Text-based Semantic Annotation Service for Multimedia Content in the Esperonto project

T. Declerck[1)], J. Contreras[2)], O. Corcho[2)], C. Crispi[1)]

[1)] Universität des Saarlandes, Germany
[2)] Intelligent Software Components S.A. (iSOCO), Spain

ABSTRACT

Within the Esperonto project, an integration of NLP, ontologies and other knowledge bases, is being performed with the goal to implement a semantic annotation service that upgrades the actual Web towards the emerging Semantic Web. Research is being currently conducted on how to apply the Esperonto semantic annotation service to text material associated with still images in web pages. In doing so, the project will allow for semantic querying of still images in the web, but also (automatically) create a large set of text-based semantic annotations of still images, which can be used by the Multimedia community in order to support the task of content indexing of image material, possibly combining the Esperonto type of annotations with the annotations resulting from image analysis.

INTRODUCTION

The Esperonto project[1] aims at upgrading the actual World Wide Web towards the emerging Semantic Web (SW). For supporting this task, a semantic annotation service is being implemented, combining natural language processing and high-level knowledge technologies, including ontologies and intelligent agents.

Since Web documents are also increasingly including image/video and audio material, the issue of the semantic annotation of such material is also being considered in the Esperonto initiative of upgrading the web towards the Semantic Web.

There has been already a considerable amount of investigation and efforts in detecting content in images on the base of image/video analysis techniques (see for example the proceedings of the WIAMIS conference 2003 in London, WIAMIS2004 in Lisbon and CIVR 2004 in Dublin), and many contributions of those conferences are stressing the need to extend the field of image/video analysis to a multimodal one, in order to improve the quality of the content analysis of image/video material, since it seems that image/video analysis is being close in reaching a ceiling in term of accuracy when detecting content on the base of low-level features only.

The research work in Esperonto can potentially contribute to this task in providing for ontology-based semantic information resulting from text analysis applied to textual documents that are part of the image (superposed text or transcripts of speech contained within the video) or adjacent to it, being caption texts or the whole text surrounding an image.

The annotation services developed in Esperonto should also be able in a longer term to (re-) use information as provided by image/video analysis techniques, making the semantic annotation services compliant to the annotation standards developed in the image/video processing area. We think here mainly at the MPEG-7 annotation framework. In the long term a real cross-media semantic annotation service should emerge.

In this paper we describe the way the Esperonto project is proposing for semantic annotation of multimedia material also present on the web, whereas we concentrate on the semantic analysis of text being adjacent to still images, not addressing sound and video.

RELATED WORK

The reported work being just a partial aspect of the Esperonto project, and at a preliminary stage, we have been looking for cooperation and information exchange with other projects and initiatives. In this section we present briefly three projects, which we think are proposing related work and with which clustering activities are on the way or should be started: SCHEMA, aceMedia and DIRECT-INFO ([10,11,12] and all the references listed in those pages).

[1] See www.esperonto.net and [1].

SCHEMA is a Network of Excellence dedicated to the issue of content-based semantic scene analysis and information retrieval.

This network is also implementing a reference platform for content-based analysis, representation indexing and retrieval of multimedia material within the framework of the MPEG-7 standard ([6]). Language comes into play in SCHEMA in form of superposed text or transcripts associated with videos. Taking into account this language data should improve content-based image retrieval.

The Esperonto project is in the meantime an affiliated partner of the SCHEMA network. The main cooperation will consist in the application of linguistic methods for the analysis of the associated language data, whereas an important contribution of Esperonto will consist in also considering language material that is not directly related to the images (e.g. speech transcripts), but also so-called "co-lateral" text, which might be indirectly related to the images to be indexed with content information.

The aceMedia project is relevant for our work in Esperonto, since a part of the project is dedicated in applying ontological framework for adding semantics to the low-level features resulting from image/video analysis ([7]). Since Esperonto is relating textual analysis to ontologies, it seems to be that via ontological descriptions an interlinking between low-level image content features and high-level linguistic content features is being possible.

The DIRECT-INFO project is proposing an integrated system combining the output of basic media analysis modules to semantically meaningful trend analysis results ([8]). DIRECT-INFO is not only going for indexing multimedia with content, but also with qualitative information, like positive or negative mentioning of entities to be seen in the multimedia documents. Clearly media analysis alone cannot reach this goal. For this DIRECT-INFO is using linguistic analysis (the same tools as those in use in Esperonto) applied to related language data (transcripts). On the top of this, a strategy has to be developed in order to detect positive and negative mentioning and so to identify trends as well.

