Beyond Multimedia Integration: corpora and annotations for cross-media decision mechanisms

Katerina Pastra

Language Technology Applications Department
Institute for Language and Speech Processing
Artemidos 6 and Epidavrou, 151-25, Maroussi, Greece
{kpastra}@ilsp.gr

Abstract

In this paper, we look into the notion of cross-media decision mechanisms, focussing on ones that work within multimedia documents for a variety of applications, such as the generation of intelligent multimedia presentations and multimedia indexing. In order for these mechanisms to go beyond the identification of semantic equivalence relations between media —which is what integration does— appropriate corpora and annotations are needed. Drawing from our experience in the REVEAL THIS project, we indicate the characteristics that such corpora should have, and suggest a number of annotations that would allow for training/designing such mechanisms. We conclude with a view on the suitability of two related markup languages (MPEG-7 and EMMA) for accommodating the suggested annotations.

1. Introduction

Multimedia integration, and in particular image-language integration, has been defined as a process of establishing associations between medium-specific representations (Pastra, 2004). The corresponding computational mechanisms are needed in intelligent multimedia systems, for a wide range of applications, such as multimedia presentation generation, multimedia dialogue, multimedia indexing (Pastra and Wilks, 2004b). It has been argued that the development of such mechanisms requires that multimedia corpora exist, in which the association between different media is explicitly annotated (Pastra and Wilks, 2004a).

There are, actually, a few data collections, which provide image-language associations for training and evaluating multimedia systems. For example, the IBM video-annotation forum has created a 62-hour video collection annotated with textual labels/categories referring to static scenes, key objects and events depicted in each shot (Lin et al., 2003); the annotation is MPEG-7 compatible and the collection was constructed mainly for multimedia integration purposes in multimedia indexing and retrieval systems. The PASCAL visual object categorization challenge (Everingham et al., 2005) made use of a collection of images associated with corresponding textual labels of the objects depicted and with object boundaries markup, for training and evaluating image categorization systems.

In both cases, the metadata representation of the initial document collection was manually enriched with a *finite set* of textual labels that were associated to a visual medium; there was no markup of *existing associations* between media that were *both present* in the multimedia document collection, as suggested in (Pastra and Wilks, 2004a) (e.g. objects in a video shot and corresponding narrative naming of the objects). Since all these attempts and suggestions facilitate the development of multimedia integration mechanisms, one would argue that whether the association to be marked-up is inherent in the data collection or manually added, is not significant. However, it becomes significant, if a system is to go beyond multimedia integration, in developing *cross-media decision mechanisms* and in particular, *within document* cross-media decision mechanisms.

In this paper, we first indicate the multimedia corpus characteristics dictated for the development of such cross-media decision algorithms, we then focus on specific types of annotations that are needed in such corpora for training/designing cross-media decision algorithms and we correlate our annotation suggestions with EMMA (Extensible MultiModal Annotation markup language) (Johnston et al., 2005) and MPEG-7 (Martinez, 2004). For doing so, we draw from our experience in REVEAL-THIS (Piperidis and Papageorgiou, 2005), an FP6 project in which —among other modules— a cross-media indexing mechanism for efficient video retrieval is built.

2. What is there beyond Multimedia Integration?

Crossing-media for achieving a task has normally taken the form of finding semantic equivalences between different (multimedia or medium-specific) documents both of which express the same information (e.g. a TV news programme and a corresponding web news article) (Boll et al., 1999); this is a form of multimedia integration across documents. However, when crossing-media within the same multimedia document (e.g. a video), one needs to go beyond the notion of semantic equivalence to other semantic interaction relations that may hold between different media/modalities. Generally put, such cross-media decision mechanisms go a step further from multimedia integration, in attempting to decide whether two medium-specific representations:

a. are associated (i.e. have a common reference) — what multimedia integration does,

b. are not associated, they actually complement each other in forming a multimodal message, or

c. are not associated or complementary, they collaborate in forming a cohesive message (semantically compatible) or they are semantically incompatible (contradict-

ing / thematically unrelated), due to e.g. errors in automatic medium-specific analysis/interpretation

Taking as an example the case of a cross-media indexing mechanism for a video retrieval system, one could say that once the mechanism takes as input the medium-specific analysis of a video segment (e.g. the output of the image analysis component, the face detection component, the speech processing component and the natural language understanding component), it has to decide which medium-specific pieces of information (medium-specific interpretations of the video) provided from these modules are more representative of the content of the segment for indexing.

