From Subtitles to Substantial Metadata: Examining Characteristics of Named Entities and their Role in Indexing

Anne-Stine Ruud Husevåg

Oslo Metropolitan University, Oslo, Norway
annesh@oslomet.no

Abstract. This paper explores the possible role of named entities extracted from text in subtitles in automatic indexing of TV-programs. This is done by analyzing entity types, name density and name frequencies in subtitles and metadata records from different genres of 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. Further analysis of the metadata records indicate an increase in use of named entities in metadata in accordance with the frequency the entities have in the subtitles. The most substantial difference was between a frequency of one or two, where the named entities with a frequency of two in the subtitles where twice as likely to be present in the metadata records. Personal names, geographical names and names of organizations were the most prominent entity types in both the news subtitles and news metadata, while persons, creative works and locations are the most prominent in culture programs. It is not possible to extract all the named entities in the manually created metadata records by applying named entity recognition to the subtitles for the same programs, but it is possible to find a large subset of named entities for some categories in certain genres. The results reported in this paper show that subtitles are a good source for personal names for all the genres covered in our study, and for creative works in literature programs. In total, it was possible to find 38% of the named entities in metadata records for news programs, 32% for literature programs, while 21% of the named entities in metadata records for talk shows were also present in the subtitles for the programs.

Keywords: Named entity recognition · Multimedia indexing · Metadata · Audiovisual archives

1 Introduction

Advances in information technology and web access over the past decade have triggered several digitization efforts by libraries, archives, and other cultural heritage institutions. Consequently, an increasing number of cultural expressions such as books, manuscripts, music, digital archaeological objects, television and radio broadcasts have become available to a broader public. However, because of low metadata quality, finding them can be difficult. . Semantic Web technology, in particular Natural Language Processing (NLP) (Chowdhury, 2003) and Linked Data (Berners-Lee, 2009), is often seen as one potential solution to this problem (Sporleder, 2010). Exploring new applications of well-known methods from library and information science is another potential solution. The purpose of this paper is to bring the two together in order to better assist users in their search for information, knowledge and entertainment.

1.1 Background and definitions

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. Using Linked Data to interlink these entities and their connections will make it easier to connect pieces of information, and publishing data from cultural heritage institutions as linked data is one way to enhance metadata quality in this field (Engels, 2010; Kobilarov et al., 2009). Most of the digitized cultural expressions are encoded in natural language meant to be read by humans, but methods from NLP allow us to extract machine-readable structures from unstructured texts, from which it is possible to retrieve entities and link these entities together.

The term “named entity” (NE) is often used as a generic term for entities referred to by a proper noun, spelled with an initial capital letter. The term is generally considered to have originated at the Sixth Message Understanding Conference (MUC-6) held in 1995 (Nadeau & Sekine, 2007). 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 of NEs in text into pre-defined entity types such as the names of persons, organizations and locations. Earlier research on NER on Norwegian texts (Haaland, 2008; Johannessen et al., 2005; Jónsdóttir, 2003; Nøklestad, 2009) uses the entity types *person*, *organization*, *location*, *work*, *event* and *miscellaneous*. The entity type *work* is used for all kinds of creative works and cultural expressions, like books, movies, TV programs, paintings and songs.

The word *indexing* is used to describe both the process that 1) facilitates the retrieval of one or more documents in a collection of documents (document retrieval), and 2) facilitates the retrieval of relevant information within a document (entity retrieval), mostly known as back-of-the-book indexing. Susan Klement (2002) refers to these processes as open-system and closed-system indexing respectively. A book is a closed system that contains a beginning and an end, and many non-fiction books use indexes in the back of the book to help the reader find information relevant to their need in a more efficient way than to read the whole book. The terms used in this kind of index would be different from the keywords used to describe one book in a large collection of books, but both processes are referred to as indexing.

The difference between open-system and closed-system indexing of books can be compared to the difference between indexing a TV program in a collection of TV programs, and indexing segments of a TV broadcast. In information retrieval, we usually focus on finding relevant documents in a collection. This is still important in a TV archive, but for some information needs, it would be useful to be able to find parts of a broadcast, such as an interview with a particular person. To distinguish between salient and non-salient NEs would improve the quality of both open-system and closed-system indexing.

In this paper, *salience* is used as a term indicating importance in a more specific manner than the term *relevance*. Traditionally, *relevance* in information retrieval has been considered as a measure of the match between a text and a query, regardless of the properties of the user. This view is now one of many views on relevance, as researchers from the information science field have recognized the dynamic and complex nature of relevance. The current study does not consider specific users and their needs, and the goal is not to identify all NEs that might be perceived as relevant to any given user at any given time. The word *salient* has been chosen to describe those NEs that are fundamental or central to the nature of the program they appear in. The identification of salient entities enables advanced entity search and retrieval, and allows users to perform more sophisticated searches than will be possible if we treat all the words in a full-text document the same. More knowledge about characteristics of entities that indexers have perceived as salient can enable systems to identify the most salient entities in texts of different genres and thus enhance precision in retrieval.

1.2 Knowledge gaps in existing research

The majority of research on named entity recognition, entity linking and entity search has been conducted on news articles and web pages (Nadeau & Sekine, 2007; Sekine & Ranchhod, 2009). The primary aim of the present research is to explore how named entities can be used to describe the content of TV programs about culture. Existing research literature reveals little information about characteristics of NEs found in culture programs compared to the characteristics of NEs found in news. Different characteristics could mean that different features should be used to select and extract NEs as content descriptors for this genre. While there is no readily available system for

Norwegian named entity recognition, NER is a large research field in the NLP community. The aim of this research is not to suggest methods for NER, but to expand the knowledge of NEs found in two little explored subsets: The Norwegian language and the genre of cultural TV-programs.

1.3 Contribution

The study reported here is a part of a PhD project that seeks to examine the role of named entities in the process of indexing and abstracting of different kinds of material. Indexing content is an integral part of any collection of material, and is vital to the information retrieval process. The Norwegian Broadcasting Corporation (NRK) are working on making old TV material available to the public, and they are continuously uploading the resulting content, providing broader access to our audiovisual heritage. These developments have heightened the need for more effective indexing and retrieval techniques. It may take a cataloguer up to three times the duration of a TV program to manually annotate keywords, depending on the genre (Gazendam et al., 2009). There is an urgent need to find ways to simplify the complex process of indexing multimedia material, and this paper contributes by appraising a possible foundation and background knowledge for automatic indexing of visually encoded information. The paper presents the findings of an exploratory study of the potential usefulness of indexing based on NER in subtitles (closed captions) in TV programs from the NRK archive. 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 a given word was uttered. This enables the subtitles to act as a source for entity retrieval.

