Named entities in indexing: A case study of TV subtitles and metadata records

Anne-Stine Ruud Husevåg

Oslo and Akershus University College of Applied Science, Oslo, Norway
Anne-Stine.Husevag@hioa.no

Abstract. This paper explores the possible role of named entities in an automatic indexing process, based on text in subtitles. This is done by analyzing entity types, name density and name frequencies in subtitles and metadata records from different TV programs. The name density in metadata records is much higher than the name density in subtitles, and named entities with high frequencies in the subtitles are more likely to be mentioned in the metadata records. Personal names, geographical names and names of organizations where the most prominent entity types in both the news subtitles and news metadata, while persons, works and locations are the most prominent in culture programs.

Keywords: Named entity recognition · Multimedia indexing · Metadata

1 Introduction

Advances in information technology and web access of the past decade have triggered several digitization efforts by libraries, archives, and other cultural heritage institutions. Consequently, a large number of cultural expressions such as books, manuscripts, music, archaeological digital objects, television and radio broadcasts have become available to a broader public, but because of low metadata quality, they are hard to find.

Cultural expressions are often connected to places, persons, things or events in different ways. It is likely that different cultural institutions have information about the same place, event, person or other entity, but these connections are not visible to users and researchers today. Semantic Web technology, in particular Linked Data, is often seen as a potential solution to this problem [1,2].

Most of the digitized cultural expressions are encoded in natural language aimed at human consumption. Using methods from Language Technology, it is possible to extract machine-readable structures from free texts, from which it is possible to retrieve entities and link these entities together.

The term 'named entity' (NE) is generally considered to have originated at the Sixth Message Understanding Conference (MUC-6) held in 1995 [3]. There is a lack of agreement on a firm definition of what a NE is, and definitions are often created to meet the needs of specific projects or campaigns for Named Entity Recognition (NER). NER is a subtask of information extraction that seeks to locate and classify single words or

multi-word expressions in text into pre-defined entity types such as the names of persons, organizations and locations. Earlier research on NER on Norwegian texts [4-7] use the entity types person, organization, location, work, event and miscellaneous. The entity type "work" contains the kind of cultural expressions that are mentioned in this paper.

In order to explore the potential usefulness of indexing based on NER in subtitles (closed captions), this paper analyzes both density and frequencies of different NEs in different material. If NEs found in culture programs have different characteristics than NEs in news, then document retrieval might benefit from a different ranking of entities in a knowledge organization system.

The paper also compares subtitles and metadata records for the same TV programs, working from the hypothesis that words mentioned in the metadata records are more salient as content descriptors than the words from the subtitles that are not a part of the metadata records. More information about salient entities enables advanced entity search and retrieval, and allows users to perform more sophisticated searches than what is possible if we treat all the words in a full-text document the same. More knowledge about salient entities can enable developers to extract the most salient entities from subtitles in order to enhance precision in document retrieval.

2 Background

In 2013, several researchers at Oslo University College, including the author, started the TORCH project - Transforming the Organization and Retrieval of Cultural Heritage [8]. The objective of this project is research and development on issues related to automatic construction and structuring of metadata to improve access to digitized cultural expressions. Much of the research activity in the TORCH project is directed towards Norwegian NER on representations of cultural expressions, and the generation of linked data. This paper is a part of the research activities in this group.

This paper will present the findings of an explorative study of the NEs occurring in subtitles and metadata records in the archive of the Norwegian Broadcasting Corporation, NRK. Automatic processing of text from subtitles is a relatively simple process, and therefore has a huge potential in terms of future implementation and actual use. The subtitles are linked to the timeframe of the broadcast, so it is possible to use words from the subtitles as locators to where in the broadcast this word was uttered. This enables the subtitles to act as a source for entity retrieval. The relationship between the subtitles and the metadata records are in many ways similar to the relationship between the full text of a book and its index, and we believe that knowledge and experiences from book indexing can be useful in guiding the focus of annotation of material in large digital collections.

3 Related work

There has been considerable work in NER, typically organized in campaigns such as MUC[1], CoNLL[2] and ACE[3], with high levels of performance, measured in precision and recall. On the named entity task at MUC-6, the majority of sites had recall and precision over 90%; the highest-scoring system had a recall of 96% and a precision of 97%. This was done on texts from the Wall Street Journal [9].

