JUMAS @ TRECVID 2010*

Bálint Daróczy[1] Daniele Falavigna[2] Roberto Gretter[2] Dávid Nemeskey[1]
Róbert Pethes[1] István Petrás[1] András A. Benczúr[1]
[1]Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute (SZTAKI)
of the Hungarian Academy of Sciences
{daroczyb, ndavid, petras, benczur}@ilab.sztaki.hu
http://datamining.sztaki.hu

[2]Human Language Technologies research unit
Fondazione Bruno Kessler
38123 Povo, Trento, Italy
{falavi, gretter}@fbk.eu

Abstract

We summarize the *fully automatic* approach to the TRECVID 2010 Known Item (KIS) and Instance Search (INS) tasks of the budapest_acad team in the JUMAS Consortium with Fondazione Bruno Kessler who provided the ASR technologies. Our submissions summarized in Table 1 use linear combinations of the following basic techniques.

- **Meta**: Text retrieval based on video metadata.
- **ASR**: Text retrieval based on ASR most likely readout.
- **SIX** and **CLEF**: Total weight of high level feature classifiers considered relevant by text based similarity to the topic. We used the publicly available Semantic Indexing (SIX) category predictions and the ImageCLEF annotations (CLEF).

Our KIS submissions were based on the linear combination of the four components. Based on experiments from last year, we expected SIX to perform best and gave high weight in our runs. This assumption proved to be wrong. It turns out that Meta contained the strongest information with little improvement possible beyond based on simple linear combination.

For Instance Search we submitted a single run where we used the Dutch translation of the queries to retrieve the ASR text. This run reached an AP sum of 0.508 for the 22 topics. The INS task proved to be very difficult. Our fairly straightforward method reached close to the 0.729 result of the best participant.

1 Introduction

In this paper we describe our approach to TRECVID 2010 Known Item Search (KIS) and Instance Search (INS) tasks using fully automatic processing. The KIS test data set consisted of 200 hours of video with approximately 96400 shots with the corresponding automatic speech recognition (ASR) transcript. The INS test data set consisted of a 118GB video footage with the corresponding automatic speech recognition (ASR) transcript.

We describe our approaches that rely on text retrieval of the automatic speech recognition (ASR) output and the metadata as well as on relevant high level feature selection (Section 3). The ASR output was given for the INS task; for the KIS task we developed our own system that we describe in Section 2. For high level visual features we used the publicly available annotations for shots produced by participants of the Semantic Indexing (SIX) task as well as our own classifiers trained for the ImageCLEF2009 [14] tasks. The combination is linear with scores

*This work was supported by the EU FP7 project JUMAS – Judicial Management by Digital Libraries Semantics and by grants OTKA NK 72845 and NKFP-07-A2 *TEXTREND*.

Run ID	MIR	matches	Meta		ASR		visual	
			cue	descr	cue	descr	SIX	CLEF
Meta cue	0.184	117	1					
Meta descr	0.193	121		1				
Meta all	0.209	141	1	1				
Meta all	0.218	141	1	3				
ASR cue	0.020	18			1			
ASR descr	0.032	24				1		
ASR all	0.029	29			1	1		
ASR all	0.032	29			1	5		
SIX	0.006	24					1	
ImageCLEF	0.000	5						1
budapest1	0.038	113	6	14	3	7	50	20
budapest2	0.026	159	6	14	3	7	100	30
budapest3	0.012	86	6	14	3	7	70	50
budapest4	0.007	153	3	7			70	50
best linear	0.223	149	20	40	1	5		

Table 1: Summary of our KIS base components and our runs submitted, in terms of the mean inverted rank (MIR) and the number of instances found in top 100 ranked (matches).

below a certain rank considered constant. Our runs, both submitted and additionally evaluated, are summarized in Table 1.

We apply text retrieval to find the relevant parts in the KIS metadata, ASR output as well as to match high level features to topics for KIS. We used the Hungarian Academy of Sciences text retrieval engine [4] that is based on Okapi BM25 [16] with proximity weights [15, 2].

Since for the INS task, the ASR output is in Dutch, we translated the queries using online dictionaries word-by-word. As a result, several synonyms were produced for every English word. For a group of alternate translations we have only kept the highest tf.idf in every document.

In our high level visual concept detection based component we did not make use of the KIS training data set. Instead, we used the ImageCLEF VCDT 2009 [14] training set that consisted of 5000 images labeled with 53 categories, a subset of the MIR Flickr 25,000 image data set. We pruned categories that were either unusable for the task or have an inaccurate detector, finally arriving at 34 concept categories. These trained concept detectors were evaluated on the IACC.A test set.

2 The Automatic Speech Recognition system

In this section we describe the main features of the FBK-irst systems developed for the TRECVID 2010 evaluation. Word transcriptions are generated in two decoding passes, after partitioning of the input audio streams into clusters of segments.

2.1 Audio partitioning

The audio partitioner consists of three main modules: the segmenter, based on a start-end-point speech activity detector, the segment classifier, based on Gaussian Mixture Models (GMMs) and the clustering module using a Bayesian Information Criterion (BIC).

