Semantic Analysis of Text Regions Surrounding Images in Web Documents

Thierry Declerck, Manuel Alcantara

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg,3
D-66123 Saarbrücken, Germany
declerck@dfki.de

Abstract

In this paper we present some on-going work and ideas on how to relate text-based semantics to images in web documents. We suggest the use of different levels of Natural Language Processing (NLP) to textual documents and speech transcripts associated to images for providing structured linguistic information that can be merged with available domain knowledge in order to generate additional semantic metadata for the images. An issue to be specifically addressed in the next future concerns the automation of the detection of relevant text/speech transcripts for a certain image (or video sequence). Beyond the time code approach, with its shortcomings, we expect from the discussion in this workshop on lexical characteristics of the language that can or should be used to describe image content an improvement of the approaches we are dealing with for the time being.

1. Introduction

We started our work within a past European project, Esperonto. The Esperonto project was dealing with annotation services for bridging the gap between the actual (html based) Web and the emerging Semantic Web. A smaller task of the project was dedicated to the investigation on how to automatically provide for semantic annotation for images present in a web page. A possible strategy we investigated was to provide for ontology-driven semantic annotation of the text surrounding an image in a web page.

This work is being continued and extended within a recently started Network of Excellence, called K-Space (Knowledge Space of Semantic Inference for Automatic Annotation and Retrieval of Multimedia Content, http://kspace.qmul.net/), in which some Labs are specifically dedicated to the contribution of Human Language Technology (HLT) for the semantic indexing (and possibly retrieval) of multimedia content. K-Space, which will be described in more details in this paper, is offering a more integrated approach for multimedia semantics, aiming at a formal integration of low-level features extracted from multimedia material on the base of state-of-the-art audio-video analysis, and high-level features resulting from text analysis coupled with semantic web technologies.

2. Background: Multimedia Semantics

The topic of Multimedia Semantics has gained a lot of interest in recent years, and large funding agencies issued calls for R&D proposals on those topics. So for example a recent call of the European Commission, the 4th call of the 6th Framework, was dedicated to the merging of results gained from R&D projects on knowledge representation and cross-media content. The goal being in making the (semantic) descriptions of multimedia content re-usable on the base of a higher interoperability of media resources, which has been so far described mainly at the level of XML syntax, as can be seen with the MPEG-7 standard for encoding and describing multimedia content.

In the line of the recent developments in the fields of Semantic Web technologies, one approach consists in looking at ways for encoding so-called low-level features, as they can be extracted from audio-video material, into a high-level features organization as one can typically find in a (domain) ontology.

The EU co-funded project aceMedia is offering a very good example of such an approach. In this project, ontologies, which are typically describing knowledge as expressed in words, are extended in order to include the low-level visual features resulting from state-of-the-art audio-video analysis systems. For the description of low-level features, the project uses as its background the MPEG-7 standard, and proposes links from the MPEG-7 descriptors to high-level (domain) ontologies (see Athanasiadis, 2005). So in a sense no full integration is proposed here, but a linkage between the MPEG-7 description scheme and ontologies represented in a Semantic Web language, and interoperability of descriptions of audio-video material is indirectly realized.

Another closely related approach (see the papers by Jane Hunter) is proposing a reformulation of the semantic metadata of MPEG-7 descriptors in machine-understandable language (MPEG-7 Description Schema being only a machine-*readable* language) and use RDFs or OWL. This step is ensuring a better interoperability of semantic multimedia descriptions. But here the cross-media aspect is missing, since no textual analysis and/or speech transcripts are taken into account.

A new iniative, the K-Space European Network of Excellence, has started recently. This project is dealing with semantic inferences for semi-automatic annotation and retrieval of multimedia content. The aim is to narrow the gap between content descriptors that can be computed automatically by current machines and algorithms, and the richness and subjectivity of semantics in high-level human interpretations of audiovisual media: the so-called *Semantic Gap*.

The project deals with a real integration of knowledge structures in ontologies and low-level descriptors for audio-video content, taking also into account knowledge that can be extracted from sources that are complementary to the audio/video stream, mainly speech transcripts and text surrounding images or textual metadata describing a video or images. The integration takes place at 2 levels: the level of knowledge representation, where features associated with various modalities (image, text/speech transcripts, audio) should be interrelated within conceptual

classes in ontololgies (from domain-specific to general purpose ontologies), and the level of processing, where high-level semantic features should be integrating for guiding (and so possibly improving) the automatic analysis of audio-video material and the corresponding extraction of semantic features.

