# Practice-oriented Evaluation of Unsupervised Labeling of Audiovisual Content in an Archive Production Environment

Victor de Boer[1,2] and Roeland J.F. Ordelman[1,3] and Josefien Schuurman[1]

[1] Netherlands Institute for Sound and Vision, Hilversum, The Netherlands
[2] The Network Institute, VU University Amsterdam, The Netherlands
[3] University of Twente, Enschede, The Netherlands

**Abstract.** In this paper we report on an evaluation of unsupervised labeling of audiovisual content using collateral text data sources to investigate how such an approach can provide acceptable results given requirements with respect to archival quality, authority and service levels to external users. We conclude that with parameter settings that are optimized using a rigorous evaluation of precision and accuracy, the quality of automatic term-suggestion are sufficiently high. Having implemented the procedure in our production work-flow allows us to gradually develop the system further and also assess the effect of the transformation from manual to automatic from an end-user perspective. Additional future work will be on deploying different information sources including annotations based on multimodal video analysis such as speaker recognition and computer vision.

**Keywords:** audiovisual access, information extraction, thesaurus, audiovisual archives, practice-oriented evaluation

## 1 Introduction

Traditionally, audiovisual content in digital libraries is being labeled manually, typically using controlled and structured vocabularies or domain specific thesauri. From an archive perspective, this is not a sustainable model given (i) the increasing amounts of audiovisual content that digital libraries ingest (quantitative perspective), and (ii) a growing emphasis on improving access opportunities for these data (qualitative perspective). The latter is not only addressed in the context of traditional search, but increasingly in the context of linking within and across collections, libraries, and media. Ultimately, search and linking is shifting from a document-level perspective towards a segment-level perspective in which segments are regarded as individual, 'linkable' media-objects. In this context, the traditional, manual labeling process requires revision to increase both quantity and quality of labels.

In earlier years, we investigated optimization of the labeling process from a "term suggestion" perspective (see e.g., [1]). Here the aim was to improve

efficiency and inter-annotator agreement by generating annotation *suggestions* automatically from textual resources related to the documents to be archived. In [2] we defined collateral data[4] to refer to data that is somehow related to the primary content objects, but that is not regarded as metadata, such as subtitles, scripts and program-guide information. Previous work at our archive emphasized the ranking of possibly relevant terms extracted from the collateral text data, leaving the selection of the most relevant terms to the archivist. The proposed term suggestion methods were evaluated in terms of Precision and Recall by taking terms assigned by archivists as 'ground-truth'. The outcome was that a tf.idf approach gave the most optimal performance in combination with an importance weighting of keywords on the basis of a Pagerank-type of analysis of keywords within the structure of the used thesaurus ($F@5 = 0.41$). One important observation of the study was that the inter-annotator agreement was limited, with an average agreement of 44%.

Although the results were promising, the evidence provided by the study was not conclusive enough to justify adaptations of the archival annotation work-flow and incorporate the suggested methodology. However, as the assumptions that drove the earlier study are still valid and have become even more clear and pressing, we recently took up the topic again. This time however from the perspective of fully *unsupervised* labeling. The main reason for this is that we expect that the efficiency gain of providing suggestions in a supervised labeling approach is too limited in the context of the increasing amounts of data that need labeling. Furthermore, instead of relying on topically condensed text sources such as program guide descriptions used in the previous study, we include a collateral text source more easily available in our production work-flow: subtitles for the hearing impaired. Finally, as inter-annotator agreement is expected to be limited given the earlier study, we wanted to investigate how this agreement relates to an unsupervised labeling scenario that aims to generate labels for improving access to audiovisual collections. This makes our task different from more generic classification or tagging tasks such as done in the MUMIS project[3].

In this paper, we present an evaluation of unsupervised labeling focusing on *the practical usage of the method in an archive production environment.* In Section 2 we overview the archival context of the labeling approach. In Section 3 we present the automatic term extraction framework that we used for the evaluations described in Section 4. Section 5 discusses and concludes the results from the evaluation, followed by some notes on future work.

