Linguistic concepts described with Media Query Language for automated annotation

Lenkiewicz, Anna
anna.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Lis, Magdalena
Magdalena@hum.ku.dk
University of Copenhagen, Denmark

Lenkiewicz, Przemyslaw
przemek.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

1. Introduction

Human spoken communication is multimodal, i.e. it encompasses both speech and gesture. Acoustic properties of voice, body movements, facial expression, etc. are an inherent and meaningful part of spoken interaction; they can provide attitudinal, grammatical and semantic information. In the recent years interest in audio-visual corpora has been rising rapidly as they enable investigation of different communicative modalities and provide more holistic view on communication (Kipp et al. 2009). Moreover, for some languages such corpora are the only available resource, as is the case for endangered languages for which no written resources exist.

However, annotation of audio-video corpora is enormously time-consuming. For example, to annotate gestures researchers need to view the video-recordings multiple times frame-by-frame or in slow motion. It can take up to 100 hours to manually annotate one hour of the recording (Auer et al. 2010). This leads to a shortage of large-scale annotated corpora which hampers analysis and makes generalizations impossible. There is a need for automatic annotation tools which will make working with multimodal corpora more efficient and enable communication researchers to focus more on the analysis of data than on the annotation of material.

The Max Planck Institute for Psycholinguistics and two Fraunhofer Institutes are working together on developing advanced audio and video processing algorithms, called recognizers (Lenkiewicz et al. 2011), which are able to detect certain human behavior in a recording and annotate it automatically. However, usage of those recognizers is rather complex and their output is limited to detecting occurrences of predefined events without assigning any semantics. They also work independently from one another delivering annotations according to their specification, like for example hand movement or specific speech characteristics, which later need to be manually analyzed by the researchers in order to find any dependencies between different features.

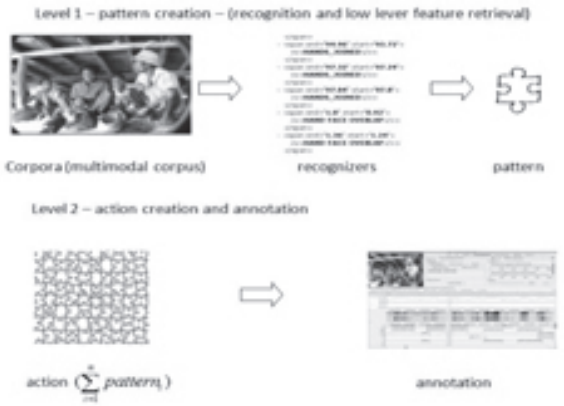

Figure 1: The general structure of the Media Query Language

The core of the work described in this paper is the development of a MQL, which can take advantage of the capabilities of the recognizers and allow expressing complex linguistic concepts in an intuitive manner.

2. Media Query Language

In the current phase of development general structure of the language is defined (Figure 1). The language contains three integrated components: *patterns, actions* and *libraries*.

2.1. Patterns

As a first step in media annotation using MQL human behavior would be decomposed to a single meaningful movement, gesture or speech element and saved for future reuse in form of a pattern as presented in Level 1 of Figure 1. A pattern for the purpose of this work is defined as a template or model, which can be used to save elements of human behavior decoded from media file under a meaningful name. It is the primary element created using MQL and it is the base for future action creation. Pattern provides a sort of architectural outline that may be reused in order to speed up the annotation process by applying search-by-example.

As a first step in pattern creation the corpora will be analyzed by recognizers and by a researcher using

specifically for this purpose designed active device application, which simplifies new pattern creation by marking interesting parts of a recording and including them in the MQL code (Figure 2). All solutions will be integrated with ELAN (Brugman & Russel 2004) a professional tool for the creation of complex annotations on video and audio resources.

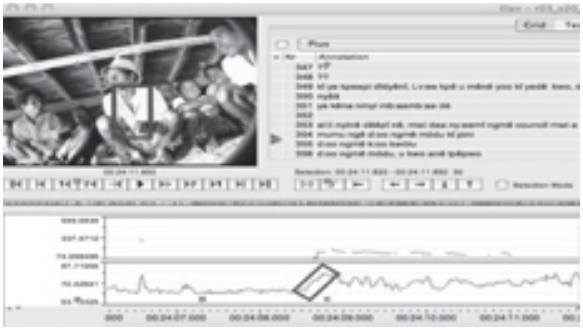

Figure 2: Example of pattern creation using active device as input collector. A to B – start and end point for hand movement. Red square – field determining tolerance in pattern matching in future search

The MQL language will be programmed in order to execute specific recognizers, depending on the nature of the request from the user and using recognizers' specification. The output of the recognizers will be collected and joined in sets of characteristic human behaviors, like for example when, which and how a given hand is moving, including information about direction (left, right, up, down) and speed. At the same time another recognizer can deliver information about speech characterization, like utterance boundaries, word separation or pitch contour. Recognizers will be therefore used to retrieve numerical description of the features mentioned above. The language will serve as a tool, which helps to convert the code received from the recognizers into patterns expressing human behavior in a language similar to natural. An example of a code fragment describing a pattern in MQL is presented in Figure 3.

