Boundary error analysis and categorization in the TRECVID news story segmentation task

Joaquim Arlandis[1], Paul Over[2], and Wessel Kraaij[3]

[1] Departament d'Informàtica de Sistemes i Computadors
Universitat Politècnica de València[*]
Cami de Vera s/n, 46022 València, Spain
jarlandi@disca.upv.es

[2] Retrieval Group
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA
over@nist.gov

[3] Department of Data Interpretation
Information Systems Division
TNO Science & Industry
2600 AD Delft, the Netherlands
wessel.kraaij@tno.nl

Abstract. In this paper, an error analysis based on boundary error popularity (frequency) including semantic boundary categorization is applied in the context of the news story segmentation task from TRECVID[4]. Clusters of systems were defined based on the input resources they used including video, audio and automatic speech recognition. A cross-popularity specific index was used to measure boundary error popularity across clusters, which allowed goal-driven selection of boundaries to be categorized. A wide set of boundaries was viewed and a summary of the error types is presented. This framework allowed conclusions about the behavior of resource-based clusters in the context of news story segmentation.

1 Introduction

Digital video indexing, retrieval, and presentation systems can require a variety of segmentation procedures. In some cases, like news videos, shots, which can be detected well automatically, can usefully be grouped into *stories.* This segmentation is more subjective as it depends more on the meaning of the video material and resource-dependent structure. Like shots, stories make for natural units of retrieval, navigation, summarization, etc.

Given a set of human judgments about where stories begin, one can test systems designed to automatically detect story boundaries. System performance

[*] Work partially supported by the PII of the Universitat Politècnica de València

[4] *TREC Video Retrieval Evaluation*, http://www-nlpir.nist.gov/projects/trecvid/

can be measured in terms of the degree to which the system finds all and only the actual boundaries. Such scoring is useful for comparison of systems' performance summarized over many test videos and stories, but it hides much information about how and why any given system or group of systems achieved a particular score. In this paper we are concerned with the details of system performance – in some of the errors systems commit and the extent to which these are predictable based on types and attributes of the data and/or the system (approach).

There is little earlier work in error analysis and categorization in video story segmentation, particularly in news. Hsu *et al.* [1] present an interesting categorization of types of transitions between stories using the TRECVID 2003 data collections, and they present percentages of error types committed by different systems and parameterizations from their own approaches. They labeled 795 story boundaries. Chua *et al.* [2] distinguish between errors found due to low-level feature misrecognition (including single objects as face, anchor, reporter, motion, audio and text) and those due to mid-level feature errors (including patterns of transitions between single objects). The former may cause the latter. Also they state that an important source of errors is related to the segmentation of stories into "substories" such as different sports within a sports section.

In this paper, an error analysis based on boundary error popularity (frequency) including semantic boundary categorization is applied in the context of the news story segmentation task from TRECVID 2003 & 2004. Clusters of systems were defined from the type of input resources they used including video, audio and automatic speech recognition. A specific index to measure and analyze boundary error popularity across clusters was defined, which allows goal-driven selection of a manageable subset of boundaries to be categorized. A wide set of boundaries was viewed and a summary of the error types along with conclusions are presented. This framework can be also applied to other segmentation tasks.

2 Story segmentation in TRECVID

TRECVID aims to assess the performance of video retrieval systems developed by the participants [3]. In 2003 and 2004 TRECVID included a specific task for story segmentation of news. The evaluation used CNN Headline News[5] and ABC World News Tonight[5] US broadcast news from 1998, in MPEG-1 format, that was collected for TDT[5] [4]. A news story was defined as a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses [4]. Non-news segments were labeled as "miscellaneous", merged together when adjacent, and annotated as one single story. The 2003 story test collection used for evaluation was composed of 52 hours of news, containing 2,929 story boundaries. In 2004, the test collection from 2003 could be used for system development and a new test collection included 59 hours and 3,105 story boundaries. The number of stories found per video varied between 14 and 42. Stories often span multiple shots but shot and story boundaries do

[5] The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

not necessarily coincide. ASR (automatic speech recognizer) output from videos was provided to participants by LIMSI [5].

