Contribution of NLP to the Content Indexing of Multimedia Documents

Thierry Declerck[1], Jan Kuper, Horacio Saggion, Anna Samiotou, Peter Wittenburg, and Jesus Contreras

Saarland University and DFKI GmbH, Stuhlsatzenhausweg 3, D66123 Saarbruecken, Germany,
declerck@dfki.de,
http://www.dfki.de/ declerck

Abstract. This paper describes the role *natural language processing* (NLP) can play for multimedia applications. As an example of such an application, we present an approach dealing with the conceptual indexing of soccer videos which the help of structured information automatically extracted by NLP tools from multiple sources of information relating to video content, consisting in a rich range of textual and transcribed sources covering soccer games. This work has been investigated and developed in the EU funded project MUMIS. As a second example of such an application, we describe briefly ongoing work in the context of the ESPERONTO project dealing with upgrading the actual web towards the Semantic Web (SW), including the automatic semantic indexing of web pages containing a combination of text and images.

1 Introduction

This paper describes the role *natural language processing* (NLP) can play in the conceptual indexing of multimedia documents which can then be searched by semantic categories instead of key words. This topic was a key issue in the MUMIS project[1]. A novelty of the approach developed in MUMIS is to exploit multiple sources of information relating to video content (for example the rich range of textual and transcribed sources covering soccer games). Some of the investigation work started in MUMIS is being currently pursued with the ESPERONTO project[2], looking at images in web pages, and trying to apply content information to the pictures on the base of the semantic analysis of the surrounding text.

In the first part of the paper (section 2) we will propose a general discussion on the role that can be played by NLP for multimedia application. This overview is widely based on [2] and [3]. In the largest part of the paper we will exemplify

[1] MUMIS was a project within the Information Society Program (IST) of the European Union, section Human Language Technology (HLT). See for more information http://parlevink.cs.utwente.nl/projects/mumis/.

[2] ESPERONTO is a project within the 5th framework within the Information Society Program (IST) of the European Union. See for more details www.esperonto.net

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 610–618, 2004.

the general topic with the presentation of the MUMIS project that is concerned with the topic of multimedia indexing and searching. The presentation of this project is an extension and update of [7]. The last part of the paper will briefly sketch the ongoing work in ESPERONTO, aiming at attaching content information in images contained in web pages, by using semantic features automatically attached by NLP tools to the surrounding texts.

2 The Role of NLP for Multimedia Applications

[2] and [3] give an overview of the role that can be played by NLP in multimodal and multimedia systems. We summarize here the central points of those studies.

2.1 Multimodal and Multimedia Systems

The terms *multimedia* and *multimodal* are often source of confusion. [2] adopts the definitions as proposed in [13], which establishes a distinction between the terms *medium*, *mode* and *code*. The term *mode* (or *modality*) refers to the type of perception concerned, being for example visual, auditory or olfactory perception. The term *medium* refers to the carrier of (CD-ROM, paper etc.), to the devices (microphone, screen, loudspeakers etc.), as well to the distinct types of information (texts, audio or video sequences). The term *code* refers to the particular means of encoding information (sign languages or pictorial languages). One can speak of a multimedia system if this allows to *generate* and/or to *analyze* multimedia/multimodal information or provide some *access* to archives of multiple media. In existing applications, often the process of analysis applies only to multimodal data, whereas generation is concerned with the production of multimedia information.

2.2 Integration of Modalities

We speak in the case of analysis of a process of *integration of modalities*, since all the available modalities need to be merged at a more abstract level in order to take the maximal advantage of every modality involved in the application. Certain representation formalisms, as they have been defined in the context of advanced NLP (see for example [12]), can play a central role in this process of fusion. So the well-defined technique of *unification* of typed feature structures, combined with a chart parser, is used in a system described in [10]. Using this formalism allows to build a semantic representation that is common to all modalities involved in the application, unifying all the particular semantic contributions on the base of their representation in typed feature structures

2.3 Media Coordination

In the case of the *generation* of multimedia material including natural language, for the purpose of *multimedia presentation*, one can speak of a process of *media coordination*: it is not enough to merge various media in order to obtain a

coherent presentation of the distinct media involved. The information contained in the various media has to be very carefully put into relation if one wants to obtain real complementarities of media in the final presentation of the global information. And since systems for *natural language generation* have been always confronted with this problem of selecting and organizing various contributions for the generation of an utterance, they can provide for a very valuable model for the coordination of media in the context of the generation of multimedia presentations. A lot of systems for natural language generation are therefore said to be *plan-based*.

