EXPLORING AUTOMATIC QUERY REFINEMENT FOR TEXT-BASED VIDEO RETRIEVAL

*Timo Volkmer**

School of CS & IT
RMIT University, Melbourne, Australia
tvolkmer@cs.rmit.edu.au

Apostol (Paul) Natsev

IBM T. J. Watson Research Center
Hawthorne, NY, USA
natsev@us.ibm.com

ABSTRACT

Text-based search using video speech transcripts is a popular approach for granular video retrieval at the shot or story level. However, misalignment of speech and visual tracks, speech transcription errors, and other characteristics of video content pose unique challenges for this video retrieval approach.

In this paper, we explore several automatic query refinement methods to address these issues. We consider two query expansion methods based on pseudo-relevance feedback and one query refinement method based on semantic text annotation. We evaluate these approaches in the context of the TRECVID 2005 Video Retrieval Benchmark using a baseline approach without any refinement. To improve robustness, we also consider a query-independent fusion approach. We show that this combined approach can outperform the baseline for most query topics, with improvements of up to 40%. We also show that query-dependent fusion approaches can potentially improve the results further, leading to 18-75% gains when tuned with optimal fusion parameters.

1. INTRODUCTION

In recent years, research in content-based video retrieval has focused on exploiting various modalities of the video content. A popular approach is leveraging the textual information that can be obtained from Closed Captions (CC), Automatic Speech Recognition (ASR), and Optical Character Recognition (OCR) sources. Closed captions are frequently unavailable, and video OCR is limited as it applies only to video segments that contain inscriptions in the video imagery. Most videos, on the other hand, carry spoken information—especially news broadcasts, which form an important domain for video retrieval. Complemented by the fact that automatic speech recognition is a well understood technique, speech-based retrieval is perhaps the most popular technique used for video search and retrieval. Typically, videos are segmented into shots, the speech track is automatically transcribed and machine translated as needed, and the resulting speech transcript is time-aligned with the video segments. Traditional text search engines can then be applied for speech-based video retrieval. Unfortunately, the high retrieval performance that text search engines achieve in pure text document retrieval is usually not observable in text-based video retrieval.

The quality of the automatic speech recognition—and machine translation for foreign sources—influences the overall retrieval performance directly. While some of these issues have been addressed in a satisfactory way for spoken text document retrieval, this is less so when applying text-based search to video retrieval. One of the main reasons is the mismatch between the semantics contained in the spoken track and the visual one. For example, when users search for video segments showing aircraft, they might use "aircraft" as a query term. The likelihood that shots depicting aircraft actually contain this term as spoken text is rather small, though. In addition, the spoken track rarely mentions the background scene or setting depicted in the video. As a result, speech-based retrieval performs well at answering *specific* queries about named people, sites, or events. It usually fails at *generic* queries involving unnamed people, objects, settings, or events.

Query expansion is a promising approach for addressing some of the above problems, such as poor recall due to missing or misaligned speech terms in regards to the visual information. In principle, the original query is expanded with additional query terms that are somehow related to the query. These may include synonyms of the original query terms, or non-synonym terms that frequently co-occur with the query terms in the same context, and are therefore topically related (e.g., "aircraft" and "airline"). Synonym or hypernym-based query expansion approaches are considered *global query expansion* since they are based on the lexical properties of the English language, and are corpus-independent. They are frequently based on dictionaries or sources such as WordNet[1] [1].

Co-occurrence based approaches, on the other hand, are considered *local* as they rely on term co-occurrence and frequency statistics, which are corpus-dependent. The typical strategy is to expand the query with terms from a number of documents that are considered relevant to the original query, as well as to adjust the query term weights based on the statistics of the relevant documents. With *pseudo-relevance feed-*

*The work was performed while this author was visiting the IBM T. J. Watson Research Center.

[1] http://wordnet.princeton.edu

back, for example, the original query is used to retrieve the top N matching documents. These are assumed pseudo-relevant, and are analyzed to select additional query terms [2]. This can improve recall—especially for short queries—by allowing document matches to additional terms related to the original query (e.g., "aircraft" expands with "airline" or "pilot"). It may also narrow down too broad queries, thereby re-ranking results and improving precision (e.g., expanding "car" with "car accident"). This of course works only as long as the refined query is indeed relevant to the original one. Experiments in text-document retrieval have shown that query expansion is highly query-dependent and bears the risk of topic drift.