DIRECT-INFO is very relevant to Esperonto and related future research work, since it implements a platform for the fusion of the results of multi-modal analysis, applied to media monitoring. The work achieved in Esperonto could help in providing for an ontological framework for this fusion work. Since the fusion work in doing using MPEG-7 as the representation language, DIRECT-INFO is also quite close to the topics of SCHEMA.

THE ESPERONTO STRATEGY FOR THE ANNOTATION OF TEXTS ASSOCIATED TO AN IMAGE

Esperonto is looking at all kind of textual information related to an image in a web page, and a specialized module of the Esperonto NLP tools has been implemented to process and annotate the different types of textual documents, which are classified among the lines of the HTML encoding of the page. So we distinguish between caption text, "alt" text (the small text displayed when the cursor is moved over an image), the source ("src") of the images (mostly a URL), the title of the document, an abstract and the normal running text.

We are looking at the source information of the image, since sometimes hints about the possible content of the image can be found in the naming of the source. We will also associate the semantic annotation generated by Esperonto to the html source information encoding of the image, providing at the same time for a unique identification of the complex annotation structure.

The Esperonto tool

Figure 1 and figure 2 below show the actual level of development of the Esperonto "multimedia" content indexing demonstrator, which we tentatively call SemmAnnot. The preprocessing module of the tool extracts from the underlying html tree the distinct text parts, like caption, alt-text etc. but allows as well to select parts of the text and to submit all those distinct part to the Esperonto NLP multilingual tools.

We show both a screen shot of the original page and a screen shot of the interface of our tool, when applied to the first image of the web page and the text parts associated to it.

The home page is about painting works by Joan Miro. The images have both captions and "alt" tags associated with them. On the left side of each picture is a comment by an expert in the work of Miro. This makes this page so interesting: the comments by an expert on art can be (automatically) included into the semantic annotation of the picture. We expect to discover similarities between the comments of the ontological resource on art delivered by the Esperonto partner "Residencia de Estudiantes". So here we see a good example where the annotation of still images in web pages will be directly influenced by available ontologies.