Segmenter. In the current version of the system, the segmenter identifies the parts of the audio stream with "high" energy values through the application of a start-end-point activity detector.

Segment classification. The goal of this module is to classify each acoustically homogeneous segment in terms of broad acoustic classes. These latter ones are modeled by Gaussian Mixture Models.

Segment clustering. Identified acoustically homogeneous segments are clustered through a Bayesian Information Criterion (BIC) hierarchical method. In this way segments in each cluster are assumed to belong to the same speaker.

A the end of the partitioning process we obtain audio segments labeled with cluster number and speaker gender.

2.2 Decoding Process

For each cluster of speech segments, the system, which makes use of continuous density Hidden Markov Models (HMMs), generates a word transcription by performing two decoding passes interleaved by acoustic feature normalization and acoustic model adaptation. Best word hypotheses, generated by the first decoding pass, are exploited for performing cluster-wise acoustic feature normalization, based on Constrained Maximum Likelihood Linear Regression (CMLLR) [7] and acoustic model adaptation. For this latter purpose, just Gaussian means of triphone HMMs are adapted through the application of up to 4 full matrix transforms estimated in the MLLR framework.

2.3 Acoustic Models

In both decoding passes acoustic models (AMs) are state-tied, cross-word, speaker-independent triphone HMMs. Each HMM is characterized by a 3 state left to right topology, with the exception of the model associated to the background noise, which has a single state. In addition to triphones, several spontaneous speech phenomena are also modeled. Output probability distributions are modeled by mixtures of Gaussian probability density functions (PDF) having diagonal covariance matrices. A phonetic decision tree is used for tying states and for defining the context-dependent allophones.

Each speech frame is parametrized into a 52-dimensional observation vector composed of 13 mel frequency cepstral coefficients (MFCCs) plus their first, second and third order time derivatives. Cepstral mean subtraction and variance normalization is performed on static features on a cluster-by-cluster basis. A projection of acoustic feature space, based on heteroscedastic linear discriminant analysis (HLDA), is embedded in the feature extraction process as follows [18]. A GMM with 1024 Gaussian components is first trained on the original 52-dimensional observation vectors. Acoustic observations in each, automatically determined, cluster of speech segments, are then normalized by applying an affine transformation estimated w.r.t. the GMM through CMLLR [7]. After normalization of training data, an HLDA transformation is estimated w.r.t. a set of state-tied, cross-word, gender-independent triphone Hidden Markov Models (HMMs) with a single Gaussian per state, trained on the normalized 52-dimensional observation vectors. The HLDA transformation is then applied to project the set of 52 normalized features into a 39-dimensional feature space. Recognition models used in the first and second decoding pass are trained on these normalized, HLDA projected, acoustic features. HMMs for the first decoding pass are trained through a conventional maximum likelihood procedure. HMMs used in the second decoding pass are trained through a speaker adaptive procedure [8]: for each cluster of speech segments an affine transformation is estimated through CMLLR exploiting as target-models triphone HMMs with a single Gaussian density per state trained on the HLDA projected acoustic features. The estimated affine transformation is then applied on the cluster data [8]. Acoustic models are finally trained from scratch on the obtained normalized acoustic data.

2.4 Language Model

The Language Model (LMs) used in the ASR system allows to estimate 4-grams probabilities. It was trained on public texts stemming from several sources (news agencies, newspapers, transcriptions of parliamentary debates, etc). In total, the training corpus consisted of 674M words. This LM included 65k unigrams, 27M bigrams, 29M 3-grams and 27M 4-grams.

3 High level video feature relevance

The KIS topics were matched against the high level features made available from the Semantic Indexing (SIX) task as well as our own classifiers for the ImageCLEF VCDT 2009 [14] task with 53 categories (CLEF). The feature descriptions served as input to text processing that provided us with a scored list of possibly relevant features for every topic. To determine the relevance of a shot to a topic, we computed the sum of these scores, weighted by the results of the high level feature extraction track for the shot.

We used the the video processing subsystem developed for TRECVID 2009 tasks [1]. We computed a grayscale HOG descriptor on a dense grid for each image. We used Bag of Words representation with k=4000

K-Means. The Bag of Words generative modeling is a well known technique from text domain [10, 9]. It also has been proved successful in image categorization and retrieval. [3, 17]. The images (shot keyframes) are represented by $D = 4000$ dimensional vectors. The BOV term frequency vector consists of the histogram of the codewords from the codebook. These L1-normalized BOV vectors were fed into linear classifiers (L2-regularized logistic regression classifier from the LibLinear package [5]) for each of the $k = 34$ concepts. This resulted in 34 concept probabilities for each image. We did this process at first on the 5000 image-sized training set, then followed similar process using the trained classifiers on the evaluation data.

Using predictions for the SIX and CLEF high level visual features, we have applied the following approach to score videos against the items:

- We weighted each high level visual features against each item (topic).
- We computed the video scores from the weight of the feature for the query item and the class predictions of the video keyframes.

