A text-based search interface for Multimedia Dialectics

Katerina Pastra
Inst. for Language & Speech Processing
Athens, Greece
kpastra@ilsp.gr

Eirini Balta
Inst. for Language & Speech Processing
Athens, Greece
ebalta@ilsp.gr

Abstract

The growing popularity of multimedia documents requires language technologies to approach automatic language analysis and generation from yet another perspective: that of its use in multimodal communication. In this paper, we present a support tool for COSMOROE, a theoretical framework for modelling multimedia dialectics. The tool is a text-based search interface that facilitates the exploration of a corpus of audiovisual files, annotated with the COSMOROE relations.

1 Introduction

Online multimedia content becomes more and more accessible through digital TV, social networking sites and searchable digital libraries of photographs and videos. People of different ages and cultures attempt to make sense out of this data and re-package it for their own needs, these being informative, educational and entertainment ones. Understanding and generation of multimedia discourse requires knowledge and skills related to the *nature* of the interacting modalities and their *semantic interplay* for formulating the multimedia message.

Within such context, intelligent multimedia systems are expected to parse/generate such messages or at least assist humans in these tasks. From another perspective, everyday human communication is predominantly multimodal; as such, similarly intuitive human-computer/robot interaction demands that intelligent systems master —among others— the semantic interplay between different media and modalities, i.e. they are able to use/understand natural language and its reference to objects and activities in the shared, situated communication space.

It was more than a decade ago, when the lack of a theory of how different media interact with one another was indicated (Whittaker and Walker, 1991). Recently, such theoretical framework has been developed and used for annotating a corpus of audiovisual documents with the objective of using such corpus for developing multimedia information processing tools (Pastra, 2008). In this paper, we provide a brief overview of the theory and the corresponding annotated corpus and present a text-based search interface that has been developed for the exploration and the automatic expansion/generalisation of the annotated semantic relations. This search interface is a support tool for the theory and the related corpus and a first step towards its computational exploitation.

2 COSMOROE

The CrOSs-Media inteRactiOn rElations (COSMOROE) framework describes multimedia dialectics, i.e. the semantic interplay between images, language and body movements (Pastra, 2008). It uses an *analogy to language discourse* analysis for "talking" about multimedia dialectics. It actually borrows terms that are widely used in language analysis for describing a number of phenomena (e.g. metonymy, adjuncts etc.) and adopts a message-formation perspective which is reminiscent of *structuralistic* approaches in language description. While doing so, inherent characteristics of the different modalities (e.g. exhaustive specificity of images) are taken into consideration.

COSMOROE is the result of a *thorough, inter-disciplinary review* of image-language and gesture-language interaction relations and characteristics, as described across a number of disciplines from computational and semiotic perspectives. It is also the result of *observation and analysis* of different types of corpora for different tasks. COSMOROE was tested for its coverage and descriptive power through the annotation of a corpus of TV travel documentaries. Figure 1 presents the COSMOROE relations. There are three main rela-

Proceedings of the EACL 2009 Demonstrations Session, pages 53–56,
Athens, Greece, 3 April 2009. ©2009 Association for Computational Linguistics

tions: semantic equivalence, complementarity and independence, each with each own subtypes.