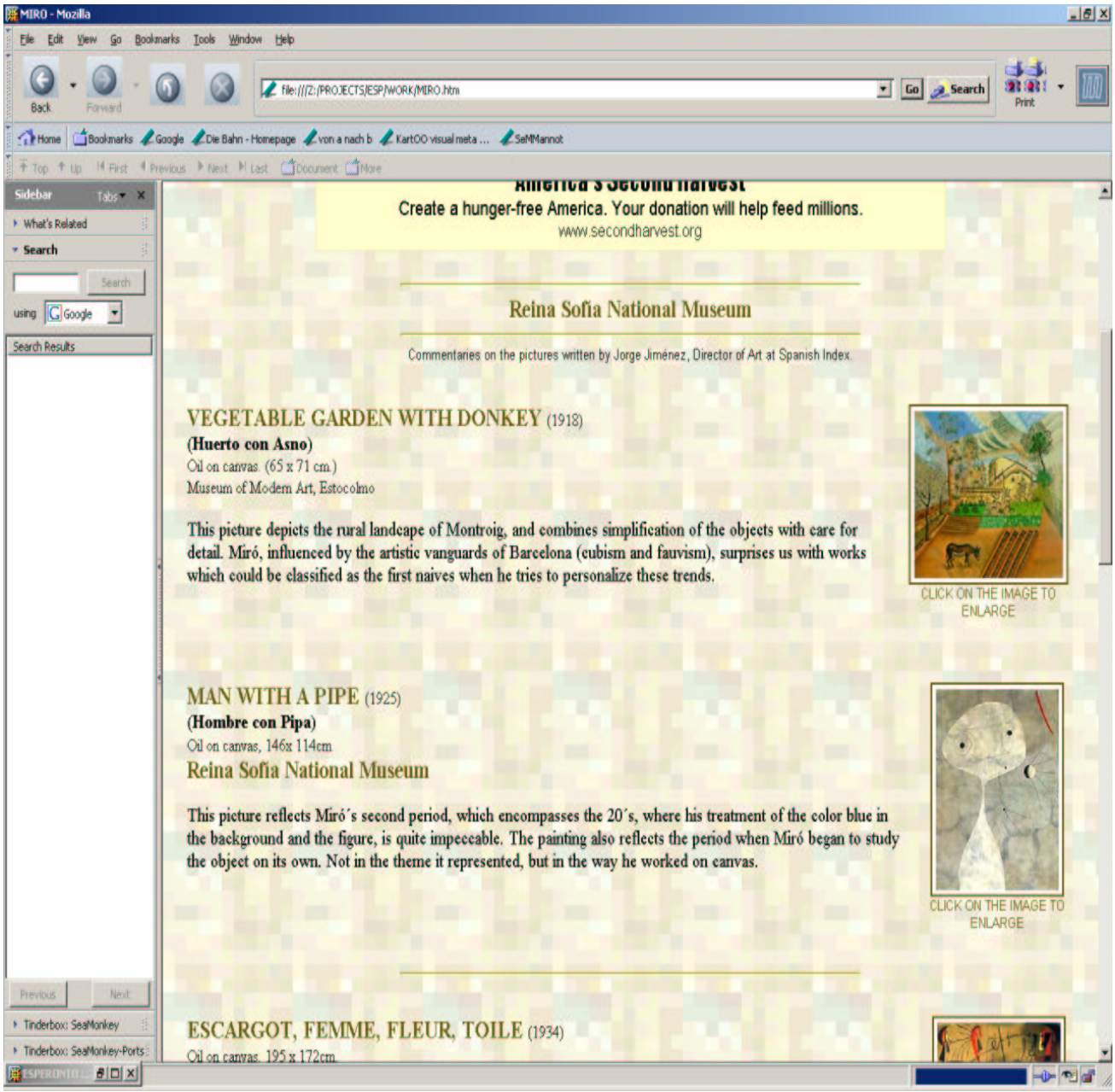

Figure 1: Screen shot of a Web page containing a commented catalogue of paintwork by Miro.

