Enriching a Thesaurus to Improve Retrieval of Audiovisual Documents

Laura Hollink, Véronique Malaisé, and Guus Schreiber

Free University Amsterdam
de Boelelaan 1081a
1081 AH Amsterdam, The Netherlands

Abstract. In many archives of audiovisual documents, retrieval is done using metadata from a structured vocabulary or thesaurus. In practice, many of these thesauri have limited or no structure. The objective of this paper is to find out whether retrieval of audiovisual resources from a collection indexed with an in-house thesaurus can be improved by anchoring the thesaurus to an external, semantically richer thesaurus. We propose a method to enrich the structure of a thesaurus and we investigate its added value for retrieval purposes.

We first anchor the thesaurus to an external resource, WordNet. From this anchoring we infer relations between pairs of terms in the thesaurus that were previously unrelated. We employ the enriched thesaurus in a retrieval experiment on a TRECVid 2007 dataset. The results are promising: with simple techniques we are able to enrich a thesaurus in such a way that it adds to retrieval performance.

1 Introduction

The objective of this paper is to investigate whether retrieval of audiovisual documents that are indexed with an in-house thesaurus can be improved by anchoring the thesaurus to an external, semantically richer thesaurus.

Many collections of audiovisual documents are indexed manually with the help of a local thesaurus. The manual indexing process is time-consuming, therefore the tendency is to only use a small set of terms to describe a document. The annotations are usually of high quality. We point out that the opposite can be said about automatic annotation using content-based feature detectors. This approach results in many annotations, but their quality is unreliable.

A low number of annotations per document can lead to low recall of search results. One way to overcome this issue is query expansion, where documents are retrieved not only with the initial query term, but also with closely related terms [18]. In the context of concept-based search, where queries are posed in terms of thesaurus concepts, query expansion depends on a rich thesaurus structure. However, local thesauri are often limited in breadth and depth. In this paper we report on an experiment in which we enrich a local thesaurus and study its added value for retrieval.

The study is performed on a dataset of the Netherlands Institute for Sound and Vision (Sound & Vision). The institute stores over 700,000 hours of Dutch broadcast video, and archives every day the daily broadcast in digital format. It has an in-house thesaurus, the GTAA, with limited structure, which is used to index and search the collection.

Our approach consists of two steps. First, we anchor the GTAA thesaurus to an external resource, the English-American WordNet [4], by searching for related concepts (*synsets* in WordNet) using a syntactic alignment procedure. The alignment is based on lexical comparison of term descriptions in the two resources. Such a mainly lexical alignment approach is bound to be incomplete and at times incorrect. However, considering the state of the art of ontology alignment, this is a realistic situation [3]. Second, based on this anchoring, we enrich the in-house thesaurus by inferring potential new relations between terms within the thesaurus.

To investigate the value of the enriched thesaurus for retrieval purposes, we perform an experiment in which we compare retrieval results achieved with the in-house thesaurus to results obtained with the enriched thesaurus. The experiment is performed on a part of the collection of Sound & Vision that was used in the TRECVid 2007 conference [15]. We use the queries and ground truth provided by TRECVid. In addition, Sound & Vision kindly provided us with the metadata of this dataset in the form of manual annotations of the audiovisual documents with GTAA terms.

Our hypothesis is that anchoring the in-house thesaurus to a rich external source will help retrieval, particularly with respect to recall: the richer semantic structure should lead to more matches. We are interested in finding out how much this approach jeopardizes precision and whether the joint effect can be judged to be positive or negative.

This paper is structured as follows. Section 2 describes the GTAA and its anchoring to WordNet. In Section 3 we describe how new thesaurus relations are inferred from the GTAA-WordNet links. Section 4 describes the setup of the retrieval experiment and the TRECVid dataset. The results of the retrieval experiment are analyzed in Section 5. We conclude with a discussion and directions for future work in Section 6.

2 Anchoring the GTAA Thesaurus to WordNet

2.1 The GTAA thesaurus

The GTAA is a Dutch, faceted thesaurus resulting from the merging of several controlled vocabularies used by audiovisual archives in the Netherlands. Its name is a Dutch acronym for “Common Thesaurus for Audiovisual Archives”. At the Netherlands Institute for Sound and Vision it is used for manual annotation of the extensive collection of broadcast video.