With TRECVID 2003/2004's story segmentation task, three types of runs (conditions) were required from participants depending on the sort of resource used: Condition 1 - using audio and video (AV), Condition 2 - using AV and ASR, and Condition 3 - using ASR only.

Participating groups submitted at least one run in each condition. A *run* is the output of a system containing a list of times at which story boundaries were expected to be found. System performances were measured in terms of precision and recall [6]. Story boundary recall (R) was defined as the number of reference boundaries detected, divided by total number of reference boundaries. Story boundary precision (P) was defined as the total number of submitted boundaries minus the total amount of false alarms, divided by total number of submitted boundaries. In addition, the F-measure, ($F = (2 * P * R)/(P + R)$), was used to compare overall performance across conditions and systems.

3 Error analysis

In the present section, an analysis of the erroneous boundaries resulting from TRECVID 2003 and 2004 evaluations was applied to the three conditions – clusters of systems – described above. First, the procedure of selection of a representative set of systems and their global results are presented. Then popularity-based indexes are described. In the two last subsections, popularity and cross-popularity indexes are used to evaluate and interpretations are presented.

3.1 System selection and overall results

Although each group participating in TRECVID could submit up to 10 runs (sets of results), at least one run per condition, in fact, 8 groups submitted a total of 41 runs in 2003, and 8 groups, as well, submitted 50 runs in 2004. According to the documentation provided by the groups[6], in almost all cases, runs from each condition and group used the same approach by combining different algorithm modules or parameterizations. Furthermore, the approaches followed by the groups were different, except for a very small number of runs. Within a group, runs from AV+ASR usually came from a combination of the approaches used in their AV and ASR runs. Because of all of that, selecting an representative subset of runs in order to get robust conclusions for error analysis and categorization was advisable.

Because the test set varies each year, one independent subset of runs from each year was selected. Within a year, the selection procedure was as follows: First, runs with similar approaches were rejected, keeping the higher F-valued ones. That included selecting a maximum of one run from each team and condition, and rejecting runs from different groups with similar approaches as documented in the papers[6] – so that independent behavior could be expected from

[6] http:///www.itl.nist.gov/iaui/894.02/projects/tvpubs/tv.pubs.org.html

Table 1. Results by condition each year. Recall and precision are averages by condition. The total number of boundaries in 2003 data was 2929 and in 2004 was 3105.

TRECVID 2004

Condition	# Sys	Recall	Misses (% truth)	Precision	FA	FA Uniques
(1)AV	6	0.566	2828 (91.1%)	0.403	33208	85.4%
(2)AV+ASR	6	0.489	2988 (96.2%)	0.550	7001	86.7%
(3)ASR	6	0.460	2984 (96.1%)	0.382	22096	78.7%
All	18	0.505	3097 (99.7%)	0.445	44710	61.2%

TRECVID 2003

Condition	# Sys	Recall	Misses	Precision	FA	FA Uniques
(1)AV	5	0.587	2405 (82.1%)	0.538	18562	93.4%
(2)AV+ASR	5	0.474	2659 (85.6%)	0.654	3350	91.6%
(3)ASR	5	0.446	2718 (92.8%)	0.478	8588	88.7%
All	15	0.502	2832 (96.7%)	0.557	23790	71.5%

different runs. So, in what follows, a run will be considered as a distinctive system. Second, systems not accomplishing a minimum of quality performances were rejected. A cutting threshold of 0.2 was applied over the F-measure so that the popularity of the boundaries was expected to capture more precisely the behavior of the most competitive systems. Finally, based on their lower F-value, two more systems were rejected, to preserve the same number in order to allow stronger conclusions from the data.

Table 1 shows the number of systems finally selected along with overall results produced by the selected systems for each condition. Recall and precision measures favor AV and AV+ASR. AV+ASR systems were more conservative than AV systems, judged by their lower recall and higher precision. Relative performance among the three conditions was the same for both years.

The missing boundaries and the false alarms (FA) shown per condition are the ones contributed by at least one system within the condition, and depend not only on the system quality (reported by recall and precision), but also on the number of systems assessed, and the boundary popularity.