## 2 Archival Context

The implementation of innovative processes for automatic content annotation in an archive production work-flow needs to be addressed critically. A key requirement with respect to this type of innovation is that the archive remains in control of the quality of the automatically generated labels. Not only because of

[4] This data is sometimes also referred to as 'context data' but as for example newspaper data can also be regarded as 'context' we prefer the term 'collateral data'.

principals of archival reliability and integrity, but also from a service-level point of view. Media professionals use a broadcast archive to search for footage that can be re-used in new productions. The probability that their search process will get disturbed due to incorrect automatic labeling is undesired, despite the fact that the overall number of entry points generated by the automatic tool will increase, potentially having a positive effect on the search process.

Authority, being in control of the quality of the annotation tool, also means having control on parameters of the tool. In the case of automatic term labeling two important variables are: (i) quality, specifically the balance between Precision and Recall –or from a classification perspective: Hits and False Positives versus Misses– that controls the relation between quantity and quality of generated labels, and (ii) the vocabulary that in an archival setting could be closely related to controlled vocabularies or thesauri that are used. In this work, the automatic labeling process is required to output terms that are defined in the Common Thesaurus for Audiovisual Archives[5] (GTAA). The GTAA closely follows the ISO-2788 standard for thesaurus structures and consists of several facets for describing TV programs: subjects, people mentioned, named entities (Corporation names, music bands etc), locations, genres, producers and presenters. The GTAA contains approximately 160.000 terms and is updated as new concepts emerge on television. For the implementation of unsupervised labeling in the archive's metadata enrichment pipeline, the balance between Precision and Recall, and the matching of candidate terms with the thesaurus have the main focus of attention.

### 2.1 Data

The general aim of the project for which the evaluation described in this paper was performed, is to label automatically the daily ingest of Radio and Television broadcasts. This data is quite heterogeneous: it contains news broadcasts, documentaries and talk shows but also sports and reality shows. As named-entity extraction tools typically perform better for common entities as opposed to less common ones, we assume that the performance will differ for different genres.

For each program that is ingested also subtitles for the hearing impaired (TT888) –a verbatim account of the (Dutch) speech present in the data– is flowing into the archive. These TT888 files are used as input for the term-extraction pipeline described in Section 3. Instead of subtitles also other collateral data such as program guide information or production scripts and auto-cues could be used. As the availability of these data is less stable as is the case for subtitles, we focus on subtitles in the current work-flow.

For evaluation purposes we selected one year of previously ingested programming for which we have manually generated labels, created by professional archivists. This set will be referred to as 'gold-standard' in our (pilot) experiments. However, as such a gold-standard implies exact matches or terminological

[5] http://datahub.io/dataset/gemeenschappelijke-thesaurus-audiovisuele-archieven

consistency, we also asked professional archivist to assess the conceptual consistency (see also [4] about consistency, [1] for the approach that was taken earlier).

As discussed above, we use the internal thesaurus as a reference for extracted terms. The GTAA is available as Linked Open Data[5] and its concepts are identified through URIs. In the production system the extracted terms end-up as URIs identifying GTAA concepts unique IDs, which in turn can also be linked to and from using RDF relations. This allows us to in the near future reuse background information in the Linked Data cloud insofar as it is linked to or from those GTAA concepts. For the evaluation described here, the term-extraction pipeline only used the "subject" and "named-entities" facets of the thesaurus for validation.

## 3 Automatic Term Extraction

An overview of the term extraction pipeline is presented in Figure 1. This shows the different steps performed in the algorithm, detailed below. The term-extraction pipeline is set up as a webservice. The webservice takes a single text, such as the subtitles for a television broadcast, as input and outputs a list of relevant thesaurus terms. This set-up allows the service to be re-used for other related tasks such as the extraction of terms from digitized program guides or other collateral data sources.

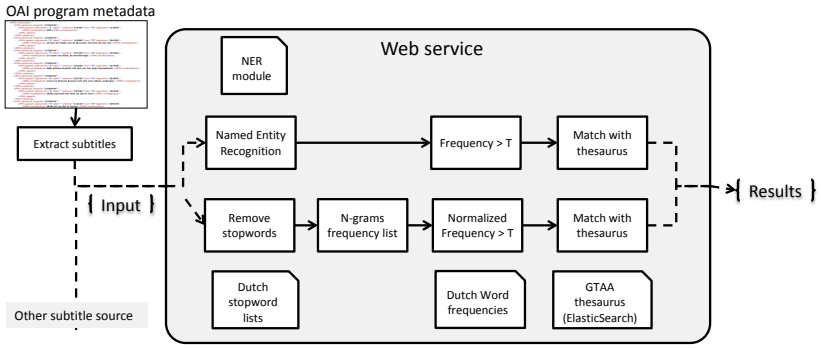


**Fig. 1.** Overview of the algorithm

The web service is called through a HTTP post request, where the input text is passed in the body as a JSON string. At the same time, parameter settings can be passed in the same HTTP request to override default values for these parameters (see Section 3.4 for the parameters).

