EmotionML: a language for embodied conversational agents

Enrico Zovato , Paolo Baggia, Davide Bonardo,

Loquendo S.p.A. – Italy

Some advanced conversational agents include the capability of handling emotions both at input and output level. They rely both on sensors that detect the user emotional state and on modules that provide coherent responses by generating the most appropriate emotional behaviour according to an affective strategy. Input processors can be based on the analysis of acustic features of the user speech, as well as on image processing and facial expression detection. At semantic level some analysis can also be performed so as to detect the valence of the user message. The emotional output can be also obtained through different modalities: text, speech, facial expressions and body gestures.
The presence of different modules dealing with emotion recognition and generation requires an efficient way to represent and exchange information. To this end, EmotionML [1] provides a suitable solution. In fact, the output of the emotion recognition modules could produce EmotionML documents that could be processed so as to obtain a single result for the emotion recognition task. On the other hand, the generation part of the system could be based on an EmotionML document which will drive modules, like for example speech synthesizers or the avatar renderer so as to simulate the intended emotional behavior.

We will consider a concrete case: the conversational agent developed within the EU funded Companions project [2]. The goal of the project is the development of a dialogue system in which emphasis is put on the affective aspects of the user-agent interaction. The Companions response to the user input is not simply an emotional backchannel, but comprises more sophisticated influencing strategies. The architecture of this system includes modules for Automatic Speech Recognition, Natural Language Understanding, Dialogue Management, Natural Language Generation, Speech Synthesis and Avatar Rendering. Beyond these modules, there are specific programs used to deal with emotions. In particular, there are a speech emotion recognizer and a semantic emotion recognizer for user input. Regarding the emotional output, the system relies on a processor for selecting the appropriate affective strategy.
As for emotion recognition, the final results is obtained through a multimodal affective fusion process between the two modalities of speech and text. The goal is to provide instantaneous emotion representation associated with an utterance but could also serve for temporal integration purposes. Emotional Speech Recognition is based on EmoVoice ™ [3] and produces as output discrete values on the basis of a valence-arousal dimensional model. The possible output labels are Positive-Active, Positive-Passive, Negative-Active, Negative-Passive and Neutral. The module that recognizes emotions at semantic level (Sentiment Analyzer), analyses the results provided by the speech recognizer and produces as output one of three labels: Positive, Negative or Neutral. The fusion module analyses these two outputs and applies some rules in order to get the most likely representation of the emotional state of the user at that particular time. The output is an EmoVoice category.

This process could be effectively managed by exploiting EmotionML. For example, the two modules involved in the emotion recognition could produce two documents that will be parsed by the fusion module:

```
<emotionml xmlns="http://www.w3.org/2009/10/emotionml"
    dimension-set="http://www.example.com/emotion/dimension/FSRE.xml">
  <emotion start=1268647200 modality="voice">
    <!-- Positive-Active -->
    <dimension name="valence" value="1.0" confidence="0.6"/>
    <dimension name="arousal" value="1.0" confidence="0.6"/>
    <reference uri="asr_output.wav#t=26,98"/>
  </emotion>
</emotionml>

<emotionml xmlns="http://www.w3.org/2009/10/emotionml"
  <emotion start=1268647200 modality="text">
    <!-- Positive -->
    <dimension name="valence" value="1.0" confidence="0.9"/>
    <reference uri="asr_output.wav#t=26,98"/>
  </emotion>
</emotionml>
```

Of course some values have to be choosen for dimensions. In simple cases like these in which only four values are considered in the arousal/valence space, corresponding to the four quadrants, only the values 0, 0.5 and 1 could be used: for example the couple 1,1 stands for Positive-Active and 0,0 for the opposite situation: Negative-Passive.
Since the output of the Sentiment Analyzer has prooven more reliable than the results provided by the emotion speech recognizer, confidence values are set accordingly. Confidence is therefore an important element and is particularly suitable for weighting different input modalities.
The two documents will be processed so as to detect the most likely emotional state. Th eoutput of this process could easily be another EmotionML document including a single emotion tag.

As far as the output of the whole system is concerned, specific modules to drive the generation of emotional behaviour are included in the architecture. The information that is sent to the multimodal output manager is actually composed of labels from three categories. The first one refers to the agent's performative like for example *suggest, propose, warn, agree, criticize, advice, confirm, incite, inform, greet, wait.* The second tag is the affective label. Affective category labels are selected from the OCC model of emotion [4] and include values like *neutral, embarrassment, happy-for, relief.* The last element, emphasis, is represented by means of one of the three lables *weak, medium* or *strong.*
These descriptors are then used by a specific processor to produce the most appropriate actions in terms of avatar rendering and speech synthesis, on the basis of available emotional features of these modalities.
Also in this case EmotionML could be effectively used to represent the target emotion, like in the example below, where a "default" vocabulary for emotion categories is choosen:

```
<emotionml xmlns="http://www.w3.org/2009/10/emotionml"
    category-set="http://www.example.com/emotion/category/OCC.xml">
  <emotion>
    <category  name="happy-for"/>
    <intensity value="0.5"/>
  </emotion>
</emotionml>
```

The three Emphasis labels can be represented though the <intensity> element of EmotionML. Although intensity is not specified through discrete values, the three labels *weak, medium* and *strong* can be easily mapped onto numerical values included in the range 0.0-1.0.
The EmotionML document could also be used in conjunction with other markup languages, like for examples SSML for speech synthesis. EmotionML could be in fact embedded in a SSML document, thus providing complementary information for producing expressive speech synthesis.

Conclusions

In this paper we have described how the EmotionML specification could be exploited in a real case, an affective dialogue system, to appropriately work with emotion representations. As from the last working draft, the EmotionML is flexible in describing emotions according to suggested vocabularies derived from existing models of emotions. There's still however a significant gap between what technologies are capable of reaching in the fields of emotion recognition and generation and the potentiality of EmotionML in terms of emotion description.

References

[1] http://www.w3.org/TR/2010/WD-emotionml-20100729/

[2] http://www.companions-project.org

[3] T. Vogt, E. André and N. Bee, "EmoVoice – A framework for online recognition of emotions from voice," in Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, 2008.

[4] Ortony, A., Clore, G. L., & Collins, A. (1988). The Cognitive Structure of Emotion. Cambridge, UK: Cambridge University Press.