False alarms are boundaries erroneously detected by systems and, as shown in Table 1, are expected to be more frequent than misses, and mostly unique. Nevertheless, compared to conditions 1 and 3 clusters, the low number of false alarms produced by systems from condition 2, along with the high percentage of uniques, suggest that the combination of AV and ASR resources contributes to increase the systems' precision compared to the single resource-based clusters.

3.2 Popularity-based indexes

The process of visual boundary categorization allows classification of boundaries into several types defined by any given set of features. Categorization of the most or least popular boundary errors in a cluster of systems can shed light on the general behavior of the systems within that cluster, and can provide valuable information for system developers. Comparisons across clusters could also be done. Boundaries that are frequently reported by most of systems in one cluster but by the fewest in other cluster are potentially interesting because they show differences across systems from different clusters. Given that, these can be considered as *target boundaries* for categorization.

In order to measure the degree of boundary error popularity across clusters a specific index was used. Further, a framework for selecting target boundaries was defined too. The *popularity* $p_c(b)$ of a boundary b in a cluster c can be defined as the number of systems from cluster c reporting the boundary b. The normalized popularity

$$P_c(b) = \frac{p_c(b)}{|c|}$$

where $|c|$ is the number of systems in the cluster c, can be used to compare popularities between clusters with different size.

Since, for a given boundary, different popularities can be reported by different clusters, the following index is named *cross-popularity* and can be used to measure the degree of high popularity of a boundary b in a cluster c versus its low popularity in another cluster d

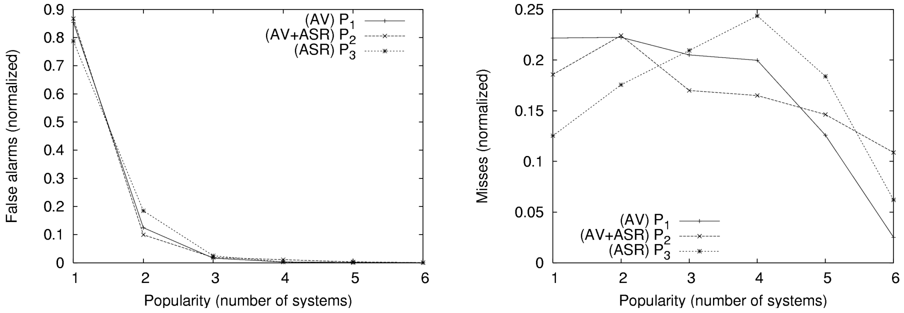
$$P_{c,d}(b) = P_c(b) - P_d(b)$$

Cross-popularity index ranges from -1 to 1. Boundaries with values over 0 are those more popular within cluster c than within cluster d, and can be represented as $P^+_{c,d}$. Negative values are assigned to the more popular boundaries within cluster d than within c, and can be represented as $P^-_{c,d}$ (notice that $P^+_{c,d} = P^-_{d,c}$). Values around 0 correspond to boundaries with similar popularities.

In error analysis, boundaries having high popularity within one cluster and low popularity within another cluster can be considered as hard for the first as well as easy for the second one. Given a set of clusters, the distribution of their *cross-popularity* values can be used to compare their behavior. Right and left tails of the distributions can be used to identify such a target boundaries for which the clusters perform in such a different way.

Given a set of erroneous boundaries $B = \{b_1, b_2, \dots, b_n\}$, a set of clusters $C = \{c_1, c_2, \dots, c_n\}$, and a number of evaluated systems belonging to some of the clusters of C, the following can be considered as a target boundary groups for error categorization:

- *Most popular boundaries within all clusters.* Those boundaries are the ones associated with the highest values of $P_C(B)$.
- *Most popular boundaries within one cluster c_i.* Those boundaries are the ones associated with the highest values of $P_{c_i}(B)$.
- *Most popular boundaries within a cluster c_i, least popular in other cluster c_j.* A number of boundaries with highest values $P_{c_i,c_j}(B)$ can be targeted.
- *Most popular boundaries within a cluster c_i, least popular in a subset of clusters $C' = \{c_k, \dots, c_m\}, C' \subset C$.* Boundaries with highest values $P_{c_i,C'}(B)$ can be targeted.



