Semantic Multimedia Information Retrieval Based on Contextual Descriptions

Nadine Steinmetz and Harald Sack

Hasso Plattner Institute for Software Systems Engineering, Potsdam, Germany,
nadine.steinmetz@hpi.uni-potsdam.de,
harald.sack@hpi.uni-potsdam.de

Abstract. Semantic analysis and annotation of textual information with appropriate semantic entities is an essential task to enable content based search on the annotated data. For video resources textual information is rare at first sight. But in recent years the development of technologies for automatic extraction of textual information from audio visual content has advanced. Additionally, video portals allow videos to be annotated with tags and comments by authors as well as users. All this information taken together forms video metadata which is manyfold in various ways. By making use of the characteristics of the different metadata types context can be determined to enable sound and reliable semantic analysis and to support accuracy of understanding the video's content. This paper proposes a description model of video metadata for semantic analysis taking into account various contextual factors.

Keywords: context model, semantic analysis, video analysis, metadata analysis, named entity recognition

1 Introduction

Context is an important factor that is mandatory for general understanding. Depending on the context, information might entail different meaning and thus, lead to different decisions. Context can be considered as the sum of available information items that put together enable unambiguous determination of the meaning of information.

For information retrieval and esp. for semantic and explorative search that take into account content-related information, it is of high importance to decide upon the various possible meaning of information. For semantic analysis, besides authoritative (textual) information supplied by experts also automatically extracted metadata or user-provided annotation contribute essential additional information about the content. However, metadata from different sources involve different characteristics and reliability.

Furthermore, due to the rich expressiveness of natural language textual information entails the problem of ambiguity. Thus, the word sense disambiguation of document metadata deserves special attention. The context needed for

disambiguating ambiguous terms within a document is provided by all the surrounding information such as further metadata or textual content related to the same document or document fragment under consideration.

The different characteristics of metadata items influence their confidence and relevance when applied as context items for the disambiguation process. So far, in computer science context is primarily discussed in the sense of *user context*. User context describes the situation of an interacting user. Here, context is used to solve a specific request in a personalized way, as e. g., in an ubiquitous computing scenario.

In this paper we present a context model that describes characteristics of metadata items. These metadata items may serve as context items for other metadata items according to their characteristics. Our context model includes a derived confidence value representing information about the anticipated ambiguity and correctness of the metadata item. This confidence value is applied to rank metadata items for a given context. Context determining metadata items, henceforth refered to as *context items*, support the subsequent semantic analysis process. As an application we apply our context model to support understanding of video metadata from various sources and improve the accuracy of semantic analysis.

The paper is organized as follows: Section 2 recapitulates related work in the field on context awareness and context definitions. The context model and a description of the identified contextual factors are presented in Section 3. In Section 4 the application of the context model within the semantic analysis process is presented. The proposed context model has been evaluated on the basis of an annotated dataset of video metadata. The evaluation results including the dataset are described in Section 5. Section 6 summarizes the achievements of this work and gives an outlook on future work.

2 Related Work

Recently, context and context-aware computing has received increasingly attention [10]. But the discussions about the influence and importance of context date far back into the past throughout various scientific fields of computer science. Mainly these discussions address the context a person is enclosed by. Therefore, characteristics of context are defined to solve personalization problems in e-commerce and ubiquitous computing, to identify life stages of a person for data mining, or to improve online marketing and management [1]. This context can be considered as user context. Although the received opinion agrees on the difficulties of defining context in general and finding a universal definition, the different disciplines identify certain characteristics for their field of interest. Lenat [7] states that for artificial intelligence context has been ignored or treated as black box for a long time. For the large knowledge base Cyc[1] he defined twelve dimensions of context to "specify the proper context in which an assertion (or

[1] `http://cyc.com/cyc/opencyc`

question) should be stated". Bazire et. al collected 150 different definitions of context from different disciplines to identify the main components of context [2]. They concluded their study by determining all definitions to the parameters constraint, influence, behavior, nature, structure, and system. In ubiquitous computing context is broadly used for two purposes: as retrieval cue and to tailor the behavior and the response type of the system [6]. Dourish has identified two different views on context: a representational and an interactional view and suggests the latter to be the more challenging for the field of interactive systems.