The research described in this paper compares text in subtitles to manually created metadata records, focusing on named entities used in the different kinds of texts. The author presupposes that NEs used for manual indexing of a TV program are more salient as content descriptors than NEs not included in the metadata records, despite their presence in the subtitles for the program. Manually created metadata records are used as a guideline for which NEs that should be included in an automatic indexing process. Further, the paper analyzes some of the differences between NEs from subtitles that have been included in metadata records and those that have not been included.

The aim of this research project is to explore ways to identify NEs that are salient as content descriptors and improve methods for automatic indexing by analyzing characteristics of named entities chosen in a manual indexing process.

2 Related work

This paper seeks to explore how NER can be used to improve indexing of broadcast material. To get a comprehensive understanding of the task at hand, it is essential to collect information from other related fields of study. This section will begin with summarizing user studies from film and broadcasting archives, continue with the role of NEs in book indexing and give an overview of relevant NER research. The section finishes with contributions from Semantic Web research where new technologies have been applied to multimedia archives.

2.1 User studies

Several studies have analyzed user requests to film and broadcasting archives. The archives discussed in this section are somewhat similar to the NRK archive, both in size and content.

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 covered, there were 695 specific requests, 451 of them were 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 (Hertzum, 2003).

Enser and Sandom (2002) analyzed a sample of 1,270 requests from 11 British film archives. They found that there were a large number of requested NEs. The footage requests included 1,143 named people, events, places or times (Enser & Sandom, 2002, p. 210). Such information was not systematically recorded in the catalogues.

Huurnink et al. (Huurnink, Hollink, Van Den Heuvel, & De Rijke, 2010) 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.

The use of smartphones has in recent years affected how people watch TV. This has led to the development of apps that provide additional information and services to users while they watch TV-programs, often referred to as second-screen apps. Subtitles bear great potential for extracting relevant information to second-screen apps, as shown in (Castillo, De Francisci Morales, & Shekhawat, 2013; Knittel & Dingler, 2016; Odijk, Meij, & de Rijke, 2013; Redondoio Garcia, De Vocht, Troncy, Mannens, & Van de Walle, 2014). The work by Redondoio Garcia et al. 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 (Nandakumar & Murray, 2014) found that about 27% of TV 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.

2.2 Indexing named entities

Most of the literature on indexing named entities has tended to focus on how to write the names: choice of name forms; disambiguation and sources for authorizing names; and the form of entries containing names. In this paper, the most important aspect of name indexing is how to determine whether a name is important enough to be included in the text that acts as a representation for a broadcast.

There is disagreement among experienced book indexers whether every name in a text should be included in the index (e.g., (Smith, 2012)) or not (e.g.,(Zafran, 2008)). Smith (Smith, 2012) assumes that authors include names in a text for a reason, while Zafran (Zafran, 2008) stress that indexers should not pick up proper names indiscriminately without considering why the author included them, and warns against this kind of use of proper names for indexing. Different types of names are treated differently in book indexing. Smith points out that personal names are especially important because book indexes provide professional and cultural recognition of people. This is also true for the NRK archive because NRK is a government-owned radio and television public broadcasting company and their content is an important part of Norwegian cultural heritage. Professional indexer Noeline Bridge claims that place names are often mentioned in passing or with very little information, especially in autobiographies and biographies. She finds it unnecessary to provide entries for such place names unless more information is provided later in the book (Bridge, 2012).

There is wider agreement on reasons to omit names in an index. Examples are names used as examples, stage-setting, time markers, attention-getting devices, comparisons and analogies (Smith, 2012; Zafran, 2008).

In their index quality study, Bishop, Liddy and Settel (Bishop, Liddy, & Settel, 1991; Liddy, Bishop, & Settel, 1991) report on a descriptive, explorative study of back-of-the-book indexes. A large proportion 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 (Bishop et al., 1991). Similar findings have been reported by Zafran, who found that most of the index entries in art books consists of personal names and titles (Zafran, 2012).

2.3 Named entity recognition (NER)

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 (Grishman & Sundheim, 1996).

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 (Johannessen et al., 2005). 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" (Jónsdóttir, 2003, p. 34). In 2015, Johansen at the University of Bergen conducted research showing that it is possible to accurately find NEs in Norwegian text by focusing only on demarcating names. He did not identify entity types (Johansen, 2015).

In the last 25 years, NER has been a popular research topic. The majority of research has been conducted on news articles and web pages (Nadeau & Sekine, 2007; Sekine & Ranchhod, 2009), but specialized systems have been developed for short, informal texts like tweets (Liu, Wei, Zhang, & Zhou, 2013), and different domains like biomedicine (Huang & Lu, 2016). 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 (Cronin, Snyder, Rosenbaum, Martinson, & Callahan, 1998; Karsdorp, Kranenburg, Meder, & Bosch, 2012; Poibeau & Kosseim, 2001; Sekine & Ranchhod, 2009) something that might vary in different genres.

2.4 Semantic Web technology and multimedia indexing

Multiple research efforts have proven the value of Semantic Web technology like NLP methods and Linked Data for multimedia indexing. Kobilarov et al. (Kobilarov et al., 2009) describe how Linked Data technologies were applied within the British Broadcasting Corporation (BBC). Ordelman et al., Tommasi et al. and Eskevich et al. all present ongoing work on multimodal video hyperlinking where anchor identification is based on the identification of named entities and used for navigating multimedia archives (Eskevich, Nguyen, Sahuguet, & Huet, 2015; R. Ordelman, Aly, Eskevich, Huet, & Jones, 2015; R. J. Ordelman, Eskevich, Aly, Huet, & Jones, 2015; Tommasi et al., 2014). Boer et al.

