Towards a Media Interpretation Framework for the Semantic Web

S. Espinosa Peraldi, A. Kaya, S. Melzer, R. Möller, M. Wessel

Hamburg University of Technology, Germany

Abstract

We present a framework for media interpretation that leverages low-level information extraction to a higher level of abstraction in order to support semantics-based information retrieval for the Semantic Web. The overall goal of the framework is to provide high-level content descriptions of documents for maximizing precision and recall of semantics-based information retrieval.[1]

1 Introduction

The Semantic Web is often envisioned as the place where 'intelligent' agents exploit the semantics of data to offer more valuable services than available today. However, for the majority of data no content description (semantic annotation) exists at present and w.r.t. vast amounts of data, it is not realistic to expect a manual annotation. Low-level information extraction (IE) has been investigated as a solution for automating the semantic annotation process. In this approach modality specific low-level features are used to extract objects from documents. For example, colors, textures and shapes are used to identify regions of images showing human faces, tokens and strings are used to extract words in texts representing person names. However, more abstract objects, e.g., an athletics awards ceremony, cannot be extracted with the same precision using low-level features only. More abstract objects can be considered as high-level descriptions of content that are composed of objects from lower level representations. Media interpretation can therefore be defined as a process that computes high-level content descriptions from lower level representations.

In this paper we present a framework for media interpretation that leverages low-level information extraction to a higher level of abstraction and, therefore, enables the automatic annotation of documents through high-level content descriptions. The availability of high-level content descriptions for documents will enable information retrieval using more abstract terms, which is crucial for providing more valuable services in the Semantic Web. The overall goal of the framework is to maximize precision and recall of semantics-based information retrieval [7].

The paper is structured as follows. Section 2 introduces the preliminaries of abduction, which is the key inference service for media interpretation and presents its formalization in the context of Description Logics (DLs). Section 3 presents the details of the media interpretation framework using an image with captioned text as example. Section 4 provides an empirical evaluation of the results of the framework on a collection of athletics images. Finally, Section 5 presents related work and concludes this work.

2 Abduction for Media Interpretation

Abduction is usually described as a form of reasoning from effects to causes. Another

[1]This work was partially supported by the EU-funded projects BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction, IST-FP6-027538) and TONES (Thinking ONtologiES, FET-FP6-7603).

Dagstuhl Seminar Proceedings 08091
Logic and Probability for Scene Interpretation
http://drops.dagstuhl.de/opus/volltexte/2008/1619

widely accepted definition of abduction considers it as inference from observations to explanations. In this view, abduction aims to find explanations for observations. In general, abduction is formalized as follows: $\Sigma \cup \Delta \models \Gamma$ where background knowledge (Σ), and observations (Γ) are given and explanations (Δ) are to be computed.

If DLs are used as the underlying knowledge representation formalism [1], Σ is a knowledge base (KB): $\Sigma = (\mathcal{T}, \mathcal{A})$ that consists of a Tbox $\mathcal{T}$ and an Abox $\mathcal{A}$. Δ and Γ are Aboxes and they contain sets of concept instance and role assertions.

We consider Abox abduction in DLs as the key inference service for media interpretation. We assume $\mathcal{A}$ to be empty and modify the previous equation to $\Sigma \cup \Gamma_1 \cup \Delta \models \Gamma_2$, by splitting the assertions in Γ into two parts: bona fide assertions (Γ_1) and assertions requiring fiats (Γ_2). Bona fide assertions are assumed to be true by default, whereas fiat assertions are aimed to be explained.

In order to compute explanations, Abox abduction can be implemented as a non-standard retrieval inference service in DLs. Different from the standard retrieval inference services, answers to a given query cannot be found by simply exploiting the knowledge base. In fact, the abductive retrieval inference service has the task of acquiring what should be added to the knowledge base in order to positively answer a query.

To answer a given query, the abductive retrieval inference service can exploit non-recursive DL-safe rules with autoepistemic semantics in a backward-chaining way. In this approach, rules are part of the knowledge base and are used to extend the expressivity of DLs. In order to extend expressivity and preserve decidability at the same time, the safety restriction is introduced for rules. Rules are DL-safe if they are only applied to Abox individuals, i.e., individuals explicitly named in the Abox [9]. In [3] we presented a detailed discussion of the abductive retrieval inference service in DLs.

