Exploring the Compositionality of Emotions in Text: Word Emotions, Sentence Emotions and Automated Tagging

Virginia Francisco and **Pablo Gervás**
Natural Interaction based on Language
Facultad de Informática
Universidad Complutense de Madrid
28040 Madrid, Spain
virginia@fdi.ucm.es, pgervas@sip.ucm.es

Abstract

This paper presents an approach to automated marking up of texts with emotional labels. The approach considers the representation of emotions as emotional dimensions. A corpus of example texts previously annotated by human evaluators is mined for an initial assignment of emotional features to words. This results in a List of Emotional Words (LEW) which becomes a useful resource for later automated mark up. An algorithm for the automated mark up of text is proposed. This algorithm mirrors closely the steps taken during feature extraction, employing for the actual assignment of emotional features a combination of the LEW resource, the ANEW word list, and WordNet for knowledge-based expansion of words not occurring in either. The algorithm for automated mark up is tested against texts from the original samples used for feature extraction to test its correctness. It is also tested against new text samples to test its coverage. The results are discussed with respect to two main issues: correctness and coverage of the proposed algorithm, and additional techniques and solutions that may be employed to improve the results.

Introduction

The task of annotating text with specific labels indicating its emotional content or inclination is fundamental for any attempt to make computer interfaces respond in some way to the affective nature of the content they are handling. This is particularly true for research attempts to produce synthesised voice with different emotional states, but it may also be applicable in other contexts, such as multimodal presentation, where colours, typography or similar means can be used to convey emotion.

A comprehensive definition of emotion must take in account the conscious feeling of the emotion, the processes that appear in the nervous system and in the brain and the expressive models of the emotion (Izard 1971). Two issues must addressed when experimenting in this field: to obtain a corpus of emotionally annotated texts to act as reference data, and to decide on a particular representation of emotion. There are different methods for representing emotions in research (Cowie & Cornelius 2003):

- Emotional categories: Description of emotions by the use of emotion-denoting words, or category labels.
- Descriptions based on psychology: The appraisal of a stimulus determines the significance of stimulus for the individual, and triggers and emotion as an appropriate response (Alter *et al.* 2000).
- Descriptions based on evaluation: These theories described the emotions from the point of view of the evaluations involves.
- Circumflex models: Emotional concepts are representd by means of a circular structure (Russell 1980). When two emotional categories are close in the circle represents the conceptual similarity of these two categories.
- Emotional dimensions: Emotional dimensions represent the essential aspects of emotion concepts. Evaluation (positive/negative) and activation (active/passive) are the main dimensions; sometimes they are complete with the power dimension (dominant/submissive).

For the particular purpose contemplated in this paper we have chosen emotional dimensions because this approach allows measurement of the similarity between different emotional states and because of the relative arbitrarily in naming the dimensions.

The aim of this work is to present an approach to emotional tagging and analizing the results obtained with it when marking up texts of a particular domain - simple version of children fairy tales. The last section discusses some ideas we are working on to improve these results.

Marking up text with emotinal dimensions

This section presents a brief review of previous work on the labelling of texts with emotions. Another important decision when annotating text with emotions is which the particular approach to be used to relate emotions and textual elements.

Existing approaches to annotation

Existing approaches can be grouped in five mainly categories: keyword spotting, lexical affinity, statistical natural language processing, approaches based on large-scale real-world knowledge and hand-crafted methods.