We weight high level visual features by matching against the visual cues and description of the search item. For example feature "male" is related to "Find the video of bald, shirtless man showing pictures of his home full of clutter and wearing headphone". We first gave a manual description of the small fixed feature set containing the name of the feature along with search terms with grammatical categories and possible Boolean relations. We label the four grammatical categories noun, plural noun, verb and adjective. As an example, a feature description looks like this:

```
female | female:n woman:n
old person | AND old:adj person:n
```

Next we process the item cue and description. We use the Stanford lexical parser [11, 12] to obtain the category of each term and WordNet [13, 6] to find synonyms and hypernyms. For a fixed search item, we iterate over the features, and consider a feature relevant for the item if

- we find a feature search word in the cue or description with the same category, in which case the feature weight is 1;
- we find a feature search word in the synonyms with the same category, in which case the feature weight is again 1;
- we find a feature search word in the hypernyms with the same category, in which case the feature weight depends on the distance of the feature word and query word in WordNet.

Unfortunately, synonyms themselves were not strong enough to find related features, while hypernyms frequently drifted. As an example, we had to manually remove the "man is a cat" along a few additional WordNet relations.

Given a video v and a search item i, we use the weight w_{if} of the relevant high level features f and the predictor p_{fv} for that feature over the video itself. In case of SIX features, we used the publicly available relevance judgment for the whole video. To compute the score of a shot against an item, we computed

$$\sum_f p_{fv} \cdot w_{if}.$$

For SIX this formula scores videos. For CLEF features, we obtained prediction for shots and aggregated for videos by averaging after skipping the first and last 5% of the video shots.

4 Conclusion

We summarize the results of our KIS submissions and post-submission experiments in Table 1. In our official runs, we relied on an incorrect assumption that visual features provide the most accurate information on the item as it was the case for the 2009 Search task. As it turns out, by simple linear combination, it is very hard to improve on text retrieval of the Metadata. We believe that cross-modal feedback techniques could have exploited information from weaker modes. Note that the official runs had higher number of items retrieved in top 100 than our most accurate combinations, showing that visual techniques could have had a role in the KIS task.

For INS our single run of sum AP 0.508 is based solely on ASR text retrieval by the Dutch translation of the queries. First, note that this corresponds to a MAP of 0.023, a result weaker than the KIS ASR retrieval result and is considered very weak. On the other hand, this result gets quite close to the AP sum value 0.729 of the best performing team, indicating the hardness of the task. Overall, we find these TRECVID tasks this year have stronger emphasis in text retrieval than in earlier campaigns.

References

[1] Bálint Daróczy and Dávid Nemeskey and István Petrás and András A. Benczúr and Tamás Kiss. SZTAKI @ TRECVID 2009. In *TREC Video Retrieval Evaluation Online Proceedings, TRECVID*, 2009.

[2] Stefan Büttcher, Charles L. A. Clarke, and Brad Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In *SIGIR '06*, pages 621–622, New York, NY, USA, 2006. ACM Press.

[3] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[4] Bálint Daróczy, Zsolt Fekete, Mátyás Brendel, Simon Rácz, András Benczúr, Dávid Siklósi, and Attila Pereszlényi. Cross-modal image retrieval with parameter tuning. In Carol Peters, Danilo Giampiccol, Nicola Ferro, Vivien Petras, Julio Gonzalo, Anselmo Peñas, Thomas Deselaers, Thomas Mandl, Gareth Jones, and Nikko Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, September 2008 (printed in 2009).

[5] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. LIBLINEAR: A library for large linear classication. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[7] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[8] D. Giuliani, M. Gerosa, and F. Brugnara. Improved automatic speech recognition through speaker normalization. *Computer Speech and Language*, 20(1):107–123, Jan. 2006.

[9] Huma Lodhi Huma, Craig Saunders, Nello Cristianini, Chris Watkins, and Bernhard Scholkopf. Classification using string kernels. *Journal of Machine Learning Research*, 2:563–569, 2002.

[10] Thorsten Joachims, Fachbereich Informatik, Fachbereich Informatik, Fachbereich Informatik, Fachbereich Informatik, and Lehrstuhl Viii. Text categorization with support vector machines: Learning with many relevant features, 1997.

[11] Dan Klein and Christopher D. Manning. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*. December 2002.

[12] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[13] George A. Miller. Wordnet: A lexical database for English. *Communications of ACM*, 38(11):39–41, 1995.

[14] Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 large scale visual concept detection and annotation task. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

[15] Yves Rasolofo and Jacques Savoy. Term proximity scoring for keyword-based retrieval systems. In *Advances in Information Retrieval*, LNCS, pages 207–218. Springer, 2003.

[16] Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. In *Document retrieval systems*, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988.

[17] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2003. IEEE Computer Society.

[18] G. Stemmer and F. Brugnara. Integration of Heteroscedastic Linear Discriminant Analysis (HLDA) into Adaptive Training. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages I–1185–1188, Toulouse, France, May 2006.