TRECVID 2007 Collaborative Annotation using Active Learning

Georges Quénot
Multimedia Information Retrieval Group

Laboratoire d'**I**nformatique de **G**renoble

November 5, 2007

Outline

- Active learning.
- Previous work: evaluation of active learning strategies.
- TRECVID 2007 collaborative annotation.
- Conclusion.

Active learning

Active learning basics

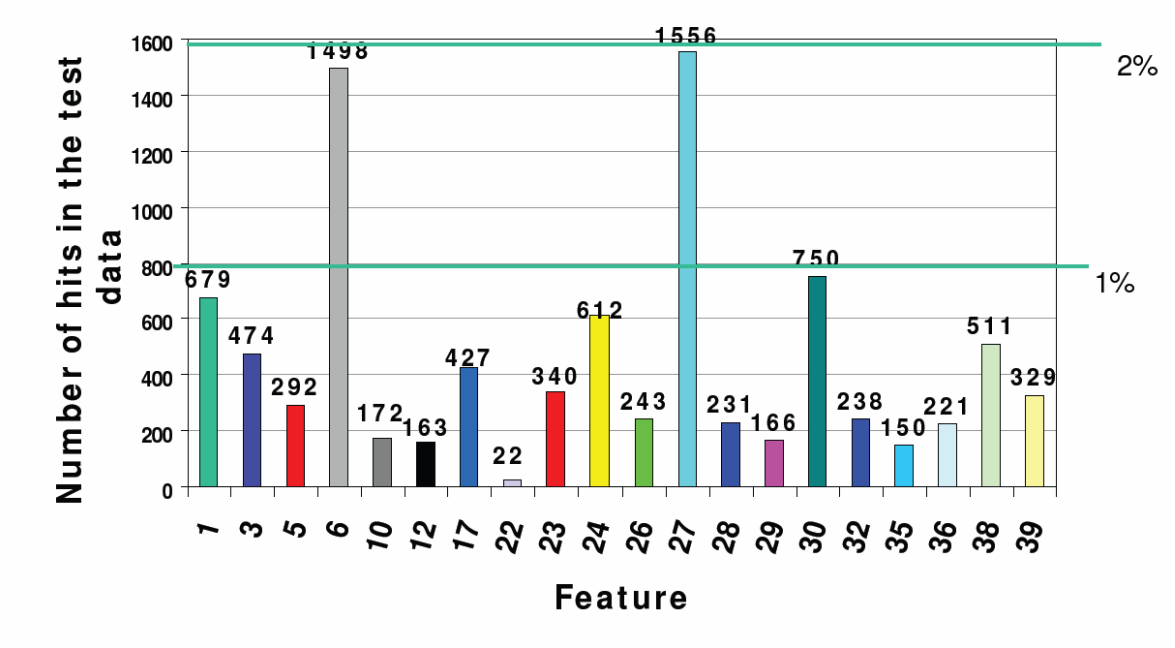
- Concept classification → “Semantic gap” problem.
- Improve classification performance ?
  - Optimize the model and the train/predict algorithm.
  - Get a large training set: quantity, quality, …
- Cost of corpus annotation:
  - Getting large corpora is (quite) easy and cheap (already there).
  - Getting annotations on it is costly (human intervention).
- Active learning:
  - Use an existing system and heuristics for selecting the samples to annotate → need of a classification score.
  - Annotate first or only the samples that are expected to be the most informative for system training → various strategies.
  - Get same performance with less annotations and/or get better performance with the same annotation count.

Active learning strategies

- Query by committee [Seung, 1992]: choose the samples which maximize the disagreement amongst systems.
- Uncertainty sampling [Lewis, 1994]: choose the most uncertain samples, tries to increase the sample density in the neighborhood of the frontier between positives and negatives → improve the system's precision.
- Relevance sampling: choose the most probable positive samples, tries to maximize the size of the set of positive samples (positive samples are most often sparse within the whole set and finding negative samples is easy).
- Choose the farthest samples from already evaluated ones, tries to maximize the variety of the evaluated samples → improve the system's recall.
- Combinations of these, e.g. choose the samples amongst the most probable ones *and* amongst the farthest from the already evaluated ones.
- Choose samples by groups which maximize the expected global knowledge gain [Souvanavong, 2004].