The output is a JSON object containing a list of thesaurus terms, on the basis of the parameter settings used (if not overridden, the default values are returned). For every term, also a *matching score* is returned (see Section 3.3).

Within the archive production workflow, the service is called when new programs are ingested. The thesaurus terms provided by the service are then added to the program metadata without manual supervision.

For the experiments described below, the subtitles are derived from an OAI-PMH interface[6] to the archive's database. We retrieves for one or more programs the subtitle information from the OAI response (the program metadata) and remove the temporal metadata and other XML markup from the subtitles so that we end up with a single subtitle text per program. These are then presented one at a time to the service. As the extraction of subject terms and named entities require an individual tuning of parameters, the textual data is processed in two parallel tracks: one for subject terms and one for named entities (see Figure 1).

### 3.1 Pre-processing and filtering

For the subject track, the first pre-processing step is to remove stopwords using a generic list of Dutch stopwords. In the next step, frequencies for 1, 2, and 3-grams are generated. For the uni-grams (single terms) also normalized frequencies are calculated using a generic list of Dutch word frequencies obtained from a large newspaper corpus.In the filtering step, candidate terms (in the form of n-grams) above a certain threshold value of frequency scores are selected. Frequency scores are based upon both the absolute frequency (how often a term occurs in the subtitles) and a relative frequency (normalized by the frequency of the term in the Dutch language, only for 1-grams). The frequency thresholds are parameters of the service. In the next phase, candidate n-gram terms are matched with terms in the thesaurus.

### 3.2 Named Entity Recognition

In the named-entity track of the algorithm, Named Entities (NEs) are extracted. Pilot studies determined that NEs –more so than non-entity terms– have a high probability of being descriptive of the program, especially if they occur in higher frequencies. For this track, we use a Named Entity Recognizer (NER). The NER is implemented as a separate module in the service and we experimented with different well-performing open-source systems NER systems for this module.

1. **XTAS**. The NER tool from the open-source xTAS text analysis suite. [7]
2. **CLTL**. An open-source NER module developed at the CLTL group[8].

In the current Web service, the NER module to be used is a parameter of the method and can be set to "XTAS" or "CLTL" for the respective services. Both modules are implemented as wrappers around existing services which take as input a text (string) and as output a JSON list of entities and their types. The types used by the web service are *person*, *location*, *organization* or *misc*. Internal NE types from the individual modules are mapped to these four types

[6] `http://www.openarchives.org/pmh/`

[7] `http://xtas.net/`. Specifically, the FROG module was used using default settings.

[8] `http://www.cltl.nl/` Here the OpenNER web service was used in combination with the CLTL POS tagger

### 3.3 Vocabulary matching

The previous phases yield candidate terms to be matched against the thesaurus of five categories: subjects (from the subject track) and persons, places, organizations, and miscellaneous (from the NE track). The next step in the algorithm identifies the concepts in the thesaurus that match these terms. As there can be many candidate terms at this stage and the GTAA thesaurus is fairly sizable with some 160.000 concepts, we need to employ a method for matching terms to thesaurus concepts that is scalable.

For this, the thesaurus has been indexed in an ElasticSearch instance[9]. ElasticSearch is a search engine that indexes documents for search and retrieval. In our case, thesaurus concepts are indexed as documents, with preferred and alternative labels as document fields. The concept schemes (facets or "axes" in the GTAA) are represented as different ElasticSearch *indices* which allows for fast search for term matches across and within a concept scheme. When searching for concepts matching a candidate term, ElasticSearch will respond with candidate matches and a *score* indicating the quality of the match between candidate term and the document. In our algorithm, we employ a threshold on this score, resulting in an additional parameter. In this final step, the different categories of candidate terms are matched to a specific concept scheme. For example, persons are matched to the "Persoonsnamen" (Person names) concept scheme in the GTAA thesaurus and both the subject terms and MISC are mapped to the "Onderwerpen" (Subject) concept scheme.