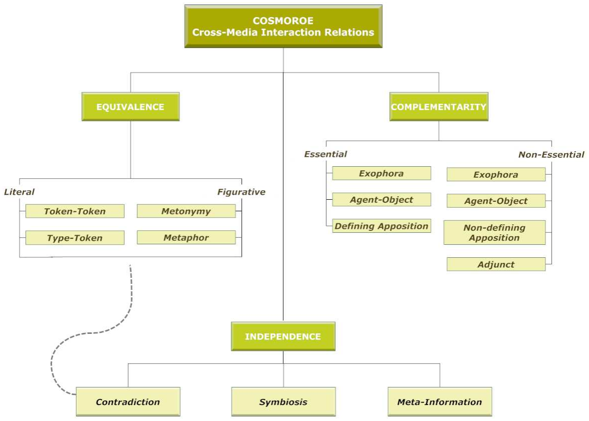


Figure 1: The COSMOROE cross-media relations

For annotating a corpus with the COSMOROE relations, a multi-faceted annotation scheme is employed. COSMOROE relations link two or more annotation facets, *i.e.* the modalities of two or more different media. Time offsets of the transcribed speech, subtitles, graphic-text and scene text, body movements, gestures, shots (with foreground and background distinction) and keyframe-regions are identified and included in COSMOROE relations. All visual data have been labelled by the annotators with one or two-word action or entity denoting tags. These labels have resulted from a process of watching only the visual stream of the file. The labelling followed a cognitive categorisation approach, that builds on the "basic level theory" of categorisation (Rosch, 1978). Currently, the annotated corpus consists of 5 hours of TV travel documentaries in Greek and 5 hours of TV travel documentaries in English. Three hours of the Greek files have undergone validation and a preliminary inter-annotator agreement study has also been carried out (Pastra, 2008).

3 The COSMOROE Search Interface

Such rich semantically annotated multimedia corpus requires a support tool that will serve the following:

- it will facilitate the active exploration and presentation of the semantic interplay between different modalities for any user, illustrating the COSMOROE theory through specific examples from real audiovisual data

- it will serve as simple search interface for general users, taking advantage of the rich semantic annotation —behind the scenes— for more precise and intelligent retrieval of audiovisual files

- it will allow for observation and educated decision-taking on how one could proceed with mining the corpus or using it as training data for semantic multimedia processing applications, and

- it will allow interfacing with semantic lexical resources, computational lexicons, text processing components and cross-lingual information resources for automatically expanding and generalising the data (semantic relations) one can mine from the corpus.

We have developed such tool, the COSMOROE search interface. The interface itself is actually a text-based search engine, that indexes and retrieves information from the COSMOROE annotated corpus. The interface allows for both simple search and advanced search, depending on the type and needs of the users. The advanced search is designed for those who have a special interest in multimedia semantics and/or ones who want to develop systems that will be trained on the COSMOROE corpus. This advanced version allows search in a text-based manner, in either of these ways:

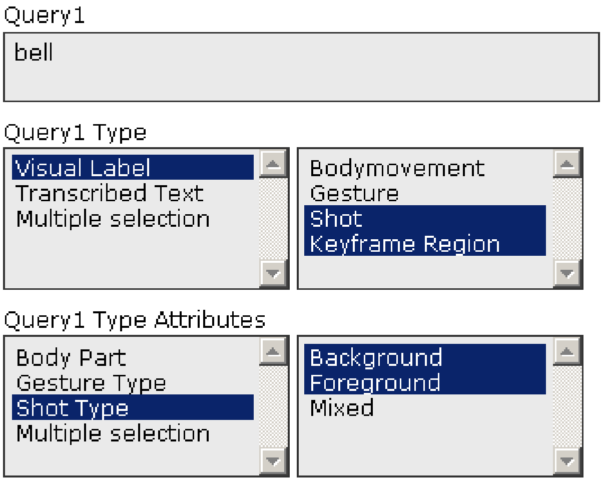
- Search using single or multiword query terms (keywords) that are mentioned in the *transcribed speech (or other text) of the video* or in the *visual labels set of its visual-units*, in order to find instantiations of different semantic relations in which they participate;

- Search using a *pair of single or multiword query terms* (keywords) that are related through (un)specified semantic relations;

- Search for specific *types of relations* and find out how these are realized through actual instances in a certain multimedia context;

- Search for specific *modality types* (e.g. specific types of gestures, image-regions etc.) and find out all the different relations in which they appear;

Figure 2 presents a search example, using the advanced interface[1]. The user has opted to search for all instances of the word "bell" appearing in the visual label of keyframe regions and/or video shots and in particular ones in which the bell is clearly shown either in the foreground or in the background. In a similar way, the user can search for concepts present in the audio part of the video, through the use of the "Transcribed Text" option or make a multiple selection. Another possibility is to use a "Query 2" set, in conjunction, disjunction or negation with "Query 1", in order to obtain the relations through which two categories of concepts are associated.