In 1931 Dewey wrote "We grasp the meaning of what is said in our language not because appreciation of context is unnecessary but because context is inescapably present." [5]. Although, this sentence addresses context in the field of psychology it is also valid for the characteristics of metadata as context items. Context is defined by the characteristics of the items included in it.

We utilize the characteristics of context items for semantic analysis, in particular for Named Entity Recognition (NER). In Natural Language Processing (NLP) the term NER refers to a method to find entities of specific types (persons, places, companies etc.) in a text. Similar to Word Sense Disambiguation (WSD) approaches we consider NER as the method to find specific entities with a unique meaning ("Berlin" as the German capital and not the town in Connecticut, U.S.) Mihalcea et. al published one of the first NER approaches using Wikipedia[2] URIs to identify specific entities [12]. This paper presents a combined approach of an analytical method comparing Wikipedia articles with contextual paragraphs and a machine-learning approach for the disambiguation process. Another machine-learning approach is presented in [3]. This approach uses different specific kernels in linear combination to disambiguate terms in a given text. The kernels are trained with surrounding words of an entity link within the paragraphs of the Wikipedia article. DBpedia Spotlight is an established NER application that applies an analytical approach for the disambiguation process. The context information of the text to be annotated is represented by a vector. Every entity candidate of a term[3] found in the text is represented as a vector composed of all terms that co-occurred within the same paragraphs of the Wikipedia articles where this entity is linked [11]. Recently, Damljanovic et. al presented an approach of combining the classic NER tagging (in terms of NLP) and entity disambiguation [4]. The terms the NER tagging tool identified as one of the expected categories (person, place, or organization) are assigned to DBpedia[4] classes. Entity candidates for this term are retrieved within the instances of the assigned ontology class.

All these NER approaches aim at the analysis of text documents. Context definitions are limited to merely structural characteristics such as word, sentence, paragraph, or full document [14]. We extend this context definition by

[2] `http://www.wikipedia.org`

[3] The authors use the expression "surface form" for a word or a word group representing an entity. Subsequently we use "term" synonymously to this definition.

[4] `http://dbpedia.org/About`

determining further specific characteristics of the metadata items pertaining to a context.

3 Context & Contextual Factors

Documents are created within a specific user context determining the purpose the document was created for. This context can also be considered as pragmatics. The metadata provided for the document as well as data automatically extracted from the document form a different context. This context determines the meaning of the given information. Therefore, we define:

Definition 1. A **context** is represented by a finite set C of context items. Each **context item** $ci \in C$ is a tuple $(term, uri, cd, c)$, where:

- $term$ denotes the value (string text) of the context item,
- uri denotes the list of (semantic) entities assigned to the $term$,
- cd denotes the contextual description $cd \in CD$ of the context item ci,
- $c \in [0...1]$ denotes the confidence value that is calculated according to cd.

Thereby we state that a context consists of context items. The context items derive from the metadata a document is provided with. The metadata items of a context belong to certain domains and thereby define the meaning of the textual information. In that way metadata items become context items.[5] Most of the metadata and automatically extracted information is provided in the form of natural language text. As already mentioned in the introduction natural language is expressive but entails the problem of ambiguity. To enable semantic annotation of documents and the documents' metadata the ambiguity of the textual information has to be removed.This is where the context comes into play. The characteristics of a context are determined by the items pertaining to it. But these context items originate from different sources, have different reliabilities and should therefore be weighted according to their significance within a context. We have defined a contextual description depicting the characteristics of these context items.

Definition 2. A **contextual description** $cd \in CD$ is a tuple (tt, st, sd, cl), where:

- $tt \in Tt$, where Tt is a finite set of text types,
- $st \in St$, where St is a finite set of source types,
- $sd \subseteq Sd$, where Sd is the set of available sources for the video,
- $cl \in Cl$, where Cl is a finite set of ontology classes,
- CD denotes the set of all contextual descriptions.

For our proposed use case, the semantic analysis of video metadata, we have restricted text types, sources, and ontology classes to the following sets:

[5] Subsequently, we use the terms metadata item and context item synonymously.

- the set of text types Tt is determined to natural language text, keywords, and tags.
- the set of ontology classes Cl is determined to place, organization, and person.
- the set of source types St is determined to authoritative and non-authoritative sources, Automatic Speech Recognition (ASR), and Optical Character Recognition (OCR).