Previous work: evaluation of active learning strategies using TRECVID 2005-2006 data and metrics

Frequency of hits by features
[from Paul Over and Wessel Kraaij, 2006]

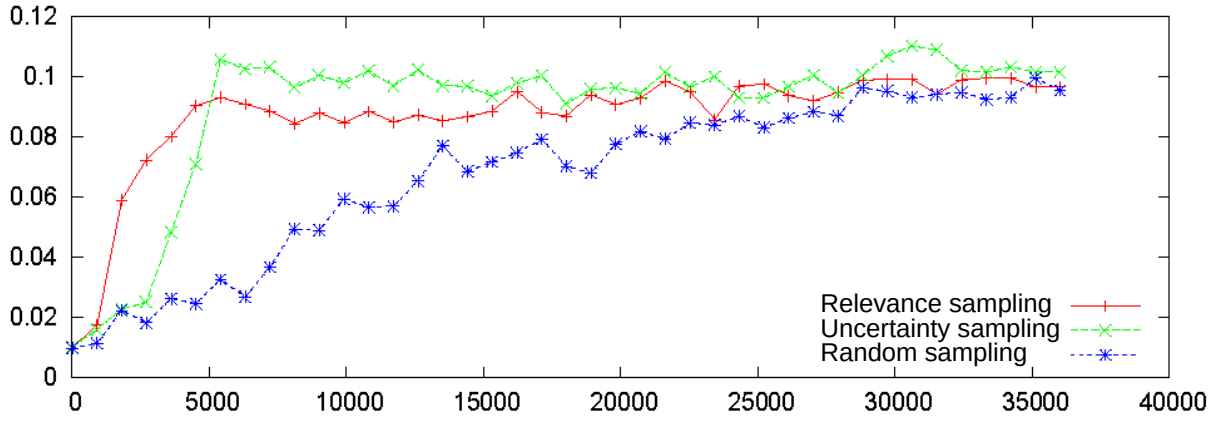


Active learning evaluations

- Use of simulated active learning.
- System: networks of SVM classifiers for multimodal fusion [Ayache, TRECVID 2006] (global performance slightly above median in TRECVID 2006 and 2007).
- The training set is restricted to the shots that contain speech → 36014 samples.
- Default step size: 1/40th of the training set → 900 samples.
- Cold start with 10 positive and 20 negative all randomly selected.
- Evaluation of:
  - Strategies: random, relevance and uncertainty sampling
  - Relation with concept difficulty
  - Effect of the step size
  - Training set size
  - Finding rates for positive samples
  - Precision versus recall compromise