Another method of query refinement attempts to prevent topic drift by disambiguating word senses using *semantic text annotation*. In this approach, the entire collection is analyzed and annotated with semantic categories. Sense ambiguity is resolved by deep parsing, part-of-speech tagging based on word context, and rule-based semantic annotation. An appropriate index including this information for all detected terms can then be built. At query time, the query terms are analyzed and annotated in the same way, and the query is refined with the relevant semantic categories [3]. For example, a query containing the term "basketball" may automatically be refined to the "SPORTS" category, "car" can be expanded to "VEHICLE", while "George Bush" can be expanded to "PRESIDENT". This approach has the potential to allow semantic refinement of query topics, while limiting topic drift. However, it is only applicable to the set of semantic categories that can be annotated robustly.

In the remainder of this paper, we describe our speech-based retrieval system and evaluate several text-based query refinement methods in the context of video retrieval. We propose query refinement using a fusion of different approaches and conclude with a discussion of our results.

2. TEXT-BASED VIDEO RETRIEVAL SYSTEM

Our speech-based search system is part of the IBM video retrieval system [4] used in the TREC Video Retrieval Benchmark (TRECVID)[2]. To study the effect of text-based query expansion, we have evaluated the speech-based retrieval system independently. It is built using the IBM Unstructured Information Management Architecture (UIMA)[3] and the JuruXML semantic search engine [5] included in the UIMA SDK[4]. In addition, we used several UIMA components developed by IBM Research for advanced text analytics. These include the RESPORATOR (RESPOnse geneRATOR) system [3] and the PIQUANT Question Answering system [6] built on top of RESPORATOR. With this setup, we evaluated the following automatic query refinement methods:

Rocchio-based query refinement: Rocchio refinement [7], a pseudo-relevance feedback method, is available natively in JuruXML. The top N documents ranked highest by the original query are assumed pseudo-relevant. This set is then analyzed to select k representative terms for query expansion, and to adjust the weights of the original query terms. While susceptible to topic drift, this approach is suitable for discovering relevant terms that do not necessarily have a lexical relationship with the original query terms but frequently co-occur with them in the pseudo-relevant documents. For example, in this fashion, "car" may be related to "BMW".

Lexical affinity-based query refinement: This approach is also based on pseudo-relevance feedback but employs an alternative term selection method, designed to minimize topic drift. It considers *lexical affinities* (LA), which are pairs of terms that frequently co-occur within a close proximity of each other—for example, within one phrase. If one of the terms in a lexical affinity appears in the query text, it is assumed that the other part of the LA is also relevant. For example, "car" may be expanded to "car accident". This method was proposed in [8], and is also available natively in JuruXML.

Semantic annotation-based refinement: In this method, the entire corpus is annotated and indexed with over 100 semantic categories using the RESPORATOR annotator [3]. It is a rule-based annotator of both named and unnamed entities, such as people, roles, objects, places, events, program categories, etc. It is used extensively by the PIQUANT question answering system [6]. Each query is analyzed by PIQUANT and annotated with one or more semantic categories. Shots would then be considered relevant not only if they contained one of the query terms, but also if they were annotated with one of the semantic categories of the query.

The performance of the above approaches depends much on the query topic, and no single approach is likely to emerge as the winner for all topics. In fact, for many topics, the best strategy is to not perform any query expansion. Such topics include named person queries, or difficult queries for which the pseudo-relevancy assumption for the top documents does not hold. We therefore considered a fusion approach in an effort to improve robustness and to combine the strengths of the individual approaches. Ideally, one should use a query-dependent method selection or weighted fusion, such as the one in [9], as it has a tremendous potential to improve performance and robustness in a query-specific way. However, for simplicity, and due to lack of a large enough independent training set of topics and ground truth, we consider only a global parameter-free fusion approach in this paper. In particular, we use simple score averaging to combine the shot ranking scores as determined by the three query refinement approaches and the original query baseline.

[2] http://www-nlpir.nist.gov/projects/trecvid/
[3] http://www.research.ibm.com/UIMA/
[4] http://www.alphaworks.ibm.com/tech/uima

3. PERFORMANCE EVALUATION

We have conducted experiments using the TRECVID 2005 test corpus and query topics[5]. This collection contains 140 broadcast news video clips from U.S., Arabic, and Chinese sources, with durations of 30 minutes to 1 hour each, and pre-segmented into 45,765 shots. Each video comes with a corresponding speech transcript obtained through automatic speech recognition, as well as machine translation for the Arabic and Chinese sources. The text search baseline is obtained by processing queries to perform part-of-speech tagging and retain only nouns, and to perform Porter stemming. The JuruXML search engine also natively identifies phrases in the form of lexical affinities, and uses them to resolve ambiguities and to obtain more accurate TF*IDF statistics. The query refinement approaches are performed with the same retrieval engine and query processing, after tuning the parameters for the two pseudo-relevance feedback methods. The number (N) of top-ranked documents to be considered relevant was set to $N = 30$ for LA-based expansion, and to $N = 12$ for the Rocchio method. The max. number (k) of terms to be added was set to $k = 30$ for both methods. Additional and original query terms were weighted with the same weight.