Useful sources for automatically extracted textual information for video data are OCR and ASR algorithms. Usually few authoritative metadata is available as e. g., a title, speaker or primary persons, publisher etc. Additionally, some video resources are provided with textual, time-related tags by non-authorative sources [6]. Therefore we have restricted the set of available source types for video metadata St to these four sources.

Metadata from ASR and OCR sources, as well as the title and description from the authoritative metadata can be considered as natural language text. Information about the speaker or the publisher are usually given as keywords. Tags form a third text type as they are mostly given as a group of single words and only subsets of the group belong together (c.f. [8] for tag processing). Tt is therefore restricted to these three text types.

To determine appropriate entities for a given textual information it helps to know the prospective ontology class the entity belongs to. Some of the provided authoritative metadata can directly be assigned to ontology classes, as e. g., the metadata item for *speaker* can directly be assigned to the ontology class *Person*. For natural language processing Conditional Random Field (CRF) classifiers[7] are used to find entities of such ontology classes in fluent text. By using a 3-class model the ontology classes *Person*, *Place*, and *Organization* can be found in a text. Therefore the set Cl is restricted to these three ontology classes.

3.1 Detailed Contextual Description and Confidence Calculation

According to the contextual description the confidence of the context item is calculated. For each of the four contextual factors (tt, st, sd, and cl) a double precision value v is calculated, where $0 < v \leq 1.0$.

Source Reliability The term *reliability* is referring to a prospective error rate concerning the source type st. Document metadata can either be created by human or computer agents. Human agents can be the author, who created the document or any user, who annotated the document with additional information. Computer agents are analysis algorithms, which extract (mostly) textual information from a multimedia document, such as OCR and ASR. All these

[6] video portals like Yovisto (www.yovisto.com) allow the videos to be tagged by any user to make time-related references to the video

[7] as used in the Stanford Named Entity Recognizer - `http://nlp.stanford.edu/software/CRF-NER.shtml`

agents provide information with different degrees of reliability. Where human agents in general can be considered more reliable than computer agents because of knowledge and experience, authoritative human agents are considered more reliable than non-authoritative human agents. According to this simple presumption the agents' reliability is ranked. The value v_{st} is set highest for authoritative ($v_{st} = 1.0$) and slightly lower for non-authoritative (human) sources ($v_{st} = 0.9$). As reliability values for computer agents we simply adopt the achieved evaluation results on precision for the considered analysis engines. Unfortunately, most video OCR evaluations base on single frame processing, which embellishes the results. Precision for video OCR on videos with equally text and non-text frames is still very low. According to [15], the error rate for news videos is up to 65%. Therefore we assume a worst case precision of 35% ($v_{st} = 0.35$) for context items with an OCR analysis as source agent. Word error rates for ASR analysis engines range between 10% and 50% (respectively an accuracy rate between 50% and 90%)[13]. We assume the worst case and determine the reliability value for context items from ASR results to $v_{st} = 0.5$.

Source Diversity Source diversity specifies how many of the available annotation sources agree on the same metadata item. The diversity ranges from a single source to all available sources. The more sources agree on the value of a context item the more reliable the item is considered. Depending on the available sources (Sd) and the set of sources that agree on the same item i (s_i), the value for the source diversity v_{sd} is calculated as follows:

$$v_{sd} = \frac{|s_i|}{|Sd|}$$

Example: The text "computer" is automatically extracted by OCR analysis from a video frame. The title of the video is "The birth of the computer". For this video the only sources of textual information are the authoritative metadata and the extracted texts by OCR. In this case $v_{sd} = \frac{2}{2} = 1.0$ for context items having $term = computer$ as the term "computer" is confirmed by both available sources.

Text Type According to the source of the metadata item the general type of the context item's values differ. Authoritative information of a document as e. g. the creator, production location, or keywords have key terms as values. These key terms usually in total depict an entity. Further authoritative textual information, such as the title or a descriptive text are given as running text in natural language. A third text type are typed literals, as e. g., "print run = 1.000 copies". It is assumed that the ambiguity of metadata items with text type 'typed literal' is lowest. Therefore the according confidence value is highest with $v_{tt} = 1.0$. But usually this text type is not representative for video metadata. The ambiguity of running text depends on the precision of the NLP algorithm used to extract key terms. We are using the Stanford POS tagger[8] to identify

[8] `http://nlp.stanford.edu/software/tagger.shtml`

word types in text. This tagger has an accuracy rate of 56% per sentence[9], which leads to $v_{tt} = 0.56$. By using this rate as reliability value for running text we have a measure independent from text length. POS tagging is not needed for context items that are given as key terms. But still, to allow an uncertainty we determine the reliability of key terms slightly lower than for typed literals as $v_{tt} = 0.9$.