The GTAA thesaurus contains approximately 160,000 terms, organised in six facets: `Location`, `Person name`, `Name`, `Maker`, `Genre` and `Subject`. `Location` describes either the place(s) where the video was shot, the places mentioned or seen on the screen or the places the video is about. `Person name` is used for people who are either seen or heard in the video, or who are the subject of the program; `Name` has the same function for named organisations, groups, bands, periods, events, etc. `Maker` and `Genre` describe the creators and genre of a TV program. The `Subject` facet is used to describe what a program is about, and aims to contain terms for all topics that could appear on TV, which makes its scope quite broad.

Since our aim is to retrieve video based on what it is about, the focus of the present paper is on the Subject facet. In the future we intend to include also other facets, such as Locations and Names. The Subject facet contains 3878 terms[1]. It is organised according to the semantic relations defined in the ISO-standard 2788 for thesauri [8], namely `Broader Term` (linking a specialized concept to a more general one), its inverse relation `Narrower Term` (linking a general concept to a more specialized one), and `Related Term` (denoting an associative link). The GTAA contains 3,382 broader/narrower relations and 7,323 associative relations between terms in the subject facet. The broader/narrower hierarchy is shallow: 85% of the hierarchy is no more than three levels deep. For integration purposes, we used a version of the GTAA that was converted to SKOS in an earlier effort [1]. SKOS provides a common data model to represent thesauri using RDF and port them to the Semantic Web [12].

2.2 Anchoring to WordNet

The archives of the Netherlands Institute for Sound and Vision are searched by broadcast professionals, who reuse material to create new television programs, and by the general public. Querying large audiovisual archives remains difficult. A recognised approach to increase recall is query expansion: retrieving documents with not only the initial query term, but also closely related terms [18]. If the aim is to use thesaurus relations for the expansion (as opposed to using lexical techniques), a rich thesaurus with many interrelated terms is necessary. To increase the structure of the GTAA, we anchor it to an external thesaurus, WordNet, and take advantage of its rich semantic structure.

WordNet is a lexical database of the English language. It currently contains 155,287 English words: nouns, verbs, adjectives and adverbs. Many of these words are polysemous, which means that one word has multiple meanings or senses. The word 'tree', for example, has three word-senses: tree#1 (woody plant), tree#2 (figure) and Tree#3 (English actor). WordNet distinguishes 206,941 word-senses.

Word-senses are grouped into synonym sets (synsets) based on their meaning and use in natural language. Each synset represents one distinct concept. An example of a synset is cliff#1, drop#4, drop-off#2, described as "a steep high face of rock". Semantic relations and lexical relations exist between word-senses and between synsets. For the purpose of this paper we will not go into details

[1] In addition to the 3,878 preferred terms, the subject facet contains around 2,000 non-preferred terms.

of all these relations, but rather explain the most common ones. The main hierarchy in WordNet is built on hypernym/hyponym relations between synsets, which are similar to superclass/subclass relations. Other frequent relations are meronym and holonym relations, which denote part-of and whole-of relations respectively. Each synset is accompanied by a 'gloss': a definition and/or some example sentences.

WordNet is freely available from the Princeton website[2]. In addition, W3C has released a RDF/OWL representation of WordNet version 2.0[3]. In this study we use this RDF/OWL version, as it allows us to use Semantic Web tools and standards to query the WordNet database.

Anchoring GTAA to WordNet is non-trivial, since the two thesauri are in different languages. As a first step, we used a Dutch lexical database (Celex) to find alternative forms of the terms in the thesaurus. In addition to the original preferred terms and non-preferred terms, we added singular forms (since WordNet words are mostly singular) and synonyms. Compound terms were split into separate words (again using Celex), which were also added. All forms, the original ones as well as the newly added ones, were used for the anchoring to WordNet as this increases the possible coverage of the anchoring.

Second, we queried an online bilingual dictionary[4] for the Dutch terms, which provided English translations and one-sentence descriptions. Third, we anchored the English GTAA terms to WordNet. In contrast to many anchoring methods (e.g. [9]), we do not compare the terms from the two thesauri, but measure the lexical overlap of their descriptions. The same approach has been followed by Knight [10]. This approach is especially well suited in our case since, much to our surprise, the one-sentence descriptions of the online dictionary *are* the WordNet glosses for 99% of the words. The anchoring process is described in more detail in [11].