As such the K-Space activities are mostly dedicated to the analysis of multimedia and cross-media data and the feature extraction out of such data. Navigation, search and retrieval in the field of semantic cross-media archives are not primarily addressed.

An interesting project with respect to K-Space is MESH, which seems to build an application scenario on the base of the multimedia and cross-media knowledge structures discussed and proposed by K-Space and aceMedia. The domain of application is given by the News domain. The project will deal with the ontology-driven semantic integration of content features extracted from video, images, speech transcripts and text. Multi- and cross-media reasoning is an important issue here, insuring consistency and non-redundancy of the integrated cross-media features. A major issue will consist in proposing an appropriate syndication of the semantically encoded material for distribution to distinct (mobile) end-user hardware, also under consideration of personalization aspects. Supporting thus the distribution of relevant multi- and cross-media content.

The 2006 edition of TRECVid is offering an interesting development, since one of its tasks is addressing searching within a multimedia database, whereas interaction with the user is also foreseen. We can expect here that the user will input his/her queries in natural language, whereas the use of certain lexical items should guide he intelligent search in large archives containing cross-media material.

3. An integrative approach in the K-Space Network of Excellence

The projects mentioned above (and some others, not listed here for reason of place), are given us important information about methodologies and technologies for the "ontologization" of low-level audio-video features extracted from multimedia content. Here we describe in some more details the K-Space project and the activities related to the use and analysis of sources complementary to audio-video material. First we describe the foreseen ontology infrastructure, which will give the base for the integration of low-, mid- and high-level features extracted from audio-video and associated text/speech transcripts.

3.1. Development of a multimedia ontology infrastructure

The multimedia ontology infrastructure of K-Space will contain qualitative attributes of the semantic objects that can be detected in the multimedia material, e.g. color homogeneity, in the multimedia processing methods, e.g. color clustering, and in the numerical data or low-level features, e.g. color models. The ontology infrastructure will also contain the representation of the top-level structure of multimedia documents in order to facilitate a full-scale annotation of multimedia documents. R&D work will be dedicated to the specification and development of a multimedia content ontology supporting the representation of the structure of the content of multimedia documents. Work will also be dedicated to research on ontologies for low-level visual features, concentrating on a model for the concepts and properties that describe visual features of objects, especially the visualizations of still images and videos in terms of low-level features and media structure descriptions. Also, a prototype knowledge base will be designed to enable automatic object recognition in images and video sequences. Prototype instances will be assigned to classes and properties of the domain specific ontologies, containing low level features required for object identification.

Partners of K-Space dealing with textual analysis will integrate into this ontology infrastructure the typical features for text analysis, also proposing ontology classes at a higher-level, that supports the modeling of interrelated cross-media features (multimedia and text). We will base our work on the proposal made by (Buitelaar et. al 2005).

3.2. Use of Textual Information and Knowledge Bases for Semantic Feature Extraction from Audio Signal

In K-Space some work will be dedicated to the extension of state-of-the-art processing and analysis algorithms to handle high-level, conceptual representations of knowledge embedded in audio content based on reference ontologies and semantically annotated associated text (including speech transcripts, when the quality of the transcripts allows it).

K-Space will consider all types of audio sources ranging from speech to complex polyphonic music signals. The description schemes of the MPEG-7 standard define how audio signals can be described at different abstraction levels: from the lowest level primitives, such as temporal or audio spectrum centroids, spectrum flatness, spectrum spread, inharmonicity, etc., to the highest level, related to semantic information. Semantic information is related to textual information on audio such as titles of songs, singers' names, composers' names, duration of music excerpt, etc.

This textual information is often encoded using the text annotation tool of the Linguistic Description Scheme (LDS) of MPEG-7. An example of such a (manual) annotation related to a video sequence is given just below:

```
<VideoSegment id="shot1_13">
    <MediaTime>
        <MediaTimePoint>T00:01:40:11008F30000</MediaTimePoint>
        <MediaDuration>PT10S26326N30000F</MediaDuration>
    </MediaTime>
    <TextAnnotation confidence="0.500000">
        <FreeTextAnnotation>
    TRACKS STOPPED ROLLING NOSE AND FORMALLY FILED A HIGHWAY WITH EIGHT DAILY NEW YORK NEWSPAPERS WHERE THE VOID OF NEWSPAPERS THE VOID OF CUSTOMERS
        </FreeTextAnnotation>
    </TextAnnotation>
</VideoSegment>
```

Interesting to note here, is that the media time is also given, so that this can be used as a way to look for alignment of the low-level features and the high-level features that can be extracted from the text.