(a) False alarms

(b) Misses

Fig. 1. Normalized histograms of popularity for (a) false alarms, and (b) misses from TRECVID 2004. The figures show a very high percentage of false alarms having low popularity versus a high percentage of misses having a clearly higher popularity. Figure (b) shows different behavior within each condition.

Boundaries with highest values of P_{c_i,c_j} are the ones for which systems from a cluster c_i work significantly worse than systems from c_j. The lowest $P_{c_i,C}(B)$ valued boundaries can also be targeted because they describe those boundaries easy for a cluster when hard for others. Histograms from popularity and cross-popularity can be used to select target boundaries.

3.3 Popularity analysis

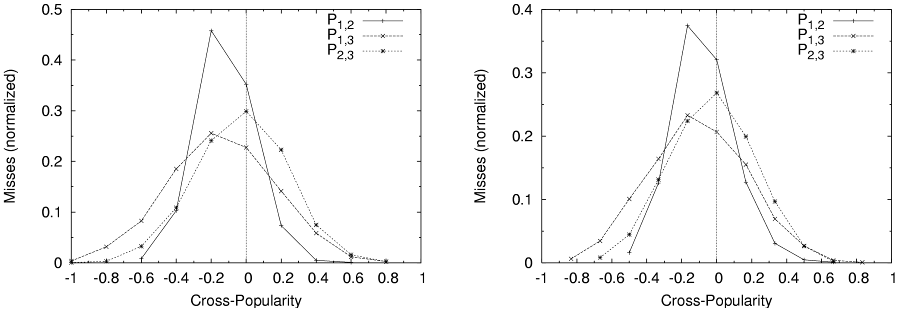
Error popularity can be used to analyze system behavior within a condition. An analysis of the popularity distributions from each condition can be made over two set of errors: missing boundaries and false alarms.

Non-significant differences observed between some false alarms led us to consider applying a clustering procedure to present consistent results: 1) some boundaries were removed to avoid boundaries from the same system closer than ±1 s; 2) one cluster around each boundary was created grouping boundaries within an interval of ±1 s; and 3) clusters containing boundaries included in other group were removed while keeping the ones with higher popularity. That ensured no boundary was contributed more than once from a system.

Figure 1 plots the popularity histograms of false alarms (cluster sizes) and misses from TRECVID 2004. Very similar results were reported from 2003 data.

Based on the data from Figure 1(a), the percentage of false alarms reported only by one or two systems was over 97.3% for all three conditions. The very low percentage of false alarms having significant high popularity means a deeper analysis is not likely to be so productive, compared with an analysis of missing boundaries, so no further analysis of false alarms was done at this time.

Regarding misses, Figure 1(b) shows significant percentages of high popularity. Different behavior between conditions AV and ASR is also evident. Condition



(a) TRECVID 2003

(b) TRECVID 2004

Fig. 2. Histograms of cross-popularity values of missing boundaries across conditions (1:AV, 2:AV+ASR, 3:ASR). The curves show the same behavior both years.

AV obtained a higher number of low-popular misses while condition ASR obtained a higher number of high-popular ones. That indicates more independent performances coming from systems within AV than systems within ASR. Systems in cluster AV+ASR reported a similar number of misses for each popularity level and the highest number with maximum popularity. The high popularity observed on misses makes this an interesting set to which to apply cross-popularity study.

3.4 Cross-popularity analysis

A cross-popularity analysis of missing boundaries was made. Figures 2(a) and (b) show the distribution of the cross-popularity values of misses across conditions for the selected sets of systems from TRECVID 2003 and 2004 evaluations, respectively. The figures show the same behavior across conditions both years.

Popularity of AV versus ASR misses is shown by $P_{1,3}$. This curve reports the higher cross-popularity values in both years, particularly in the left tails where the more difficult boundaries for ASR and less difficult ones for AV are located. The $P_{1,2}$ curve is the sharpest one. That means that the misses' popularity was more similar across AV and AV+ASR than across any other resources because of the high number of values close to zero. Thus, $P_{1,2}$ and $P_{2,3}$ curve shapes indicate that AV+ASR systems shared a higher number of misses with AV-only systems than with ASR-only systems, what suggests that AV resource had more weight than ASR in the AV+ASR algorithms.