GTAA terms that were found to correspond to multiple WordNet synsets were anchored to all those synsets. There were three reasons why we didn't attempt to do sense disambiguation. First, we are aiming for an increased recall so our primary focus is finding correct correspondences rather than avoiding incorrect correspondences. Second, disambiguation of terms with little context (which is the case for the GTAA terms) is difficult. In the future, we intend to take into account broader terms for disambiguation purposes. The third and most important reason is that linking to more than one synset is often correct because WordNet makes finer distinctions than the GTAA. For example, WordNet distinguishes four meanings for the GTAA term "chicken", described by the glosses: 'adult female chicken', 'the flesh of a chicken used for food', 'a domestic fowl bred for flesh or eggs' and 'a domesticated gallinaceous bird thought to be descended from the red jungle fowl'. This fine-grained distinction is absent in the GTAA.

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

[3] http://www.w3.org/TR/wordnet-rdf/

[4] http://lookwayup.com

In total, 1,855 GTAA terms were anchored to WordNet. 885 of those corresponded to only one synset, 464 corresponded to two synsets, 242 to three synsets, 121 to four, 76 to five, 35 to six and 32 corresponded to seven or more synsets. Some WordNet synsets are linked to more than one GTAA term. For example, the WordNet synset "Studio" was an anchor for both the GTAA terms "Atelier" and "Studio". An informal evaluation of a small sample of the correspondences suggests that the number of synsets that is aligned with a particular GTAA term is not an indication of the quality of the matches; GTAA terms that are matched to multiple synsets are equally well matched as GTAA terms that are matched to only one synset.

Correspondences based on split compound words are less good than those based on original preferred terms or singular forms. However, we estimate that only 10% is actually incorrect. The majority anchors a term to a related or broader synset.

3 Thesaurus Enrichment

3.1 Approach

We used the anchoring to WordNet to infer new relations within the GTAA. Using SeRQL [2] queries we related pairs of GTAA subject terms that were not previously related. Figure 1 illustrates how a relation between two terms in the GTAA, t_1 and t_2, is inferred from their correspondence to WordNet synsets w_1 and w_2. If t_1 corresponds to w_1 and t_2 corresponds to w_2, and w_1 and w_2 are closely related, we infer a relation between t_1 and t_2. The inferred relation is symmetric, illustrated by the two-way arrow between t_1 and t_2.

Two WordNet synsets w_1 and w_2 are considered to be 'closely related' if they are connected though either a direct (i.e. one-step) relation without any intermediate synsets or an indirect (i.e. two-step or three step) relation with one or two intermediate synsets. The latter situation is shown in Figure 1. From all WordNet relations, we used only meronym and hyponym relations, which roughly translate to part-of and subclass relations, and their inverses holonym and hypernym. A previous study demonstrated that other types of WordNet relations do not improve retrieval results when used for query expansion [7]. Both meronym and hyponym can be considered hierarchical relations in a thesaurus. Only sequences of two relations are included in which each has the same direction, since previous research showed that changing direction, especially in the hyponym/hypernym hierarchy, decreases semantic similarity significantly [7, 5]. For example, w_a hypernym of w_b hyponym of w_c is not included.

3.2 Newly inferred relations

A total of 1039 pairs of GTAA terms was newly related: 404 with one step between WordNet synsets w_1 and w_2, 362 with 2 steps and 273 with three steps between w_1 and w_2. Around 90% of the relations were derived from hyponym

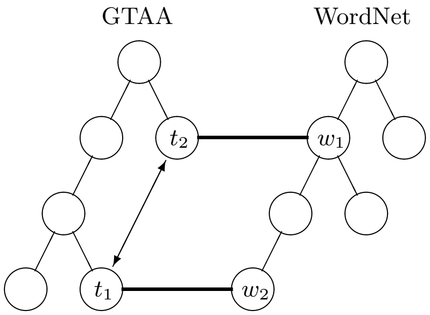

Fig. 1. Using the anchoring to WordNet to infer relations within the GTAA.

relations and only 10% from meronym relations, which is a more rare relation in WordNet.