2.4 Natural Language Access to Multimedia

Multimedia repositories of moving images, texts, and speech are becoming increasingly available. This together with the needs for 'video-on-demand' systems require fine-grain indexing and retrieval mechanisms allowing users access to specific segments of the repositories containing specific types of information.

It turns out that natural language can play a multiple role. It is first easier to access information contained in the multimedia archive using queries addressed to (transcript of) audio sequences or to the subtitles (if available) associated to the videos as to analyze the pictures themselves. It is further more appealing to access visual data by means of natural language, since the latter supports more flexible and efficient queries as the query based on image features. And ultimately natural language offers a good means for condensing visual information. The selected list of projects concerned with video indexing we give below is stressing this fact: at some point always some language data will be considered to support retrieval of images or videos.

In order to support this kind of natural language access, video material was usually manually annotated with 'metadata' such as people involved in the production of the visual record, places, dates, and keywords that capture the essential content of what is depicted. Still, there are a few problems with human annotation. First, the cost and time involved in the production of "surrogates" of the programme is extremely high; second, humans are rather subjective when assigning descriptions to visual records; and third, the level of annotation required to satisfy user's need can hardly be achieved with the use of mere keywords.

2.5 NLP Techniques for Indexing Multimedia Material

Many research projects have explored the use of parallel linguistic descriptions of the images (either still or moving) for automatic tasks such as indexing [42], classifying [43], or understanding [44] of visual records, instead of using only content-based (or visually-based) methods in use [40]. This is partly also due to the fact that NLP technologies are more mature for extracting meaning as the technologies in use in the field of image. Content-based indexing and retrieval of visual records is based on features such as color, texture, and shape. Yet visual understanding is not well advanced and is very difficult even in closed domains. For example, visual analysis of the video of a football match can lead to the

identification of interesting "content" like a shooting scene (i.e., the ball moving towards the goal) [45], but this image analysis approach will hardly ever detect who is the main actor involved in that scene (i.e., the shooter). For accessing visual information with the help of natural language, certain systems make use of a shallow analysis of linguistic data associated with pictures, like the transcripts of audio comments or subtitles. In most of the case this is already enough in order to provide for a first classification and indexing of the visual data (see for example [11], [32] or [33]). Other systems use more sophisticated linguistic analysis, like *information extraction* (IE): the detection of *named entities* and of standard linguistic patterns can help the multimedia retrieval systems to filter out non-relevant sequences (for example the introduction of speakers in news broadcasting). An example of such systems is given in the "Broadcast News Navigator" developed at MITRE (see [15]) . The MUMIS project, described in details below, is going even further, since full IE systems are analyzing a set of so-called "collateral" (parallel) documents and produce unified conceptual annotations, including metadata information that is used for indexing the video material and supports thus concept-based queries on a multimedia archive. A similar approach is described in [23], where the domain of application is classical dance. In this work only a small set of textual documents is considered, in a monolingual setting.

3 MUMIS: A Multimedia Indexing and Searching Environment

MUMIS has been proposing an integrated solution to the NLP-based multimedia content indexing and search. The solution consists of using information extracted from different sources (structured, semi-structured, free, etc.), modalities (text, speech), and languages (English, German, Dutch) all describing the same event to carry out data-base population, indexing, and search. MUMIS makes an intensive use of linguistic and semantic based annotations, coupled with domain-specific information, in order to generate formal annotations of events that can serve as index for videos querying. MUMIS applies IE technologies on multilingual and multimedia information from multiple sources.

The novelty of the project was not only the use of these 'heterogeneous' sources of information but also the combination or cross-source fusion of the information obtained from each source. Single-document, single-language information extraction is carried out by independent systems that share a semantic model and multi-lingual lexicon of the domain. The result of all information extraction systems is merged by a process of alignment and rule-based reasoning that also uses the semantic model.

For this purpose the project makes use of data from different media (textual documents, radio and television broadcasts) in different languages (Dutch, English and German) to build a specialized set of lexicons and an ontology for the selected domain (soccer). It also digitizes non-text data and applies speech recognition techniques to extract text for the purpose of annotation. Audio material has been analyzed by Phicos [46], an HMM-based recognition system, in order

to obtain transcriptions of the football commentaries (spontaneous speech). It uses acoustic models, word-based language models (unigram and bigram) and a lexicon. For Dutch, English, and German different recognition systems have been developed. i.e. different phone sets, lexicons, and language models are used. Transcriptions for 14 German, 3 Dutch, and 8 English matches have been produced. [25] gives more details on the *automatic speech recognition* (ASR) and the transcription work done in the context of MUMIS.