Class Cardinality The contextual factor of class cardinality corresponds to the number of instances the assigned ontology class contains. In general a descriptive text does not refer to a specific ontology class, if a CRF classifier does not find any classes in the text. The entities found in such a text can be of any type. In that case the context items found in this natural language text are assigned to the most general class, $\top$ class of the ontology[9] and the class cardinality is highest. According to the ontology class cl assigned to the metadata item and its known cardinality the value v_{cl} is calculated proportional to the overall number of all known entities ($|\top|$), where $\top$ denotes the most general class containing all individuals of the knowledge base, and $|\top|$ denotes the number of all instances pertaining to this class. A high class cardinality entails a high ambiguity. Therefore, the value v_{cl} is inverted to reflect a reverse proportionality regarding the amount of the value and the ambiguity:

$$v_{cl} = 1 - \frac{|cl|}{|\top|}$$

Example: A context item of a video might be identified as Person (by uploading author or by an automatic NER tagging tool). Using the DBpedia Version 3.8.0 as knowledge base, the class "Person" contains 763,644 instances. owl:Thing as top class of the DBpedia ontology holds 2,350,907 instances. Accordingly, the confidence value $v_{cl} = 1 - \frac{763,644}{2,350,907} = 0.67$ for a context item assigned to the DBpedia ontology class "Person".

The number of entity candidates of a term can also be a measure for the prospective ambiguity of the term. However, evaluations showed better results for the approach on class cardinality. Details on the evaluation results are described in Section 5.

After calculating each confidence value for the four constituents of the contextual description the total confidence value for a context item calculates as follows:

$$c = \frac{v_c + v_{sd} + v_{sr} + v_{tt}}{4}$$

3.2 Exemplary Confidence Calculation for Context Items

Let an example video have the following authoritative metadata information:

[9] which means, all entities of the knowledge base have to be considered and the amount cannot be restricted to a certain class

Table 1. Example values for contextual factors and the according confidence

term	*tt*	v_{tt}	*cl*	v_{cl}	*st*	v_st	*sd*	v_{sd}	*c*
TED	keyword	0.9	Organiz.	0.96	auth.	1.0	auth.	0.5	0.84
George Dyson	keyword	0.9	Person	0.85	auth.	1.0	auth.	0.5	0.81
computer	nat. language	0.56	$\top$	0.0	auth.	1.0	auth., OCR	1.0	0.64
birth	nat. language	0.56	$\top$	0.0	auth.	1.0	auth.	0.5	0.52
computer	nat. language	0.56	$\top$	0.0	OCR	0.35	auth., OCR	1.0	0.48
alamogordo	nat. language	0.56	$\top$	0.0	OCR	0.35	OCR	0.5	0.35

– Title: “The birth of the computer.”
– Speaker: “George Dyson”
– Publisher: “TED”

Additionally, “computer” and “alamogordo” were extracted from the video via OCR analysis.

Speaker and publisher information are considered as keywords. The title and the OCR texts are considered as natural language text. Speaker is assigned to the DBpedia ontology class “Person” and publisher is assigned to the DBpedia ontology class “Organization”. The NER tagger did not find any class types in the title or the OCR information. After NLP pre-processing six context items are generated from the given metadata. The contextual factors and the calculated confidence value of the six context items are shown in Table 1.

3.3 Context Items Views

As shown in Figure 1, the identified contextual factors and dimensions influence different superordinate characteristics and can be aggregated in two different views: the confidence view and the relevance view on context items. The confidence view aggregates the characteristics of a context item described above. But context items also have characteristics regarding their context relevance within the video.