Although we intend to only implicitly evaluate the quality of the inferred relations by looking into their value for retrieval, a manual inspection of a portion of the new relations suggests that many of them have the potential to be beneficial for retrieval. At the same time, many others seem so trivial that we don't expect much added value. Only very few seem wrong. We did not detect a difference in quality between relations inferred from hyponyms and those inferred from meronyms. Examples of new relations that we consider valuable are:

`squid`	- `colouring (pigment)`	*(hyponym 1 step)*
`pharmacy`	- `medicine`	*(hyponym 1 step)*
`barbecues`	- `picknicks`	*(hyponym 2 steps)*
`national anthems`	- `music`	*(hyponym 3 steps)*
`pearls`	- `jewelery`	*(hyponym 3 steps)*
`fjords`	- `seas`	*(meronym 1 step)*
`cement`	- `concrete`	*(meronym 2 steps)*
`flour`	- `meal`	*(meronym 3 steps)*

Relations that we consider trivial are, for example:

`cigarettes`	- `cigars`	*(meronym 1 step)*
`computers`	- `machines`	*(hyponym 1 step)*
`coffeeshops`	- `restaurants`	*(hyponym 1 step)*

Examples of incorrect new relation are:

`acupuncture`	- `negotiation`	*(hyponym 1 step)*
`banknotes`	- `copies`	*(hyponym 1 step)*
`apples`	- `foetusses`	*(hyponym 2 step)*

Inferred relations between pairs of GTAA terms that were already each others `Broader Term`, `Narrower Term` or `Related Term` were not included in the retrieval experiment, nor in the above numbers, since they do not *add* to the structure of the GTAA. The considerable overlap between what we inferred and what was already present in the GTAA is, however, an indication that the inferred relations make sense.

4 Retrieval with the enriched thesaurus

We employed the enriched thesaurus for retrieval of television programs from the archives of the Netherlands Institute for Sound and Vision. The programs were annotated with subject terms from the GTAA. Our aim is twofold. First, we want to know the value of the inferred relations for retrieval, and compare that to retrieval with existing GTAA relations. Second, we are interested in the added value of the inferred relations when we use them in combination with the existing GTAA relations.

4.1 Experimental setup

We query the collection in nine runs, each using a different type of relation or combination of relations:

Exact Only programs annotated with the query term are returned. This run is used as a baseline.

GTAA bro Programs annotated with the query term or broader terms are returned.

GTAA nar Programs annotated with the query term or narrower terms are returned.

GTAA rel Programs annotated with the query term or related terms are returned.

GTAA all Programs annotated with the query term or terms that are related through (a combination of) GTAA relations (narrower, broader, related) are returned.

Via WN 1 step Programs annotated with the query term or terms related through a one-step inferred relation are returned.

Via WN 2 step Programs annotated with the query term or terms that are related through a two-step inferred relation are returned.

Via WN 3 step Programs annotated with the query term or terms that are related through a three-step inferred relation are returned.

Via WN all Programs annotated with the query term or with a term that is related through a (combination of) one-, two- or three-step inferred relations are returned.

All Programs annotated with the query term or terms that are related in any of the above ways are returned.

At present, we allowed three steps between the query term and the target term. More than three steps resulted in an explosion of the number of returned documents.

Of each run, we measure the precision (Prec), recall (Rec) and the harmonic mean of the two, called F_1-measure:

$$\text{Prec} = \frac{|\text{Retrieved\&Relevant}|}{|\text{Retrieved}|} \quad \text{Rec} = \frac{|\text{Retrieved\&Relevant}|}{|\text{Relevant}|}$$

$$F_1 = 2 \cdot \frac{\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$$

where |Retrieved| is the number of programs a run retrieved, and |Relevant| is the number of programs that is relevant for a query.

4.2 TRECVid data: corpus and queries

In order to determine the added value of the inferred relations, a dataset and a ground truth are needed that are large enough to distinguish if there are any significant differences between runs. In the current study, we used the TRECVid 2007 dataset for the high-level feature extraction task. This dataset consists of 50 hours of news magazine, science news, news reports, documentaries, educational programming, and archival video from the Netherlands Institute for Sound and Vision, 36 queries ('features') and a manually constructed ground truth. Sound & Vision kindly provided us with the metadata of this dataset in the form of manual annotations of the television programs with GTAA terms.