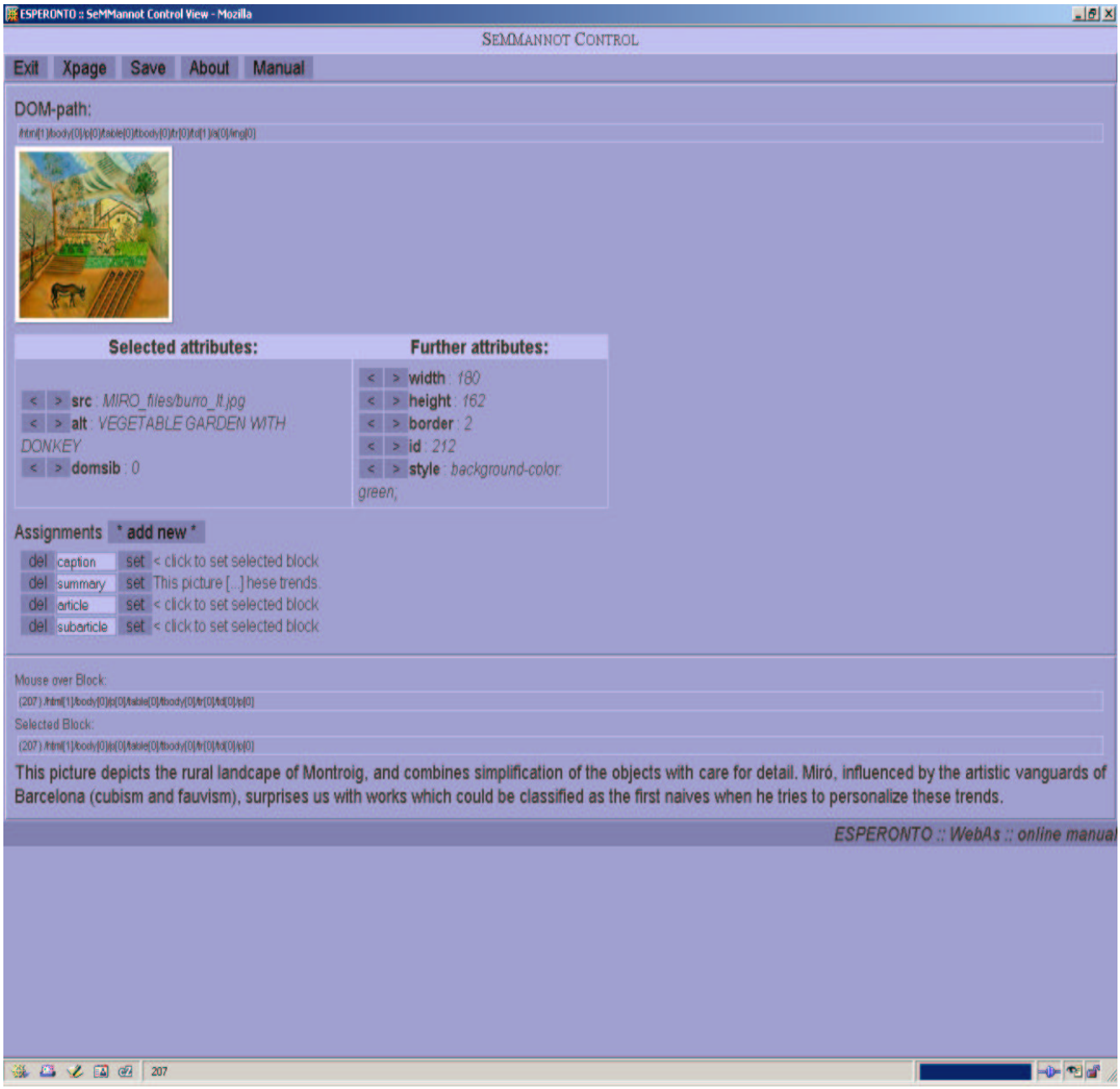

Figure 2: Screen shots of the actual version of the Esperonto Multimedia automatic content annotation demonstrator, applied to the first image displayed in Figure 1. Here the comment by the expert has been selected by the user and is sent to the NLP tools and the merging component of the Esperonto prototype.

The page chosen is also interesting with respect to another property: the caption text is not relevant to the content of the picture. It just states, for every picture on the page, that it is possible to zoom on the picture by clicking on it! So, our tool will have to provide for a kind of filter of the textual content of caption or other textual parts of the texts.

In the second screen shot, the reader can see that the first image of the Web page has been selected by the user, whereas automatically the "alt" encoded information is selected, as well as the information on the source of the image (VEGETABLE GARDEN WITH DONKEY and http://www.spanisharts.com/reinasofia/miro/burro_lt.jpg respectively). The user has then been selecting the text written by the specialist, situated directly to the left of the picture, and classified it as a "summary". This summary, the alt, the source, and the whole text are outputted in various files containing Metadata tracking their provenience and the type of text they contain (caption, summary etc.). The Esperonto NLP tools subsequently process those files.

Linguistic Processing of the various text types

When available at all, the caption and alt text types need a specialized linguistic treatment, since those texts are very often quite short and sometimes contain no full sentence. The NLP tools should be able therefore able to provide for an incomplete linguistic analysis, and an accurate ontology-based semantic annotation can not always take place on those strings, since knowledge context is missing. But the words and phrases used in those textual documents can serve as an anchor for detecting the relevant passages in the surrounding text in the web page.

So for example the image displayed in Figure 1: The text included in the "alt" html tag "'VEGETABLE GARDEN WITH DONKEY'" is not a full sentence, but just one nominal phrase (NP). The NLP analysis delivers the linguistic dependency of these NP. In our example we are lucky, since this string is exactly the same as the first string of the "summary" selected by the user. So the correlation between text and picture is strongly supported. When we access the art ontology of Residencia de Estudiantes, after the linguistic analysis, and there is an instance of a painting called with this name, then we can annotate the text parts associated with the picture with this ontology-based semantic information.

In any case, the textual analysis system can detect that this image is about a painting due to the usage of typical terms in the summary: "Oil on canvas. (65 x 71 cm.)", which are also encoded in the ontology. Also the first words of the normal text in the "summary" (This picture …) give hints that the image in the page is about a picture. And here again the use of information classified in the ontology on paintings will allow the specific semantic annotation about "title of picture", "dimension of the artwork" and "material of the artwork" etc.

On the base of the linguistic analysis (delivering linguistic dependency structures), the semantic annotation service can infer the topic of the picture, which his expressed in the linguistic direct object of the predicate "depicts" and also, for example, the artistic movement to which the artwork may be attached to the linguistic adverbial complementation of the verb "classified".