Query1

bell

Query1 Type

Visual Label
Transcribed Text
Multiple selection

Bodymovement
Gesture
Shot
Keyframe Region

Query1 Type Attributes

Body Part
Gesture Type
Shot Type
Multiple selection

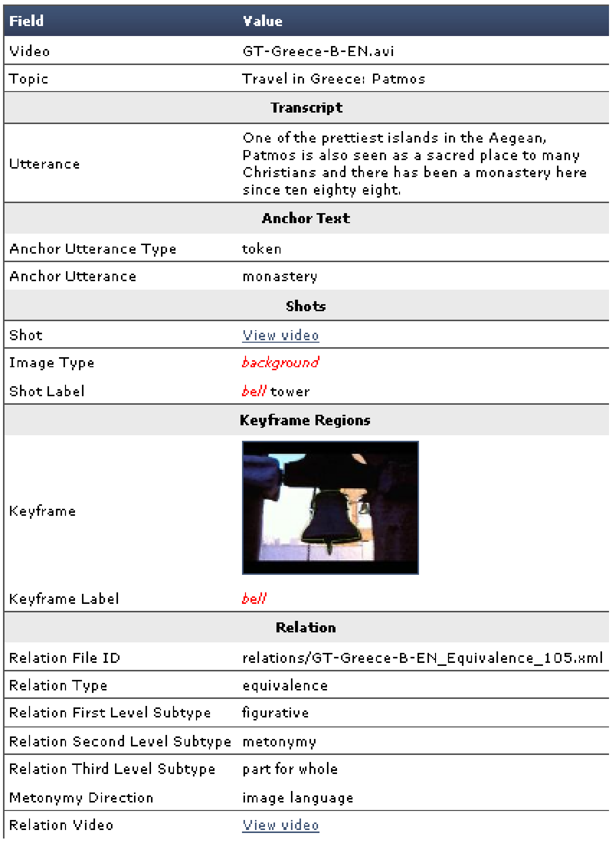
Background
Foreground
Mixed

Figure 2: Search example

Multimedia relations can also be searched independently of their content, simply denoting the desired type. Finally, the user can search for special types of visual-units, such as body movements, gestures, images, without defining the concept they denote.

After executing the query, the user is presented with the list of the results, grouped by the semantic relation in which the visual labels —in the example case presented above— participate. Each hit is accompanied by its transcribed speech. Indication of the number of results found is given and the user has also the option to save the results of the specific search session. By clicking on individual hits in the result list, one may investigate the corresponding relation particulars.

Figure 3 shows such detailed view of one of the results of the query shown in Figure 2. All relation components are presented, textual and visual ones. There are links to the video file from which the relation comes, at the specified time offsets. Also, the user may watch the video clips of the modalities that participate in the relation (e.g. a particular gesture) and/or a static image (keyframe) of a participating image region (e.g. a specific object) with the contour of the object highlighted.



Field	Value
Video	GT-Greece-B-EN.avi
Topic	Travel in Greece: Patmos
	Transcript
Utterance	One of the prettiest islands in the Aegean, Patmos is also seen as a sacred place to many Christians and there has been a monastery here since ten eighty eight.
	Anchor Text
Anchor Utterance Type	token
Anchor Utterance	monastery
	Shots
Shot	View video
Image Type	*background*
Shot Label	*bell* tower
	Keyframe Regions
Keyframe	
Keyframe Label	*bell*
	Relation
Relation File ID	relations/GT-Greece-B-EN_Equivalence_105.xml
Relation Type	equivalence
Relation First Level Subtype	figurative
Relation Second Level Subtype	metonymy
Relation Third Level Subtype	part for whole
Metonymy Direction	image language
Relation Video	View video

Figure 3: Example result - relation template

In this example, one may see that the word "monastery", which was mentioned in the transcribed speech of the file, is grounded to the video sequence depicting a "bell tower" in the background and to another image of a "bell", through a metonymic relation of type "part for whole". What is actually happening, from a semantic point of view, is that although the video talks about a "monastery", it never actually shows the building, it shows a characteristic part of it instead. In this page, the option to save these relation elements as a text file, is also provided.