The queries consist of a single or moderately complex query term, such as **Sports** or **Explosion_Fire**. This corresponds to the types of queries that are posed in the online search interface of Sound & Vision, where the majority of queries consist of a single term, sometimes completed with a broadcast date. Simple, unequivocal queries are a requirement in this type of study, as complex queries could obscure the results.

We manually translated the high-level features to get queries in terms of GTAA subjects. Features that consisted of two subjects were interpreted as the union of both and we queried for programs containing one and/or the other. This was clearly the intended semantics of the features as can be seen from descriptions such as the one for **Walking_Running**: `Shots depicting a person walking or running`. Of the initial 36 features, three did not have a satisfactory translation, and were therefore discarded.

All TRECVid tasks are at the level of shots, while the GTAA subject annotations are at the level of television programs. We adapted the given ground truth to be on program-level. In the resulting ground truth, nine queries appeared in more than 2/3 of the programs and were therefore discarded. `Person` and `Face`, for example, appeared in each program. Six programs were not usable in the present experiment since they did not have a subject annotation and could therefore never be retrieved.

After adaptation, the dataset consisted of 104 television programs annotated with on average 3.6 GTAA subject terms, 25 queries and a ground truth that listed on average 27 relevant programs for each query.

We stress that although we use the TRECVid dataset, our results are not comparable to those of systems that participated in the TRECVid 2007 conference. We retrieve programs based on metadata and the structure of a thesaurus, while TRECVid participants retrieve shots based on the audiovisual signal.

5 Results and Interpretation

Table 1 and Figure 2 summarize the results. Please note that although the range of the y-axis of the plots in Figure 2 differ, the height of the bars represents the same value in all three plots.

Throughout this section we use Students paired t-test to compare the performance of runs[5]. Significance levels, t-values, degrees of freedom and the appropriate version of the test (one or two tailed) will be reported as, for example, ($t =$, $p =$, $df =$, one-tailed).

5.1 Existing GTAA relations

The results of the runs using existing thesaurus relations merely confirm what was known about thesaurus based retrieval. We discuss them since they form a baseline against which we can compare the performance of the inferred relations. The human entered subject terms are reliable, and using them gives high precision, in our case even 100% (the 'exact' run). We suspect that the level of correctness of our annotations was higher than usual thanks to the special attention the Netherlands Institute for Sound and Vision gave to the collection they prepared for TRECVid. In many cases, of course, human annotators do err and disagree [17]. The time-consuming nature of human annotation causes the number of subject terms per program to be low, much lower than the number of topics that is visible in the video. This makes the recall of the run that relies solely on these human annotations unacceptably low: 2% on average.

Including terms that are broader than the query does not add to recall. This is partly due to the fact that our queries are all fairly general, and many don't have a broader term. Still, it is a confirmation of what was found in an earlier study [7]. Narrower terms, on the other hand, do seem to add a little to recall, although the result is not statistically significant (t =1.51, p =0.07, df =24, one-sided), and they maintain a high precision. This is what we would expect from the definition of narrower terms: "the scope (meaning) of one falls completely within the scope of the other" [13]. Related terms are less reliable:

[5] The t-test requires a normal distribution. Normality was assessed with Quantile-Quantile plots. Although for some of the smaller samples - the exact run, for example, returned programs for only seven queries and had therefore only 7 precision values - normality could not be proven, we assume that precision and recall are normally distributed quantities given a large number of queries.