[1] Message Understanding Conference

[2] Conference on Natural Language Learning

[3] Automatic Content Extraction

(Boer, Ordelman, & Schuurman, 2016) report on the evaluation of automatic labeling of Dutch Radio and TV-broadcasts, based on NER in subtitles. On average, they reached an accuracy level of 0.75. They found that the accuracy for names was the highest and accuracy for topics the lowest. The best scoring programs were a sports program and a news program, as well as "magazine"-like programs with reports and interviews on current events. For these programs, the accuracy was above 0.73. They reached sufficiently high precision as to not disturb the archival quality requirements, but recall was low, as professional archivists used labels that were not found by the automatic approach.

3 Data

This research is a part of the TORCH project (Transforming the Organization and Retrieval of Cultural Heritage) (Tallerås, Massey, Husevåg, Preminger, & Pharo, 2014). 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. The project group has gained access to Norwegian subtitles from 11,048 TV shows in different genres, and 780,278 metadata records from both TV and radio, from NRK. The subtitles are from recent years, 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 contain an unstructured description field named *content*, into which indexers working at NRK have written an abstract containing all relevant search terms. Valuable entities, such as the names of people, places and events are hidden within the ambiguity of these natural language descriptions. From this big dataset, we have manually annotated subtitles and matching metadata records from 34 literature programs, 18 talk shows and 8 news programs. This resulted in 6,272 NEs from subtitles and 3,573 NEs from metadata. Several criteria were applied to select the programs. The first criterion was genre. We wanted to analyze different culture programs and compare them to typical news programs. We then had to make sure that both the subtitles and metadata files were complete. In this process, we discovered that some programs only had subtitles for foreign language segments, so we had to choose programs that included subtitles for the whole program. Because different program series often have a unique tone and style, we wanted to include programs from different talk show hosts and cultural programs. We examined records from programs about movies and music as well, but found that the files from the literature programs were the most complete and diverse, and would serve as the best example for the genre.

There is no readily available NER system for Norwegian, and because we wanted to avoid all the possible sources of error in a new system, we performed our analysis on manually annotated NEs. The research discussed in this paper assumes a functional NER system for Norwegian text.

4 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 our results to the results of earlier research on Norwegian NER, we have chosen to use the same top categories and followed the annotation guidelines outlined in (2003). The NER community in Norway is very small, and using the same categories and guidelines as Nøklestad, Jónsdottir and Haaland (Haaland, 2008; Jónsdóttir, 2003; Nøklestad, 2009) enables us to see common features over different data sets and draw conclusions about Norwegian texts with higher certainty. *Persons*, *organizations* and *locations* are the most common categories used in NER research, and these have proven to be very functional for NER on news. To explore if other categories are more important for different genres, both we and the other Norwegian NER researchers decided to use six top categories: *Person*, *organization*, *location, creative work*, *event* and *other*. In TORCH we also included subclasses to *person* and *creative work*. In this paper, the subcategory *fictional character* is separated from *person* when the author found it relevant to show the difference. The TORCH project and annotating tools is discussed in detail in (Hoff & Preminger, 2015; Tallerås et al., 2014).

Manual annotation requires considerable time and effort, and annotation has been done on a carefully selected subset of the data where we had complete files of good quality for both subtitles and metadata. We wanted to explore the possibilities in linking representations of cultural expressions from different cultural institutions. Therefore, we chose to annotate two different types of programs about literature, and two different kinds of talk shows that consists of in-depth interviews with artists, politicians, writers and other celebrities, often discussing a newly released book, movie or album. The selection of programs used in this paper is not statistically representative. The programs were chosen for their characteristics as typical for their genre, and because they contain mentions of entities that are useful to connect with other collections in a linked data network. These programs are examples of programs from the cultural heritage domain, and this paper compares them to programs that has a typical news structure with various news stories about current affairs.

The possibility of entity linking have great value from a library and information science point of view. The Norwegian Broadcasting Corporation is a cultural heritage institution that have collected archive material of different genres since 1933. There exists a great body of knowledge on NER on news texts, but notably less so on art and culture, and few NER projects have a category for creative works like titles for books and movies. These entities are not common in news text, but they are important for cultural heritage institutions and collections. Recognizing and linking salient NEs could help utilize parts of the collection that are practically hidden for users today because of inadequate indexing and search possibilities.

We have measured the inter-annotator agreement to evaluate the quality of the manual annotation conducted by members of the TORCH research group. Four metadata records consisting of 274 annotations, and five subtitle texts consisting of 459 annotations were randomly selected from the corpora and annotated by two annotators. The rest of the corpora is only annotated once, by the author of this paper or one other member from the TORCH project. When we measured the inter-annotator agreement, the overall average F-measure was 0.90 for the metadata and 0.96 for the subtitles. In the literature about inter-coder agreement for computational linguistics, most researchers agree that values above 0.80 are necessary to ensure an annotation of reasonable quality (Artstein & Poesio, 2008; Bayerl & Paul, 2011). Perfect agreement among annotators is very rare because of different domain expertise, personal biases or mistakes due to slips of attention or misinterpretations. In our data set, the agreement was very high for the main categories, but we sometimes disagreed on subtypes of the category *creative works* due to the lack of contextual information in the data material.

5 Results and discussion

The lack of good methods for automatic indexing is an issue that has grown in importance due to recent digitalizing efforts. Computers use statistics to compensate for the lack of human language interpretation, and the aim for this paper is to try to identify some statistical patterns in the way named entities occurs both in natural language and in manually prepared summaries. The results presented in this section represent one way of examining the descriptive value of named entities in different genres, in an attempt to translate human assessments to numbers.

The first part of this section deals with the occurrence of named entities in text, it then continues by comparing occurrences of named entities in subtitles with named entities chosen by indexers to describe content. The last part of the results section is concerned with frequency as an indicator for salience.

5.1 Entity types in different genres

To evaluate the perceived salience of NEs, this paper has analyzed and compared the percentage of NEs in different types of text. To determine whether NEs found in culture programs have different characteristics than NEs in news, entity types and name frequencies from text in different genres have been analyzed.

The two charts below illustrate the breakdown of the different types of named entities in subtitles and metadata.