There is currently no publicly available NER system for processing Norwegian text. The major research in this area was carried out at the University of Oslo within The Nomen Nescio Named Entity Recognition project between 2001 and 2003 [4]. Our research group decided to use the same definition and entity types as the Nomen Nescio project, in order to be able to compare results. They defined NEs to be "entities that have an initial capital letter both when they do and do not occur in the initial position of a period" [6, p. 34]. In 2015, Johansen at the University of Bergen conducted research that shows that it is possible to accurately find the names in Norwegian text by focusing only on demarcating names. He did not identify entity types [10].

There has been a lot of research on NER the last 25 years. The majority of research has been conducted on news articles and web pages [3, 11], but specialized systems have been developed for short, informal texts like tweets [12], and different domains like biomedicine [13]. Specialized systems are necessary when the texts are substantially different from the news-wire genre. Researchers seems to disagree on whether methods based on frequency counts would find the most important entities or not [11, 14-16], something that might vary in different genres.

In their index quality study, Bishop, Liddy and Settel [14,15] reports on a descriptive, explorative study of back-of-the-book indexes. A number of books (42% of those that had an index) in their study contained indexes that consisted only of proper nouns, i.e. NEs. Bishop et al. found that the percentage of proper names in the indexes they examined were 60 % in humanities, 69 % in fine arts, 50 % in social sciences and 30 % in science and technology. The authors point out that there might be differences among specific disciplines; not all humanity books are alike [17]. Similar findings have been reported by Zafran, who found that most of the index entries in art books consists of names and titles [19].

The use of smartphones have recent years affected how people watch TV. This have led to the development of so-called second-screen apps, apps that provide additional information and services to users while they watch TV programs. Subtitles bear great potential for extracting relevant information to second-screen apps, as shown in [20-23]. The work by Redondoio Garcia et al. [21] is especially relevant in the context of this paper, as they have performed named entity recognition on subtitles for news broadcasts and expanded them with structured data from DBpedia to generate context aware metadata for a TV news show. In a survey about television viewing habit and the use of second screens, Nandakumar and Murray [24] found that about 27 % of TV

[1] Message Understanding Conference
[2] Conference on Natural Language Learning
[3] Automatic Content Extraction

show-related searches is about the characters and their relations, 23 % about the plot, 16 % about location/events, 14 % about trivia, 9% about products and 11 % about other.

Enser and Sandom [25] analysed a sample of 1,270 requests from 11 British film archives. They found that there was a large number of requests NEs. The footage requests included 1,143 named people, events, places or times [25, p. 210]. Such information was not systematically recorded in the catalogues.

A case study carried out at the Deutsches Filminstitut (DIF) in 2000, examined how and what users requested from a comprehensive multimedia collection. In the 275 e-mails, there were 695 specific requests, 451 of them was regarding NEs. This study revealed that many of the requests entailed information regarding attributes of films that had not been indexed, and that further development of indexing procedures was needed in order to increase information retrieval efficiency [26].

Huurnink et al. [27] report on a study of transaction logs from an audiovisual broadcast archive in The Netherlands. They found that queries predominantly consist of (parts of) broadcast titles and of proper names.

4 Data

The TORCH project group has gained access to Norwegian subtitles from 11048 TV shows in different genres, and 780278 metadata records from both TV and radio, from NRK. The subtitles are quite recent, while the metadata records cover a time period from 1990, when the system was implemented, to 2013 when the data was exported. The metadata records contains an unstructured description field named content, where librarians working at NRK have written an abstract containing all relevant search terms. Valuable entities, such as the name of people, places and events are hidden within the ambiguity of these natural language descriptions.

5 Method

In the TORCH project, we chose to develop our own annotation tool to allow annotations on specific levels adapted to our projects. In order to be able to compare the results, we have followed the annotation guidelines outlined in [6] for the main entity types. The TORCH project and annotating tools is discussed in detail in [8, 28].

From the material of matching couples of subtitles at metadata records, the project group chose to annotate the TV programs Bokprogrammet (The book program) and Filmbonanza (a program about movies and TV series). These programs are quite similar. These programs where chosen because they contain mentions of entities that are useful to connect with other collections in a linked data network. These two programs are examples of programs from the cultural heritage domain, and this paper compares them to a program that has a typical news structure with various news stories about current affairs. The selection of programs described in this paper is not statistically representative; the programs are chosen for their characteristics as typical for their genre.

