Towards a Cross-Media Analysis of Spatially Co-located Image and Text Regions in TV-News

Thierry Declerck[1] and Andreas Cobet[2]

[1] DFKI GmbH, Language Technology Lab,
Stuhlsatzenhausweg.3, 66123 Saarbrücken, Germany
declerck@dfki.de
[2] Technische Universität Berlin, Communication Systems Group,
Straße des 17. Juni, 10623 Berlin, Germany
cobet@nue.tu-berlin.de

Abstract. We describe in this poster/short paper on-going work on the extraction and semantic interpretation of text regions in television news programmes. We present some of the data we consider in this work, the actual technologies in use and where they have to be improved. Finally we briefly discuss a possible innovative and valuable approach to the establishment of a cross-media analysis framework.

1 Introduction

The European Network of Excellence "K-Space"[1], which started in 2006, is dealing with semantic inferences for semi-automatic annotation and retrieval of multimedia content. The aim of the project is to contribute in narrowing 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 the integration of knowledge structures, as encoded in high-level representation languages, 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, or even text included in the images. These complementary resources typically fall into two groups: primary resources, which are directly attached to multimedia, and secondary resources, which are more loosely coupled with the audio-video material.

We concentrate in this position paper on the primary complementary resources, and investigate the possible use of text extracted from images by means of detection of textual regions in images and optical character recognition (OCR) processes for adding semantics to images and also to support audio/video analysis. We describe a first experiment on text extraction from news videos (news programmes of the German broadcaster "ARD"), for which we identified a list of relevant patterns of

[1] "K-Space" stays for "Knowledge Space of Shared Technology and Integrative Research to Bridge the Semantic Gap"; see also www.k-space.eu.

B. Falcidieno et al. (Eds.): SAMT 2007, LNCS 4816, pp. 288 – 291, 2007.

textual information appearing on the TV screen during those news programmes, and their relations to displayed images. We show 2 examples of such patterns (out of 6 we identified), and explain what can be gained from those patterns, which are particular in the sense that the text belonging to one semantic unit might be distributed around an image. The use of the extracted text for supporting the semantic annotation of their containing images implies the applications of linguistic analysis of the extracted text and an appropriate detection images regions. We concentrate in this paper on a short description of such patterns, showing results of the detection of textual regions in images, of the OCR procedure and of linguistic analysis applied to the extracted text.

2 The Data

In the following we just present 2 typical examples of presenting news information to the TV public[2], showing how the broadcaster (here the German public broadcaster ARD) combines image and text to convey information.

Fig. 1. In this pattern, we can see the speaker and a background image, directly surrounded by two textual contributions (ignoring here the name of the News programme and the date)

In the first case above, we consider the two textual regions being close to the background image, just above and below it. The text analysis tool applied to the extracted text can detect the topic (decision of the Parliament about election) and also where it takes place (in Kiew). Interesting here: there is no linguistic hint, that this "decision" is being discussed in Kiew: We can infer this only on the base of heuristics applied to the distribution of words around the image. On the base of world

[2] As already mentioned above, they are more such patterns, which cannot be described here due to space limitation.

knowledge, we can also infer that the Parliament presented here is the Ukrainian one (this information being most probably given by the speaker). Other information we can recognize: the voice to be heard in the audio streaming is in this pattern belonging to the news speaker. In case we know her name, this information can help in improving speaker recognition. In other patterns of information display, we can assume that the voice being heard is not from the speaker, but from someone presented in the image (or video).

A more complex pattern, with respect to semantic interpretation is shown in Fig. 2.

Fig. 2. We can see above the background picture a short phrase and below the picture the name of a person. The text should be read as "Accusations against the son of Annan".

In Fig. 2 the person name below the image is pointing to the content of the image. But the information on the top of the image is mentioning: "complains against son". So here we need also some inferences to get the point that the accusations are addressed against the son of the person shown in the image, but that the son is not shown here.

3 First Results

The textual region detection tools used in our experiment perform quite good[3], but the OCR tools applied have still to be improved[4], and below we can see one error generated by the OCR procedure: the name of Kiew being represented as "Krew". But

[3] This information on the base of a small informal evaluation done on the data. A formal evaluation is still to be proposed.

[4] First steps have been done in this respects, which we will present in a next version of this short paper.

here when we know that we deal with Named Entities, a list (or gazetteer) of such Entities can be given to the OCR mechanisms for matching/correcting their results against it. An example of an output of the system (for the example given in Fig. 1) is:

Krew 4 78 452
Parlamentsbeschluss 4 84 102
Zur 4 84 150
Wahl 4 130 143

Here we have a good result, with only 1 error ("Krew" instead of "Kiew"). We can extract out of this data structure the different text contributions. We cannot propose for an analysis of the string "Krew" for the time being (unknown word). The (automatic) linguistic dependency analysis of the textual contribution at the top of the image is giving:

[NP Parlamentsbeschluss *(head_noun)* [PP zur Wahl] *(noun_modifier)*]

The identified head noun is the main topic. And in fact this corresponds to the image, showing a parliament. So the key frame is not about the „election", but about a parliament decision about the election. The dependency analysis allows thus to reduce considerably the number of key words that can be used for indexing, replacing them by structured textual fragment. We can map here the head noun onto an ontology as well, so that the image might be also annotated with concepts like "political institutions" or the like.

4 First Conclusions

On the base of the patterns described above, we start to provide a textual analysis, which has to take into account the non-sequential list of words distributed over the screen. For this, we have to take into account the region information of the text parts, also including information about the region of the image to which the text parts are "belonging".

The linguistic and semantic annotation resulting from the analysis of those textual parts can then support the semantic annotation of the image as well. Very interesting in this context, is the fact that the image itself contributes to the semantic content of the text, since the image plays sometimes the role of filling the gap of omitted textual elements in the image displayed in the actual news. So for examples we noticed that only short nominal phrases are used (very seldom verbs are used in this context), and that images are acting as linking information between the textual part. This is very well documented in the figure 2 above, where the "complains against son" and "Annan" are linked by the image of Annan. In normal natural text, we would expect the sentence "Complains against the son of Annan".

If this sentence would have been associated with an image, we would certainly expect to have not Annan himself shown in the image, but rather the son of Annan. In this sense the patterns we recognized can really give extended support to person detection in images, in a cross-media fashion.