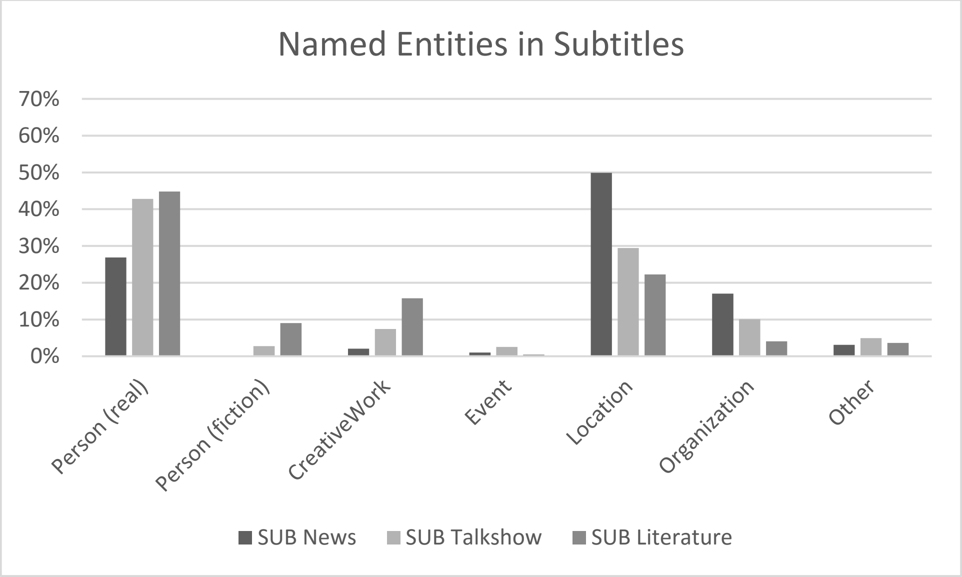

Figure 1: The distribution of different categories of named entities in subtitles.

These numbers suggest that different genres have a different distribution of entity types. Personal names, locations and organizations are not surprisingly the most important categories in news. The event category is rarely used, neither in this data set nor in other research on Norwegian NER (Nøklestad, 2009). A closer examination of the texts show that events often are mentioned in the form “[common noun] at [location]”. The findings of this study are consistent with those of Nøklestad, (Nøklestad, 2009) who found that names of persons, locations and organizations are the most prominent NEs in Norwegian texts from newspapers, magazines and fiction. He also found that organizations have more mentions in newspapers (29% than in magazine articles (10%) and fiction (2.5%. The vast majority of NEs in Norwegian works of fiction were personal names, constituting 77%. Nøklestad and the current study have produced results which corroborate the basic assumptions in a great deal of the previous work in named entity recognition, namely that persons, locations and organizations are the important NEs in news. *Personal names* and *locations* are the most important categories in culture programs as well, but we see a shift towards focusing more on individual people and less on actual locations where events have occurred. Similar to *locations*, *organizations* are less prominent in cultural programs. As a whole, the *creative works* category has a similar level of occurrence as the *organizations* category, but while *organizations* are more prominent in news than in literature programs, the numbers are almost opposite for *creative works*, which are almost absent in news and prominent in literature programs. With regard to the number of named entities from different categories, the program genre *talk show* is positioned between news and literature programs for all the mentioned categories.

The NEs in metadata records for the same programs have a slightly more polarized distribution, as we see in fig. 2.

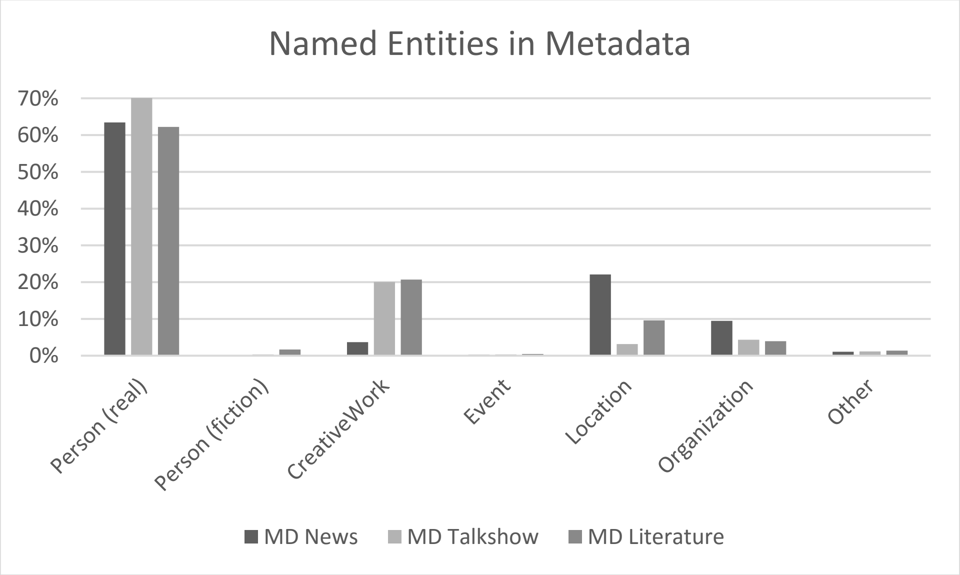

Figure 2: The distribution of different categories of named entities in metadata.

Compared to fig. 1, the relative order of the entity types in fig. 2 is nearly equal. Personal names make up an even larger percentage of the whole in the metadata, and locations are less frequently mentioned. One of the limitations of the descriptive nature of the research described in this paper is that it does not identify solid explanations for the observed findings. The metadata creators have used very general guidelines for the indexing of this material. Their task has been to write text descriptions that can be an answer to the general questions of *who*, *what*, *where* and *when*, questions that are important both in journalistic practices and in classification and indexing theory. The observed results are likely a consequence of an indexing practice developed over time in order to accommodate the institution's information need.

5.2 Density of named entities (NE)

In fig. 1 and 2, NEs that consist of several words are counted as one NE, e.g., 'Barack Obama' is counted as one. Every mention of a NE is counted. 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* and *literature* have a NE density of 5% and *talk shows* have a NE density of 3%. The subtitles of different programs in the same genre show little variations in terms of name density.

For the metadata records, *news* have an average of 21% NEs, *literature programs* have 25% NEs and *talk shows* have 39% NEs. The data set shows large differences from program to program in all genres. The standard deviation is 13 for *news* and *talk shows* and 8.76 for the *literature* programs. However, despite this large variation, it is evident that the name density is substantially higher in the metadata records than in the subtitles for all programs analyzed. Numbers from Nøklestad (Nøklestad, 2009) 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 to 11.7% of the tokens in 30 news stories from English papers (Coates-Stephens, 1992, 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 (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999, p. 124).

