for use in research and educational settings. The *Academia* collection currently contains about 7,000 hours of video.

In general, each video in the training and test set is an individual episode of a television show. Their length varies widely with the average length being around 30 minutes. Note that the VideoCLEF 2009 Subject Classification set excludes several videos in the TRECVid collection for which archival metadata was not available. We would also like to explicitly point out that participants were not required to make use of the training data set, but were free to collect their own training data, if they wished.

***Beeldenstorm* data set** For both the Affect Task and Linking Task a data set consisting of 45 episodes of the documentary series *Beeldenstorm* (Eng. Iconoclasm) was used. The *Beeldenstorm* series consists of short-form Dutch-language video documentaries about the visual arts. Each episode lasts approximately eight minutes.

Beeldenstorm is hosted by Prof. Henk van Os, known and widely appreciated, not only for his art expertise, but also for his narrative ability[9]. This data set is also supplied by *Beeld & Geluid*, but it is mutually exclusive with the TRECVid 2007/2008 data set. The narrative ability of Prof. van Os makes the Beeldenstorm set an interesting corpus to use for affect detection and the domain of visual arts offers a wide number of possibilities for interesting multimedia links for the linking task.

Finally, the fact that each episode is short makes it possible for assessors to watch the entire episode when creating the ground truth. Knowledge of the complete context is important for relevance judgments for cross-language related resources and also for defining narrative peaks.

The ground truth for the Affect Task and Linking Task was created by a team of three Dutch-speaking assessors during a nine-day assessment and annotation event at Dublin City University referred to as *Dublin Days*. The videos were annotated with the ground truth with the support of the Anvil[10] Video Annotation Research Tool [8]. Anvil makes it possible to generate frame-accurate video

genocide (genocide), geschiedenis (history), gezinnen (families), havens (harbors), hersenen (brain), illegalen (undocumented immigrants), journalisten (journalist), kinderen (children), landschappen (landscapes), media (media), militairen (military personnel), musea (museums), muziek (music), natuur (nature), natuurkunde (physics), ouderen (seniors), pers (press), politiek (politics), processen (lawsuits), rechtszittingen (court hearings), reizen (travel), taal (language), verkiezingen (elections), verkiezingscampagnes (electoral campaigns), voedsel (food), voetbal (soccer), vogels (birds), vrouwen (women), wederopbouw (reconstruction), wetenschappelijk onderzoek (scientific research), ziekenhuizen (hospitals).

[8] http://www.academia.nl/

[9] http://www.avro.nl/tv/programmas_a-z/beeldenstorm/

[10] http://www.anvil-software.de/

annotations in a graphic interface. Particularly important for our purposes was the support offered by Anvil for user-defined annotation schemes. Details of the ground truth creation are included in the discussions of the individual tasks in the following section.

2 Subject Classification Task

2.1 Task

The goal of the Subject Classification Task is automatic subject tagging. Semantic-theme-based subject tags are assigned automatically to videos. The purpose of these tags is to make the videos findable to users who are searching and browsing the collection. The information needs (i.e., queries) of the users are not specified at the time of tagging. In VideoCLEF 2009, the Subject Classification Task had the specific goal of reproducing the subject labels that were hand assigned to the test set videos by archivists at *Beeld & Geluid*. Since these subject labels are currently in use to archive and retrieve video in the setting of a large archive, we are confident about their usefulness for search and browsing in real-world information retrieval scenarios. The Subject Classification Task was introduced during the VideoCLEF 2008 pilot [11]. In 2009, the number of videos in the collection was increased from 50 to 418 and the number of subject labels increased from 10 to 46.

2.2 Evaluation

The Subject Classification Task is evaluated using Mean Average Precision (MAP). This choice of score is motivated by the popularity of techniques that approach the subject tagging task as an information retrieval problem. These techniques return, for each subject label, a ranked list of videos that should receive that label. MAP is calculated by taking the mean of the Average Precision over all subject labels. For each subject label, precision scores are calculated by moving down the results list and calculating precision at each position where a relevant document is retrieved. Average Precision is calculated by taking the average of the precision at each position. Calculations were performed using version 8.1 of the `trec_eval`[11] scoring package.