In case (a), the association indicates a highly significant piece of information, which could be used as an indexing term. In case (b), a conjunction of different medium-specific pieces of information forms the indexing term. In case (c), all medium-specific pieces of information are used as indexing terms (i.e. they curry different pieces of information which are equally important and which are thematically related) or, in the case of semantic incompatibility, the indexing mechanism will have to choose a piece(s) of information from a specific medium that is to be trusted as more reliable for expressing the video segment content. Training or designing such cross-media decision mechanism is far from trivial. Appropriate corpora and annotations are needed and it is exactly this notion of appropriateness that we are going to elaborate in the following sections.

3. Multimedia Corpus Characteristics for Cross-media Mechanisms

The REVEAL THIS prototype is envisioned to provide users with real time, personalized access to multimedia and multilingual data coming from a variety of sources (Piperidis and Papageorgiou, 2005). Naturally, the development of such system is demanding in terms of not only medium-specific processors and cross-media decision algorithms needed, but also of training and testing data to assist in their development. From our experience in building the corpus for the REVEAL THIS prototype, we concluded that a corpus built for training and evaluating within-document cross-media decision algorithms must have specific characteristics that are dictated from the very nature of the algorithms, i.e., the fact that:

a. they work within documents rather than across, which imposes constraints on the nature of documents to be included in the corpus, and

b. they rely on the output of multiple medium-specific processors, each of which has different training needs (with regards to both quantity and content of the data collected).

In particular, the documents collected must be **multimedia** (e.g. video/TV programmes or/and illustrated web articles) as opposed to:

- single-medium documents covering a range of different modalities, for example, a corpus of thematically associated photographs, 3D graphics and sketches, all considered different modalities (types of images) of the same medium (vision)[1], and
- single-medium documents covering a wide range of different media, for example, a corpus of thematically associated radio programmes, text-only webpages and photographs.

A corpus of multimedia documents provides the cross-media decision algorithm with a number of medium-specific representations per document (or document part) to associate, combine or reject as erroneous for indexing.

On the other hand, the corpus must be **multi-genre**, so that the needs of the medium-specific processors are covered; for instance, a speech-identification component requires data with a variety of speakers, speaking for about 10 minutes each, a face identification component requires data presenting the same face from different viewpoints with no occlusions, an image analysis component requires data rich in images of the objects/object classes that are to be identified, not to mention the read vs. spontaneous speech and the formal vs. colloquial language characteristics that a well-balanced corpus should generally have[2].

Part of the REVEAL-THIS corpus features videos of the plenary sessions and press conferences of the European Parliament (EP) in English and Greek, evening TV news programmes in the same languages that were aired the same day as the EP events (so that the possibility of the general news referring to what was discussed in the EP is maximized), and videos of travel documentaries and TV travel magazines in both languages.

This data satisfies the above mentioned criteria; there is visual variety needed for the development of the face detection/identification and image analysis sub-components, since there is a variety of viewing angles used in the EP plenary transmissions (i.e., the speaker is seen from various angles, in various postures), there is a variety of number of people depicted (e.g. shots of the whole audience, close ups of the speaker and the MP sitting beside the speaker etc.). In the press conferences, there is a variety of shots depicting the panel and the crowd of journalists, or one journalist asking a question and the speaker who answers.

In the travel documentary files, it is mainly scenery that is being presented, with faces/people being present only in the form of a journalist/narrator of the documentary or as an example of the locals of a region. There is a variety of scenery though, such as archaeological sites, mountains, seaside, examples of local architecture, the local cuisine, the night life (clubs, restaurants etc.) for training an image analysis/categorization module. The variety of genres satisfies, also, the read vs. spontaneous speech distinction, as in the case of e.g. European Parliament plenary speeches vs. press-conferences. The formal vs. colloquial language distinction is satisfied —to an extent— through the inclusion

[1] Similarly, corpora of texts and audio/speech files capture two modalities (text and speech) of the same medium (language); they are multimodal corpora but not multimedia (they do not include modalities of different media).

[2] The last two characteristics actually facilitate the development of speech and text processing systems that can generalise beyond such distinctions.

of politics-news programmes (in which language is less formal than the one used in the European Parliament), but also through the *multi-domain* character of the corpus, i.e. the politics (mainly formal language use) vs. travel (mainly informal language use) data included. As in most cases of software development, training/designing an algorithm on more than one domains renders it able to generalise; in our case, the use of multiple domains provides for different media-combinations within multimedia documents, i.e. for a variety of multimedia relation cases to train the algorithm on (cf. for example, the rich imagery of travel documentaries and the corresponding narration as opposed to the face-people rich only imagery of the parliamentary sessions and the corresponding audio streams).