The curves $P_{1,2}$, $P_{1,3}$ show a negative asymmetry and thus a higher density is located under zero[7]. That indicates that a higher number of boundaries were more difficult (more popular) for conditions AV+ASR and ASR than for condition AV and this is directly related to the higher average recall reported by AV systems (Table 1).

[7] For more clarity, the polarity of the cross-popularity was chosen in order to show the highest dissimilarities as negative values.

The tails $P^-_{1,3}$ and $P^-_{2,3}$ show higher values than any other tails. That means that more boundaries were harder for condition 3 than other boundaries were for other conditions. That behavior becomes significant over 0.4 and under -0.4 indexes – suggesting ASR by itself to be the most limited resource, i.e., the probability that a given boundary was missed by at least 40% more of systems from a condition than from other condition is significantly higher for ASR-only.

Concerning misses, very similar observations were made for both years even though test set and evaluated systems were different. Boundaries on the right and left tail of each curve are potentially interesting candidates for visual examination and categorization as the hardest in one condition compared to another condition.

4 Boundary error categorization

Boundary error categorization has to be driven by the pursued goals. On the one hand, selecting unsuitable boundaries can lead to partial conclusions. On the other hand, selecting more boundaries than needed can turn out as unnecessary time spent when handling large amounts of video data. Selecting error boundaries from popularity and cross-popularity indexes can lead to a specific categorization based on outstanding performance differences across resource-based system clusters. Given the three predefined conditions $C = \{1, 2, 3\}$, the following groups of errors were considered to select candidate misses: 1) P_C , 2) P_i , $i \in C$, and 3) $P_{i,j}$, $i, j \in C$.

Boundary categorization can be made at different levels: on low-level features, e.g., transitions involving presence or absence of faces, sorts of camera motions or background sounds, on mid-level features like segment or dialog structures found in story transitions, as well as, on high-level semantics concerning news content like the characterization of type and subtype of the linked stories. In this paper, the categorization level was in terms of semantic content by classifying the stories into four groups: regular news, weather, briefs, and miscellaneous, and by defining the following categories of transitions:

- NN: Regular news followed by regular news.
- NW: Regular news followed by CNN weather section.
- NM: News followed by misc or misc followed by news. It includes beginnings, ends, breaks, and changes of sections in the news show.
- BB: Briefs section. Transitions between short pieces including headlines, CNN and ABC financial briefs and brief segments in CNN sports section.

Furthermore, three binary features were evaluated:

- Trigger: Trigger phrases. A binary feature indicating the presence of a standard news trigger phrase denoting a change of story.
- Shot: Shot boundary overlaps along with story boundary.
- CNN: Boundaries from CNN videos. The remaining boundaries correspond to ABC broadcast videos.

Table 2. Results of the popularity-based categorization for each boundary group. The table shows number of misses categorized (Viewed), percentage each category, and percentages having specific binary features. Number of boundaries with popularity=1.0 and averages of cross-popularity of the selected boundaries are also shown.

	Categories (%)				Binary features (%)			Totals and averages	
Popularity	NN	NW	NM	BB	Trigger	Shot	CNN	Viewed	Popularity=1.0
P_C	75	0	10	15	15	25	65	20	20
P_1	66	0	32	10	44	48	64	50	71
P_2	52	0	16	34	20	62	78	50	325
P_3	46	0	8	52	8	64	70	50	185
Cross-pop	NN	NW	NM	BB	Trigger	Shot	CNN	Viewed	Cross-Popularity
$P^{+}_{1,2}$	12	35	53	0	94	100	76	17	0.53
$P^{-}_{1,2}$	76	0	14	10	2	88	80	49	0.50
$P^{+}_{1,3}$	40	8	52	0	92	88	62	50	0.55
$P^{-}_{1,3}$	72	0	34	0	10	100	80	50	0.73
$P^{+}_{2,3}$	66	0	28	14	52	92	72	50	0.52
$P^{-}_{2,3}$	40	0	60	0	34	98	58	50	0.58

Table 2 shows the type of errors and frequencies found for each boundary group for a number of viewed boundaries from TRECVID 2004 data. The proportion of boundary types in the test collection was unknown. A maximum of 50 boundaries in each group were selected for categorization. For cross-population targeted boundaries only those over 0.5 were selected. For each cross-population group, $P_{i,j}$, the average of the cross-population index of the selected boundaries is shown. Categorization was made by viewing clips from 20 seconds before to 20 after the truth boundary.