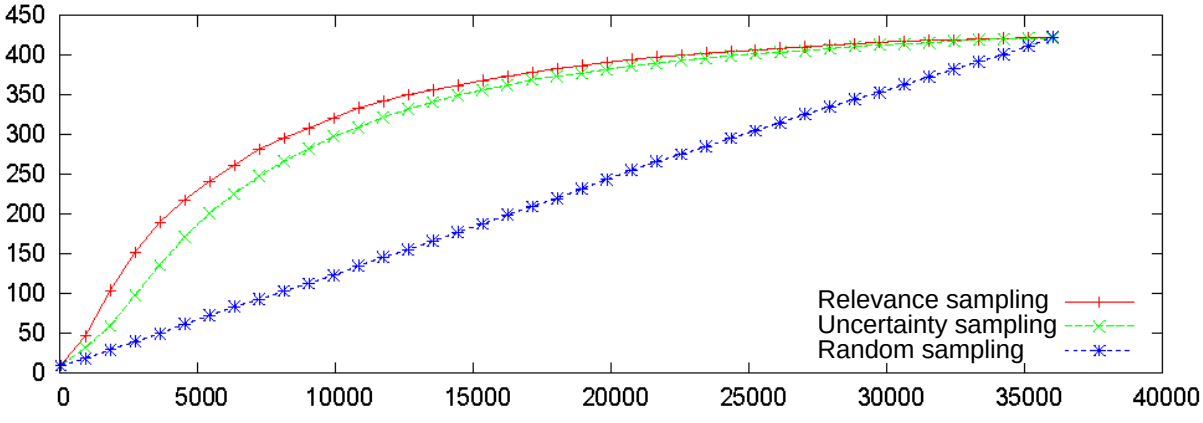
Three evaluated strategies

- Significant level of fluctuations: smooth increase would be expected.
- Random sampling: baseline (linear scan is actually worse).
- Relevance sampling is the best one when a small fraction (less than 15%) of the dataset is annotated.
- Uncertainty sampling is the best one when a medium to large fraction (15% or more) of the dataset is annotated.



Finding positive and negative samples

- Number of positive samples found along iterations.
- Relevance sampling founds positives more rapidly but this is not related to better performance, except close to the beginning.



Simulated active learning conclusions

- The maximum performance is reached when 12 to 15% of the whole dataset is annotated (for 36K samples).
- The optimal fraction to annotate depends upon the size of the training set: it roughly varies with the square root of the training set size (25 to 30% for 9K samples).
- Random sampling is not the worst baseline, linear scan is even worse.
- Simulated active learning can improve system performance even on fully annotated training sets.
- Uncertainty sampling is more “precision oriented”.
- Relevance sampling is more “recall oriented”.
- “Cold start” not investigated.
- Details in our SPIC 2007 paper.

TRECVID 2007
Collaborative Annotation

TRECVID 2007 collaborative annotation

- Follows TRECVID 2003 and 2005 collaborative annotations.
- Annotations are done by TRECVID participants.
- A tool is provided to the annotators. It is Web-based since 2005.
- Images are displayed to the annotators one concept at a time, one or several images at a time, possibility to play the shot.
- The user marks each image as either positive, negative or unsure (default is negative).
- New in 2007: active learning.

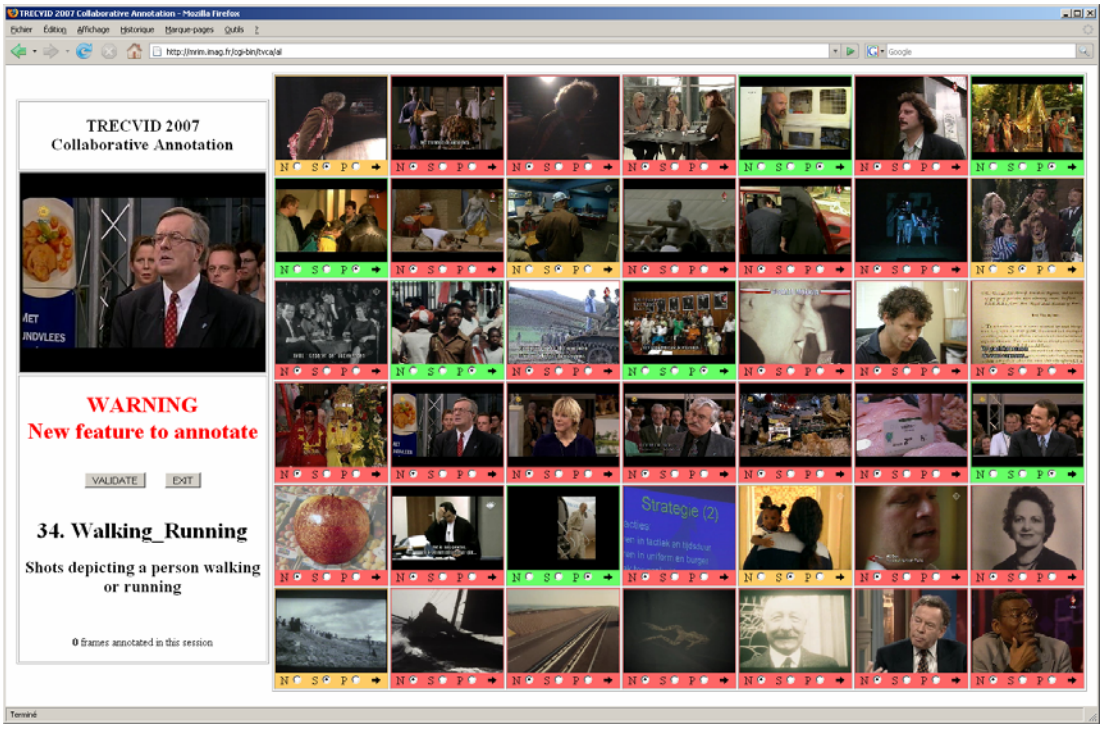
Active learning in the collaborative annotation