6 Results and discussion

The following table shows the distribution of entity types in subtitles. The paper then compares the results with results of earlier NER on Norwegian texts.

Table 1. Percentage of entity types. The full names of the columns are Documents, Named entities, Persons, Organizations, Locations, Work, Event and Other. Numbers from newspaper articles, magazine articles and works of fiction are obtained from Nøklestad [5, p. 71].

Corpus	Docs	NEs	Per	Org	Loc	Work	Event	Other
Subtitles Bokprogrammet	12	1252	45 %	3,90 %	27,20 %	18,30 %	0,80 %	4,70 %
Subtitles Filmbonanza	7	733	53,80 %	3,10 %	15,00 %	22,10 %	1,80 %	4,20 %
Subtitles news	8	775	26,80 %	17,20 %	49,80 %	2,10 %	1,00 %	3,10 %
Newspaper articles	210	4545	40,60 %	29,10 %	25,50 %	2,00 %	0,80 %	2,00 %
Magazine articles	46	1926	51,00 %	9,90 %	28,90 %	2,50 %	0,20 %	7,60 %
Works of fiction	9	1119	76,80 %	2,40 %	17,80 %	0,60 %	0,09 %	2,30 %

These numbers suggest that different entity types are more prominent in different genres. The NEs in metadata records for the same programs have a slightly more polarized distribution, as we see in table 2.

Table 2. Distribution of entity types in metadata records.

Corpus	Docs	NEs	Per	Org	Loc	Work	Event	Other
Metadata Bokprogrammet	12	873	60 %	5 %	13 %	19 %	0,1 %	2 %
Metadata Filmbonanza	7	48	71 %	0 %	6 %	15 %	4 %	4 %
Metadata news	8	434	61 %	10 %	23 %	4 %	0 %	2 %

Compared to table 1, the relative order of the entity types in table 2 is nearly equal. The personal names make up an even larger percentage of the whole in the metadata, and locations are less frequently mentioned.

In table 1 and 2, NEs that consist of several words are counted as one NE, e.g., 'Barack Obama' is counted as one. In order to measure the density of NEs in the different texts, all the words in compound NEs are counted as separate words. For the subtitles, news has a NE density of 5 % and Bokprogrammet and Filmbonanza both have a NE density of 6%. For the metadata records, Bokprogrammet has a NE density of 19%, Filmbonanza 20% and news texts 21%.

Numbers from Nøklestad [5] on Norwegian text shows a NE density of 6% for news articles, 4% for magazine articles and 2% for fiction. English news texts have a higher density of NEs. Coates-Stephens found that NEs amounted for 11.7 % of the tokens in 30 news stories from English papers [29, p. 171]. Goldstein et al. found that NEs represented 16.3% of the words in summaries, compared to 11.4% of the words in non-

summary sentences. 71% of summaries had a greater NE density than the non-summary sentences [30, p. 124]. The fact that the proportion of NEs is so much higher in summaries and the metadata records analyzed in this paper, confirms that NEs should play an important role in knowledge organization systems.

In fig. 1, we see the percentage of NEs of different entity types in the subtitles that this study has found in the metadata records.