### 3.4 Parameters

The algorithm parameters are shown in Table 1. This table shows the parameter name, the default value and the description. All default values can be overridden in the HTTP POST request. These default values were determined in pilot experiments (Section 4.1) and the experiment described in Section 4.2 was used to determine optimal values for a number of these parameters for a specific task.

## 4 Experiments

### 4.1 Pilot experiments

We performed a number of pilot experiments to fine-tune the setup of the main experiment. In one of these pilot experiments, we compared the output of an earlier version of the algorithm to a gold-standard of existing manual annotations (see Section 2.1). The results showed that although there was some overlap[10], comparing to this gold standard was not deemed by the experts to be an informative evaluation, since many "false positives" identified by the algorithm were identified to be interesting nonetheless. Therefore in subsequent experiments,

[9] `http://www.elastic.co/products/elasticsearch`

[10] For this non-optimized variant, recall was 21%.

| nr. | Parameter name | Default | Description |
|---|---|---|---|
| P1 | tok.min.norm.freq | $4 \times 10^{-6}$ | threshold on normalized freq for 1-gram |
| P2 | tok.max.gram | 3 | Maximum N for topic N-grams |
| P3 | tok.min.gram | 2 | Minimum N for topic N-grams (excl. 1) |
| P4 | tok.min.token.freq | 2 | threshold on absolute freq for 1-gram |
| P5 | repository | cltl | NER module (xtas or cltl) |
| P6 | ne.min.token.freq | 2 | Threshold on absolute freq for all NEs |
| P7 | ne.organization.min.score | 8 | Threshold on ElasticSearch matching score |
| P8 | ne.organization.min.token.freq | 2 | Threshold on absolute freq for |
| P9 | ne.person.min.score | 8 | Threshold on matching score for persons |
| P10 | ne.person.min.token.freq | 1 | Threshold on absolute freq for persons |
| P11 | ne.location.min.score | 8 | Threshold on matching score for locations |
| P12 | ne.location.min.token.freq | 2 | Threshold on absolute freq for locations |
| P13 | ne.misc.min.score | 8 | Threshold on matching score for misc |
| P14 | ne.misc.min.token.freq | 2 | Threshold on absolute frequency for misc |

**Table 1.** Parameters and default values for the service

we presented the extracted terms to domain experts for evaluation. In this way, only precision of the suggested terms can be determined (no "recall"). This pilot also suggested that the correctness of suggested terms should be determined on a scale rather than correct or incorrect.

In a second pilot experiment, we presented extracted terms for random programs to four in-house experts and asked them to rate this on a five point Likert-scale [6]. The results were used to improve the matching algorithm and to focus more on the named entities rather than the generic terms since the matching here seemed to result in more successful matches. Lastly, in feedback to this pilot the experts indicated that for some programs the term extraction was considerably less useful than for others. This was expected but in order to reduce the amount of noise from programming that from an archival perspective has a lesser degree of annotation priority, we selected programs with a high priority[11]. For the main experiment we sampled from this subset rather than from the entire collection. From this evaluation we derived default parameter values shown in Table 1 which provided 'reasonable' results (not too many obvious errors) including for example the value for P4, P6, P8, P10, P12 and P14 (minimum frequencies for terms to be considered a candidate term).

In the main experiment, the goal was twofold: (i) to determine the quality of the algorithm and (ii) to determine optimal values for other system parameters.

### 4.2 Experimental Setup

For the main experiment, we randomly selected 18 individual broadcasts from five different Dutch television shows designated as being of high-priority by the archivist. These shows are the evening news broadcast (4 videos), two talk shows

[11] This prioritization is done by archivists independently of this work. It is in use throughout the archive and mostly determined by potential (re)use by archive clients.