2.3 Techniques

Computer Science, Chemnitz University of Technology, Germany (see also [9]) The task was treated as an information retrieval task. The test set was indexed using an information retrieval system and was queried using the subject labels as queries. Documents returned as relevant to a given subject label were tagged with that label. The number of documents receiving a given label was controlled by a threshold. The submitted runs varied with respect to whether or not the

[11] http://trec.nist.gov/trec_eval

archival metadata was indexed in addition to the speech recognition transcripts. They also varied with respect to whether expansion was applied to the class label (i.e., the query). Expansion was performed by augmenting the original query with the most frequent term occurring in the top five documents returned by an initial retrieval round. If fewer than two documents were returned, queries were expanded using a thesaurus.

SINAI Research Group, University of Jaén, Spain The SINAI[12](see also [13]) The group approached the task as a categorization problem, training SVMs using the training data provided. One run, `SINAI_svm_nometadata`, extracted feature vectors from the speech transcripts alone and one run, `SINAI_svm_withmetadata`, made use of both speech recognition transcripts and metadata.

Computer Science, Alexandru Ioan Cuza University, Romania (see also [2]) A training set was created by using subject category labels to select documents from Wikipedia and also from the Web at large (using Google). The training set was used to create a category file for each subject category containing a set of informative terms representative of that category. Two categorization methods were applied, one made use of information retrieval techniques to match the speech recognition transcripts of the videos to the category files and the other made use of a Naive Bayes multinomial classifier to classify the videos into the classes represented by the category files.

2.4 Results

The MAP of the results of the task are reported in Table 1. The results confirm the viability of techniques that approach the Subject Classification Task as an information retrieval task. Such techniques proved useful in VideoCLEF 2008 [11] and also provide the best results in 2009 where the size of the collection and the label set increased. Also, consistent with VideoCLEF 2008 observations, performance is better when archival metadata is used in addition to speech recognition transcripts.

In the wake of VideoCLEF 2008, we decided that we wanted to provide a training data set of videos accompanied by speech transcripts in 2009 to see whether training classifiers on data from the same domain as the test data would improve performance. The runs submitted this year demonstrate the efficacy of an approach that combines Web data and information retrieval techniques. A supervised approach which uses same-domain training data cannot easily achieve the same level of performance. These results leave open the question of how much training data is necessary in order for a supervised approach to compete with the information retrieval approach.

In all cases, runs that make use of metadata outperform runs that make use of ASR transcripts only. This performance differences demonstrate the high

[12] SINAI stands for *Sistemas Inteligentes de Acceso a la Información*

Table 1. Subject Classification Results Test Set

run ID	MAP
cut1_sc_asr_baseline	0.0067
cut2_sc_asr_expanded	0.0842
cut3_sc_asr_meta_baseline	0.2586
cut4_sc_asr_meta_expanded	0.2531
cut5_sc_asr_meta_expanded	**0.3813**
SINAI_svm_nometadata	0.0023
SINAI_svm_withmetadata	0.0028

value of using metadata, if available, to supplement ASR transcripts in order to generate class labels for videos.

The Alexandru Ioan Cuza University (UAIC) team reported results on the training set only. Note that because they train using data that they have collected themselves, the training set constitutes for the purposes of their experiments a separate, unseen test set. The results are not, however, directly comparable to those given in Table 1. We do not repeated them here, but rather refer the interested reader to the UAIC team paper [2]. Here, we include the comment that the best UAIC run involved using both general Web and Wikipedia training data and then combining the output Information-Retrieval approach (which they find improves the quality of the first-best label) and the output of a Naive Bayes-classifier (which they find contributes to the overall label quality).

The UAIC results are consistent with our overall conclusion that it is better to collect training data from external sources rather than to use the training set. We believe that there are two possible sources to which we can attribute the failure of the training set to allow the training of high quality classifiers. First, the training set was relatively small, including only 212 videos. Although some semantic categories are represented by a fair number of video items, other categories may have as few as two items associated with them in the training set. Second, the transcripts of the training set contain a high level of speech recognition errors, which means that important terms might be mis-recognized and thus fail to occur in the transcripts at all, or fail to occur with the proper distribution.