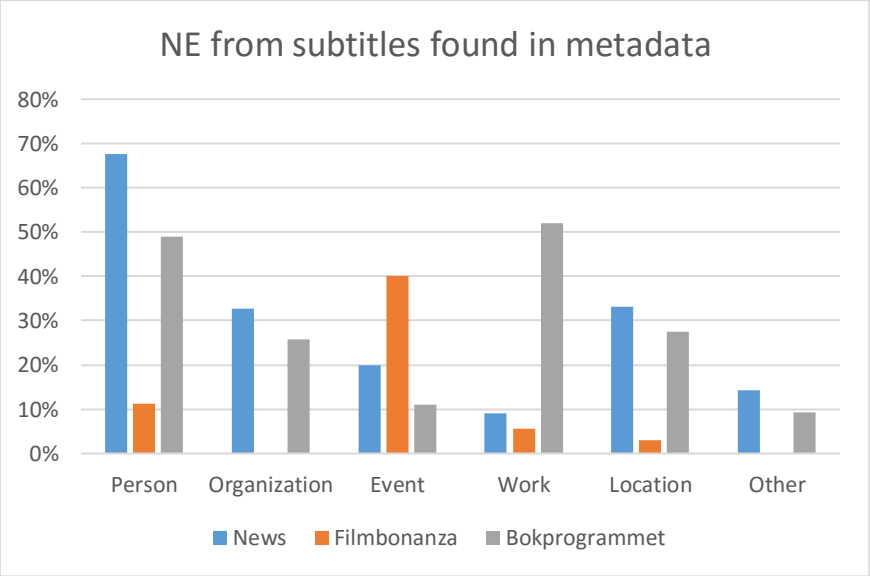

Fig. 1. Percentage of NEs from subtitles from different TV shows found in metadata, arranged by entity type.

Fig. 1 shows what entity types the indexers of the different programs have assessed as important. The considerations for news and Bokprogrammet are in many ways similar, with the big exception of the entity type "work". The metadata records for Filmbonanza are much shorter and contain fewer named entities. The actual number of the entity type "event" are very small and should not be emphasized too much.

When we look at the metadata records, we see that they also contain other NEs in addition to NEs from subtitles. The variations are large in this material, from one additional NE to 20 (of a total of 33 NEs in the record). In average, a little more than half of the NEs in the metadata records from news (54 %) and Bokprogrammet (55 %) was also in the subtitles, for the short records in Filmbonanza the average was 69 %. The additional NEs were mostly personal names. This was to a large extent people who either spoke in the program or was working behind the camera.

In order to find out if the frequency of the NEs in the subtitles was of importance for the likelihood of librarians choosing the NEs to represent the program in the metadata records, fig. 2 separates the NEs occurring three times or more from the less frequently mentioned NEs.

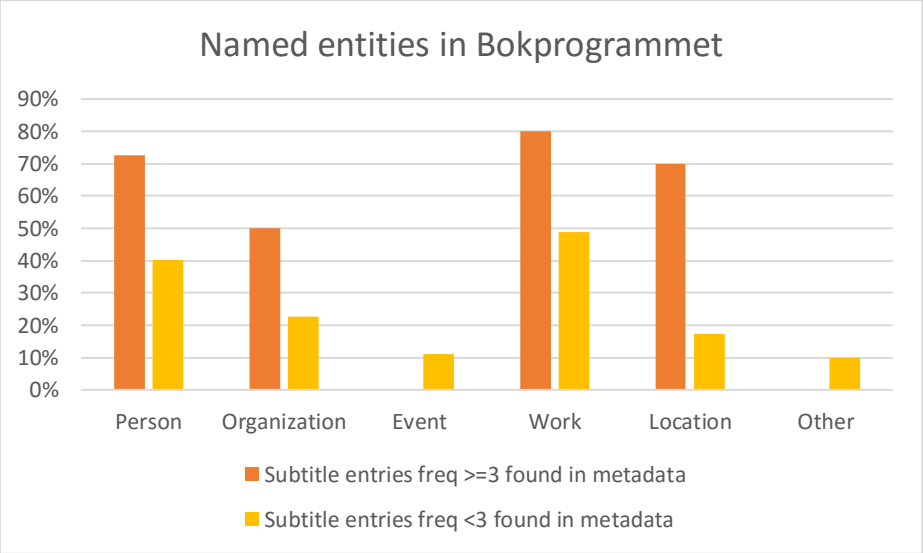

Fig. 2. Percentage of NEs from subtitles from Bokprogrammet (a literature program) found in metadata, arranged by frequency and entity type.

In fig. 2, we see that the entity types person, work and location are more likely to be found in the metadata records if they are mentioned three times or more in the subtitles. Because of the low numbers in the data material, it is hard to draw any definite conclusions about the entity types organization, event and other.

The structure and content in Bokprogrammet and Filmbonanza is quite similar, so the numbers were expected to be quite similar, but the librarians that created the metadata records have chosen to describe these two programs differently due to intern guidelines. The metadata records for Bokprogrammet have a much fuller text description in the content field than the records for Filmbonanza. This is not ideal when we are attempting to facilitate methods for automatic indexing, but it gives a realistic picture of what we can expect from a large archive of cultural heritage material that extends over a long time span.