This data amounts to approximately 80 hours of politics in each language (same audio streams/interpreted), with 1/8 of this data being press-conferences, 10 hours of TV news in English and 20 hours in Greek, 87 hours of travel videos in English and 53 hours in Greek (a small part of this data consists of same audio stream in both languages and another small part consists of audio stream in one language and subtitles in the other). The data is available in a variety of formats (MPEG-2, WMV and TS for the politics data). Unfortunately, only the politics videos are free for research purposes (as long as the original source/copyright of the Europe By Satellite (EBS) channel is acknowledged).

4. Annotating Cross-media Relations

While a multimedia corpus with the above-mentioned characteristics is *necessary* for training/designing cross-media decision mechanisms, it is by itself not *sufficient*; proper annotations are also needed, which, actually, bares the question of what does *proper* refer to. As mentioned earlier, a within-document cross-media mechanism needs to be able to identify the relation that holds between medium-specific pieces of information, not only ones that hold in a non-processed corpus (i.e. ones in which the annotator detects and identifies the medium-specific units), but also ones that hold between automatically interpreted information units.

4.1. Multimedia Relations

We have identified three relations, which may hold between medium-specific pieces of information. In this paper we present these relations briefly, so that we focus on the annotations they entail, and illustrate their functionality from the scope of a cross-media indexing mechanism:

- *Equivalence*: the information expressed refers to the same entity (object, state, event, or property). It is a case of association between media or of what is commonly referred to as "multimedia integration". In such a case, a cross-media indexer may keep either or both pieces of information for indexing the document.
- *Complementarity*: the information expressed in one medium is (an essential or not) complement of the information expressed in another medium. Association signals (e.g. textual indexicals pointing to an image or image part) indicate cases of essential complementarity, while non-essential complementarity is characterized by one medium modifying or playing the role of an adjunct for the other (e.g. an image showing —among others— the location of a speaker whose speech is captured by the corresponding audio stream of the document).
- *Independence*: each medium curries an independent message, which is, however, coherent with the document topic, e.g. video footage showing the European Parliament and its plenary sessions and corresponding audio referring to the reformations of the European Constitution. In case an automatic speech recognition module fails to output the phrase "European Constitution" giving instead the phrase "Euro Pin contest", no coherence with the output of an image analyser that has identified the European Parliament building correctly exists. In such a case, an indexer must be able to detect the incoherence and decide which medium-specific information is correct, correlating each of these with the output of other medium-specific modules (e.g. an image categorizer or/and a text categorizer).

In order to further illustrate these multimedia relations, let's consider the following cases from figures 1 to 4, all of which come from a travel documentary (TV programme in English). Figure 1 depicts a case of *equivalence*: the narrator refers to yellow taxi-boats that are being used for island hopping in some Greek islands and the corresponding keyframe of the video actually depicts these boats. The image of the boat and the corresponding noun phrase refer to the same entity, the former through visual means, the latter through language.

Figure 1: A case of *equivalence*: "...the yellow taxi-boats..."

Figure 2 illustrates a case of *non-essential complimentarity*, and in particular a case in which the image functions as an adjunct of manner to the main verb of the corresponding clause: the narrator refers to the next destination of her trip in Greece saying the she is heading to the island of Patmos. The corresponding keyframe of the video depicts the high-speed boat by which the narrator is travelling to

the island. Therefore, the image depicts the means through which the narrator is "heading to" the island which is not known through the use of the verb "head to". In this case, the contribution of the image to the multimedia message is complimentary to the information provided by the speech, though not essential.

Figure 2: A case of *non-essential complimentarity*: "...we are heading to Patmos..."

On the contrary, figure 3 depicts a case of *essential complimentarity*; the narrator talks about pollution in Athens and concludes saying that this pollution has taken its toll on something that she points to using at the same time the indexical "this". Both the pointing gesture and the use of the indexical render the extra information provided by the corresponding keyframe —which is the reference of the indexical and the target of the gesture, i.e. the Acropolis— essential. In this case, visual information needs to compliment the utternace for the multimedia message to be comprehensive.

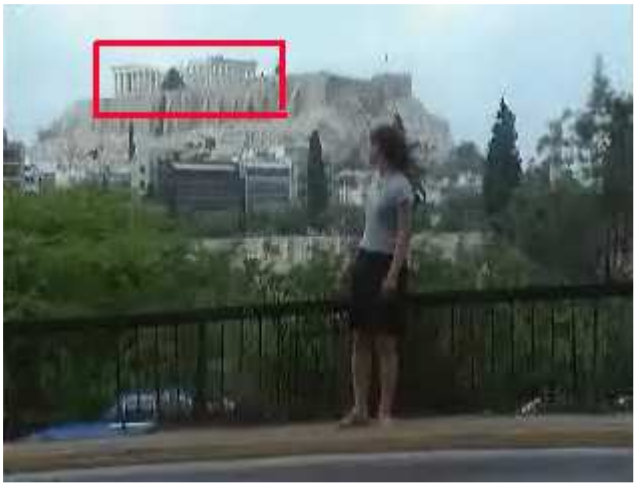

Figure 3: A case of *essential complimentarity*: "...[pollution has taken its toll]...on this..."