Earlier research on Norwegian text from newspapers shows a similar density of NEs as the findings reported in this paper. Earlier research on English summaries shows that NEs make up a larger part of the words in summaries, compared to non-summary sentences. Therefore, we anticipated a higher name density in the shorter content descriptive text in the metadata records than in the subtitles, but the difference was even larger than expected. The data reported in this paper support the assumption that NEs should play an important role in descriptive metadata and knowledge organization systems.

5.3 Automatic extraction of NEs from subtitles

This paper seeks to investigate the underlying support for automatically extracting salient NEs from subtitles for programs where no descriptive metadata exists. The following calculations are based on the notion that NEs in the metadata records have been chosen by the indexers because they are the most salient NEs as content descriptors for that particular program.

All the annotated NEs from subtitles and metadata were manually compared with each other so that it was possible to match NEs with different name forms, parts of names, abbreviations and misspellings. In this chart, only unique NEs are counted. Multiple mentions of the same NE are only counted as one.

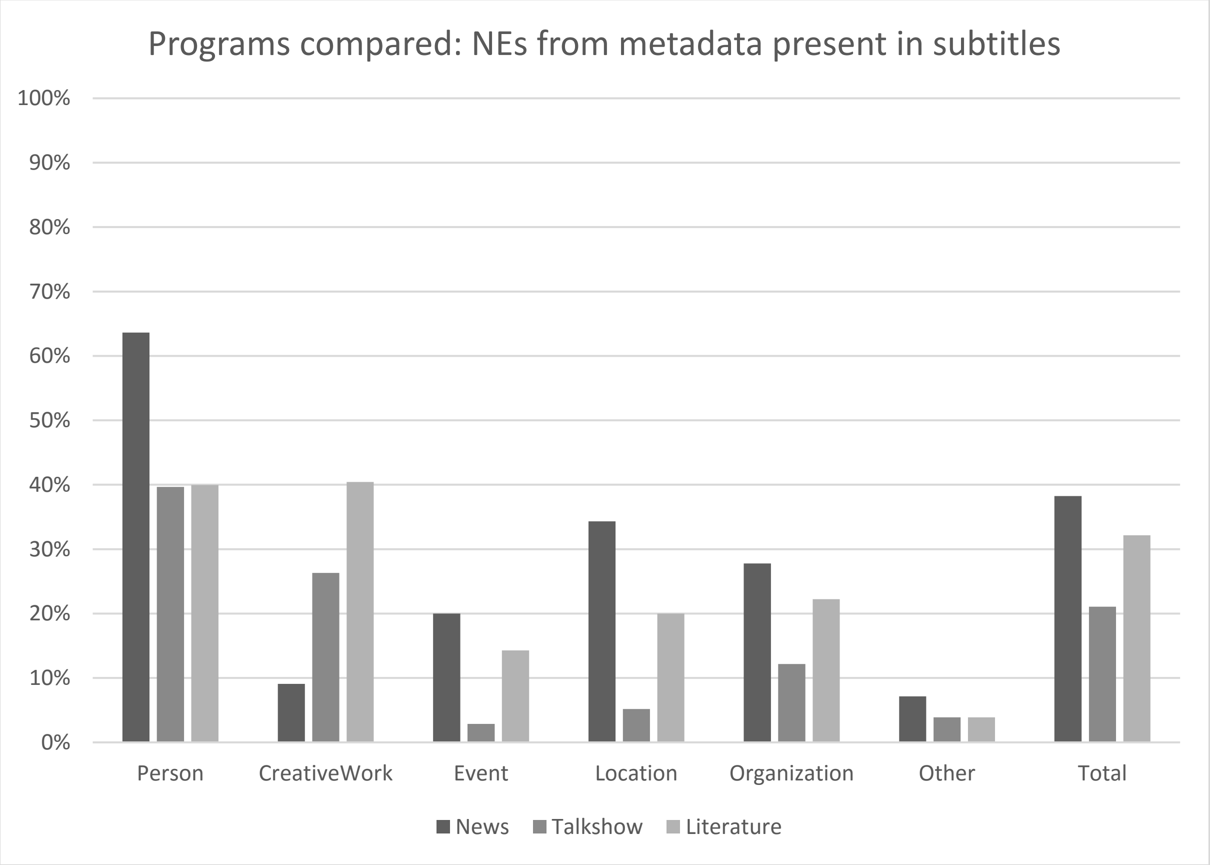

Figure 3: Overview showing the percentage of NEs from metadata also present in subtitles for different genres.

Fig. 3 shows that 38% of the NEs used in the metadata descriptions of the news programs are present in the programs subtitles. 32% of the NEs in metadata for literature programs and 21% of the NEs in metadata for talk shows are possible to extract from subtitles from these programs. This means that a substantial amount of NEs in the metadata records cannot be extracted from the subtitles. Different categories of NEs were more prominent for the different genres. What is interesting in this chart is the general pattern of a better match between NEs in metadata and subtitles for news stories than for talk shows and literature programs. In this data collection, we see that 64% of the personal names needed to describe news programs in the metadata records were also present in the subtitles for the programs, compared to 40% for talk shows and literature programs. 34% of the locations that librarians found it necessary to refer to in order to describe the news stories could have been extracted from the subtitles, while the subtitles for talk shows would be a bad source for this information, containing only 5% of the essential names of locations.

When we examine all the NEs present in subtitles, as shown in fig. 4, we see that indexers have considered 50% of the NEs from news and literature subtitles to be salient as content descriptors. For talk shows the total is 30%, and it is clear that there is a big difference between the language used by the program itself and the words used to describe the program in the metadata. The long term-goal for this research is to enable automatic extraction of the salient NEs, and this chart shows that NEs of certain entity types are more prone to be salient than others, like events, locations and organizations for news. Compared with the other entity types, the categories event and other consist of few actual NEs, which can cause each NE to have a big impact on those categories in this chart.