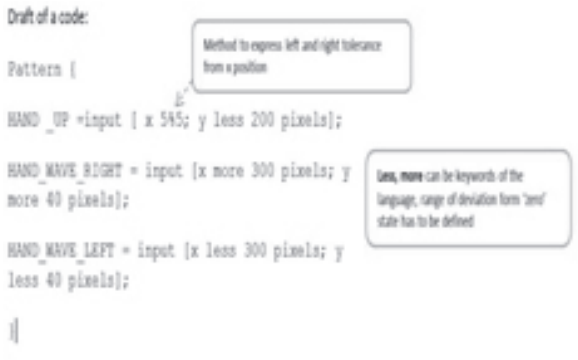

Figure 3: Example of a pattern defined in MQL code

2.2. Libraries

In order to simplify the feature annotation process and assure that already created patterns can be reused, each pattern created by the user needs to be added to a pattern library. Meaningful name and description needs to be given. MQL allows users to name the patterns and actions in a not imposed way. Formulation of the query is also free, as long as it will be in conformity with the language syntax and will include keywords and pattern names. An example of such library can be a set of patterns describing hand movement patterns like hands overlapping, hands rising, one hand movement with specific speed, etc. It is assumed that a pattern library will be dedicated to single actions occurrences. Moreover, it is planned that the system would inform the user whenever a currently created pattern already exists under different name and in which place of the hierarchy. The main goal of this functionality is the creation of well-developed pattern libraries, which in the future can limit the need for new pattern creation. With proper libraries available for the end-users a lot of media annotation work can be carried out using only already existing patterns. Prototype of such a library is shown in Figure 4.

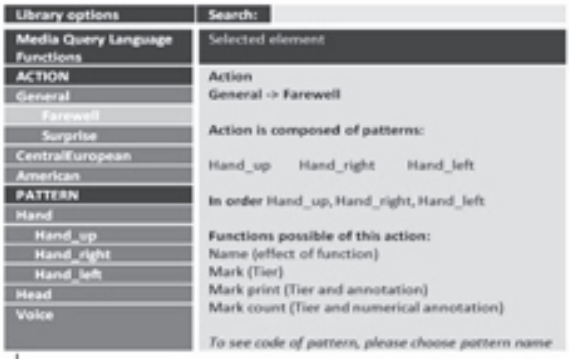

Figure 4: Example of a library defined in MQL code

2.3. Actions

The second level of the language will be dedicated to action creation and file annotation. It is called the executing phase of programming with MQL. On this level the user will be able to combine patterns into actions.

Action in MQL is defined as a set of predefined patterns (human movements, gestures or speech elements) composed together. Advancement of a single pattern to an action is also possible. Actions may be composed of patterns that may require execution of more than one recognizer in order to detect features of interest. An example of a MQL code fragment describing an action is presented in Figure 5.

```
Action {
FAREWELL = [HAND_UP + HAND_WAVE_RIGHT + HAND_WAVE_LEFT]
}
```

Figure 5: Example of an action defined in MQL code

A good example to illustrate the concept of action is annotation of utterance type. In signaling whether an utterance is a statement, a question or a command, different communicative behaviors (e.g. both speech prosodic cues and face expression) can work together. With MQL researchers will be able to aggregate such acoustic and gestural patterns together and save them in the form of an action in an action library according to the same rules as apply to pattern library creation. Thanks to this solution the decision about event classification into specific human behavior and creation of annotation can be taken by researchers rather than an automatic system. Work of the recognizers can be limited to detecting audio and body movement as detailed as possible.

3. Conclusions and future work

Thanks to the proposed MQL solution, together with the recognizers, the major problem of researchers, which is the time needed to manually annotate data, will decrease significantly. The expertise of the researchers will be used better by transferring their focus from highly laborious basic feature annotation to adding meaning to retrieved data and semantics to MQL language components.

Using MQL in the annotation process is going to significantly change the way in which the annotation work is carried out and how it can be reused. Using common ways to describe meaningful features and creating reusable libraries will contribute to creation of the interoperability standards.

The feedback on the current version of MQL received from researchers is encouraging. Flexibility and sufficient expressiveness of complex linguistic concepts is seen as a great advantage.

In the future work the problems of semantic gap coverage and interoperability standards will be tackled thanks to the possibility of expressing complex linguistic concepts using MQL.

Acknowledgment

The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA).

References

Auer, E., P. Wittenburg, H. Sloetjes, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel (2010). Automatic annotation of media field recordings. In C. Sporleder and K. Zervanou (eds.), *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*. Lisbon: University de Lisbon, pp. 31-34.

Brugman, H., and A. Russel (2004). Annotating multi-media / multimodal resources with ELAN *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, pp. 2065-2068.

Kipp, M., J.-C. Martin,P. Paggio, and F. K. J. Heylen, **ed.** (2009). Multimodal Corpora: From Models of Natural Interaction to Systems and Applications. *Lecture Notes in Computer Science 5509*. London: Springer.

Lenkiewicz, P., P. Wittenburg, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel (2011). Application of audio and video processing methods for language research. *Proceedings of the conference Supporting Digital Humanities 2011* (SDH 2011), Copenhagen, Denmark, November 17-18, 2011.