Last, figure 4 presents a case of *independence*. The narrator makes a general comment on the place she has visited and the corresponding keyframe shows her sitting on a rock at a beach. The information provided by each modality (speech and image respectively) is completely independent, in the sense that one does not refer in any way to something expressed in the other[3]. Still, what is being said is semantically compatible (coherent) to what is depicted in the keyframe, since a quiet place with no other people than the narrator is depicted. This would not be the case, if the keyframe depicted a beach full of people; in the latter case, the "antithesis" that one would evidence, would either lead one to take this as an ironic comment, or as an editorial mistake during the production of the documentary.

Figure 4: A case of *independence*: "...I have finally found a place that's not overun by tourists..."

4.2. Annotation types

Based on such a multimedia relations framework, one may identify a number of annotation types that are needed for training/designing a cross-media indexer:

a. *Association* - A(X,Y): a link between medium-specific representations, e.g. a linguistic expression (of an object, property, spatial relation etc.) and its visual reference (Pastra and Wilks, 2004a).

b. *Partial Association* - PA(X,Y): a link between e.g. a visual feature and a corresponding linguistic expression which denotes only the type of the feature and not its value e.g. the word "coloured" is linked to the corresponding visual feature of an object, but not to the specific value of the feature (e.g. red). An image always specifies feature values, whereas language may be more elusive.

c. *Association Signal* - AS(X,Y): annotation of the association signals, i.e. indexicals (e.g. linguistic or pictorial, or even gestures) and the object of signalling. The signals indicate a relation between e.g. image and text,

[3]The word "place" in the utterance (cf. the caption of figure 4) does not refer to the specific place depicted in the image, but to the island —in general— that the narrator visited in the last part of the documentary.

but do not express any commonly expressed piece of information. For example, the word "this" may point to something pictorial, but its function is just that, to point to something. It does not express what the thing pointed to is. The latter is information that is provided only by the image.

d. *Adjunct* - AJ(X,Y): adverbial-type modification (location, time, manner) (cf. the example given for the *complementarity* relation).

e. *Apposition* - AP(X,Y): in this case both media refer to the same entity but in different conceptual levels, e.g. Y stands as a type ("The president"), and X as a token (the photograph of Bill Clinton). What makes this situation different from the "association" annotation is that this case is tied to the specific context and should not be generalised (i.e. the word president is not used only for Bill Clinton, but for a number of other people too). It is not the same case as in e.g. Association(furniture,imageRegion:chair), i.e. the association of the word "furniture" with the photograph of a chair, which though in different conceptual levels, it is not tied to a specific context of discourse.

f. *Coherence* - CH(X,Y): indication of coherence or lack of between medium specific representations (cf. examples in the *independence* relation above).

The first two annotation types allow the identification of *equivalence* relations between medium-specific pieces of information, (c), (d), and (e)[4] indicate *complementarity* relations and (f) allows for the identification of *independence* relations.

4.3. Annotation language requirements

One could go on elaborating on these types of annotations and applying them on a multimedia corpus, however this goes beyond the scope of this paper. It is significant, though, to note that annotating a multimedia corpus with the above mentioned relations and annotation types, requires that the markup language used has two specific characteristics:

- it allows for modular description of the structure of a multimedia document and of the individual media it consists of, facilitating correlation of parts of the individual media that reflect different levels of granularity. For instance, it should allow theencoding of a relation between an audio stream down to the level of the token and an image region, or between a token and a keyframe or set of keyframes
- secondly, it allows for the creation of a multimedia unit, one that consists of different media (media parts or media elements) from within the document, the properties of which are re-defined accordingly. The unit goes beyond the structural decomposition of the document or the individual media, and is actually a multimedia semantic unit (cf. for example the case of essential complementarity, in which two medium-specific pieces of information must be interpreted together for a complete message to be formulated).

Do commonly used multimedia markup languages have such features? We conclude this paper, with a brief discussion of this issue.

5. Suitability of commonly used Markup-Languages

MPEG-7 is a standard for describing multimedia content developed by ISO (Martinez, 2004); it is an elaborate markup language which allows for —among others— low-level feature description (e.g. colour, motion activity, sound timbres), high-level feature description (e.g. objects, events etc.), structural information (e.g. image regions, keyframes, audio segments etc.), relations between these structural units (spatial, temporal, aggregation), and textual annotations for each unit (free or structured text, syntactic dependencies, or even externally-defined controlled vocabulary and classification schemes).