The core linguistic processing for the annotation of the multimedia material consists of advanced information extraction techniques for identifying, collecting and normalizing significant text elements (such as the names of players in a team, goals scored, time points or sequences etc.) which are critical for the appropriate annotation of the multimedia material in the case of soccer. One system per language has been used or developed.

Each system delivers an XML output, an example being shown in figure 1 which serves as the input of a *merging component*, whose necessity in the project is due to the fact that MUMIS is accessing and processing multiple sources from distinct media in distinct languages. The merging tool is combining the semantically related annotations generated from those different data sources, and detect inconsistencies and/or redundancies within the combined annotations. The merged annotations are then stored in a database, where they will be combined with relevant metadata that are also automatically extracted from the textual documents.

Those annotations are delivered to the process of indexing key frames from the video stream. Key frames extraction from MPEG movies around a set of predefined time marks - result of the information extraction component - is being carried out to populate the database. JPEG key frames images are extracted that serve for quick inspection in the user interface.

Within the MUMIS user interface, the user first interacts with a web-portal to start a query session. An applet is being down-line loaded, which mainly offers a query interface. The user then enters a query that either refers to metadata, formal annotations, or both. The on-line system searches for all formal annotations that meet the criteria of the query. In doing so it will find the appropriate meta-information and/or moments in some media recording. In case of meta-information it simply offers the information in scrollable text widgets. This is done done in a structured way such that different type of information can easily be detected by the user. In the case that scenes of games are the result of queries about formal annotations the user interface first presents selected video key frames as thumbnails with a direct indication of the corresponding metadata. The user can then ask for more metadata about the corresponding game or for more media data. A snapshot of the demonstrator is show in figure 2 above.

4 Multimedia Indexing in the Esperonto Project

Within the Esperonto project, NLP-based annotation strategies, combining with ontologies and other knowledge bases, are applied in order to upgrade the actual Web towards the emerging Semantic Web. In this project, experiments are

7. Ein Freistoss von Christian Ziege aus 25 Metern geht ueber das Tor.
(7. A 25-meter free-kick by Christian Ziege goes over the goal.)

```
<EVENTS>
   <TYPE>Free-kick</TYPE>
   <DISTANCE>Meter-from_(25)</DISTANCE>
   <1_PLAYER>Ziege</1_PLAYER>
   <CLASS>goal_scene_fail</CLASS>
   <ARTEFACT>Goal</ARTEFACT>
   <TIME>7:00</TIME>
</EVENTS>

<META> DOM_NAME="SOCCER"</META>

<PLAYER>
   <PLAYER_NAME>Ziege</PLAYER_NAME>
   <PLAYER_NOTE>#(3,5)</PLAYER_NOTE>
   <PLAYER_POS>#4 ##3</PLAYER_POS>
   <PLAYER_NUMBER>##17</PLAYER_NUMBER>
   <PLAYER_AGE>##28</PLAYER_AGE>
   <PLAYER_CLUB>##FC Middlesbrough</PLAYER_CLUB>
   <PLAYER_NUMB_PLAYS>##52</PLAYER_NUMB_PLAYS>
</PLAYER>
```

Fig. 1. Example of the XML encoding of the result of the automatic extraction from a sentence of a relevant event, with its associated relations entities and relations. The information about the player is dynamically included from the processing of 2 structured texts reporting on the same game, marked with # and ## respectively.

also conducted on the use of Semantic Web annotation structures, eventually consisting of complex ontology-based frames (or templates), that are associated to parts of text surrounding pictures to the indexing of the pictures themselves. Work is still not advanced enough in order to be reported in detail this paper, but actual results show that the main problem will consist in automatically detecting the parts of text related to the pictures. Here the caption of the picture, as well as the name of the image in the html document can offer a support for filtering the relevant semantic annotation from the surrounding texts.

5 Conclusions

The MUMIS experience has shown that NLP can contribute in defining semantic structures of multimedia contents, at the level proposed by domain-specific IE analysis. The full machinery of IE, combined with ASR (and in the future with Image Analysis, so for example with the actual results of the Schema Reference platform[3] can be used for multimedia contents development and so efficiently

[3] See http://www.schema-ist.org/SCHEMA/ and also the contribution on the SCHEMA reference platform in this volume.