There is general awareness shared by VideoCLEF participants that although MAP is a useful tool, it may not be the ideal evaluation metric for this task. The reader can refer to the papers of the Chemnitz [9] and SINAI [13] for additional discussion and results reported with additional performance metrics.

The ultimate goal of subject tagging is to generate a set of tags for each video that will allow users to find that video while searching or browsing. The utility of a tag assigned to a given video is therefore not entirely independent of the other tags assigned. Under the current formulation of the task, the presence or absence of the tag is the only information that is of use to the searcher. The ranking of a video in a list of videos that are assigned the same tag is for

this reason not directly relevant to the utility of that tag for the user. Future work must necessarily involve developing appropriate metrics for evaluating the usefulness to the uses of sets of tags assigned to multimedia items.

3 Affect Task

3.1 Task

The goal of the Affect Task at VideoCLEF 2009 was to automatically detect narrative peaks in documentaries. Narrative peaks were defined to be those places in a video where viewers report feeling a heightened emotional effect due to dramatic tension. This task was new in 2009. The ultimate aim of the Affect Task is to move beyond the information content of the video and to analyze the video with respect to characteristics that are important for viewers, but not related to the video topic.

Narrative peak detection builds on and extends work in affective analysis of video content carried out in the areas of sports and movies, cf. e.g., [4]. Viewers perceive an affective peak in sports videos due to tension arising from the spontaneous interaction of players within the constraints of the physical world and the rules and conventions of the game. Viewers perceive an affective peak in a movie due to the action or the plot line, which is carefully planned by the script writer and the filmmaker.

Narrative peaks in documentaries are a new domain in so far as they cannot be considered to fall into either category. Documentaries convey information and often have storylines, but do not have the all-dominating plot trajectory of a movie. Documentaries often include extemporaneous narrative or interviews, and therefore also have a spontaneous component. The affective curve experienced by a viewer watching a documentary can be expected to be relatively subtly modulated.

It is important to differentiate narrative peak detection from other cases of affect detection, such as hotspot detection in meetings. Hotspots are moments during meetings where people are highly involved in the discussion [16]. Hotspots can be self-reported by meeting participants or annotated in meeting video by viewers. In either case, it is the participant and not the viewer whose affective reaction is being detected.

We chose the the *Beeldenstorm* series for the narrative peak detection task in order to make the task as simple and straightforward as possible in its initial year. *Beeldenstorm* features a single speaker, the host Prof. van Os, and covers a topical domain, the visual arts, that is rich enough to be interesting, yet is relatively constrained. These characteristics help us to control for the effects of personal style of the host and of viewer familiarity with topic in the affect and appeal task. Further, as mentioned above, the fact that the documentaries are short makes it possible for annotators to watch them in their entirety when annotating narrative peaks.

3.2 Evaluation

For the purposes of evaluation, as mentioned above, three Dutch speakers annotated the *Beeldenstorm* collection by each identifying the three top narrative peaks in each video. Annotators were asked to mark the peaks where they felt the dramatic tension reached its highest level. They were not supplied with an explicit definition of a narrative peak. Instead, all annotators needed to form independent opinions of where they perceived narrative peaks. In order to make the task less abstract, they were supplied with the information that the *Beeldenstorm* series is associated with humorous and moving moments. They were told that they could use this information to formulate their notion of what constitutes a narrative peak. Peaks were required to be a maximum of ten seconds in length.

Although the annotators did not consult with each other about specific peaks, the team did engage in discussion during the definition process. The discussion ensured that there was underlying consensus about the approach to the task. In particular, it was necessary to check that annotators understood that a peak must be a high point in the storyline as measured by their perceptions of their own emotional reaction. Dramatic objects or facts in the spoken or visual content that were not part of the storyline as it was created by the narrator/producer were not considered narrative peaks. Regions in the video where the annotator guessed that the speaker or producer had intended there to be a peak, but where the annotator did not feel any dramatic tension were not considered to be peaks. An example of this would be a joke that the annotator did not understand completely.