Confidence view The **correctness** of a context item is influenced by the source diversity as well as by the source reliability. The more sources agree on an item and the higher the reliability of the item’s source is, the higher is the reliability that this item is correct. The **ambiguity** of a context item is influenced by the text type and the class cardinality assigned to the item. Natural language text needs NLP technologies to identify important key terms. Due to the possible number of potential errors the ambiguity of natural language text is considered higher than for simple restricted key terms. For key terms no further processing is needed. Also, the lower the amount of instances of the assigned ontology class the lower is the item’s potential ambiguity.

Both, ambiguity and correctness influence the confidence of a context item. With the term confidence we aim at the trust level we assign to the item for

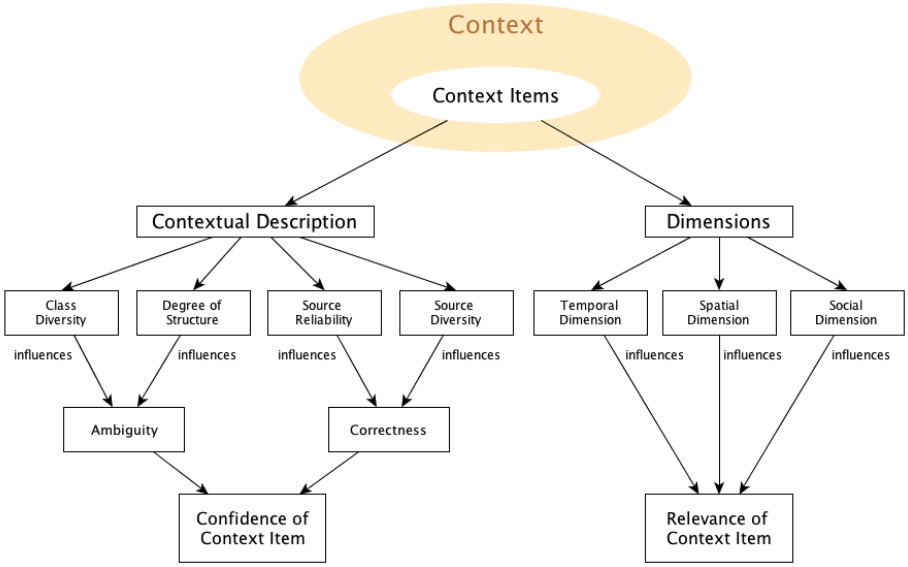

Fig. 1. Contextual factors of context items

further analysis steps. A high correctness rate and a low ambiguity rate entail a high confidence for the context item. The confidence view is used to order context items according to their correctness and ambiguity. The higher the confidence the higher is the probability that the context item is analyzed correctly and the accurate entity is assigned to the item.

Relevance view The spatial, temporal, and social dimension specify the relevance of a context item in relation to other context items of the document. With the help of the dimensions the divergence of the context items w.r.t. the document's content can be identified. By creating a context for a semantic analysis of context items the relevance view is important to aggregate the amount of all context items related to a document to smaller groups of stronger content-related coherence. In this way the semantic analysis of context items can be performed within more accurate and therefore also more meaningful contexts.

Metadata items of time referenced documents, such as video or audio files can be assigned to document fragments or the full document. The **temporal dimension** reflects the reference period of the item. The values of this dimension have a range between the smallest unit of the document (e. g., a frame for a video) and the full document. The **spatial dimension** assigns the metadata item to a specific region respectively to the entire document. E. g., for a video document the values starts with a single pixel within a frame over "geometrically determined region within a frame" to the full frame. The **social dimension** plays a special role within the characteristics of context items. It takes into account information

about social relationships of the user who created the metadata item as well as the user, who accesses the document. Therefore, this dimension is dependent of the user and covers a personal perspective.

4 Using Context for Semantic Analysis

We apply the proposed context description model to the semantic analysis of video metadata and the annotation of textual information with semantic entities. Subsequently, we will refer to our semantic analyzing engine as *conTagger*. The semantic analysis of metadata items of a video consists of three main steps:

- Collecting metadata items and defining contextual description
- Calculating the confidence value and sorting the list of metadata items according to their confidence
- Disambiguating every item using dynamically created context