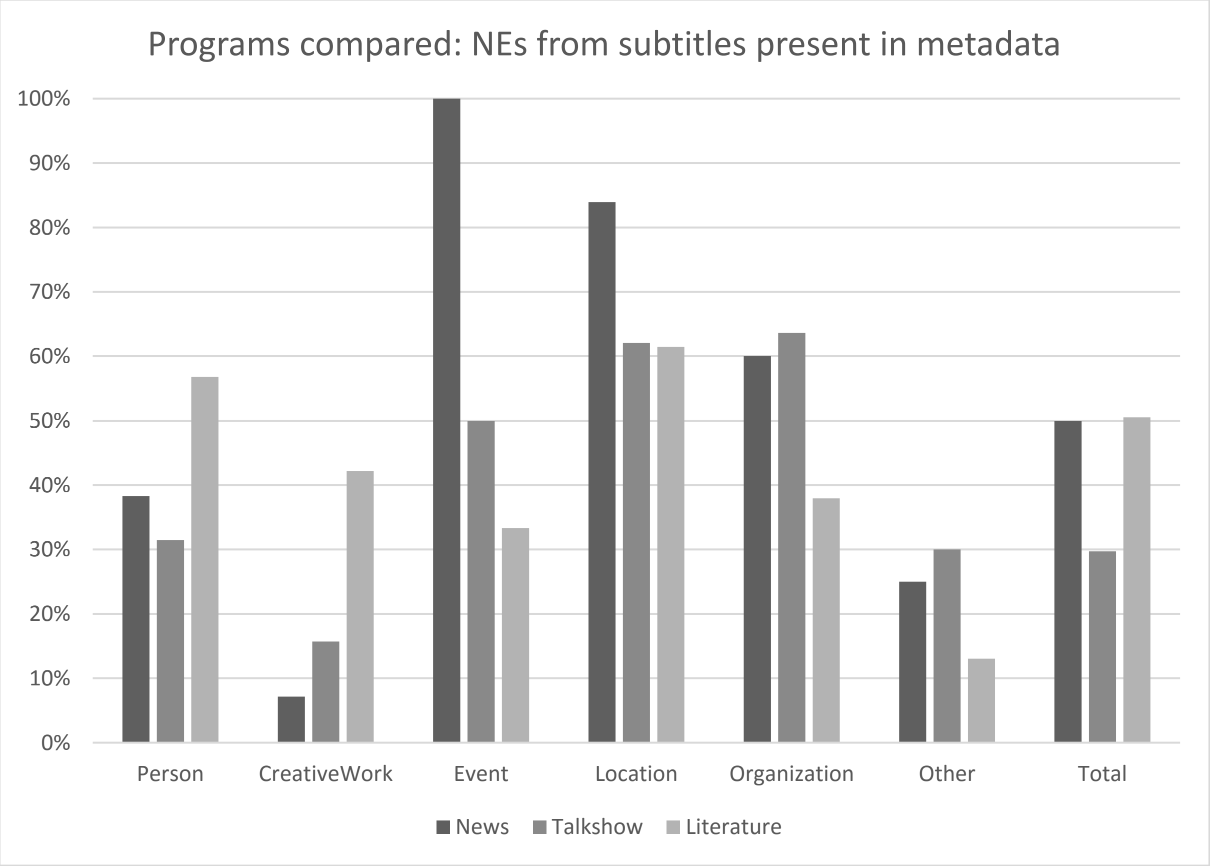

Figure 4: Overview showing the percentage of NEs from subtitles also present in metadata for different genres.

From earlier research that primarily have focused on news text, we know that the categories *person*, *location* and *organization* are important. This chart tells us that different NEs are important in literature programs. Of all the organizations mentioned in the literature programs, a lower percentage was used to describe the program in the metadata than what was the case for news programs and talk shows. *CreativeWork*, a category that contains titles for books, movies and other creative works, is much more important in literature programs. In this chart, we see that 42% of all creative works mentioned in the program also were present in the programs' metadata. The combination of entity type and genre can be used to predict the salience of a given NE to some extent, but it is not enough to draw conclusions. Of all the creative works mentioned in the program, we can assume that approximately 40% should be used in the metadata description, but how can we tell which ones that should be chosen? The frequency of how many times a NE is mentioned in the subtitles could provide valuable information.

5.4 Frequency as indicator of salience

In information retrieval, *tf–idf*, short for *term frequency–inverse document frequency*, is a numerical statistic widely used to determine how relevant a word is to a document in a collection. If a word is frequently used in a document, but not in the collection as a whole, that document will be regarded as the most relevant result for a search for that word. Because frequency is used by search engines as an expression for salience, it is interesting to investigate the impact of high-frequency and low-frequency NEs from the subtitles on the metadata records. One of the objectives of the current study is to determine whether frequency is an important factor. In a preliminary analysis of a smaller subset of the annotated data, Husevåg (Husevåg, 2016) showed that NEs that were mentioned three times or more in the subtitles were more likely to also be present in the metadata.

In a larger data set where the analyzed texts have different lengths, it is more difficult to set a threshold for salience at an exact number. Table 1 shows the relationship between frequency of NE in subtitles and their presence in metadata.

Frequency in subtitles	NEs in SUBnews	Match in MDnews	News	NEs in SUBlit	Match in MDlit	Literature	NEs in SUBtalk	Match in MDtalk	Talk shows
1	176	43	24%	1098	233	21%	732	89	12%
2	53	30	57%	270	108	40%	185	53	29%
3	30	15	50%	110	63	57%	69	21	30%
4	11	4	36%	50	30	60%	40	16	40%
5	4	2	50%	43	28	65%	33	22	67%
6	4	3	75%	23	16	70%	29	17	59%
7	6	5	83%	21	16	76%	11	8	73%

Table 1: Percentage of NEs found in metadata, arranged according to frequency of mentions in subtitles.

This table shows the percentage of NEs from subtitles also present in metadata records, arranged according to frequency in the subtitles. For NEs mentioned more than seven times in the subtitles, the numbers are so low that coincidences may have a big impact on the results. This is also to some extent the case for the figures presented for the news genre. This table shows a clear trend of increasing presence in the metadata for NEs with higher frequency in the subtitles, with a substantial difference between a frequency of one or two.

The most striking result to emerge from table 1 and fig. 3 and fig. 4 is that NEs in subtitles for talk shows seem to be less salient as content descriptors than NEs in other genres. This result may be attributed to the casual conversations in this genre, where names and places might be mentioned only in passing.

6 Conclusions and recommendations for further work

NRK's archived materials are becoming increasingly available on-line. NRK have made a major digitization effort to make Norwegian cultural heritage from the last century of radio and TV available to the public. This gives Norwegians the opportunity 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 indexers with a tool to automatically locate indexable entities and facilitate information retrieval.