Run	Precision	Recall	F_1
GTAA exact	1.00 ± 0.00	0.03 ± 0.06	0.17 ± 0.11
GTAA broader	0.81 ± 0.35	0.03 ± 0.06	0.16 ± 0.10
GTAA narrower	0.89 ± 0.30	0.04 ± 0.07	0.20 ± 0.12
GTAA related	0.42 ± 0.26	0.33 ± 0.23	0.29 ± 0.15
GTAA all	0.39 ± 0.25	0.46 ± 0.27	0.34 ± 0.17
Via WN one-step	0.70 ± 0.40	0.05 ± 0.07	0.18 ± 0.10
Via WN two-step	0.76 ± 0.30	0.07 ± 0.11	0.20 ± 0.12
Via WN three-step	0.81 ± 0.38	0.04 ± 0.06	0.16 ± 0.10
Via WN all	0.58 ± 0.40	0.13 ± 0.17	0.26 ± 0.14
All	0.38 ± 0.25	0.57 ± 0.29	0.38 ± 0.19

Table 1. Precision, recall and F_1-measure of the nine runs, summarized by the mean ± the standard deviation.

precision halves compared to using only exact matches (t =7.14, p <0.01, df =7, two-sided), but recall increases to 33% (t =6.63, p <0.01, df =24, one-sided).

Combining the hierarchical broader/narrower relations with the related terms, only slightly (but significantly) lowers precision further compared to using only the related terms (t =1.9, p =0.03, df =24, two-sided). It does, however, raise recall to 46% (t =4.3, p <0.01, df =24, one-sided). This suggests that also sequences of different types of relations are beneficial to retrieval.

5.2 Newly inferred relations

The one-, two- and three-step inferred relations perform equally well. We observe a small difference in precision (three-step is higher), but this is not significant (t =1.3, p =0.24, df = 7, two-sided). This suggests that the notion of relatedness can be interpreted in a broad sense and does not need to be restricted to only one step in the WordNet hierarchy.

When the one-step, two-step and three step inferred relations are combined ('Via WordNet All') precision remains relatively high: 58%. Recall, on the other hand, is low: 13%. On the whole, the inferred relations give results that are comparable to the results of existing relations in the GTAA that were created by experts. When comparing them to GTAA narrower terms, they score better on recall but worse on precision. When comparing them to GTAA related terms we observe the opposite effect: the inferred relations score higher on precision but lower on recall. This difference in recall can in part be explained from the fact that there are 7 times as many related terms as inferred relations. Also with respect to F_1-measures, the performance of the inferred relations is between that of GTAA narrower and GTAA related terms. There were no significant differences between the F_1 of inferred relation and GTAA narrower terms (t =0.9, p =0.40, df = 8, two-sided) or GTAA related terms (t =2.0, p =0.07, df = 12, two-sided). These results show that the inferred relations are valuable for retrieval in situations where there is no other structure in the vocabulary.