The first two episodes for which the annotators defined peaks were discarded in order to assure that the annotators perception of a narrative peak had stabilized. This warm-up exercise was particularly important in light of the fact that at the end of the annotation effort, assessors reported that it was necessary to become familiar with the style and allow an affinity for the series to develop before they started to feel an emotional reaction to narrative peaks in the video.

The peaks identified by the assessors were considered to be a reflection of underlying "true" peaks in the narrative of the video. We assumed that the variation between assessors is the result of noise due to effects such as personal idiosyncracies. In order to generate a ground truth most highly reflective of "true" peaks, the peaks identified by the assessors were merged. The assessment team consisted of three members who each identified three peaks in 45 videos for a total of 405 marked peaks. The assessors were able to give a rough estimate of the minimum distance between peaks and on the basis of their observations, it was decided to consider two peaks that overlapped by at least two seconds to be the same peak. After merging the peaks, 293 of the 405 peaks turned out to be distinct. The merging process was carried out by fitting a 10 second window to overlapping assessor peaks in order to ensure that merged peaks could never exceed the specified peak length of 10 seconds.

Evaluation involved the application of two scoring methods, the *point-based approach* and the *peak-based approach*. Under point-based scoring, the peaks

chosen by each assessor are assessed without merging. A hypothesized peak receives a point in every case in which it falls within eight seconds of an assessor peak. The run score is the total number of peaks returned by all peak hypotheses in the run. A single episode can earn a run between three points (assessors chose completely different peaks) and nine points (assessors all chose the same peaks). There are no episodes in the set that fall at either of these extremes. The distribution of the peaks in the files is such that the best possible run would earn 246 points. Under peak-based scoring, a hypothesis is counted as correct if it falls within an 8 second window of a peak representing a merger of assessor annotations. Three different types of merged reference peaks are defined for peak-based scoring. Three different peak-based scores are reported that differ in the number of assessors required to agree in order for a region in the video to be considered a peak. Of the 293 total peaks identified, 203 peaks are "personal peaks" (peaks identified by only one assessor), 90 are "pair peaks" (peaks that are identified by at least two assessors) and 22 are "general peaks" (peaks upon which all three assessors agreed).

3.3 Techniques

Narrative peak detection techniques were developed that used the visual channel, the audio channel and the speech recognition transcript. Each group took a different approach.

Computer Science, Alexandru Ioan Cuza University, Romania (see also [2]) Based on the hypothesis that speakers raise their voices at narrative peaks, three runs were developed that made use of the intensity of the audio signal. A score was computed for each group of words that involved a comparison of intensity means and other statistics for sequential groups of words. The top three scoring points were hypothesized as peaks.

Computer Vision and Multimedia Laboratory, University of Geneva, Switzerland (see also [7]) The assumption was made that dramatic peaks correspond to the introduction of a new topic and thus correspond to change in word use as reflected in the speech recognition transcripts. Additionally, the video and audio channel effects assumed to be indicative of peaks were explored. Finally, a weighting was deployed that gave more emphasis to positions at which peaks were expected to occur based on the distribution of peaks in the development data. The weighting is used in `unige-cvml1`, `unige-cvml2` and `unige-cvml3`. Run `unige-cvml1` uses text features alone. Run `unige-cvml3` uses text plus elevated speaker pitch. Run `unige-cvml2` uses text, elevated pitch and quick changes in the video. Run `unige-cvml4` uses text only and no weighting. Run `unige-cvml5` sets peaks randomly to provide a random baseline for comparsion.

Delft University of Technology and University of Twente, Netherlands (see also [10]) Only features extracted from the speech transcripts were exploited. Run `duotu09fix` predicted peaks at fixed points chosen by analyzing the development

data. Run `duotu09ind` used indicator words as cues of narrative peaks. Indicator words were chosen by analyzing the development data. Run `duotu09rep` applied the assumption that word repetition, reflecting the use of an important rhetorical device, would indicate a peak. Run `duotu09pro` used pronouns as indicators of audience directed speech and assumed that high pronoun densities would correspond to points where viewers feel maximum involvement. Run `duotu09rat` exploited the affective scores of words, building on the hypothesis that use of affective speech characterizes narrative peaks.