The output of the abductive retrieval inference service should be a set of explanations Δ that are consistent w.r.t. Σ and Γ. This set, which is called Δs, is transformed into a poset according to a preference score. We propose the following formula to compute the preference score of each explanation: $S(\Delta) := S_i(\Delta) - S_h(\Delta)$ where S_i and S_h are defined as follows:

$$S_i(\Delta) := |\{i | i \in inds(\Delta) \text{ and } i \in inds(\Gamma_1)\}|$$
$$S_h(\Delta) := |\{i | i \in inds(\Delta) \text{ and } i \in newInds\}|$$

The set *newInds* contains all individuals that are hypothesized during the generation of an explanation (new individuals). The function *inds* returns the set of all individuals found in a given Abox or a set. The preference score reflects the two criteria proposed by Thagard for selecting explanations [11], namely simplicity and consilience. In fact, the less hypothesized individuals an explanation contains (simplicity) and the more observations an explanation involves (consilience), the higher its preference score gets.

3 The Media Interpretation Framework

The media interpretation framework aims to compute high-level content descriptions of media documents from lower level information extraction results.

For this purpose, it exploits conceptual and contextual knowledge (see Figure 1). Here, the contextual knowledge refers to specific prior knowledge relevant for the high-level interpretation, which we will discuss later. The conceptual knowledge is represented in a formal ontology that consists of a TBox and a set of non-recursive DL-safe rules about the domain of interest. The formal representation of the conceptual knowledge enables the framework to compute interpretations using various reasoning services such as the abductive retrieval inference service presented in Section 2.

The high-level interpretation of a media document requires an Abox as input (analysis Abox), which contains the results of the low-level semantics extraction. It produces another Abox as output (interpretation Abox), which contains high-level content descriptions. The analysis Abox corresponds to Γ in the abduction formula (see Section 2). The interpretation Abox is computed in a cyclic process, and at the end of this process it contains

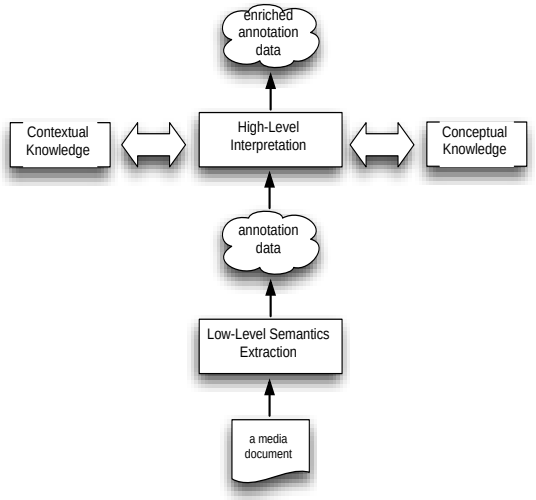

Figure 1. Architecture of the media interpretation framework

all possible interpretations of the media document. Each cycle of the interpretation process consists of the following steps:

First, Γ is split into bona fide and fiat assertions. Currently, all role assertions in the analysis Abox are selected as fiat assertions (Γ_2), and all other assertions as bona fide ones (Γ_1). Second, each assertion from Γ_2 is transformed into a corresponding query to exploit the abductive retrieval inference service. Consequently, the abductive retrieval inference service returns all possible consistent explanations. Third, for each explanation it is checked whether new information can be inferred through deduction.

The interpretation process selects new assertions as fiat assertions from each generated explanation, and repeats these steps until no new explanation can be generated.

Additionally, contextual knowledge can be used to enhance the results obtained by the interpretation process: A set of aggregate concepts can be defined as target concepts. Target concepts serve as an additional termination criteria to omit the computation of interpretations which are useless in practice. Consequently, the framework terminates the cyclic interpretation process, once a generated explanation contains an instance of the target concepts.

In the future the contextual knowledge can be extended. E.g., more appropriate (probably domain-specific) strategies for identifying fiat assertions can be developed and integrated into the framework.