These features render it ideal for expressing multimedia relations between medium-specific units of information (e.g. an imageRegion and an audioSegment). Describing a multimedia unit with a number of re-defined properties of the kind explained in the previous section, could be done, in an extension of the MPEG-7 description scheme. The types of descriptors that could be needed, in such a case, can be exemplified through the exploration of another markup language, EMMA.

EMMA (Extensible Multimodal Annotation Markup Language) is an XML markup language for the interpretation of user input, an initiative under the auspices of W3C, which is still in a working-draft format (Johnston et al., 2005). The language provides mainly high-level description elements for the annotation of automatically generated interpretations of multimodal user input. In focussing on the output of medium-specific and multimedia integration mechanisms, it allows for the description of the relations between such output in forming a multimedia message. The most representative —for our purposes— examples, are the emma:hook element and the notion of "composite derivation" (which is supported through a number of description elements).

The "hook" element denotes the incomplete semantics of a medium-specific piece of information, it is actually an indicator of association between medium-specific pieces of information, equivalent to what we called "association signal" in the previous section; its value is the type of medium from which complementary information is needed/expected (e.g. video, ink etc.).

The notion of composite derivation refers to the combination of medium-specific pieces of information (of different media) into one multimedia unit. For this unit, the markup language determines the units it consists of (e.g. the word "Destination" + pen-input pointing to an image region that stands for "Boston") and determines the scope of

[4] Though a special case of association, apposition is considered indicative of a complementarity relation between media, because it indicates that the two media should better be considered together as forming one message, rather than as equivalent (and therefore mutually exclusive).

the annotations (time-stamps, confidence values etc.) of its constituents. These description elements can be extended to accommodate the multimedia relation framework presented in the previous sections and its corresponding annotation types for accommodating cross-media decision elements.

6. Conclusion

In this paper, we presented the notion of *within-document cross-media decision mechanisms* and argued that the development of such mechanisms requires corpora of *multimedia*, *multi-genre* and *multi-domain* documents, annotated with such information that will allow the mechanism to identify *equivalence*, *complementarity* and/or *independence* relations that may hold between medium-specific pieces of information in a multimedia document.

We suggested a number of such annotations and noted that, in order to facilitate these annotations, a multimedia content markup language should allow the description of relations between levels of medium analysis of different granularity and the creation of new multimedia units emerging from the interrelations between medium-specific pieces of information within a document. The former is easily accommodated within MPEG-7, while the latter can be covered more easily in EMMA. A cooperation between the two schemes seems appropriate and timely, so that the description of cross-media relations within (and even across) documents is systematically dealt with.

7. Acknowledgements

This work is carried out in the framework of the REVEAL THIS project (FP6-IST-511689 grant). The author would like to thank the project participants for fruitful discussions on the notion of cross-media decision mechanisms.

8. References

Boll, S., W. Klaus, and J. Wandel, 1999. A cross-media adaptation strategy for multimedia presentations. In *Proceedings of ACM Multimedia*.

Everingham, M., L. Van Gool, C. Williams, and A. Zisserman, 2005. Pascal visual object classes challenge results. World Wide Web (http://www.pascal-network.org/challenges/VOC/voc).

Johnston, M., W. Chou, D. Dahl, G. McCobb, and D. Raggett, 2005. Emma: Extensible multimodal annotation markup language. Technical report, World Wide Web Consortium (W3C). Working Draft: 16 September 2005.

Lin, C., B. Tseng, and J. Smith, 2003. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. TRECVID Proceedings.

Martinez, J., 2004. Mpeg-7 overview. Technical Report ISO/IECJTC1/SC29/WG11 N6828, International Organisation for Standardisation. Version 10.

Pastra, K., 2004. Viewing vision-language integration as a double-grounding case. In *Proceedings of the AAAI Fall Symposium on "Achieving Human-Level Intelligence through Integrated Systems and Research"*.

Pastra, K. and Y. Wilks, 2004a. Image-language multimodal corpora: needs, lacunae and an ai synergy for annotation. In *Proceedings of the 4th Language Resources and Evaluation Conference*.

Pastra, K. and Y. Wilks, 2004b. Vision-language integration in ai: a reality check. In *Proceedings of the 16th European Conference in Artificial Intelligence*.

Piperidis, S. and H. Papageorgiou, 2005. Reveal this: Retrieving video and language for the home user in an information society. In *Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies (EWIMT)*.