This paper has analyzed subtitles and metadata records from different TV programs. Compared to earlier research on NER on Norwegian text, this paper shows similar findings with respect to entity types and name density for news text. Personal names, geographical names and names of organizations were the most prominent entity types both in the news subtitles and news metadata in this paper, and in the newspaper articles in (2009). The analysis suggests that entities of the entity types person, creative work and location are important as salient content descriptors of culture programs. Another important finding was that the density of NEs in metadata records is much higher than the NE density in subtitles, implying that NEs are suitable words to describe this kind of material. This finding is coherent with findings from book indexes which 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 (Coates-Stephens, 1992; Goldstein et al., 1999). The descriptive texts in the Norwegian metadata records presented in this paper have an average NE density of 21-39%, which is higher than the news-article summaries in (Goldstein et al., 1999), 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 (Hertzum, 2003). Recognition of NEs in subtitles is a possible solution to this challenge. This study found that 38% of the NEs used in the manually examined metadata descriptions of news stories could be found in the programs subtitles, and 32 % for literature programs. For talk shows, only 21% of the NEs in the metadata records could be extracted from the subtitles.

These findings are rather disappointing, and show that even with manual corrections of misspellings and different name forms, it is not possible to extract all the NEs that have been registered in the metadata records from the subtitles. This is consistent with results of Boer et al. who found that recall in their study was rather low as professional archivists label content with some labels that are not found by the automatic approach. They also found that term extraction from subtitles was most successful for news-type titles and other often-occurring program types (e.g., sports shows) (Boer et al., 2016). The results from the current study cannot be directly compared to the findings in Boer et al. due to different outcome measures, but their accuracy levels of 0.75 were more in line with what could be expected than the results presented in this paper. The difference between expected and obtained results may be due to the content and the quality of the metadata records. Further research should be done to investigate the metadata records, to determine if the NEs mentioned are directly related to the content of the program, or if they also contain information about production staff and copyright holders that will be of little interest to the average user.

This paper has examined the significance of frequency of NEs mentioned in subtitles. The results show that NEs with higher frequency in the subtitles are more likely to be present in the metadata records. The prevalence of NEs in metadata increases in accordance with the frequency in the subtitles, with a substantial difference between a frequency of one or two. Further analysis revealed that NEs that are mentioned one, two or three times in a talk show seem to be considered as less salient than NEs in other genres with similar frequency.

This study set out to explore the potential usefulness of indexing based on NER in subtitles. The study was designed to give an account of features that can be used to select and extract NEs that will be salient as content descriptors for culture programs from subtitles. If successful, this could be a cost-effective method to enhance precision and recall in retrieval of TV programs. The results of this investigation show that it is not possible to recreate all the NEs in the manually created metadata records by applying NER and automatic methods to the subtitles, but it is possible to find a large subset of NEs for some NE categories in certain genres. The results reported in this paper show that subtitles are a good source for personal names for all genres, and for creative works in literature programs. The study has gone some way towards enhancing our understanding of the importance of document genres, entity categories and frequencies to determine the salience of a given NE, but further work needs to be done to establish whether the NEs missing in the subtitles are important content descriptors. An important issue that was not addressed in this study, was the needs and behavior of users. It is recommended that further research be undertaken to compare search terms from actual users with the terms found in subtitles to get a more complete comprehension of the NEs potential to improve retrieval.

7 Acknowledgements

An early version of this work was presented at the 15th European Networked Knowledge Organization Systems (NKOS) Workshop in association with the International Conference on Theory and Practice of Digital Libraries in 2016, and the author is grateful for all the useful comments and discussions that have contributed to the extended version of this paper. Thanks are due to NRK who have made their internal archive available for research, and to David Massey and the other members of the TORCH project group who have helped with annotation and data collection.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, *37*(4), 699–725.

Berners-Lee, T. (2009, June 18). Linked Data - Design Issues. Retrieved February 17, 2013, from http://www.w3.org/DesignIssues/LinkedData.html

Bishop, A. P., Liddy, E. D., & Settel, B. (1991). Index quality study, Part I: Quantitative description of back-of-the-book indexes. In *Indexing Tradition and Innovation* (pp. 15–51). Port Aransas, TX: American Society of Indexers.

Boer, V. de, Ordelman, R. J. F., & Schuurman, J. (2016). Evaluating unsupervised thesaurus-based labeling of audiovisual content in an archive production environment. *International Journal on Digital Libraries*, *17*(3), 189–201. https://doi.org/10.1007/s00799-016-0182-6

Bridge, N. (2012). Geographic Names. In N. Bridge (Ed.), *Indexing names* (pp. 299–315). Information Today.

Castillo, C., De Francisci Morales, G., & Shekhawat, A. (2013). 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* (pp. 1115–1116). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2484204

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Coates-Stephens, S. (1992). *The analysis and acquisition of proper names for robust text understanding* (Ph.D.). City University London. Retrieved from http://openaccess.city.ac.uk/8015/

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science*, *49*(14), 1319–1328. https://doi.org/10.1002/(SICI)1097-4571(1998)49:14<1319::AID-ASI9>3.0.CO;2-W

Engels, R. (2010). *Åpen og samordnet tilgang til kulturarven: anbefalinger for en vellykket tilstedeværelse i den digitale kulturelle verden*. Oslo: ABM-utvikling. Retrieved from http://www.esis.no/people/robert.engels/papers/engels-abm-skrift66.pdf

Enser, P. G., & Sandom, C. J. (2002). Retrieval of archival moving imagery-CBIR outside the frame? In *Image and Video Retrieval* (pp. 206–214). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-45479-9_22

Eskevich, M., Nguyen, H., Sahuguet, M., & Huet, B. (2015). Hyper Video Browser: Search and Hyperlinking in Broadcast Media. In *Proceedings of the 23rd ACM International Conference on Multimedia* (pp. 817–818). New York, NY, USA: ACM. https://doi.org/10.1145/2733373.2812618

Gazendam, L., Wartena, C., Malaisé, V., Schreiber, G., Jong, A. de, & Brugman, H. (2009). Automatic Annotation Suggestions for Audiovisual Archives: Evaluation Aspects. *Interdisciplinary Science Reviews*, *34*(2–3), 172–188. https://doi.org/10.1179/174327909X441090