After the presentation of the media interpretation framework, we discuss the details of the underlying interpretation process using an image and an athletics ontology. The athletics ontology that serves as the background knowledge Σ consists of a Tbox and a set of non-recursive DL-safe rules. Some axioms of the Tbox, which are relevant for our example are shown below:

$$
\begin{array}{lcl}
Person & \sqsubseteq & \exists hasPart.PersonFace \sqcap \\
 & & \exists hasPart.PersonBody \sqcap \\
 & & \exists hasName.Name \sqcap \\
 & & \exists hasGender.Gender \sqcap \\
 & & \neg PersonFace \sqcap \ldots \\
Jumper & \sqsubseteq & Person \\
SportsTrial & \sqsubseteq & \exists hasPerformance. \\
 & & Performance \sqcap \\
 & & \exists hasRanking.Ranking \sqcap \\
 & & \exists hasParticipant.Person \\
 & & \neg Person \sqcap \ldots \\
JumpingEvent & \sqsubseteq & SportsTrial \sqcap \\
 & & \exists_{\leq 1} hasParticipant.Jumper \\
PoleVault & \sqsubseteq & JumpingEvent \sqcap \\
 & & \exists hasPart.Pole \sqcap \\
 & & \exists hasPart.Bar \\
HighJump & \sqsubseteq & JumpingEvent \sqcap \\
 & & \exists hasPart.Bar
\end{array}
$$

In this Tbox, some concepts such as *Person* are more abstract than others, and are designed as aggregates, which consist of parts such as *Name* and *PersonFace*. Note that some of the aggregate's parts can be extracted through text analysis (e.g., *Name*, *Gender*, *Performance*, etc.), while others through image analysis (e.g., *PersonFace*, *Bar*, *Pole* etc.). Furthermore, the Tbox contains several disjointness axioms between concepts, which are not shown here completely for brevity. The disjointness axioms are necessary to avoid 'awkward' explanations, which would otherwise be generated.

Additionally, the background knowledge contains a set of non-recursive DL-safe rules that are used to model several characteristic constellations (relations) of objects in the athletics domain as follows:

$adjacent(Y, Z) \leftarrow Person(X), hasPart(X, Y), PersonFace(Y), hasPart(X, Z), PersonBody(Z)$

$adjacent(Y, Z) \leftarrow PoleVault(X), hasPart(X, Y), Bar(Y), hasPart(X, W), Pole(W), hasParticipant(X, Z), Jumper(Z)$

$adjacent(Y, Z) \leftarrow HighJump(X), hasPart(X, Y), Bar(Y), hasParticipant(X, Z), Jumper(Z)$

$adjacent(X, Z) \leftarrow hasPart(X, Y), adjacent(Y, Z)$

$s2o(Y, Z) \leftarrow Person(W), hasName(W, Y), Name(Y), SportsTrial(X), hasPerformance(X, Z), Performance(Z), hasParticipant(X, W)$

Some of these rules such as the *adjacent* rules can be extracted from images, whereas others are derived from text (such as the subject-to-object (*s2o*) rule).

To better illustrate the interpretation process and the use of the background knowledge, we continue with the stepwise interpretation of an athletics image. The image below shows a pole vault trial, and is captioned with a text:

Yelena Isinbayeva goes over 5.01m but knocks off the bar on her descent (Hasse Sjögren)

Assume that for this image low-level image analysis delivers an analysis Abox with the following concept instance and role assertions:

$\Gamma = \{pface_1 : PersonFace, pole_1 : Pole, bar_1 : Bar, pbody_1 : PersonBody, (pface_1,\ pbody_1) : adjacent, (pbody_1,\ bar_1) : adjacent\}$

To begin with the interpretation, all role assertions are selected as fiat assertions and, therefore, Γ_2 becomes:

$\Gamma_2 = \{(pbody_1,\ bar_1) : adjacent, (pface_1,\ pbody_1) : adjacent\}$

In the second step, the role assertions are transformed into corresponding queries and the abductive retrieval inference service is asked for explanations. Only the query derived from the role assertion $(pface_1, pbody_1) : adjacent$ results in the generation of an explanation. It explains the adjacency of the face and the body by hypothesizing a person instance to whom they both belong to (see the first *adjacent* rule). Note that other *adjacent* rules are considered as well, however they cause the generation of explanations that are inconsistent (due to the disjointness axioms in the Tbox). The interpretation process discards such explanations. Assume that the newly inferred person instance is named new_ind_1. In the third step, the interpretation process applies the rules forwards to check whether new information can be deduced. This yields the following assertions: $(bar_1,\ new\_ind_1) : adjacent$, $(pbody_1,\ new\_ind_1) : adjacent$.[2] At this state, the interpretation process defines a new Γ_2 by selecting all newly inferred role assertions as fiat assertions and repeats the whole cycle. Here, only the query derived from the role $(bar_1,\ new\_ind_1) : adjacent$ results in the generation of explanations:

- $\Delta_1 = \{new\_ind_2 : PoleVault,$
  $(new\_ind_2,\ bar_1) : hasPart,$
  $(new\_ind_2,\ pole_1) : hasPart,$
  $(new\_ind_2,\ new\_ind_1) : hasParticipant,$
  $new\_ind_1 : Jumper\}$
- $\Delta_2 = \{new\_ind_3 : HighJump,$
  $(new\_ind_3,\ bar_1) : hasPart,$
  $(new\_ind_3,\ new\_ind_1) : hasParticipant,$
  $new\_ind_1 : Jumper\}$

At this point, no further explanations can be generated and the interpretation process terminates. Observe that both explanations are consistent and represent possible interpretations of the image. However, in practice one would like to get 'preferred' explanation(s) only. For this purpose, the preference score presented in Section 2 can be used. The preference score of Δ_1 is calculated as follows: Δ_1 incorporates the individuals bar_1, $pole_1$ and new_ind_1, and therefore $S_i(\Delta_1)$=3. Furthermore, it hypothesizes only one new individual, namely new_ind_2, such that $S_h(\Delta_1)$=1. The preference score of Δ_1 is therefore $S(\Delta_1)$

[2] See the fourth *adjacent* rule

= $S_i(\Delta_1)$ - $S_h(\Delta_1)$ = 2. Analogously, the preference score of the second explanation is $S(\Delta_2)$=1. Consequently, Δ_1 becomes the 'preferred' explanation for the image. In fact, the result is plausible, since this image should better be interpreted as showing a pole vault and not a high jump, due to the fact that image analysis could detect a pole, which should not be ignored as in the high jump explanation (consilience).

We continue with the interpretation of the captioned text to show the results of the interpretation process in the text modality. Assume that for the sentence *'Yelena Isinbayeva goes over 5.01m but knocks off the bar on her descent'*, text analysis delivers an analysis Abox with the following assertions:

$\Gamma = \{name_1 : Name, performance_1 : Performance,$
$(name_1, \text{'Yelena Isinbayeva'}) : hasValue,$
$(performance_1, \text{'5.01'}) : hasValue,$
$(name_1,\ performance_1) : s2o\}$

Given this analysis Abox, the role assertion $(name_1, performance_1) : s2o$ is the only fiat assertion that is exploited for generating explanations. It generates a single explanation:

$\Delta_1 = \{new\_ind_1 : Person, new\_ind_2 : SportsTrial,$
$(new\_ind_2,\ performance_1) : hasPerformance,$
$(new\_ind_2,\ new\_ind_1) : hasParticipant$
$(new\_ind_1\ , name_1) : hasName\}$

which infers a person with the given name, who participates in a sports trial with the given performance.

Due to space limitations, we discuss only a simplified example for text interpretation here. In fact, similar to the *adjacent* rules several *s2o* rules exist in the ontology. However, due to the disjointness axioms in the Tbox, they do not result in the generation of further explanations.

4 Evaluation

The overall goal of the framework is to provide high-level content descriptions of media documents for maximizing precision and recall of semantics-based information retrieval. In this section, we provide an empirical evaluation of the results of the framework on a collection of athletics images in order to analyze the utility of the framework.

For this purpose, we implemented the media interpretation framework shown in Figure 1. The core component of this implementation is the DL-reasoner RacerPro [4] that supports various inference services. The abductive retrieval inference service, which is the key inference service for media interpretation, is integrated into the latest version of RacerPro. The framework gets analysis Aboxes, exploits various inference services of RacerPro, and returns interpretation Aboxes as high-level content descriptions. For the time being, the computation of preference scores is not implemented and, therefore, interpretation Aboxes contain all possible explanations.