3.4 Results

The results of the task are reported in Table 2. The results make clear that it is quite challenging to effectively support the detection of narrative peaks using audio and video features. Recall that `unige-cvml5` is a randomly generated run. Most runs failed to yield results appreciably better than this random baseline. The best scoring approaches exploited the speech recognition transcripts, in particular, the occurrence of pronouns reflecting user directed speech and the use of words with high effective ratings.

Table 2. Narrative peak detection results

run ID	**point-based**	**peak-based** $>$ **1 assessor** (“personal peaks”)	**peak-based** $>$ **2 assessors** (“pair peaks”)	**peak-based** $>$ **3 assessors** (“general peaks”)
`duotu09fix`	47	28	8	4
`duotu09ind`	55	38	12	2
`duotu09rep`	30	21	7	0
`duotu09pro`	**63**	**44**	17	4
`duotu09rat`	59	33	**18**	**6**
`unige-cvml1`	39	32	6	0
`unige-cvml2`	41	30	11	2
`unige-cvml3`	42	31	8	0
`unige-cvml4`	43	31	9	0
`unige-cvml5`	43	32	8	3
`uaic-run1`	33	26	7	2
`uaic-run2`	41	29	10	3
`uaic-run3`	33	24	7	2

Because of the newness of the Narrative Peak Detection Task, the method of scoring is still a subject of discussion. The scoring method was designed such that algorithms were given as much credit as possible for agreement between the peaks they hypothesized and the peaks chosen by the annotators. See the papers of individual participants [7] [10] for some additional discussion.

4 Linking Task

4.1 Task

The Linking Task, also called "Finding Related Resources Across Languages," involves linking episodes of the *Beeldenstorm* documentary (Dutch language) to Wikipedia articles about related subject matter (English language). This task was new in 2009. Participants were supplied with 165 *multimedia anchors*, short (ca. 10 seconds) segments, pre-defined in the 45 episodes that make up the *Beeldenstorm* collection. For each anchor, participants were asked to automatically generate a list of English language Wikipedia pages relevant to the anchor, ordered from the most to the least relevant.

Notice that this task was designed by the task organizers such that it goes beyond a named-entity linking task. Although a multimedia anchor may contain a named entity (e.g., a person, place or organization) that is mentioned in the speech channel, the anchors have been carefully chosen by the task organizers so that this is not always the case. The topic being discussed in the video at the point of the anchor may not be explicitly named. Also, the representation of a topic in the video may be split between the visual and the speech channel.

4.2 Evaluation

The ground truth for the linking task was created by the assessors. We adapted the four graded relevance levels used in [6] for application in the Linking Task. Level 3 links are referred to as *primary links* and are defined as "highly relevant – the page is the single page most relevant for supporting understanding of the video in the region of the anchor." There is only a single primary link per multimedia anchor representing the one best page to which that anchor can be linked. Level 2 links are referred to as *secondary links* and are defined as "fairly relevant – the page treats a subtopic (aspects) of the video in the region of the anchor." The final two levels: Level 1 (defined as: "marginally relevant, the page is not appropriate for the anchor") and Level 0 (defined as "irrelevant, the page is unrelated to the anchor"), were conflated and regarded as irrelevant. Links classified as Level 1 are generic links, e.g., "painting," or links involving a specific word that is mentioned, but is not really central to the topic of the video at that point.

Primary link evaluation For each video, the primary link was defined by consensus among three assessors. The assessors were required to watch the entire episode so as to have the context to decide the primary link. Primary links were evaluated using recall (correct links/total links) and Mean Reciprocal Rank (MRR).

Related resource evaluation For each video, a set of related resources was defined. This set necessarily includes the primary link. It also includes other secondary

links that the assessors found relevant. Only one assessor needed to find a secondary link relevant for it to be included. However, the assessors agreed on the general criteria to be applied when chosing a secondary link. Related resources were evaluated with MRR. The list of secondary links is not exhaustive, for this reason, no recall score is reported.