(3+4 videos), a documentary show (4 videos) and a sports news show (3 videos). For these videos, we presented four evaluators with (a) the video, (b) the existing metadata (which did not include descriptive terms) ad (c) the terms generated by the algorithm using different parameter settings. The evaluators were asked to indicate the relevance of the terms for the video on a five-point Likert scale:

0: Term is totally irrelevant or incorrect,
1: Term is not relevant,
2: Term is somewhat relevant,
3: Term is relevant,
4: Term is very relevant

**Parameter Settings** Parameters P1-P4 were set to their default values as listed in table 1. These were established in the pilot experiments and proved reasonable for this specific task. For P5, we used both values, so both NER modules are evaluated. Some terms were found by both modules, and other terms were found by only one of the two. Evaluators did not see the origin of the terms. P6 was fixed to 2, as were the thresholds on the NE specific frequencies (P7, P9, P11, P13). For the Elasticsearch matching scores, we used a bottom threshold of 9.50 and presented all terms with a score higher than that value to the evaluators. We retain the scores so that in the evaluation we can compare the quality for threshold values of 9.50 and higher. The pilot studies showed that with thresholds below 9.50, mostly incorrect terms were added. The scores were also not available to the evaluators to avoid an evaluation bias. For the 18 videos, a total of 289 terms for XTAS and 222 terms for CLTL were presented to the evaluators.

### 4.3 Results

One of the evaluators (Eval4) finished all 18 videos. Table 2 shows the statistics for the four evaluators including the average score given for all terms. This shows that there is quite some disagreement among the averages. To measure inter-annotator agreement, we calculated the Pearson-coefficient between the pairs of evaluators. The results are shown on the right in Table 2. The agreement between Eval1 and Eval2 is rather low at 0.45, but for the other pairings it is on a more acceptable level. For most of the subsequent evaluations, we use the average score for an extracted term given by the evaluator.

| | | | Agreement | | |
|---|---|---|---|---|---|
| Evaluator | Nr. evaluated | Avg. score | Eval2 | Eval3 | Eval4 |
| Eval1 | 8 | 1.31 | 0.45 | 0.66 | 0.69 |
| Eval2 | 14 | 2.21 | 0.63 | 0.74 | |
| Eval3 | 6 | 1.57 | 0.73 | | |
| Eval4 | 18 | 1.64 | | | |

**Table 2.** Evaluator results (left) and inter-annotator agreement matrix (right)

**Named Entity Modules** To determine the difference in quality of the two NER modules, we separated the scores for the two values (CLTL and XTAS) and determined the average score. If all terms are considered (respectively 289 and 222 terms for XTAS and CLTL), the average score for XTAS is 1.79 and that for CLTL is slightly higher at 1.94. We can also plot the average scores of the two modules given a single threshold on the matching scores for the terms (in this case we use a single value for the threshold parameters P7, P9, P11 and P13). This is shown in Figure 2.

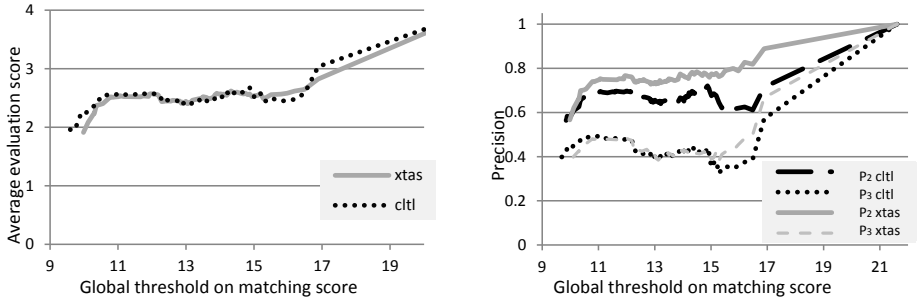


**Fig. 2.** Average scores (left) and precision graphs (right) for the global threshold values on matching score for the two NER modules

This figure shows that the performance of the two modules is very comparable. It shows that at very low thresholds ($< 10$), the performance for both modules indeed drops considerably. Investigation of the data shows that below 10, mostly terms with average score 0 are added, which corresponds with findings from the pilot study. Furthermore, the graph shows that increasing the threshold, increases the average evaluation score for both modules. However, there is only a slight gain between 10 and 16. Based on these results, we concluded that the choice of NER module is of no great consequence to the overall quality of the results