To test the implementation, we used an ontology about the athletics domain and an image corpus. The corpus consists of images showing either a pole vault or a high jump event. The images have been manually annotated with annotation tools in order to train low-level feature extractors for prospective athletics corpora. I.e., using the annotation tools, annotators manually annotated regions of images (as visual representations of concepts), with corresponding concepts from the ontology such as *Pole*, *Bar* and *PersonFace*. Afterwards, annotated images have been analyzed automatically to detect relations between concept instances. Finally, for each image in the corpus an analysis Abox with corresponding assertions has been generated.

We tested the implementation in the following setup: the aggregate concepts *PoleVault* and *HighJump* from the domain ontology are defined as target concepts. Analysis Aboxes of pole vault and high jump images are used as input for high-level media interpretation. The results obtained for pole vault and high jump images are shown in Figure 2 and 3, respectively. To analyze the usefulness of the results for information retrieval, in both figures interpretation Aboxes are categorized w.r.t. the existence (or absence) of aggregate concept instances: *A*) contains no aggregate concept instances at all *B*) contains an aggregate concept instance but no target concept instance *C*) contains a *HighJump* and a *PoleVault* instance *D*) contains a *PoleVault* instance *E*)

contains more than one *PoleVault* instances and one or no *HighJump* instances

At first sight, only interpretation Aboxes that fall into the category D in Figure 2 look like 'good' interpretation results for pole vault images, because the corresponding images are annotated with a single *PoleVault* instance. However, if the implementation would be enhanced to include preference scores, as discussed in Section 3 for an example pole vault image, all interpretation Aboxes of category C and E would include the most 'preferred' explanation only (in this case a single *PoleVault* instance), and hence fall into the category D, too.

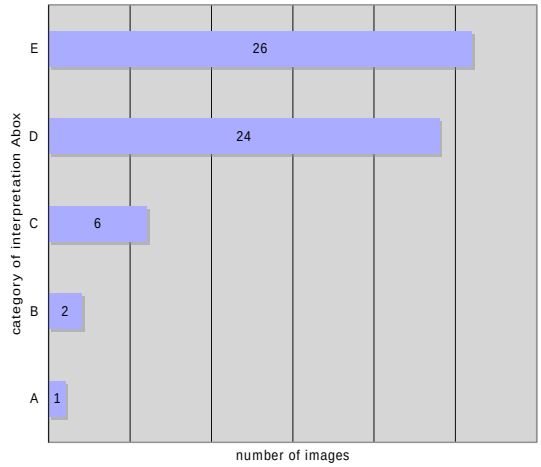

Figure 2. Results for pole vault images.

Both in Figure 2 and 3, category A interpretation Aboxes are identical to the corresponding analysis Aboxes and indicate that no new knowledge could be inferred through high-level interpretation. For other images (category B interpretation Aboxes) high-level interpretation infers new knowledge (including an aggregate concept instance) but fails to derive an instance of the target concepts.

In fact, category B interpretation Aboxes contain a *Person* instance to explain the existence of *PersonBody* and *PersonFace* instances and their constellation in the image. Deeper analysis of category A and B interpretation Aboxes showed that insufficient interpretation results are caused by the failure of image analysis to extract some of the existing relations in the corresponding images. Taking into account the ambiguity and uncertainty involved in the image analysis process, this information (the failure of adequate interpretation) can be used to create a valuable feedback for the image analysis tools.

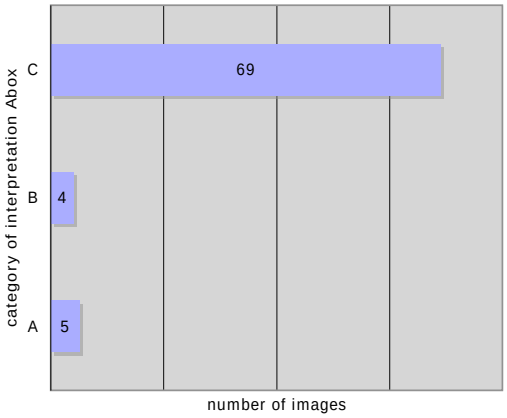

Figure 3. Results for high jump images.