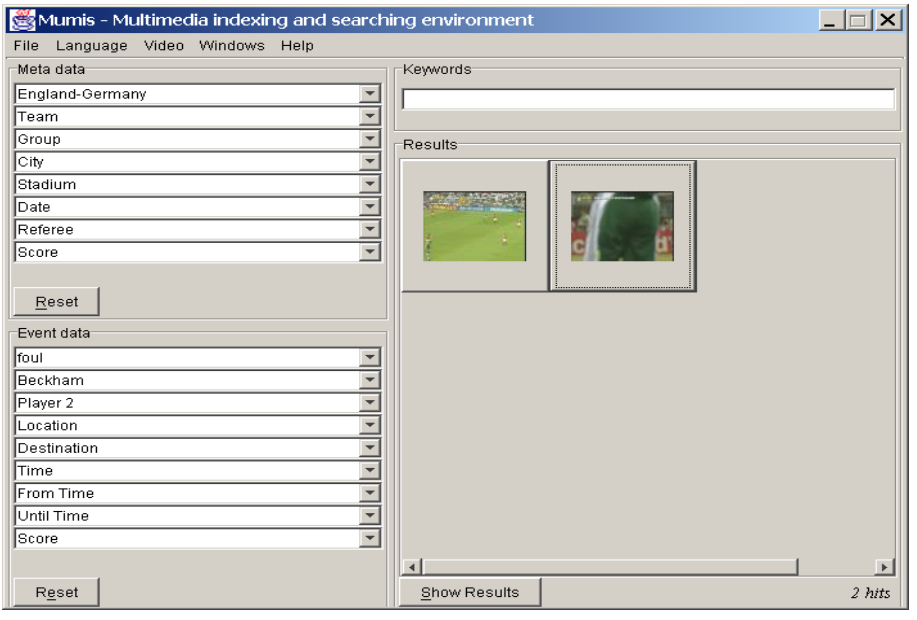

Fig. 2. MUMIS User Interface. Thumbnails proposed for the query "Show the fouls comited by Beckham in the game England-Germany"

support cross-media (and cross-lingual) information retrieval and effective navigation within multimedia information interfaces, thus simplifying the access to content (and knowledge) distributed over multiple documents and media, which are also increasingly available on the Web. The ESPERONTO project is currently porting some of the experiences gained in former projects on content indexing of multimedia documents in the Semantic Web framework. We will try to integrate part of the work described above within the SCHEMA reference platform.

Work that remains to do consist in fully integrating results of image/video content analysis with the semantic analysis of text/transcripts, towards a full Semantic Web annotation services for multimedia documents.

Acknowledgements. This research has in part been supported by EC grants IST-1999-10651 for the MUMIS project and IST-2001-34373 for the ESPERONTO project.

References

1. Adani N., Bugatti A., Leonardi R., and Migliorati P. Semantic description of multimedia documents: the Mpeg-7 approach. In Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001, Brescia, 2001.
2. André E. Natural Language in Multimedia/Multimodal Systems. In Mitkov R. (ed.), Handbook of Computational Linguistics, Oxford, 2000.

3. André E. The Generation of Multimedia Presentations. In Handbook of Natural Language Processing, Marcel Dekker, 2000.
4. Assfalg J., Bertini M., Colombo C., and Del Bimbo A. Semantic annotations of sports videos. In Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001, Brescia, 2001
5. Cunningham H. Information Extraction: A user Guide, Research Report CS-99-07, Department of Computer Science, University of Sheffield, May 1999.
6. Day N. MPEG-7 Applications: Multimedia Search and Retrieval. In Proceedings of the First International Workshop on Multimedia Annotation, MMA-2001, 2001.
7. Declerck T., Wittenburg P., Cunningham H. The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment. Proceedings of the Workshop on Human Language Technology and Knowledge Management, ACL-2001, 2001.
8. Declerck T. A set of tools for integrating linguistic and non-linguistic information. Proceedings of SAAKM 2002, ECAI 2002, Lyon.
9. Djoerd H., de Jong F., Netter K. (Eds). 14th Twente Workshop on Language Technology, Language Technology in Multimedia Information Retrieval, TWLT 14, Enschede, Universiteit Twente, 1998.
10. Johnston M. Unification-based Multimodal Parsing, In Proceedings of the 17th International Conference on Computational Linguistics, COLING-98, 1998.
11. de Jong F., Gauvin J., Hiemstra D., Netter K. Language-Based Multimedia Information Retrieval. In Proceedings of the 6th Conference on Recherche d'Information Assistee par Ordinateur, RIAO-2000, 2000. Indexing Workshop (CBMI2001), 2001.
12. Krieger H.-U., Schaefer U. TDL – a type description language for constraint-based grammars. In Proceedings of the 15th International Conference on Computational Linguistics, COLING-94, 1994.
13. Maybury M. Multimedia Interaction for the New Millenium. In Proceedings of Eurospeech 99, 1999.
14. McKeown K. Text generation, Cambridge University Press, 1985.
15. Merlino A., Morey D., Maybury M. Broadcast News Navigation using Story Segments. ACM International Multimedia Conference, 1997.
16. Miller, G.A. WordNet: A Lexical Database for English. Communications of the ACM 11. 1995.
17. Moore J., Paris C. Planning Text for Advisory Dialogues. In Proceedings of the 27th ACL, Vancouver, 1989.
18. Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann, 1995.
19. Seventh Message Understanding Conference (MUC-7), http://www.muc.saic.com/, SAIC Information Extraction, 1998.
20. Naphade Milid R. and T.S. Huang. Recognizing high-level concepts for video indexing. In Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001, Brescia, 2001. Extraction and Navigation System. In Proceedings of the 6th Conference on Recherche d'Information Assistee par Ordinateur, RIAO-2000, 2000.
21. Saggion H. , Cunningham H., Bontcheva K., Maynard D, Ursu C. Hamza O. and Wilks Y. Access to Multimedia Information through Multisource and Multilanguage Information Extraction. 7th Workshop on Applications of Natural Language to Information Systems (NLDB 2002), 2002.
22. Salembier P. An overview of Mpeg-7 multimedia description schemes and of future visual information challenges for content-based indexing. In Proceedings of the Conference on Content-Based Multimedia Indexing, CBMI-2001, Brescia, 2001.