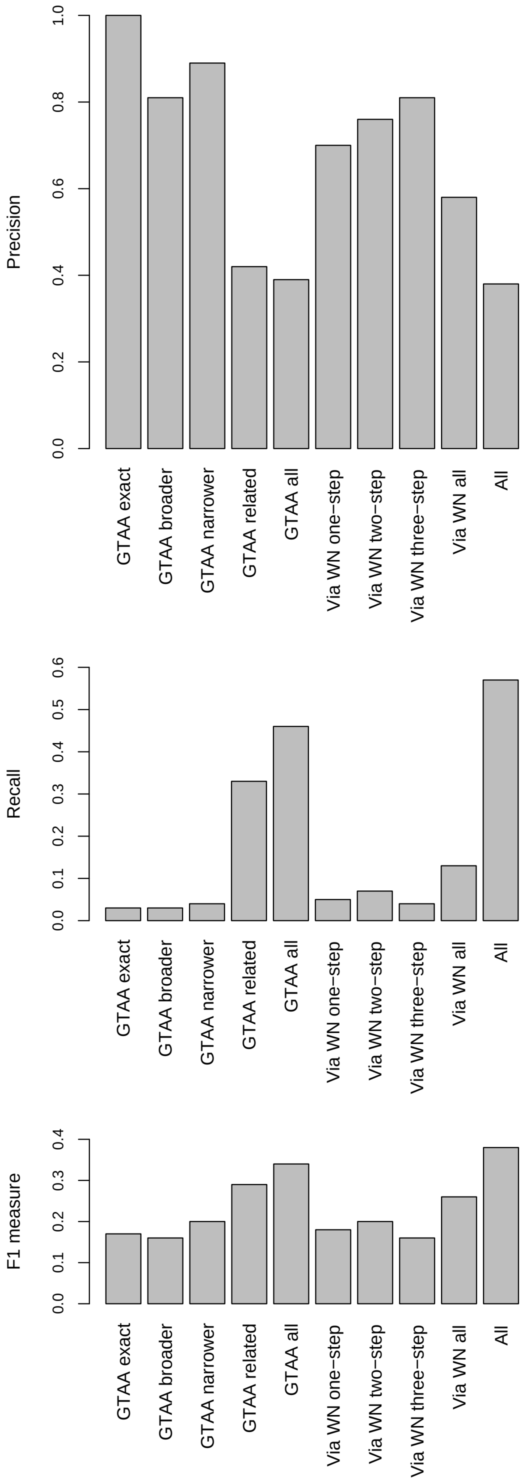

Fig. 2. Retrieval results with different thesaurus relations

Using all relations together improves the recall significantly over using only the existing GTAA relations (t =4.0, p <0.01, df =24, one-sided). Again, this suggests that combination of different types of relations is beneficial to the retrieval results. It also suggests that enrichment of a weakly structured thesaurus has added value to the retrieval results.

The mean increase in recall from the 'GTAA all' run to the 'All' run was 0.11. This increase could in part be attributed to the higher number of retrieved programs. However, the increase in recall was significantly more than we would expect if the additionally retrieved programs were randomly taken from the collection (t =2.40, p =0.02, df =27, two-sided). We calculated the expected increase in recall $\mathbb{E}_{incr}$ for each query as follows:

$$\mathbb{E}_{incr} = \frac{(R_{All} - R_{GTAAall}) \cdot (C - RC_{GTAAall})}{N - R_{GTAAall}} \cdot \frac{1}{C}$$

where R_{All} and $R_{GTAAall}$ are the number of retrieved programs in the 'All' and 'GTAA all' runs respectively, C is the number of correct programs for the query in the collection, $RC_{GTAAall}$ is the number of correctly retrieved programs in the 'GTAA all' run and N is the total number of programs in the collection (104 in our case).

6 Discussion and future work

In this paper we experimented with retrieval using a thesaurus that was enriched by anchoring it to an external resource. We have shown that with simple techniques new relations can be inferred that are valuable for retrieval purposes.

We investigated both the effect of only using the newly inferred relations and using them in combination with existing thesaurus relations. Retrieval with only the inferred relations yielded an F_1-measure of around 0.2. This is comparable to the intuitive and widely used approach of using `Narrower Term` thesaurus relations. This finding suggests that it is possible to use an external resource to enrich an otherwise unstructured vocabulary and base the retrieval on the inferred relations. For example, we see possibilities to enrich lexicons of high-level feature detectors that are used in content-based video retrieval. LSCOM is such a vocabulary for annotation and retrieval of video, containing concepts that represent realistic video retrieval problems, are observable and are (or will be) detectable with content-based video retrieval techniques [14]. In a recent effort, LSCOM was manually linked to the CyC knowledge base[6], thus creating structure within LSCOM. We would be interested to compare and combine this manually added structure with an enriched version of LSCOM using the methods proposed in the present paper.

When the inferred relations are used in combination with existing thesaurus relations, it appears that the use of the enriched set of relations increases recall moderately (from 0.46 in the 'GTAA all' run to to 0.57 in the 'all' run) with

[6] `http://www.cyc.com/`

comparable precision (0.39 vs. 0.38). This indicates that it is beneficial to enrich an already structured thesaurus. However, the number of additionally retrieved documents was too low to draw any definite conclusions about the added value of the inferred relations over an already structured thesaurus. We suspect that a higher number of inferred relations will be needed to confirm this finding. Future research directions therefore include alternative methods of thesaurus enrichment.

When looking at the F-measure scores it is interesting to note that mixing relationships increases performance, both in the non-enriched ("GTAA all") and in the enriched ("all") case. This suggests that the nature of the relationship (broader, narrower, related) is not a big issue, at least not in this case. It would be worthwhile to study this in more detail in future experiments.

In future work we want to consider the effect of the inferred relations in more detail. Further experiments could reveal what the effect of different WordNet relations is: hyponyms, meronyms, but also other relations that were not yet used. A further distinction between types of inferred relations, which could start at the anchoring phase, will give more insight into the optimal ranking strategies for semantic search results.