As shown in Table 2, just 20 boundaries were missed by all systems from all conditions. This is 0.65% of the total truth boundaries. Those were mostly regular transitions between news stories, with low percentages of trigger phrases and shot transitions overlapped. This behavior could be expected and no relevant conclusions can be obtained from this.

Results in the group of the most popular boundaries within a condition shown in Table 2 suggest some differences between the three resource-based conditions. AV-only systems failed to find a significant number of NM-boundaries while systems using ASR, particularly ASR-only, revealed a weakness in detecting BB boundaries instead of NM or even NN.

From the viewpoint of cross-popularity, which focuses on boundaries which discriminate maximally among the three conditions, Table 2 shows that the percentages of NN-boundaries in $P^{+}_{1,2}$ (12%) and $P^{+}_{1,3}$ (40%) are clearly lower than the ones in $P^{-}_{1,2}$ (76%) and $P^{-}_{1,3}$ (72%). That means that AV got a lower number of high-valued cross-popularity misses than ASR and AV+ASR, so that systems from AV identified these NN-boundaries better than the remaining ones. Taken into account the very low percentages of trigger phrases (2% and 10%) from $P^{-}_{1,2}$ and $P^{-}_{1,3}$, and the high ones from $P^{+}_{1,2}$ and $P^{+}_{1,3}$, the use of ASR clearly leds to increase missing NN-boundaries more than other boundaries when no trigger phrases are available.

Also looking at cross-popularity, NM-boundaries seem significantly easier for systems using AV+ASR (14% and 28%) when harder for any other (53% and 60%). That indicates that, for these boundaries, the combination of ASR and AV resources improved performance compared to using a single resource. On other hand, due to the fact that BB-boundaries usually include change of shot, this feature probably helps AV systems to be more precise than others using ASR on BB-boundaries. Also notice that NW-boundaries were found very frequently in the tail of the distribution $P^{+}_{1,2}$ (35%).

The data shown in Table 2 and the conclusions extracted should be considered as relative to precision and recall measured from the system results (Table 1) which could be affected by systems tuning. Nevertheless, it can be assumed the systems were designed to maximize precision and recall and represent a real sample of the state-of-the-art.

5 Conclusions

Results of boundary error popularity from the TRECVID 2003 & 2004 news story segmentation task were analyzed. The analysis was targeted to find behavior patterns in clusters of systems defined by the input resource they used, and included semantic categorization of news boundary errors. An error cross-popularity index was defined and used to draw conclusions. Very similar observations were made both years even though test set and evaluated systems were different. Finally, categorization provided information about what kind of boundaries were harder for a cluster while easier for other and suggested that behavior can be predicted as a function of the input resources used. That can point out opportunities for system improvements.

References

[1] Hsu, W.H-M., Chang, S-F.: Generative,Discriminative, and Ensemble Learning on Multi-modal Perceptual Fusion toward News Video Story Segmentation. IEEE International Conference on Multimedia and Expo (2004)

[2] Chua, T. S., Chang, S. F., Chaisrn, L., Hsu, W.: Story Boundary Detection in Large Broadcast News Video Archives - Techniques, Experience and Trends. Proceedings of the 12th annual ACM conference on Multimedia (MM'04) (2004) 656–659

[3] Kraaij, W., Smeaton, A. F., Over, P., Arlandis, J.: TRECVID 2004 - An Overview. TREC Video Retrieval Evaluation Online Proceedings, http://www-nlpir.nist.gov/projects/trecvid/tv.pubs.org.html (2003)

[4] Wayne, C.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. Language Resources and Evaluation Conference (LREC) (2000) 1487–1494

[5] Gauvain, J. L., Lamel, L., Adda, G.: The LIMSI Broadcast News Transcription System. Speech Communication, **37(1-2)** (2002) 89–108

[6] Voorhees, E. M., Harman, D. K.: Common Evaluation Measures. Proceedings of the Tenth Text Retrieval Conference (TREC) A-14, http://trec.nist.gov/pubs/trec10