Our work will consist here in proposing a linguistic and semantic analysis of all the available free text annotations used in the semantic representantion of audio signal, and mapping this onto either the structured annotation scheme of LDS (specifying the "who", the "what", the "why", the "when" etc in an explicit way), or to provide for an ontology based semantic annotation (in term of instances of ontology classes).

We will also use TRECVid data, using aligned speech transcripts and video shots, and looks for ways to extracts high-level semantics from the transcripts (which are attached to the audio-video stream using also the LD scheme). For sure the quality of transripts is often bad, and here we will use robust NLP methods and limits ourself to the detection of basic textual chunks.

For improving the alignment of text/transcripts with the audio (or video) signal, we try to identify typical lexical items that link directly such text/transcripts to the signal ("here you can see" etc.).

3.3. Analysis of Complementary Textual Sources for adding Semantic Metadata to Multimedia Content

The human understanding of multimedia resources is often facilitated by usage of complementary sources. In order to simulate this attitude, K-Space will implement mining methods and tools for such complementary resources in order to reduce the semantic gap by deriving annotations from those sources, and so to reach a more complete annotation of (sequences of) images.

The project will address mining and analysis for semantic features extraction within two different types of resources:

- Mining and analysing primary resources: Analysis of the primary resources that are attached to the multimedia data, e.g. texts around pictures, subtitles of movies, etc.
- Mining and analysis of secondary and tertiary resources: Analysis of data and text related to the multimedia data under consideration, e.g. a programme guide for a TV broadcaster or a web site displaying similar pictures.

4. Linguistic Analysis of relevant Text Regions

We report on a first experiment made within the Esperonto project, where also a small ontology on artworks has been made availble to the project parners. In this ontology, typical terms were associated to every class (so for example the terms "surrealism" and "cubism" are associated to the class "artistic_movement".

In the Esperonto scenario, we first defined the possibly relevant text regions for the semantic annotation of the image (see below in Figure 1 the example of such an image, in a web page dedicated to the painter Miro, the first image being the base for our indexing prototype tool). We identified following text regions (in both the text and in the html code):

- Title of the document
- Caption text: „Click on the image to enlarge" (a non relevant item, to be filtered by the tools, also on the base of lexical properties of the words).
- Content of the HTML „Alt" tag: "'VEGETABLE GARDEN WITH DONKEY'"
- Content of the HTML „Src" tag: *http://www.spanisharts.com/reinasofia/miro/burro_lt.jpg*
- Abstract text
- Running text

On the base of this, we wrote a tool that supports the manual selection of such textual regions, and send those to a linguistic processing engine. The linguistic processing engine has been augmented with metadata sepcifying the type of text to be processed (we expect for example the Title and the "Alt" text to consist mostly of phrases.)

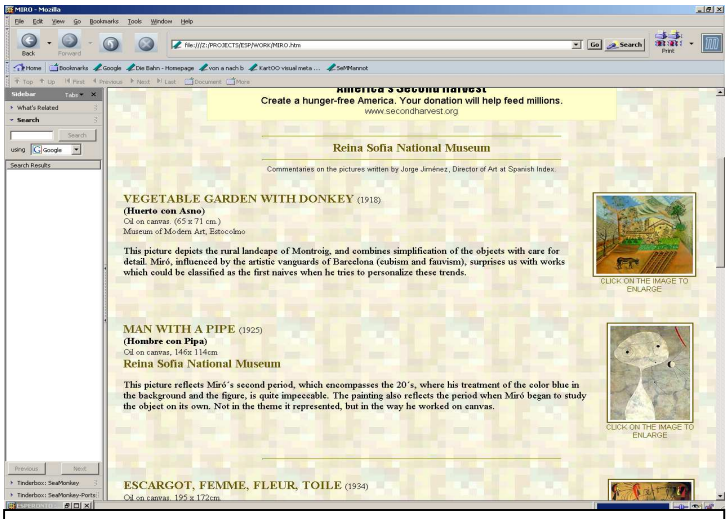

Figure 1 Example of a web page with images of paintings. Various text regions are offering different kind of "metadata" to the

5. The Linguistic Analysis of the Various Text Regions

In the following lines, we show some of the (partial) results of the linguistic analysis, as applied to the various text segments. Our tools are delivering a dependency annotation:

- „Alt" text: 'VEGETABLE GARDEN WITH DONKEY'

<NP HEAD="garden" PRE_MOD="vegetable" <POST_MOD CAT= "PP" HEAD="with" NP_COMP_HEAD="donkey"</POST_MOD></NP>