We tuned these parameters based on performance on the TRECVID 2003 corpus and queries; they are likely to be suboptimal for the 2005 corpus. The 2003 and 2005 TRECVID collections are both based on broadcast news but differ in many other aspects, such as including different channels with different production rules, the use of non-English sources in 2005, and the use of different ASR engines. We have used the machine translations for the non-English sources that NIST has provided for 2005 without further processing. The quality of these is inferior to the quality of the native English sources because of the error rate of the machine translation.

To evaluate performance, we executed blind runs on the TRECVID 2005 test set using the 24 search topics as specified for the 2005 search task evaluation. We use Average Precision to measure performance on a specific topic, and Mean Average Precision (MAP) to aggregate performance results across multiple topics. Average Precision is the official performance metric adopted by TRECVID, and essentially represents the area under the precision-recall curve.

Table 1 shows a comparison of the three query refinement approaches and the baseline, as evaluated on both 2003 and 2005 datasets and topics. We note that most of the results previously reported at TRECVID were produced using some form of query refinement. Even though the two sets of results are based on two different sets of the query topics, and are therefore not directly comparable, we still note a significant performance loss on the 2005 corpus. This is most likely due to the poor quality of the machine-translated non-English sources, and to suboptimal parameters for the 2005 data. More interesting, however, is the discrepancy in relative performance of different approaches on the two corpora. Our results on the 2003 collection show that query expansion can yield significant (50%) improvements on the "clean" sources when properly tuned. The opposite is true, however, on the "noisier" data—only one of the query expansion approaches actually outperforms the no-expansion baseline.

[5] http://www-nlpir.nist.gov/projects/tv2005/tv2005.html

Query Refinement/ Expansion Method	Training Set TRECVID-2003	Testing Set TRECVID-2005
No refinement	0.0831	0.0558
Semantic refinement	0.1237	0.0546
LA-based expansion	0.1275	0.0578
Rocchio expansion	0.1291	0.0413

Table 1. Mean Average Precision scores of text search baseline and three query expansion approaches, evaluated on two different corpora and two sets of search topics. Parameters were tuned to optimize TRECVID 2003 performance and were applied blindly on TRECVID 2005 data and topics.

To gain further insight, we analyzed the performance broken down into subsets of topics grouped by query class. We considered the 5 classes *Named People*, *Unnamed People*, *Object*, *Scene/Setting*, and *Event/Action*. Since one query can belong to more than one class, we grouped the 24 topics from TRECVID 2005 into 7 Person-X queries, 5 People queries, 6 Object queries, 10 Setting queries, and 7 Event queries. Figure 1 shows the query-class specific performance of the 3 query expansion approaches and the baseline. This figure confirms our hypothesis that query expansion is highly topic-dependent and no single method is likely to outperform the others on all topics. It also gives a possible explanation why query expansion hurts overall performance for two of the methods. Since the overall MAP score is influenced mostly by top-performing queries, the Person-X query class dominates all other query classes due to the much higher scores it generates. Any approach that does not fare well on Person-X topics is therefore likely to have poor overall performance. Incidentally, Person-X queries work just as well, or better, without query expansion.

As no single approach works best for all topics, we wanted to minimize query-dependency—and improve robustness—by combining our query expansion techniques together with the baseline approach. We used a score averaging fusion scheme (global parameter-free fusion) to combine all 4 methods, and also considered a query-specific *Oracle* fusion. The latter serves to measure *potential* performance gains with optimally tuned query-specific fusion parameters. For the Oracle evaluation, we considered 5 different score normalization methods, along with 3 non-weighted fusion methods (AVG, MAX, and PRODUCT), and chose the optimal combination for each query as observed on the test set. Query-dependent weighted fusion approaches are likely to perform better, and are subject of ongoing work [9].