**Global Precision Values** Other than averages, we also determined precision values by setting cutoff points to the average score. Specifically, we calculate $P_N$ which we define as the *precision, given that a term with a score of N or higher is considered "correct"*. We calculate this for $N = 2$ and $N = 3$, which corresponds to minimum scores of "somewhat relevant" and "relevant" respectively. Figure 2 shows these values for the different global threshold values. Here, we can see that the $P_2$ values are around 0.7 to 0.8 for most threshold values (not considering very high values where very few terms are added. The more strict version of $P_3$ hovers around 0.4, which is considerably low. To get an even better insight in the hits and misses of the two versions of the algorithm, for different values of the threshold we list the number of terms evaluated in four bins (0-1, 1-2, 2-3,

3-4) . These are shown in Table 3 for both CLTL and XTAS. This table shows for example that given a threshold on the matching score of 11, the algorithm extracts a total of 155 terms when using the XTAS tool. In that case, 18 extracted terms receive an evaluation between 0-1 and 116 receive an average evaluation between 2 and 4 (41+75).

| | Threshold | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | | 10.5 | | 11 | | 12 | | 14 | | 16 | |
| Score bin | cltl | xtas | cltl | xtas | cltl | xtas | cltl | xtas | cltl | xtas | cltl | xtas |
| 0-1 | 42 | 62 | 26 | 31 | 21 | 23 | 18 | 18 | 8 | 5 | 2 | 1 |
| 1-2 | 16 | 20 | 15 | 19 | 13 | 16 | 12 | 16 | 8 | 13 | 3 | 4 |
| 2-3 | 40 | 48 | 37 | 42 | 37 | 41 | 37 | 41 | 22 | 26 | 10 | 10 |
| 3-4 | 81 | 88 | 73 | 78 | 68 | 75 | 62 | 70 | 29 | 33 | 9 | 14 |
| Total | 179 | 218 | 151 | 170 | 139 | 155 | 129 | 145 | 67 | 77 | 24 | 29 |

**Table 3.** Frequencies of terms in average evaluation bins for six threshold values.

**Individual Score Parameters** In the previous paragraphs, we have used a global value for the parameters P7, P9, P11 and P13. We now look at optimal values for each of these. For this, we weigh the Precision for each axis (Named Entity class corresponding to one of the four parameters) against an estimated recall. For this estimated Recall we assume that the total number of correct items for that NE class is the total number to be found. This means that the maximum Recall is 1.0 (which is found at threshold values 9.5). This is of course an incorrect assumption but it does give us a gradually increasing Recall when the threshold is lowered. Given that there are not significantly more false negatives than are found, this is a reasonable estimate for the true Recall. After calculating the Recall, we then calculated the F1 measure, which is the weighted average between Precision and Recall. All three values are calculated with the assumption that an average evaluation of 2 or higher is "correct", we therefore get $P_2$, $R_{est,2}$ and $F1_{est,2}$. The maximum value for $F1_{est,2}$ is an indication for the optimum value of the threshold. These optimal values are presented in Table 4. This shows that the optimal threshold values are approximately 10 for person and 12 for locations and miscellaneous (regardless of the NER module). For organizations, the two modules present different values. This might reflect an artifact in the data

### 4.4 Result Summary

The evaluation results indicate that the agreement between evaluators is not very high but at least acceptable. Using their assessments as ground-truth we saw that precision values of around 0.7 to 0.8 are obtained in a less strict evaluation where terms should minimally "somewhat relevant" ($P_2$). When we apply a stricter evaluation that requires a term to be "relevant", performance drops to

| | threshold | | $P_2$ | | $R_{est,2}$ | | $F1_{est,2}$ | |
|---|---|---|---|---|---|---|---|---|
| | cltl | xtas | cltl | xtas | cltl | xtas | cltl | xtas |
| P7 (person) | 10.12 | 10.12 | 0.58 | 0.54 | 0.88 | 0.83 | 0.7 | 0.65 |
| P9 (organization) | 10.56 | 12.05 | 0.8 | 0.76 | 0.89 | 0.85 | 0.84 | 0.8 |
| P11 (location) | 12.19 | 12.19 | 0.82 | 0.79 | 1.00 | 1.00 | 0.90 | 0.88 |
| P13 (misc) | 12.15 | 12.15 | 0.75 | 0.83 | 1.00 | 1.00 | 0.86 | 0.91 |

**Table 4.** "Optimal" values for the threshold parameters for the four NE categories for both NER modules. At these values the $F1_{est,2}$ is maximized.

around 0.4. Concerning parameter settings, thresholds in the range of 10 for person and 12 for locations and miscellaneous provides optimal results. With respect to the two NER modules we have seen that the choice of NER module is of no significant consequence to the overall quality of the results.