- Statistical natural language processing: This method involves feeding a machine learning algorithm a large training corpus of text marked-up with emotions (Goertzel *et al.* 2000). These techniques are usually semantically weak because lexical elements which not are obvious affect keywords have little predictive value.
- Hand-Crafted method: This method involves modelling emotional states in terms of hand-crafted models of affect based on psychological theories about human needs, goals, and desires. This requires a deep understanding and analysis of the text. The difficulty with this approach is that it is very difficult to generalize. An example of this approach is Dyer's DAYDREAMER (Dyer 1987).
- Keyword spotting: Text is marked up with emotions based on the presence of emotional words like ”angry”, ”sad” . . . The disadvantages of this approach are two: errors in the marked up when negation is involved and reliance on obvious emotional words. Examples of this approach are the Ortony's Affective Lexicon (Ortony, Clore, & Collins 1988) or the ANEW word list (Bradley & Lang 1999).
- Lexical affinity: This method is more sophisticated than the previous. This technique not only detects affective words but also assigns arbitrary words a probability of indicating different emotions. These probabilities are usually obtained from a corpus. The weaknesses of this approach are mainly two: it can easily have problems with negation and it is difficult to develop a reusable model becuase words and affinity are obtained from a corpus.
- Approach based on large-scale real-world knowledge: Rather than looking at surface features of the text this method evaluates the affective qualities of the underlying semantic content of text. Relies on having large-scale real-world knowledge about people's common affective attitudes toward situations, things, people and actions (Liu, Lieberman, & Selker November 2002). There are three large-scale generic knowledge bases of commonsense which can be used for this purpose: Cyc (Lenat 1995), Open Mind Common Sense (OMCS) (Singh 2002) and Thought Treasure (Mueller 1998). An example of this approach is the Liu's Model of Textual Affect Sensing using Real-World Knowledge (H.Liu, Lieberman, & Selker 2003).

Parts of the texts annotated

On deciding the parts of the text which are going to be marked with emotions there are different options: word, phrase, paragraph, chapter. The easier approach is classified sentences into one of the emotions, the emotional structures in this case are phrases. Another approach more sophisticated is to combine into large units the affectively marked sentences using an algorithm. Boundaries between larger regions of text can be determined using layout structure (paragraph, scene, chapter breaks . . .) or discourse cues (keywords and phrases which denote a break in the discourse).

EmoTag

EmoTag mark up texts with the three emotional dimensions: valence, arousal and dominance. EmoTag relies on a dictionary of word to emotion assignments. This is obtained from a corpus of human evaluated texts by applying language analysis techniques. Similar techniques are later applied to assign emotions to sentences from the assignments for the words that compose them.

Construction of the Dictionary

This section deals with the process of building two basic resources for emotional mark up: a corpus of fairy tale sentences annotated with emotional information, and a list of emotional words (LEW). Both the corpus and the list of emotional words are annotated with emotional dimensions (valence, arousal and dominance).

The method we are going to use for the mark up follows an approach which mixes keyword spotting and lexical affinity in the hope that the weaknesses of each individual approach are reduced by their combination. Based on a large corpus of marked up emotional text we have obtained a list of words and their relation with emotions (LEW). When we mark up a text we look for every word in this first list. If the word is not in our list we look in the ANEW word list (Bradley & Lang 1999), which is a list obtained by means of a keyword spotting. If the word is in none of the two lists we try to obtain from an ontology (WordNet) a word related to it which is in one of our emotional word lists.

In the following sections we describe in detail how we have obtained the list of emotional words (LEW) and how our approach works.

Corpus Annotation Method If we want to obtain a program that marks up texts with emotions, as a human would, we need first a corpus of marked-up texts in order to analyze and obtain a set of key words which we will use in the mark up process. Each of the texts which forms part of the corpus may be marked by more than one person because assignment of emotions is a subjective task so we have to avoid ”subjective extremes”. In order to do that we obtain the emotion assigned to a phrase as the average of the mark-up provided by different persons. Therefore the process of obtaining the list of emotional words involves two different phases: first several persons mark up some texts of our corpus, then from the mark-up texts of the previous phase we obtain emotional words.

As a working corpus, we selected eight popular tales, with different lengths, written in English. Tales are split into sentences and evaluators are offered three boxes for each sentence in which to put the values of the emotional dimensions: valence, arousal and dominance. In order to help people in the assignment of values for each dimension we provide them with the SAM standard. SAM figures comprise the bipolar scales of each emotional dimension (Lang 1980) as can be seen in the Figure 1.