4.1 Semantic Analysis Based on Ranked Context Items

The conTagger disambiguates context items based on term-entity co-occurrence as well as on the Wikipedia link graph. According to the degree of integration of the context item to be disambiguated within the context (which is a list of context items) the highest ranked entity candidate is chosen[10]. A context item initially includes a list of entity candidates for the item's textual term - if the term is ambiguous. After the disambiguation process the entity candidates are replaced by the resulting entity. The resulting entity features a disambiguation score. This disambiguation score has a range of [0.0 ... 1.0] and represents a trust value of the disambiguation process. The higher the value the higher the probability of a correct disambiguation. If an already disambiguated context item is added to influence the disambiguation of another item, the assigned entity is used as a fix point for the context creation. Otherwise the entire list of all entity candidates of the context item is used for the disambiguation process. Non-ambiguous context items initially contain only one entity featuring a disambiguation score of 1.0. Subject to these conditions the following hypothesis is put forward:

Hypothesis 1. The disambiguation results of context items are improved, if context items with higher confidences are disambiguated first.

4.2 Dynamically Creating Context for Disambiguation

The context of a context item determines the meaning of ambiguous textual information of the item to a single entity. The more specific the context the higher the probability of a correct disambiguation. Usually documents of any type are structured according to content-related segments. The more segments

[10] For more information on the disambiguation process, please cf. [8]

Table 2. Evaluation of Hypothesis 2

	ASR		OCR		Tags	
	Recall	Precision	Recall	Precision	Recall	Precision
conTagger, Segment-Based	**55.0**	**61.0**	**56.0**	**24.0**	**71.0**	**69.5**
conTagger, Video-Based	53.0	46.0	51.0	21.0	69.0	68.0

are aggregated as context the more general the contextual information is considered. Ambiguous textual information is hard to be disambiguated using a rather general context, because a general context probably contains more heterogenous information. Thus, the document should be segmented into fragments of coherent content to be able to create more accurate contexts. Considering this presumption the following hypothesis is put forward:

Hypothesis 2. The context of context items within a document should be restricted to segments of coherent content.

Following hypotheses 1 and 2 the context for the disambiguation of each item is created dynamically. Only context items of the same segment and with a defined minimum confidence value are added to the context and thereby influence the disambiguation process. The context items from authoritative metadata are added as context for the disambiguation of all time-related context items – but also only if their confidence value exceeds a certain threshold. This threshold can be set dynamically for each context item type and is discussed in Section 5. The same applies for the disambiguation score of a disambiguated context item. For the dynamic context creation the score has to exceed a defined threshold. This threshold is also discussed in Section 5. By using thresholds for confidence value and disambiguation score the precision of the disambiguation process is aimed to be high without decreasing the recall. Using the dynamically created context each context item is disambiguated and the highest ranked entity is assigned as determining entity for the textual information instead of the list of entity candidates. Analysis and evaluation results for hypotheses 1 and 2 are discussed in the following section.

5 Evaluation

To evaluate NER algorithms a ground truth consisting of a text and a list of correct entities assigned to terms in this text is needed. Few datasets of simple texts and according entities are available in order to compare different NER algorithms. The creation of such a dataset is costly and time-consuming. Mendes et. al published such a dataset for the evaluation of the NER tool *DBpedia Spotlight* [11].

For the evaluation of *conTagger* a dataset consisting of different types of video metadata including the correct entities assigned to all the available textual information is needed. As far as we know, no dataset of that structure and for

that purpose is available. Therefore, we have created a dataset of annotated video metadata in order to be able to evaluate our approach.

The evaluation dataset consists of metadata from five videos. The videos are live recordings of TED[11] conference talks covering the topics physics, biology, psychology, sociology, and history science. The metadata for each video consist of authoritative metadata (including title, speaker, providing organization, subject, keywords, descriptive text, and a Wikipedia text corresponding to the speaker), user-generated tags, and automatically extracted text from OCR and ASR. The videos have been partitioned into content-related video segments via automatic scene cut detection. The time-related metadata (tags, ASR, and OCR) is assigned to the related video segments. Overall the dataset consists of 822 metadata items, where an item can be a key term or fluent text consisting of up to almost 1000 words[12].