The results of the combination hypothesis approach are

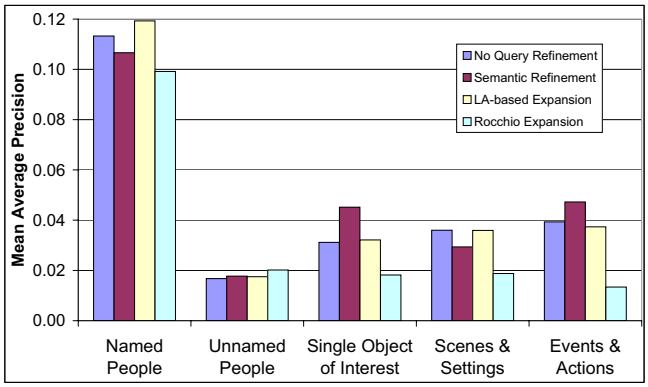

Fig. 1. Query-class specific performance evaluation.

listed in Table 2. The simple AVG fusion approach does improve robustness, as it outperforms the baseline for 4 out of 5 query classes. It leads to 11% overall improvement and a Mean Average Precision within 10% of the best performance (0.067) reported for text-based automatic search in TRECVID 2005. Furthermore, the *Oracle* method demonstrates significant potential gains for all query classes, ranging from 18% to 75%, with an overall improvement of 27% over all topics. This clearly attests to the promise of this combination hypothesis approach for query expansion.

Based on the Wilcoxon signed rank test, both the results for AVG fusion and *Oracle* fusion are statistically significant at the 5% level.

Query Class (#topics/class)	No query Expansion	Query Expansion Fusion	
		AVG (gain)	Oracle (gain)
Person-X (7)	0.1133	0.1238 (9%)	0.1341 (18%)
People (5)	0.0167	0.0240 (43%)	0.0266 (59%)
Object (6)	0.0312	0.0403 (29%)	0.0545 (75%)
Setting (10)	0.0360	0.0373 (4%)	0.0440 (22%)
Action (7)	0.0393	0.0382 (-3%)	0.0566 (44%)
All Topics (24)	0.0558	0.0617 (11%)	0.0711 (27%)

Table 2. Query-class specific Mean Average Precision scores for no-expansion baseline, AVG fusion-based query expansion, and an *Oracle* method with test set-optimized fusion parameters.

4. CONCLUSIONS AND FUTURE WORK

We have investigated three complementary automatic query refinement approaches and shown that these have excellent potential for improving speech-based video retrieval. While query expansion performance is query specific, and no single approach emerges as a clear winner across all topics, we observed consistent performance patterns within 5 query classes, including named and unnamed people, objects, settings, and events. In particular, each class exhibited different behavior with respect to the optimal query expansion method. A simple combination hypothesis approach was able to improve robustness, leading to performance gains for 4 out of the 5 query classes, including 30-40% gains on the *Objects* and *People* classes, and 11% improvement over all topics.

In future work, we will consider query-class dependent fusion approaches, such as the one presented in [9]. Query-class dependent method selection and fusion have a considerable potential for further improvements, as shown by the *Oracle* fusion method and its potential gains in all query classes, ranging from 18% to 75% with a 27% overall gain. Weighted fusion approaches are likely to yield even higher gains.

5. ACKNOWLEDGMENTS

This material is based on work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

6. REFERENCES

[1] E. M. Vorhees, "Query expansion using lexical-semantic relations," in *Proc. of ACM-SIGIR 1994*, Dublin, Ireland, August 1994, pp. 61–69.

[2] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proc. of ACM-SIGIR 1996*, Zurich, Switzerland, 18–22 August 1996, pp. 4–11.

[3] J. Prager, E. Brown, A. Coden, and D. Radev, "Question answering by predictive annotation," in *Proc. of ACM-SIGIR 2000*, Athens, Greece, 24–28 July 2000, pp. 184–191.

[4] A. Amir, J. O. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M. Naphade, A. Natsev, J. R. Smith, J. Tešić, and T. Volkmer, "IBM Research TRECVID-2005 video retrieval system," in *TRECVID 2005 Workshop Notebook Papers*, Gaithersburg, MD, USA, 14–15 November 2005.

[5] Y. Mass, M. Mandelbrod, E. Amitay, D. Carmel, Y. Maarek, and A. Soffer, "JuruXML—an XML retreival system," in *Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, Schloss Dagstuhl, Germany, 9–11 December 2002, pp. 73–80.

[6] J. C. Carroll, K. Czuba, J. Prager, A. Ittycheriah, and S. B. Goldensohn, "IBM's PIQUANT II in TREC2004," in *NIST Special Publication 500-261: Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburgh, MD, USA, 16–19 November 2004, pp. 184–191.

[7] J. J. Rocchio, *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14. Relevance Feedback in Information Retrieval, pp. 313–323, Prentice-Hall Inc., Englewood Cliffs, NJ, USA, 1976.

[8] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer, "Automatic query refinement using lexical affinities with maximal information gain," in *Proc. of ACM-SIGIR 2002*, Tampere, Finland, 11–15 August 2002, pp. 283–290.

[9] L. Kennedy, A. Natsev, and S.-F. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *ACM Multimedia 2005*, Singapore, Nov. 2005.