## 5 Discussion and Conclusion

In this paper we reported on an evaluation of automatic labeling of audiovisual content in an archive production environment. The aim was to evaluate if an unsupervised labeling approach based on subtitles using off-the-shelf NER tools and a baseline thesaurus matching approach would yield results that are acceptable given archival production requirement with respect to quality, authority and service levels to external users. We conclude that results are acceptable in this context, with parameter settings that are optimized using a strict evaluation approach, allowing only terms when they are relevant as opposed to somewhat relevant. Precision given these parameter settings are sufficiently high to not disturb the archival quality requirements but the downside is that Recall is rather low: professional archivists will typically label content with more labels then the automatic approach. However, given the pressure on manual resources in the traditional work-flow, the current automated set-up is a useful starting point. Furthermore, having a stable production work-flow running allows us to (i) monitor the longitudinal behavior of the approach, among others by asking for feedback from external users, allowing us to assess the effect of the change also from an end-user perspective, and (ii) work on incremental improvements, gratefully deploying the experimentation framework that was set-up during the research described here. We have seen that the NER modules used do not differ much so that considerations such as stability, speed and resource use may be the most important factors for choosing a module. However, we note that we only tested two modules and there are many others around such as the Stanford NLP toolkit [7] or GATE [8]. It is likely that NER modules that are trained specifically on the type of input (in our case speech transcriptions) would improve performance both in terms of recall and precision.

One of the first items on our future work list will be an analysis of results over different genres of programming. In the current experiment we took 'annotation priority' as a selection mechanism, but from a quality perspective it makes

more sense to select input to the term-extraction pipeline based on expected performance. This will also allow us to investigate more effectively how improvements can be obtained. Based on observations in the field, we expect that there is room for improvement especially for non-common terms in the named-entity track and aiming for a better capturing of global document features to improve disambiguation and subject term assignment.

Other improvements in recall can be achieved through clustering of synonyms, using (external) structured vocabularies or by improving the named entity reconciliation (identifying the occurrence of the same entity in a text even though spelling variants are used). Finally, we will also look into the use of other collateral data sources such as program guides and scripts, and combinations of data sources, potentially also coming from multimodal analysis components such as speaker recognition and computer vision [9].

*Acknowledgments* This research was funded by the MediaManagement Programme at the Netherlands Institute for Sound and Vision, the Dutch National Research Programme COMMIT/ and supported by NWO CATCH program (`http://www.nwo.nl/catch`) and the Dutch Ministry of Culture.

## References

1. Gazendam, L., Wartena, C., Malaisé, V., Schreiber, G., de Jong, A., Brugman, H.: Automatic annotation suggestions for audiovisual archives: Evaluation aspects. Interdisciplinary Science Reviews **34**(2-3) (2009) 172–188
2. Ordelman, R., Heeren, W., Huijbregts, M., de Jong, F., Hiemstra, D.: Towards affordable disclosure of spoken heritage archives. Journal of Digital Information **10**(6) (2009)
3. Declerck, T., Kuper, J., Saggion, H., Samiotou, A., Wittenburg, P., Contreras, J.: Contribution of nlp to the content indexing of multimedia documents. In Enser, P., Kompatsiaris, Y., OConnor, N., Smeaton, A., Smeulders, A., eds.: Image and Video Retrieval. Volume 3115 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2004) 610–618
4. Iivonen, M.: Consistency in the selection of search concepts and search terms. Information Processing & Management **31**(2) (1995) 173–190
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**(3) (2009) 1–22
6. Likert, R.: A technique for the measurement of attitudes. Archives of psychology (1932)
7. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. (2014) 55–60
8. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving gate to meet new challenges in language engineering. Natural Language Engineering **10** (9 2004) 349–373
9. T. Tommasi, R. Aly, K. McGuinness, K. Chatfield, R. Arandjelovic, O. Parkhi, R. Ordelman, A. Zisserman, T.T.: Beyond metadata: searching your archive based on its audio-visual content. In: IBC 2014, Amsterdam, The Netherlands (2014)