Table 3. Evaluation results (R – Recall, P – Precision, and F_1-Measure) of the *conTagger* compared to simple segment-based NER, DBpedia Spotlight and the Wiki Machine.

	conTagger			*Simple NER*			*Wiki Machine*			*Spotlight*		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
Authorative	**60.0**	54.5	57.0	52.0	46.0	49.0	59.5	**56.5**	**58.0**	50.0	44.0	47.0
Tags	**71.0**	**69.5**	**70.0**	61.0	60.0	60.5	44.0	62.0	51.5	60.0	59.0	59.5
ASR	55.0	**61.0**	**58.0**	56.5	38.0	45.5	**61.5**	50.0	55.0	56.0	34.0	42.5
OCR	**56.0**	**24.0**	**34.0**	44.0	17.5	25.0	24.5	18.0	21.0	47.0	18.0	26.0
Segments	54.0	**58.0**	**56.0**	57.0	39.0	46.5	57.0	49.0	52.5	**59.0**	35.0	43.5
Video	56.0	**48.0**	**52.0**	57.0	30.0	39.5	**58.0**	43.0	49.5	54.0	31.0	39.5

For evaluating Hypothesis 2 the context items were disambiguated using the entire video as context as well as for only context items of the same segments for comparison. Evaluation results are shown in Table 2. Recall values state how many of the entities of the ground truth are found by the respective analysis approach. Precision states how many of the extracted entities are present in the ground truth. As anticipated, the disambiguation results are improved using content based segments as context. Especially results for ASR metadata items differ in recall and precision for both variants. This probably follows from the fact that in our dataset there are much more ASR metadata items and because speech usually comprehends wider spread content in terms of context information. However, recall and precision are not significantly different, which results from the homogeneous character of the single videos of our dataset and their video segments.

To evaluate the *conTagger* regarding Hypothesis 1 we have compared the evaluation results to our own simple segment-based NER, NER by DBpedia

[11] `http://www.ted.com`

[12] For downloading the dataset and the ground truth please cf. the readme file at `http://tinyurl.com/cztyayu`

Spotlight[11] and NER by the Wiki Machine[3]. For the analysis of the video metadata using DBpedia Spotlight, all metadata items assigned to a video segment – constituting a context – have been processed together via the Spotlight Webservice. The Wiki Machine results are achieved by disambiguating each metadata item on its own.

The evaluation results according to the different sources as well as video and segment-based evaluation are depicted in Table 3. The results are aggregated according to different sources and different relevance views. For the different sources the recall and precision values are calculated per video and averaged over all five videos. For segments the recall and precision values are calculated for every segment over all sources and averaged over all segments. The evaluation results for videos are calculated respectively. Most notably, the *conTagger* achieves significantly good results on the metadata items with lower confidence, as OCR and ASR results. The overall evaluation of annotated entities per segment and video confirms the very good results. *ConTagger* achieves very good precision and F_1-measure results compared to the other NER approaches. As described in Section 3.1 the ambiguity of a context item can be defined by the number of entity candidates. We evaluated the disambiguation process using the inverted normalized number of entity candidates instead of the class cardinality measure. Better evaluation results were achieved by using the class cardinality. F_1-measures for all source types were lower at an average of 5% when using the ambiguity measure based on entity candidates. Obviously a low number of entity candidates does not necessarily mean that the correct entity is amongst the few candidates. Therefore, the ambiguity measure is set according to class cardinality of an assigned class.

For the dynamic context creation we have processed exhaustive test runs to determine the best suited thresholds for the confidence value and the disambiguation score when adding items to the context for a disambiguation process. The values for both parameters range between 0 and 1. Therefore, the context creation and subsequent disambiguation process has been performed with all combinations of confidence and disambiguation score values increasing the parameters in steps of 0.05, resulting in 441 runs. Subsequently, the parameters settings achieving the best recall and precision values aggregated over different source types have been identified. The best recall and precision results for metadata items from OCR and ASR analysis (featuring lowest confidence values) are achieved by creating the context from context items with a minimum confidence value of $c = 0.7$. Authoritative metadata items (featuring highest confidence values) are disambiguated using context items with highest confidence values in any case, because no time-referenced items must be used. Therefore the identified minimum threshold for the dynamic context creation is comparatively low with $c = 0.25$. The minimum threshold for the disambiguation of time-referenced tags is determined mid-range. This means that some of the other time-referenced metadata items (from OCR or ASR analysis) are used as context items, but not all of them as the lowest calculated confidence value for time-referenced metadata was calculated with $c = 0.285$. Apparently the disambiguation score is not

as important as the confidence value for the context items used as influential items for a disambiguation process.