Figure 3 shows that every high jump image is interpreted as either showing a high jump or a pole vault event (category C), besides incompletely analyzed ones, which fall into the categories A or B. Different from pole vault images, interpretations of high jump images cannot be disambiguated through preference scores. This result indicates that necessary rules are missing in the background knowledge due to the fact that, currently, image analysis cannot extract distinctive features of high jump images. Note that appropriate fusion of information from different modalities can help to solve this problem, even if image analysis results cannot be improved. For example, if an image is captioned with text, text analysis can extract additional information that can be used to disambiguate the interpretation of the image.

Our experiments showed that, if provided with an appropriate ontology and low-level annotations, the existing implementation of the media interpretation framework delivers promising results for images and can be used for maximizing precision and recall of semantics-based information retrieval systems.

5 Related Work and Conclusion

The idea of formalizing interpretation as abduction is investigated in [5] in the context of text interpretation. In [10], Shanahan presents a formal theory of robot perception as a form of abduction, where low-level sensor data is transformed into a symbolic representation of the world in first-order logic and abduction is used to derive explanations. In [2] a detailed discussion of abductive reasoning tasks in DLs including Abox abduction is presented. The authors consider the development of algorithmic techniques based on semantic tableaux for employing abductive inference in expressive DLs as the most promising approach. However, a solution to the Abox abduction problem is formally presented, but for the time being it is not shown how to derive solutions.

The abduction approach we follow in this work is based on the combination of the works in [5], [10] and [8]. In contrast to approaches such as [6], which use abduction in the context of rules in logic programming only, we combine existing DL reasoning mechanisms and rules in a coherent framework and consider abduction as a new type of non-standard retrieval inference service, which is integrated into existing DL reasoners.

In this paper we presented a media interpretation framework that leverages low-level information extraction to a higher level of abstraction and, therefore, enables the automatic annotation of documents through high-level content descriptions. The availability of high-level content descriptions for media documents will enable semantics-based information retrieval using more abstract terms, which is essential for the Semantic Web. The key inference service used by this framework is the abductive retrieval inference service that generates explanations for observations.

The empirical evaluation presented in this work indicates that a coherent framework incorporating appropriate ontology design, dedicated low-lewel IE and reasoning in DLs delivers promising results for media interpretation. Further analysis of test results showed that the implementation of the proposed preference score will enhance the results of media interpretation and, therefore, contribute to the maximization of precision and recall of semantics-based information retrieval.

Currently, we are investigating fusion of information from different modalities to enhance media interpretation results. In our future work, we will investigate inductive learning of DL-safe rules for abduction using training data.

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[2] C. Elsenbroich, O. Kutz, and U. Sattler. A Case for Abductive Reasoning over Ontologies. In *Proc. OWL-2006: OWL Experiences and Directions Workshop*, 2006.

[3] S. Espinosa, A. Kaya, S. Melzer, R. Möller, and M. Wessel. Multimedia Interpretation as Abduction. In *Proc. DL-2007: International Workshop on Description Logics*, 2007.

[4] V. Haarslev, R. Möller, and M. Wessel. RacerPro User's Guide and Reference Manual Version 1.9.1, May 2007.

[5] J. R. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. *Artificial Intelligence Journal*, Vol. 63, 1993.

[6] A. Kakas and M. Denecker. Abduction in logic programming. In *Computational Logic: Logic Programming and Beyond. Part I*. Springer, 2002.

[7] R. Möller, V. Haarslev, and B. Neumann. Semantics-based information retrieval. In *Proc. IT&KNOWS-98: International Conference on Information Technology and Knowledge Systems*, 1998.

[8] B. Neumann and R. Möller. On Scene Interpretation with Description Logics. In *Cognitive Vision Systems: Sampling the Spectrum of Approaches*. Springer, 2006.

[9] B. Neumann and R. Möller. Ontology-based reasoning techniques for multimedia interpretation and retrieval. In *Semantic Multimedia and Ontologies : Theory and Applications*. 2007. To appear.

[10] M. Shanahan. Perception as Abduction: Turning Sensor Data Into Meaningful Representation. *Cognitive Science Journal*, 2005.

[11] R. P. Thagard. The best explanation: Criteria for theory choice. *The Journal of Philosophy*, 1978.