Last, a user may get a *quantified profile* of the contents of the database (the annotated corpus) in terms of number of relations per type, per language, per file, or even per file producer, number of visual objects, gestures of different types, body

[1] Only part of the advanced search interface is depicted for the screenshot to be intelligible

movements, word tokens, visual labels, frequencies of such data per file/set of files, as well as co-occurrence statistics on word-visual label pairs per relation/file/language and other parameters.

For the novice or general user, a simple interface is provided that allows the user to submit a text query, with no other specifications. The results consist of a hit list with thumbnails of the video-clips related to the query and the corresponding transcribed utterance. Individual hits lead to full viewing of the video clip. Further details on the hit, i.e. information an advanced user would get, are available following the advance-information link. The use of semantic relations in multimedia data, in this case, is hidden in the way results are sorted in the results list. The sorting follows a highly to less informative pattern relying on whether the transcript words or visual labels matched to the query participate in cross-media relations or not, and in which relation. Automating the processing of audiovisual files for the extraction of cross-media semantics, in order to get this type of "intelligence" in search and retrieval within digital video archives, is the ultimate objective of the COSMOROE approach.

3.1 Technical Details

In developing the COSMOROE search interface, specific application needs had to be taken into consideration. The main goal was to develop a text-based search engine module capable of handling files in the .xml format and accessed by local and remote users. The core implementation is actually a web application, mainly based on the Apache Lucene[2] search engine library. This choice is supported by Lucene's intrinsic characteristics, such as high-performance indexing and searching, scalability and customization options and open source, cross-platform implementation, that render it one of the most suitable solutions for text-based search.

In particular, we exploited and further developed the built-in features of Lucene, in order to meet our design criteria:

- The relation specific .xml files were indexed in a way that retained their internal tree structure, while multilingual files can easily be handled during indexing and searching phases;
- The queries are formed in a text-like manner by the user, but are treated in a combined way by the system, that enables a relational search, enhanced with linguistic capabilities;
- The results are shown using custom sorting methods, making them more presentable and easily browsed by the user;
- Since Lucene is written in Java the application is basically platform-independent;
- The implementation of the Lucene search engine as a web application makes it easily accessible to local and remote users, through a simple web browser page.

During the results presentation phase, a special issue had to be taken into consideration, that is video sharing. Due to performance and security reasons, the Red5[3] server is used, which is an open source flash server, supporting secure streaming video.

4 Conclusion: towards computational modelling of multimedia dialectics

The COSMOROE search interface presented in this paper is the first phase for supporting the computational modelling of multimedia dialectics. The tool aims at providing a user-friendly access to the COSMOROE corpus, illustrating the theory through specific examples and providing an interface platform for reaching towards computational linguistic resources and tools that will generalise over the semantic information provided by the corpus. Last, the tool illustrates the hidden intelligence one could achieve with cross-media semantics in search engines of the future.

References

K. Pastra. 2008. Cosmoroe: A cross-media relations framework for modelling multimedia dialectics. *Multimedia Systems*, 14:299–323.

E. Rosch. 1978. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, chapter 2, pages 27–48. Lawrence Erlbaum Associates.

S. Whittaker and M. Walker. 1991. Toward a theory of multi-modal interaction. In *Proceedings of the National Conference on Artificial Intelligence Workshop on Multi-modal Interaction*.

[2] http://lucene.apache.org/

[3] http://osflash.org/red5/