4.3 Techniques

Centre for Digital Video Processing, Dublin City University, Ireland (see also [3]) The words spoken between the start point and the end point of the multimedia anchor (as transcribed in the speech recognition transcript) were used as a query and fired off against an index of Wikipedia. For `dcu_run1` and `dcu_run2` the Dutch Wikipedia was queried and the corresponding English page was returned. Stemming was applied in `dcu_run2`. Dutch pages did not always have corresponding English pages. For `dcu_run3`, the query was translated first and fired off against an English language Wikipedia index. For `dcu_run4` a Dutch query expanded using psuedo-relevance feedback was used.

TNO Information and Communication Technology, Netherlands (see also [14]) A set of existing approaches were combined in order to implement a sophisticated baseline to provide a starting point for future research. A wikify tool was used to find links in the Dutch speech recognition transcripts and in English translations of the transcripts. Particular attention was given to proper names, with one strategy giving preference to links to articles with proper-name titles and another strategy ensuring that proper name information was preserved under translation.

4.4 Results

The results of the task are reported in Table 3 (primary link evaluation) and Table 4 (related resource evaluation). The best run used a combination of different strategies, referred to by TNO as a "cocktail." The techniques applied by DCU achieved a lower overall score, but demonstrate that in general it is better not to translate the query, but rather to query Wikipedia in the source language and then cross over to the target language by using Wikipedia's own links article-level links between languages. Note that the difference is in reality not as extreme as suggested by Table 3 (i.e., by `dcu_run1` vs. `dcu_run3`). A subsequent version of the `dcu_run3` experiment (not reported in Table 3) that makes use of a version of Wikipedia that has been cleaned up by removing clutter (e.g., articles scheduled for deletion and meta-articles containing discussion) achieves a MRR of 0.171 for primary links. Insight into the difference between the DCU approach and the TNO approach is offered by an analysis that makes a query-by-query comparison between specific runs and average performance. DCU runs provide an improvement over average performance for more queries than TNO run [14].

Table 3. Linking results: Primary link evaluation. Raw count correct and MRR

run ID	raw	MRR
dcu_run1	44	0.182
dcu_run2	44	0.182
dcu_run3	13	0.056
dcu_run4	38	0.144
tno_run1	57	0.230
tno_run2	55	0.215
tno_run3	**58**	**0.251**
tno_run4	44	0.182
tno_run5	47	0.197

Table 4. Linking results: Related resource evaluation. MRR

run ID	MRR
dcu_run1	0.268
dcu_run2	0.275
dcu_run3	0.090
dcu_run4	0.190
tno_run1	0.460
tno_run2	0.428
tno_run3	**0.484**
tno_run4	0.392
tno_run5	0.368

5 Conclusions and Outlook

In 2009, VideoCLEF participants carried out three tasks, Subject Classification, Narrative Peak Detection and Finding Related Resources Across Languages. These tasks generate enrichment for spoken content that can be used to provide improvement in multimedia access and retrieval.

With the exception of the Narrative Peak Detection Task, participants concentrated largely on features derived from the speech recognition transcripts and did not exploit other audio information or information derived from the visual channel. Looking towards next year, we will continue to encourage participants to use a wider range of features.

We see the Subject Classification Task as developing increasingly towards a tag recommendation task, where systems are required to assign tags to videos. The tag set might not necessarily be known in advance. We expect that the formulation of this task as an information retrieval task will continue to prove useful and helpful, although we wish to move to metrics for evaluation that will better reflect the utility of the assigned tags for real-world search or browsing.

In 2010, VideoCLEF will change its name to MediaEval[13] and its sponsorship will be taken over by PetaMedia,[14] a Network of Excellence dedicated to research and development aimed to improve multimedia access and retrieval. In 2010, several different data sets will be used. In particular, we introduce data sets containing creative commons data collected from the Web (predominantly English language) that will be used in addition to data sets from *Beeld & Geluid* (predominantly Dutch data). We will offer a tagging task, and affect task and a linking task as in 2009, but we will extend our task set to include new tasks, in particular: geo-tagging and multimodal passage retrieval. The goal of MediaEval is to promote cooperation between sites and projects in the area of the benchmarking, moving towards the common aim of "Innovation and Education via Evaluation."