- Abstract/Running text: "…This picture depicts the rural landcape of Montroig …"

<SENT SUBJ="This picture" PRED="depicts OBJ="the rural lansdscape of Montroig"</SENT>

- Detailled annotation of the direct_object: <NP HEAD="landscape" PRE_MOD="rural" <POST_MOD CAT="PP" HEAD="of" NP_COMP_HEAD="Montroig"</POST_MOD> </NP>

6. The Semantic Annotation

On the base of a mapping between the linguistic dependency and the terms associated to the classes of the ontology (whereas we accomodated the classes of the ontology to be associated with patterns (for coping for example with date expressions), we could provide for a semantic annotation of the texts associated with the picture.

6.1. The (Toy) Art Ontology (schematized)

- Object > Artork > Painting [has_creator, has_name, has_subject, has_dimension, has_material, has_genre, has_date...]
- Person > Artist > Painter [has_name, has_birth_date, part_of_artistic_movement …]

6.2. The Instantiation of Classes

- Title: Vegetable garden with donkey
- Creator: Miro
- Date: 1918
- Genre: naïve (if correctly extracted by some reasoning on the linguistically and semantically annotated text)
- Subject: rural landscape of Montroig + garden and donkey (if the association between the title and the explanation given by the art expert can be grouped).
- Dimension: 65x71
- Material: Oil on canvas

6.3. Some remarks

This result was possible due to various facts. First, the system "knew" that the text was about art, and we assumed that the text is related to the picture. Second, we had an ad-hoc relation of terms to the concepts of the ontology (for example "Oil"). Third we had defined typical patterns realising some concepts (date, material etc.). But our focus was more on syntactic analysis (in fact dependency analysis). So the Subject of the sentence "This picture" together with the typical verb "depicts" and its DirectObject allowed here to "map" the whole DierctObject to the "subject" of the picutre (what the picture is about). The dependency analysis of the DirectObj allows us to further precise the topic of the picture: it is a rural (mod) Lanscape (head) of Montroig (post_nom_mod), thus introducing quite fine granularity in the indexing of the image.

The missing point here: there is no principled relation between the terms in the ontology and the results of the image analysis (in term of low-level features). We think here that a domain ontology taking into account the specific features for the multi-modal analysis components could help in establishing this relationship, not only at lexical level but also maybe at the syntactic level (the dependency relations in linguistic fragments of texts refering to images could give some hints about the distribution of objects in the picture).

But clearly one has to think first of a specific classification of lexical items in terms of possible indices of multimedia content, before looking a syntactic properties of text related to images.

7. Conclusions

We have described some approaches that take advantages of so-called complementary sources (text/transcripts) for automatically adding semantic metadata to image material. Till now we concentrated on the linguistic processing aspect, with a very small lexical base. Lexical consideration would allow to extend our approach and to really evaluate it. More principled lexical information would also support the automatic detection of text parts that are referring directly to the content of the image under consideration, and not to metadata related to this image (in which museum is the picture, wo made it etc.) or on topics not related to the image at all.

We will also have to think at principled ways for integrating the lexical knowledge into the multimedia infrastructure. At the beginning we would follow a similar approach that has been proposed for the integration of lexical information in domain specific ontologies, and proposed in the SmartWeb project.

8. Acknowledgments

"The research programm described in this paper is supported by the European Commission, contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space."

9. References

Buitelaar P., Sintek M., Kiesel M (2005).. Feature Representation for Cross-Lingual, Cross-Media Semantic Web Applications. In: *Proceedings of the ISWC 2005 Workshop "SemAnnot"*..

Athanasiadis T., Tzouvaras V., Petridis K., Precioso F., Avrithis Y. and Kompatsiaris Y. (2005). Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. *In proceedings of the ISWC 2005 Workshop "SemAnnot"*.

Jane Hunter: Enhancing the semantic interoperability of multimedia through a core ontology. IEEE Trans. Circuits Syst. Video Techn. 13(1): 49-58 (2003)

Jane Hunter: Adding Multimedia to the Semantic Web: Building an MPEG-7 ontology. SWWS 2001: 261-283

AceMedia project: http://www.acemedia.org/aceMedia

BUSMAN project: http://busman.elec.qmul.ac.uk/

Esperonto Project: http://www.esperonto.net

K-Space Project: http://kspace.qmul.net

SmartWeb Project: http://www.smartweb-projekt.de

TRECVid: http://www-nlpir.nist.gov/projects/trecvid/