The use of TRECVid data enabled us to experiment on a dataset of reasonable size. However, it also raises some issues. The TRECVid ground truth is based on the pooled results of TRECVid 2007 participants: only items returned by at least one of the participants are judged, while items not returned by anyone are considered incorrect. This could in theory lead to a negative image of our results, since we did not contribute to the pool. It is been argued that this is a negligible problem. Zobel [19], for example, demonstrated that the difference in rating between a system in- and outside the pool is small. However, all systems in his test were content-based image retrieval systems. The retrieval approach under consideration in the present paper is concept-based. Since we use another type of information (metadata instead of the audiovisual signal), it is well possible that we retrieve a set of documents that is disjoint from the set that was retrieved by the content-based systems that contributed to the pool. Therefore, the effect of being outside the pool is potentially larger.

The translation of TRECVid topics to GTAA terms was done manually. In a final application this translation would be done either automatically, which is done in [16], or by the searcher, as in [6]. However, in the present paper our goal was not to build an application but to investigate the possibilities of retrieval with an automatically enriched thesaurus.

References

1. M. Assem, van, V. Malaisé, A. Miles, and A. Th. Schreiber. A method to convert thesauri to skos. In *In Proceedings of the Third European Semantic Web Conference*, pages 95–109, Budvar, Montenegro, 2006.
2. J. Broekstra and A. Kampman. SeRQL: A second generation RDF query language. In *Proceedings of the SWAD-Europe Workshop on Semantic Web Storage and Retrieval*, pages 13–14, Amsterdam, The Netherlands, November 2003.

3. J. Euzenat, A. Isaac, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, O. Šváb, V. Svátek, W. R. van Hage, , and M. Yatskevich. First results of the ontology alignment evaluation initiative 2007. In B. Ashpole, M. Ehrig, J. Euzenat, and H. Stuckenschmidt, editors, *Ontology Matching*, CEUR Workshop Proc., 2007.
4. C. Fellbaum, editor. *WordNet: an electronic lexical database*, Cambridge, MA, USA, 1998. MIT press.
5. G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. The MIT Press, 1998.
6. L. Hollink, A. Th. Schreiber, J. Wielemaker, and B. J. Wielinga. Semantic annotation of image collections. In *Proceedings of the K-Cap 2003 Workshop on Knowledge Markup and Semantic Annotation*, October 2003.
7. L. Hollink, G. Schreiber, and B. Wielinga. Patterns of semantic relations to improve image content search. *Journal of Web Semantics*, 5:195–203, 2007.
8. International Organization for Standardization. *ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri.* ISO, Geneva, 1986.
9. L. R. Khan and E. Hovy. Improving the precision of lexicon-to-ontology alignment algorithm. In *AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA, USA, October 1997.
10. K. Knight and S. Luk. Building a large-scale knowledge base for machine translation. In *the AAAI-94 Conference*, 1994.
11. V. Malaisé, A. Isaac, L. Gazendam, and H. Brugmann. Anchoring dutch cultural heritage thesauri to wordnet: two case studies. In *ACL 2007 Workshop on Language Technology for Cultural Heritage Data*, 2007.
12. A. Miles and S. Bechhofer. SKOS simple knowledge organization system reference. W3C working draft, 25 January 2008. Electronic document. Accessed April 2008. Available from: http://www.w3.org/TR/skos-reference/.
13. A. Miles and D. Brickley. SKOS core guide. W3C working draft, November 2005. Electronic document. Accessed February 2008. Available from: http://www.w3.org/TR/swbp-skos-core-guide/.
14. M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.
15. P. Over, W. Kraaij, and A. F. Smeaton. TRECVID 2007 - an introduction. In *TREC Video Retrieval Evaluation Online Proceedings*, 2007.
16. C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9(5):975–986, August 2007.
17. T. Volkmer, J. A. Thom, and S. M. M. Tahaghoghi. Exploring human judgement of digital imagery. In *ACSC '07: Proceedings of the thirtieth Australasian conference on Computer science*, pages 151–160, Darlinghurst, Australia, Australia, 2007. Australian Computer Society, Inc.
18. E. Voorhees. Query expansion using lexical-semantic relations. In W. B. Croft and C. J. Rijsbergen, van, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag.
19. J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314, New York, NY, USA, 1998. ACM.