[13] http://www.multimediaeval.org/
[14] http://www.petamedia.eu/

6 Acknowledgements

We are grateful to TrebleCLEF,[15] a Coordination Action of European Commission's Seventh Framework Programme for a grant that made possible the creation of a data set for the Narrative Peak Detection Task and the Linking Task. Thank you to the University of Twente for supplying the speech recognition transcripts and to the Netherlands Institute of Sound and Vision (*Beeld & Geluid*) for supplying the video. Thank you to Dublin City University for providing the shot segmentation and keyframes and also for hosting the team of Dutch-speaking video assessors during the Dublin Days event. We would also like to express our appreciation to Michael Kipp for use of the Anvil Video Annotation Research Tool. The work that went into the organization of VideoCLEF 2009 has been supported, in part, by PetaMedia Network of Excellence and has received funding from the European Commission's Seventh Framework Programme under grant agreement no. 216444.

References

1. J. Calic, S. Sav, E. Izquierdo, S. Marlow, N. Murphy, and N. O'Connor. Temporal video segmentation for real-time key frame extraction. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.
2. T.-A. Dobrilă, M.-C. Diaconaşu, I.-D. Lungu, and A. Iftene. UAIC: Participation in VideoCLEF task. In *Experiments Multilingual Information Access Evaluation Vol. II Multimedia Experiments (this volume)*. Springer Berlin/Heidelberg, 2010.
3. Á. Gyarmati and G. J. F. Jones. When to cross over? Cross-language linking using Wikipedia for VideoCLEF 2009. In Carol Peters et al., editor, *Experiments Multilingual Information Access Evaluation Vol. II Multimedia Experiments (this volume)*. Springer Berlin/Heidelberg, 2010.
4. A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154, Feb. 2005.
5. M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of the International Conference on Semantic and Digital Media Technologies (SAMT)*, 2007.
6. J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.
7. J.J.M. Kierkels, M. Soleymani, and T. Pun. Identification of narrative peaks in video clips: Text features perform best. In *Experiments Multilingual Information Access Evaluation Vol. II Multimedia Experiments (this volume)*. Springer Berlin/Heidelberg, 2010.
8. M. Kipp. Anvil – a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech*, pages 1367–1370, 2001.
9. J. Kürsten and M. Eibl. Video classication as IR task: Experiments and observations. In *Experiments Multilingual Information Access Evaluation Vol. II Multimedia Experiments (this volume)*. Springer Berlin/Heidelberg, 2010.

[15] http://www.trebleclef.eu/

10. M. Larson, B. Jochems, E. Smits, and R. Ordelman. A cocktail approach to the VideoCLEF09 linking task. In *Experiments Multilingual Information Access Evaluation Vol. II Multimedia Experiments (this volume)*. Springer Berlin/Heidelberg, 2010.
11. M. Larson, E. Newman, and G. J. F. Jones. Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual content. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access. Lecture Notes in Computer Science*, volume 5706, pages 906–917. Springer Berlin/Heidelberg, 2009.
12. P. Pecina, P. Hoffmannová, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF-2007 Cross-Language Speech Retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 674–686, Berlin, Heidelberg, 2008. Springer-Verlag.
13. J.M. Perea-Ortega, A. Montejo-Ráez, M.T. Martín-Valdivia, and L.A. Ureña López. Using Support Vector Machines as learning algorithm for video categorization. In *Experiments Multilingual Information Access Evaluation Vol. II Multimedia Experiments (this volume)*. Springer Berlin/Heidelberg, 2010.
14. S. Raaijmakers, C. Versloot, and J. de Wit. A cocktail approach to the VideoCLEF09 linking task. In *Experiments Multilingual Information Access Evaluation Vol. II Multimedia Experiments (this volume)*. Springer Berlin/Heidelberg, 2010.
15. A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval (MIR)*, pages 321–330, New York, NY, USA, 2006. ACM.
16. B. Wrede and E. Shriberg. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proceedings of Eurospeech*, pages 2805–2808, 2003.