These evaluation results support our premise that the characteristics and the use of contextual factors of different metadata items support the semantic analysis process.

6 Ongoing & Future Work

Major contributions of our work include the definition of contextual information of video metadata for the purpose of NER and calculating a confidence value. This value is used to bring metadata items in a specific order and to use them as context items for the disambiguation process. Based on this information and the temporal, spatial and social dimension metadata items influence the results of semantic analysis as context items. Current NER approaches miss to identify specific characteristics of document metadata. We have presented an extensible context description model that determines the important facts of document metadata items in a context. The characteristics of the context items are exemplary applied for semantic analysis on video metadata. Moreover, the context model is also applicable to any document type where metadata is harvested from different sources.

Ongoing and future work concentrates on the further refinement of a context. The social dimension plays an important role from the users' perspective of metadata. Metadata endorsed by friends or colleagues can be helpful for the user as additional descriptive information. This dimension also can be used to represent the pragmatics of a user when editing or creating metadata. In this way the context might change over time.

Future work includes the consideration of the influence of this additional context dimension on the context model and its application. A context can also further be refined by sample low level adjustments as white and black lists. Whitelisting can either be achieved statically by applying a specific knowledge base that only "knows" relevant entities and reduces the ambiguity of terms or by logical constraining rules. E. g., a document produced in 1960 does most likely only reference persons born before this date. So only a constrained number of entities qualifies for the analysis of this document. While persons naturally have a time reference, other real world entities may be hard to classify. Ongoing work includes the definition of a time-related scope for various entity types. Blacklisting on the other hand disqualifies particular entities for the analysis process. This can also either be achieved manually or automatically. Automatic blacklisting can be achieved by adding the previously deselected entity candidates (those that were not selected by disambiguation) of a disambiguated context item to a context restriction. With every disambiguated context item this "negative context" grows and a dynamic blacklist is achieved. Entity candidates related to this negative context will receive a penalty.

With the presented work we point out the importance of contextual factors of metadata. The proposed context model enables the characterization of

metadata items from different sources and of various structure. By using the example of video metadata we were able to show how to support the (automatic) comprehension of a document's content with the help of its metadata.

References

1. G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 335–336, New York, NY, USA, 2008. ACM.
2. M. Bazire and P. Brézillon. Understanding context before using it. In *Proceedings of the 5th international conference on Modeling and Using Context*, CONTEXT'05, pages 29–40, Berlin, Heidelberg, 2005. Springer-Verlag.
3. V. Bryl, C. Giuliano, L. Serafini, and K. Tymoshenko. Supporting natural language processing with background knowledge: coreference resolution case. In *Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I*, ISWC'10, pages 80–95, Berlin, Heidelberg, 2010. Springer-Verlag.
4. D. Damljanovic and K. Bontcheva. Named entity disambiguation using linked data. In *9th Extended Semantic Web Conference (ESWC2012)*, May 2012.
5. J. Dewey. Context and thought. *University of California publications in philosophy*, 12, 1931.
6. P. Dourish. What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8(1):19–30, Feb. 2004.
7. D. Lenat. The dimensions of context-space. Technical report, Cycorp, 1998.
8. N. Ludwig and H. Sack. Named entity recognition for user-generated tags. In *Proceedings of the 2011 22nd International Workshop on Database and Expert Systems Applications*, DEXA '11, pages 177–181, Washington, DC, USA, 2011. IEEE Computer Society.
9. C. D. Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, CICLing'11, pages 171–189, Berlin, Heidelberg, 2011. Springer-Verlag.
10. P. Mehra. Context-aware computing: Beyond search and location-based services. *IEEE Internet Computing*, 16:12–16, 2012.
11. P. N. Mendes, M. Jacob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proc. of 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria*, 2011.
12. R. Mihalcea and A. Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.
13. C. K. Mostefa Djamel, Hamon Olivier. Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR : Results from the first evaluation campaign. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC06)*, Genoa, Italy, May 2006. ELRA.
14. P. Sen. Collective context-aware topic models for entity disambiguation. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 729–738, New York, NY, USA, 2012. ACM.
15. H. D. Wactlar, A. G. Hauptmann, M. G. Christel, R. A. Houghton, and A. M. Olligschlaeger. Complementary video and audio analysis for broadcast news archives. *Commun. ACM*, 43(2):42–47, Feb. 2000.