The descriptions of Filmbonanza typically mentions a few persons, a movie and maybe a place or an event where interviews takes place. The entity type event is used more both in subtitles and metadata for Filmbonanza, this is due to the coverage of film festivals. Organizations are lacking in the material from Filmbonanza, and rarely occurs in Bokprogrammet. In the descriptions of Bokprogrammet, we observe an attempt to name all the persons and locations we see in the program, and the works mentioned.

The analysis of the news material shows that the personal names are included in the metadata records regardless of frequency in the subtitles. Organizations appear more frequently in news, and 75% of high frequent organizations in subtitles are found in the metadata. Few works are mentioned in the news, and the works that are mentioned are

usually very different from the works mentioned in culture programs. The works mentioned in the news material studied in this paper were laws, treaties, declarations and conventions, in addition to mentions of the name of the news program itself.

7 Conclusion

NRK's archive materials are becoming increasingly available on-line. They have made a major digitization effort in order to make Norwegian cultural heritage from the last century of radio and TV available to the public. This gives Norwegians the possibility to relive nostalgic moments, and for new generations to take part in historical experiences. This, however, presupposes that the users are able to find specific items. It is impossible to manually go through and index all the digitized material, but the use of new technology can provide librarians with a tool to automatically locate indexable entities and facilitate information retrieval.

This paper has analyzed subtitles and metadata records from different TV programs. The analysis shows that the density of NEs in metadata records is much higher than the NE density in subtitles, implying that NEs are more important than other parts of speech in the descriptions of this kind of material. This finding is coherent with findings from book indexes who have an even higher density of NEs. Compared to studies of English texts, Norwegian texts have a significant lower name density: 4-6 % for non-fiction texts including news, compared to 11.4 % and 11.7 % in English non-summary news texts [29, 30]. The descriptive texts in the Norwegian metadata records presented in this paper have a NE density of 19-21 %, which is higher than the news-article summaries in [30], where NEs represented 16.3 % of the text.

User studies on multimedia collections reveal that named people, events, organizations, works and locations are common search requests, and that further development of indexing procedures is needed in order to be able to respond to these requests [26]. Recognition of NEs in subtitles is a solution to this challenge.

This paper has looked closer at the different entity types used to describe different kind of programs. The findings indicate that NEs with high frequencies in the subtitles are more likely to be mentioned in the metadata records, and that differences in frequencies have a higher discriminatory value for some entity types. Compared to earlier research on NER on Norwegian text, this paper shows similar findings for news text. Personal names, geographical names and names of organizations where the most prominent entity types both in the news subtitles and news metadata in this paper, and in the newspaper articles in [5]. The analysis suggest that high frequent entities of the entity types person, work and location are important as salient content descriptors of culture programs. The material contained few events, but for Filmbonanza, the events were also found in the metadata. That was not the case for the literature program Bokprogrammet. Personal names are often considered important regardless of frequency, especially in news material. The research presented in this paper have been conducted on a small sample, and these findings should be examined more closely on a larger sample to obtain results that are more reliable.

References

[1] R. Engels, Åpen og samordnet tilgang til kulturarven: anbefalinger for en vellykket tilstedeværelse i den digitale kulturelle verden. Oslo: ABM-utvikling, 2010.

[2] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee, "Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections," in The Semantic Web: Research and Applications, L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, Eds. Springer Berlin Heidelberg, 2009, pp. 723–737.

[3] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investig., vol. 30, no. 1, pp. 3–26, 2007.

[4] J. B. Johannessen, K. Hagen, Å. Haaland, A. B. Jónsdottir, A. Nøklestad, D. Kokkinakis, P. Meurer, E. Bick, and D. Haltrup, "Named Entity Recognition for the Mainland Scandinavian Languages," Lit. Linguist. Comput., vol. 20, no. 1, pp. 91–102, Mar. 2005.

[5] A. Nøklestad, "A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection," Doctoral thesis, University of Oslo, Oslo, 2009.

[6] A. B. Jónsdóttir, "ARNER, what kind of name is that? : an automatic rule-based named entity recognizer for Norwegian," Master thesis, University of Oslo, Oslo, 2003.

[7] Å. Haaland, "A Maximum Entropy Approach to Proper Name Classification for Norwegian," Doctoral thesis, University of Oslo, Oslo, 2008.