23. Salway A., Talking Pictures: Indexing and Representing Video with Collateral Texts. In Hiemstra D., de Jong F., Netter K. (Eds), Language Technology in Multimedia Information Retrieval (Proceedings of the 14th Twente Workshop on Language Technology, TWLT 14), Enschede, Universiteit Twente, 1998.
24. Staab S., Maedche A., Handschuh S. An Annotation Framework for the Semantic Web. In The First International Workshop on Multimedia Annotation, Tokyo, Japan, 2001.
25. Wester M., Kessens J.M. and Strik H. Goal-directed ASR in a Multimedia Indexing and Searching Environment (MUMIS). Proceedings of the 7th International Conference on Spoken Language Processsing (ICLSP2002), 2002.
26. EUR: http://www.foyer.de/euromedia/
27. GDA: http://www.csl.sony.co.jp/person/nagao/gda/
28. INF: http://www.informedia.cs.cmu.edu/
29. ISI: http://www.wins.uva.nl/research/isis/isisNS.html
30. ISLE: http://www.ilc.pi.cnr.it/EAGLES/ISLE_Home_Page.htm
31. NSF: http://www.nsf.gov./od/lpa/news/press/pr9714.htm
32. OLI: http://twentyone.tpd.tno.nl/olive
33. POP: http://twentyone.tpd.tno.nl/popeye
34. SUR: http://www-rocq.inria.fr/ nastar/MM98/node1.html
35. THI: http://www.dcs.shef.ac.uk/research/groups/spandh/projects/thisl
36. UMA: http://ciir.cs.umass.edu/research/
37. UNL:http://www.ias.unu.edu/research_prog/science_technology/ universalnetwork_languge.html
38. VIR: http://www.virage.com/
39. COL: http://www.cs.columbia.edu/ hjing/sumDemo
40. Veltkamp R. and Tanase M. Content-based Image Retrieval Systems: a survey. Technical report UU-CS-2000-34, Utrecht University, 2000.
41. Chang S.F., Chen, W., Meng H.J., Sundaram H. and Zhong D. A Fully Automated Content-based Video Search Engine Supporting Spatio Temporal Queries. IEEE Transactions on Circuits and Systems for Video Technology, 1998.
42. Netter K. Pop-Eye and OLIVE. Human Language as the Medium for Cross-lingual Multimedia Information Retrieval. Technical report, Language Technology Lab. DFKI GmbH, 1998.
43. Sable C. and Hatzivassiloglou V. Text-based approaches for the categorization of images. Proceedings of ECDL, 1999.
44. Srihari R.K. Automatic Indexing and Content-Based Retrieval of Captioned Images, Computer 28/9, 1995.
45. Gong Y., Sin L.T., Chuan C.H., Zhang H. and Sakauchi M. Automatic Parsing of TV Soccer Programs. Proceedings of the International Conference on Multimedia Computing and Systems (IEEE), 1995.
46. Steinbiss V., Ney H., Haeb-Umbach R., Tran B.-H., Essen U., Kneser R., Oerder M., Meier H.-G., Aubert X., Dugast C. and Geller D. The Philips Research System for Large-Vocabulary Continuous-Speech Recognition. In Proc. of Eurospeech '93, 1993.