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). 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* (pp. 121–128). New York, NY, USA: ACM. https://doi.org/10.1145/312624.312665

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. In *COLING* (Vol. 96, pp. 466–471). Retrieved from http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/C96-1079.pdf

Haaland, Å. (2008). *A Maximum Entropy Approach to Proper Name Classification for Norwegian* (Doctoral thesis). University of Oslo, Oslo. Retrieved from https://www.duo.uio.no//handle/123456789/26307

Hertzum, M. (2003). Requests for information from a film archive: a case study of multimedia retrieval. *Journal of Documentation*, *59*(2), 168–186. https://doi.org/10.1108/00220410310463473

Hoff, K., & Preminger, M. (2015). Usability Testing of an Annotation Tool in a Cultural Heritage Context. In E. Garoufallou, R. J. Hartley, & P. Gaitanou (Eds.), *Metadata and Semantics Research* (Vol. 544, pp. 237–248). Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-24129-6_21

Huang, C.-C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics*, *17*(1), 132–144. https://doi.org/10.1093/bib/bbv024

Husevåg, A.-S. R. (2016). Named Entities in Indexing: A Case Study of TV Subtitles and Metadata Records. In *CEUR Workshop Proceedings* (Vol. 1676, pp. 48–58). Retrieved from http://ceur-ws.org/Vol-1676/paper6.pdf

Huurnink, B., Hollink, L., Van Den Heuvel, W., & De Rijke, M. (2010). Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *Journal of the American Society for Information Science and Technology*, *61*(6), 1180–1197.

Johannessen, J. B., Hagen, K., Haaland, Å., Jónsdottir, A. B., Nøklestad, A., Kokkinakis, D., … Haltrup, D. (2005). Named Entity Recognition for the Mainland Scandinavian Languages. *Literary and Linguistic Computing*, *20*(1), 91–102. https://doi.org/10.1093/llc/fqh045

Johansen, B. (2015). Named-Entity Chunking for Norwegian Text using Support Vector Machines. *Norsk Informatikkonferanse (NIK)*. Retrieved from http://ojs.bibsys.no/index.php/NIK/article/view/248

Jónsdóttir, A. B. (2003). *ARNER, what kind of name is that? : an automatic rule-based named entity recognizer for Norwegian* (Master thesis). University of Oslo, Oslo. Retrieved from https://www.duo.uio.no//handle/123456789/26385

Karsdorp, F. B., Kranenburg, P. van, Meder, T., & Bosch, A. (2012). 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)*. Edicoes Cilibri. Retrieved from http://depot.knaw.nl/12956/1/karsdorp_et_al2012b.pdf

Klement, S. (2002). Open-system versus closed-system indexing. *Indexer*, *23*(1), 23–31.

Knittel, J., & Dingler, T. (2016). 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* (pp. 93–103). New York, NY, USA: ACM. https://doi.org/10.1145/2932206.2932217

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., … Lee, R. (2009). Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, … E. Simperl (Eds.), *The Semantic Web: Research and Applications* (pp. 723–737). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-02121-3_53

Liddy, E. D., Bishop, A. P., & Settel, B. (1991). Index quality study, Part II: Publishers survey and qualitative assessment. In *Indexing Tradition and Innovation* (pp. 53–79). Port Aransas, TX: American Society of Indexers.

Liu, X., Wei, F., Zhang, S., & Zhou, M. (2013). Named entity recognition for tweets. *ACM Trans. Intell. Syst. Technol.*, *4*(1), 3:1–3:15. https://doi.org/10.1145/2414425.2414428

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26. https://doi.org/10.1075/li.30.1.03nad

Nandakumar, A., & Murray, J. (2014). 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* (pp. 3–10). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2602317

Nøklestad, A. (2009). *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection* (Doctoral thesis). University of Oslo, Oslo. Retrieved from https://www.duo.uio.no//handle/123456789/26326

Odijk, D., Meij, E., & de Rijke, M. (2013). Feeding the Second Screen: Semantic Linking Based on Subtitles. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (pp. 9–16). Paris, France, France: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. Retrieved from http://dl.acm.org/citation.cfm?id=2491748.2491751

Ordelman, R., Aly, R., Eskevich, M., Huet, B., & Jones, G. J. F. (2015). Convenient Discovery of Archived Video Using Audiovisual Hyperlinking. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia* (pp. 23–26). New York, NY, USA: ACM. https://doi.org/10.1145/2802558.2814652

Ordelman, R. J., Eskevich, M., Aly, R., Huet, B., & Jones, G. (2015). Defining and evaluating video hyperlinking for navigating multimedia archives. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 727–732). ACM.

Poibeau, T., & Kosseim, L. (2001). Proper name extraction from non-journalistic texts. *Language and Computers*, *37*(1), 144–157.

Redondoio Garcia, J. L., De Vocht, L., Troncy, R., Mannens, E., & Van de Walle, R. (2014). Describing and contextualizing events in tv news show. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 759–764). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2579326

Sekine, S., & Ranchhod, E. (2009). *Named entities: recognition, classification and use*. Amsterdam: John Benjamins.

Smith, S. (2012). Name problems: Dispelling the Simplicity Myth. In N. Bridge (Ed.), *Indexing names* (pp. 239–260). Information Today.

Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains. *Language and Linguistics Compass*, *4*(9), 750–768. https://doi.org/10.1111/j.1749-818X.2010.00230.x

Tallerås, K., Massey, D., Husevåg, A.-S. R., Preminger, M., & Pharo, N. (2014). Evaluating (linked) metadata transformations across cultural heritage domains. In *Metadata and semantics research* (pp. 250–261). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-13674-5_24

Tommasi, T., Aly, R., McGuinness, K., Chatfield, K., Arandjelovic, R., Parkhi, O., … Tuytelaars, T. (2014). Beyond metadata: searching your archive based on its audio-visual content (pp. 1.3-1.3). Presented at the IBC 2014. https://doi.org/10.1049/ib.2014.0003

Zafran, E. (2008). Pick It Up or Pass It Up? *Key Words*, *16*(2), 45–55.

Zafran, E. (2012). Names in Art Books. In N. Bridge (Ed.), *Indexing names* (pp. 219–226). Information Today.