[8] K. Tallerås, D. Massey, A.-S. R. Husevåg, M. Preminger, and N. Pharo, "Evaluating (linked) metadata transformations across cultural heritage domains," in Metadata and semantics research, Springer, 2014, pp. 250–261.

[9] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History.," in COLING, 1996, vol. 96, pp. 466–471.

[10] B. Johansen, "Named-Entity Chunking for Norwegian Text using Support Vector Machines," Nor. Inform. NIK, 2015.

[11] S. Sekine and E. Ranchhod, Named entities: recognition, classification and use. Amsterdam: John Benjamins, 2009.

[12] X. Liu, F. Wei, S. Zhang, and M. Zhou, "Named entity recognition for tweets," ACM Trans Intell Syst Technol, vol. 4, no. 1, p. 3:1–3:15, Feb. 2013.

[13] C.-C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: success, failure and the future," Brief. Bioinform., vol. 17, no. 1, pp. 132–144, Jan. 2016.

[14] T. Poibeau and L. Kosseim, "Proper name extraction from non-journalistic texts," Lang. Comput., vol. 37, no. 1, pp. 144–157, 2001.

[15] B. Cronin, H. W. Snyder, H. Rosenbaum, A. Martinson, and E. Callahan, “Invoked on the Web,” J. Am. Soc. Inf. Sci., vol. 49, no. 14, pp. 1319–1328, Jan. 1998.
[16] F. B. Karsdorp, P. van Kranenburg, T. Meder, and A. Bosch, “Casting a Spell: Identification and Ranking of Actors in Folktales,” in Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), 2012.
[17] A. P. Bishop, E. D. Liddy, and B. Settel, “Index quality study, Part I: Quantitative description of back-of-the-book indexes,” Index. Tradit. Innov., pp. 15–51, 1991.
[18] E. D. Liddy, A. P. Bishop, and B. Settel, “Index quality study, Part II: Publishers survey and qualitative assessment,” Index. Tradit. Innov., pp. 53–79, 1991.
[19] E. Zafran, “Names in Art Books,” in Indexing names, N. Bridge, Ed. Information Today, 2012, pp. 219–226.
[20] C. Castillo, G. De Francisci Morales, and A. Shekhawat, “Online matching of web content to closed captions in IntoNow,” in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 1115–1116.
[21] J. L. Redondoio Garcia, L. De Vocht, R. Troncy, E. Mannens, and R. Van de Walle, “Describing and contextualizing events in tv news show,” in Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 759–764.
[22] D. Odijk, E. Meij, and M. de Rijke, “Feeding the Second Screen: Semantic Linking Based on Subtitles,” in Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, Paris, France, France, 2013, pp. 9–16.
[23] J. Knittel and T. Dingler, “Mining Subtitles for Real-Time Content Generation for Second-Screen Applications,” in Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video, New York, NY, USA, 2016, pp. 93–103.
[24] A. Nandakumar and J. Murray, “Companion apps for long arc TV series: supporting new viewers in complex storyworlds with tightly synchronized context-sensitive annotations,” in Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video, 2014, pp. 3–10.
[25] P. G. Enser and C. J. Sandom, “Retrieval of archival moving imagery-CBIR outside the frame?,” in Image and Video Retrieval, Springer, 2002, pp. 206–214.
[26] M. Hertzum, “Requests for information from a film archive: a case study of multimedia retrieval,” J. Doc., vol. 59, no. 2, pp. 168–186, Apr. 2003.
[27] B. Huurnink, L. Hollink, W. Van Den Heuvel, and M. De Rijke, “Search behavior of media professionals at an audiovisual archive: A transaction log analysis,” J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 6, pp. 1180–1197, 2010.
[28] K. Hoff and M. Preminger, “Usability Testing of an Annotation Tool in a Cultural Heritage Context,” in Metadata and Semantics Research, vol. 544, E. Garoufallou, R. J. Hartley, and P. Gaitanou, Eds. Cham: Springer International Publishing, 2015, pp. 237–248.
[29] S. Coates-Stephens, “The analysis and acquisition of proper names for robust text understanding,” Ph.D., City University London, 1992.

[30] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, “Summarizing Text Documents: Sentence Selection and Evaluation Metrics,” in Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 1999, pp. 121–128.