The linguistic processing of textual elements attached to a still image is thus a cascaded and incremental one. Starting with the alt/caption type of text, linguistic annotations are created that can serve as an anchor for aggregating the linguistic information gained so far with linguistic annotation resulting from the analysis of selected parts of the surrounding text or from the whole text. So in a sense the NLP tools are merging and aggregating linguistic annotations from various text types. In this, we can say that the Esperonto NLP tools are proposing a cross-document textual summarization, which is guided by the detection of recurring key words and linguistic structures in the various types of textual documents that are associated with a still image in the web. On the top of this consolidated linguistic base, ontology-based semantic annotation can take place.

The semantic annotation procedure

The linguistic dependency structure offered to the semantic annotation component is offering a good base for ontology-based semantic annotation, since a kind of linguistic consolidation has taken place: only the textual elements (and associated linguistic annotations) are taking into account, which seem to really refer to the picture.

Then the next step consists in (automatically) accessing the available ontologies (in our case an art ontology), and tries to map linguistic expression to concepts and relations described in the ontology. Using the ontological information, the textual expressions can then be marked up, in a process that people call "knowledge mark-up". But also the available ontology can be further "populated", in the

sense that textual parts that can be conceptually annotated can define a new instance of concepts (or relations) described in the ontology. And also the image described can be included in the ontology as a particular instance of an artwork.

CONCLUSIONS AND OUTLOOK

We have shown, that up to a certain degree, the semantic annotation strategy applied to text can be extended to the annotation of still images present in web pages. This information, if understood as semantic metadata for images, can probably be used for the purpose of improving the quality of the "traditional" analysis of images by means of so-called low-level features.
We expect that this kind of semantic annotations (metadata) we provide for still images in the web can be useful for the Multimedia community in the sense that we are contributing to a multimodal approach to content indexing of images. Ideally the collaborative effort of analysis strategies applied to various media (including text), should lead to shared ontogical description of the different types of features playing a role in the analysis: low-level features for images, linguistic and semantic features for text etc, and their contribution to a real cross-media analysis and generation infrastructure.

ACKNOWLEDGEMENTS

The R&D work carried out for the Esperonto project is funded under the 5th Framework Programme of the European Commission (IST-2001-34373). Thanks also to Oliver Schönleben, research assistant at Saarland University, for his work on the Esperonto tool for the annotation of still images.

REFERENCES

1. Benjamins R., J. Contreras, A. G. Pérez, H. Uszkoreit, T. Declerck, D Fensel ,Y. Ding, M. J. Wooldridge and V. Tamma. Esperonto Application: Service Provision of Semantic Annotation, Aggregation, Indexing and Routing of Textual, Multimedia, and Multilingual Web Content, Proceedings of WIAMIS, 2003.

2. Buitelaar P., Declerck T., 2003, Linguistic Annotation for the Semantic web, in: Siegfried Handschuh, Steffen Staab (eds.) Annotation for the Semantic Web, IOS Press.

3. Declerck T., A set of tools for integrating linguistic and non-linguistic information, in Proceedings of SAAKM (ECAI Workshop) 2002.

4. Declerck T., J. Kuper , H. Saggion, A. Samiotou, P. Wittenburg, J. Contreras. Contribution of NLP to the Content Indexing of Multimedia Documents, Lecture Notes in Computer Science Volume 3115/2004 Pages 610-618, Springer-Verlag Heidelberg, 6 2004

5. Buitelaar P., D. Olejnik , M. Hutanu , A. Schutz , T. Declerck , M. Sintek, Towards Ontology Engineering Based on Linguistic Analysis, Proceedings of LREC 2004.

6. Mezaris V. , H. Doulaverakis, R. Medina Beltran de Otalora, S.Herrmann, I. Kompatsiaris, M. G.Strintzis, A test-bed for region-based image retrieval using multiple segmentation algorithms and the MPEG-7 eXperimentation Model: The Schema Reference System, in Proceedings of CIVR 2004, Dublin.

7. Dasiopoulou S., V. K. Papastathis, V. Mezaris, I. Kompatsiaris, M. G. Strintzis, An Ontology Framework For Knowledge-Assisted Semantic Video Analysis and Annotation, in Proceedings of the ISWC 2004 Workshop "SemAnnot", Hiroshima.

8. Rehatschek H., N. Diakopoulos, G. Kienast, V. Hahn, T. Declerck, DIRECT-INFO: A Distributed Multimodal Analysis System for Media Monitoring Applications, in Proceedings of EWINT 2004, London.

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

10. SCHEMA Network: http://www.iti.gr/SCHEMA/

11. ACEMEDIA: http://www.acemedia.org/

12. DIRECT-INFO: http://www.direct-info.net/