- Minimize annotation effort: target less than 2 full days of work per participant (~13 hours @ 2s / annotation).
- (Almost) transparent to the annotator.
- Cold start using TRECVID 2005 annotations (same concepts but significantly different collection contents).
- Annotation driven active learning.
- Implementation of active cleaning (use of active learning to double or triple check the annotated already samples that conflicts with the system prediction during cross-validation).
- Neighbor sampling (select shots just before and just after a positive shot).

TRECVID 2007 Collaborative annotation

Sequential annotation interface

TRECVID 2007 Collaborative annotation



Parallel annotation interface

Annotation driven active learning

- Two engines running in parallel:
  - The web-based annotation engine,
  - The active learning sample ranking engine.
- Active learning is continuously running, cycling on the 36 concepts (in approximately 18 hours, 16 processors).
- The next concept to retrain is the one that received the highest number of annotations since its last training.
- When a user connects, he is asked to annotate the concept with the fewer number of annotations in total ; when he has annotated at least 100 samples, another concept is proposed to him.
- Active learning is transparent to the user except that he has to switch quite often from one concept to another.

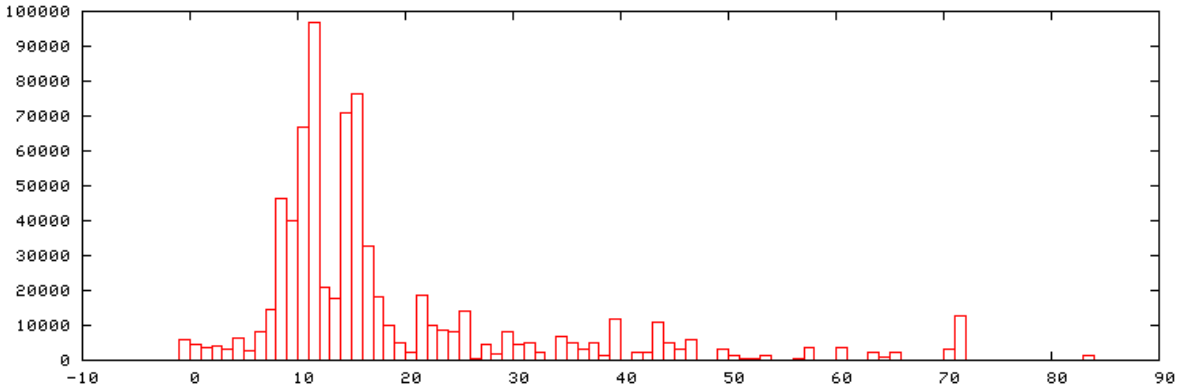
TRECVID 2007 collaborative annotation

- 21,532 subshots to annotate with 36 concepts.
- 32 participating teams.
- Each team was asked to annotate 3% of the subshots × concepts.
- About 92% once-equivalent annotation.
- Some annotated several times due to active cleaning:

	Annotated	% Annotated	Negative	Skipped	Positive	% Positive
Pass 1	641223	82.7	578299	13163	49761	7.76
Pass 2	46864	6.05	11904	7478	27482	58.6
Pass 3	21987	2.84	9383	4040	8564	39.0
Pass 4	1492	0.19	324	940	228	15.3
Synthesis	641223	82.7	578683	15348	47192	7.36

TRECVID 2007 collaborative annotation

- About two-month effort.
- Main effort during two weeks (second an third weeks).
- First week not open to public to guarantee a small step size during the first iterations.

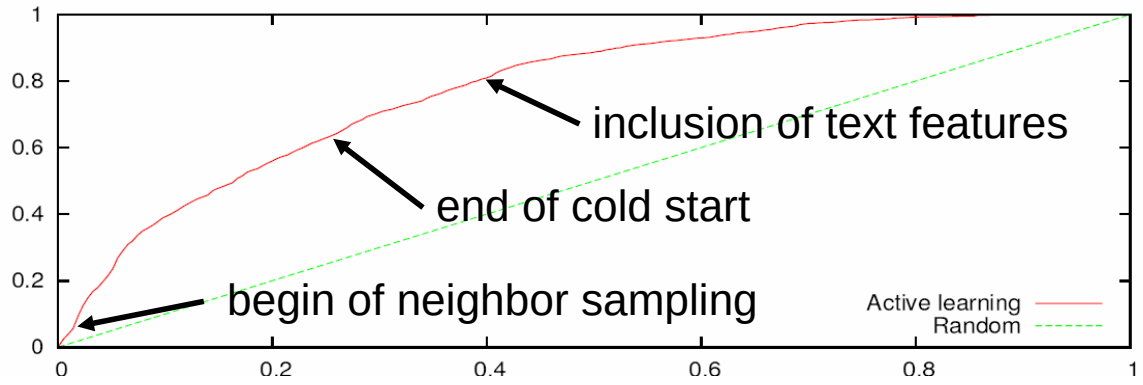


Daily annotations in the collaborative annotation project (GMT time, May 2007 days).

Collaborative Annotation Evaluation

Finding positive and negative samples

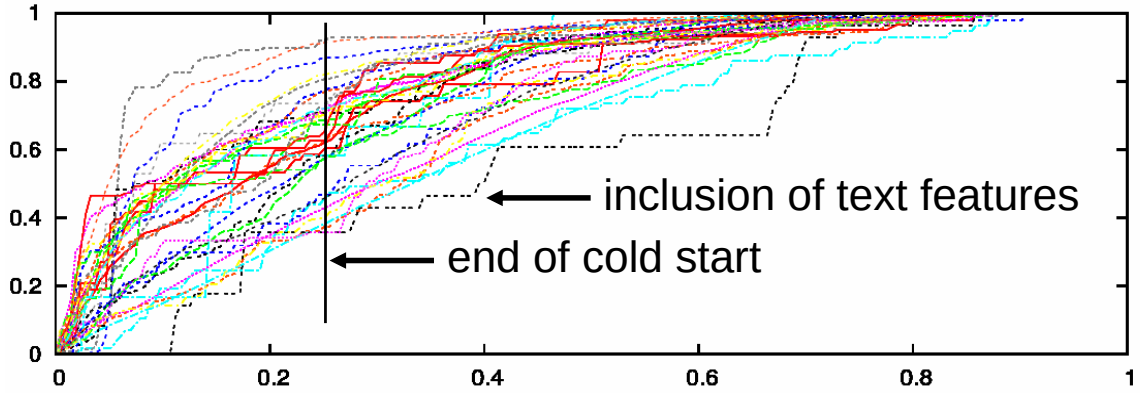
- Similar to experiments with simulated active learning on TRECVID 2005-2006 data.



Evolution of the fraction of positive samples found with the fraction of annotated samples; comparison between active learning and random annotation, all concepts.

Finding positive and negative samples

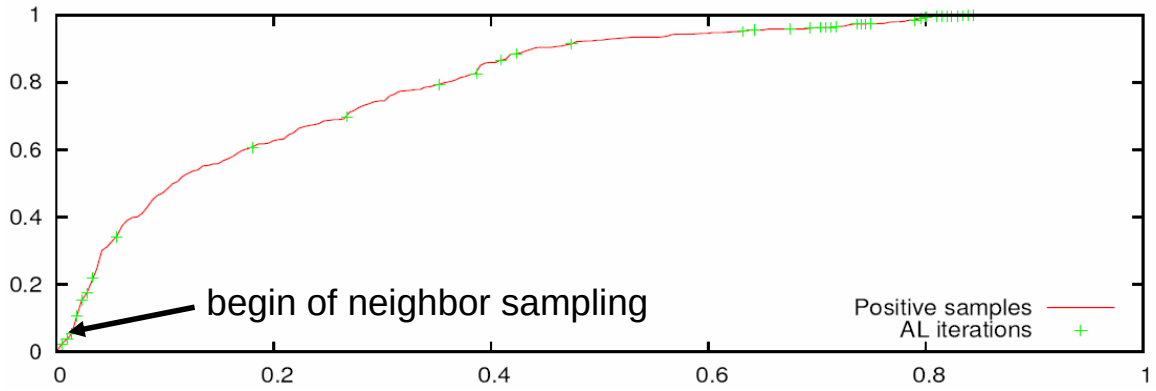
- Evolution by concept is very variable.
- A few do worst than random close to the beginning.



Evolution of the fraction of positive samples found with the fraction of annotated samples for the 36 concepts individually.

Annotation driven active learning

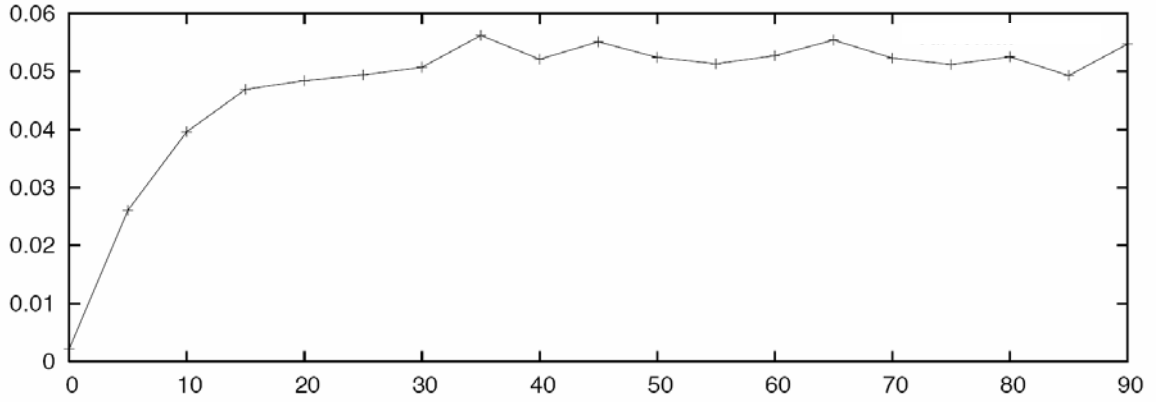
- Small step size at the beginning: training driven
- Larger step size afterwards: annotation driven



Evolution of the fraction of positive samples found with the fraction of annotated samples for the “Animal” concept with marking of the active learning iteration occurrences

Evolution of performance with the annotated fraction of the training set

- Consistent with results obtained in the study with simulated active learning.
- Maximum performance reached when 35% of the training set is annotated.
- Reproducible with another system?



Conclusion

Conclusion

- Active learning based collaborative annotation:
  - Heterogeneous cold start (TV 2005 → TV 2007).
  - Annotation driven active learning.
  - Neighbor sampling.
  - Active cleaning.
- Significant global benefit compared to random sampling but possibly not as high as could have been due to the small collection size.
- Similar behavior in the finding of positive samples and performance evolution though the strategies and cold start conditions differ.
- Neighbor sampling significantly improve the finding rate.
- Difficult to quantify but active cleaning significantly improve the annotation quality.

Future work

- Improve quality check: interface and strategy.
- Better key frame selection:
  - Increase diversity,
  - Multiple key frames per (sub)shot,
  - Increase the positive / negative ratio.
- Use of ontology structure.
- Collaborative annotation 2008?
- Annotate much more concepts?
- Back to local annotation?

Many thanks to all for annotation, feedback and more.