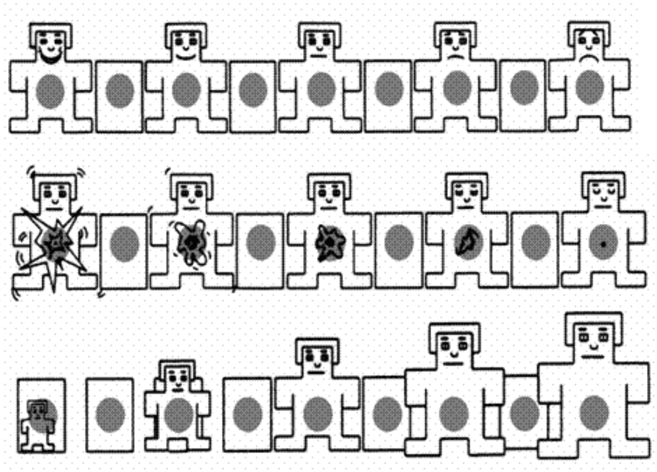

Figure 1: Dimensional scales according with SAM: valence, arousal and dominance

Extraction method for a list of emotional words Based on the tales marked up by different persons we obtain a data base of words and their relation with emotional dimensions.

Firstly we split text into phrases and we obtain for every phrase the representation of the emotional content. Phrases are processed with the qtag [1] tagger which for each word returns the part-of-speech (e.g. noun, verb, etc.). Every sentence is divided into words and with every word and its label we carry out the following process:

- Check if the label is in the list of stop labels, if is in there we leave it out. Our stop list is composed of the following labels: verbs ”to be”, ”to do”, ”to have” and all their conjugations, conjunctions, numbers, determiners, existential there, prepositions, modal auxiliary (might, will), possessive particles, pronouns, infinitive marker (to), interjections, adverbs, negative markers (not, n’t), quotation mark, apostrophe. . .
- If the label is not in the stop list we proceed to extract the stem of the word using a slightly modified version of the Porter stemming algorithm (Porter). The modifications are intended to ensure that no information relevant to the current process is lost during the stemming process.
- Once we have the stem of the word it is inserted into our word data base. If the word was already in our list we add up the new values to the ones we had.
- Once all the tales have been processed we carry out a normalization and expansion process of our list of words. We extend our list with synonyms and antonyms of every word which are looked up in WordNet (Miller 1995). This process looks up all the synonyms and antonyms for every word in the list, and for each of them the additional words are inserted in our list. We divide the numeric value we have for each of the three dimensions: valence, arousal and dominance, by the number of appearances of the word in the texts, to work out the average value of each dimension for each word. For inserting related words into the database, the same values of dimensions as the original word are used for synonyms and the opposite value is used in the case of the antonyms (9 - original value).

A Method for Automated Mark Up of Emotions

Our process classifies sentences into emotions, the first step is to split the tales into sentences and split each sentence into words in order to carry out our process based in the relation between words and different emotions. In order to obtain the value of each of the emotional dimensions of the sentence we look up every word of the sentence and assign to it a value for the three dimensions as given by our lists. Based on these values of the words we obtain the final value of the sentence. The process is explained next:

- By means of the tagger qtag, mentioned in the previous section, we obtain the tag for every word in the sentence; based on these tags and words we decide the emotion of the sentence.
- If the tag associated to the word is in our label stop list we leave it out.
- If the tag is not in our stop list we get the stem of the word by means of the modified Porter stemming algorithm mentioned before.
- Once we have the stem and the tag of the word that we want to classify, we look it up in the lists of emotional words(LEW). If the word is present we get the value of each of the three dimensions.
- If the word is not in any of the lists available we obtain the hypernyms of the word from WordNet, and we look them up in the available lists (first LEW, then ANEW); the first appearance of a hypernyms is taken and the emotional content associated to the hypernyms is associated to our original word.
- If none of the hypernyms appear in the available lists, the word does not take part in the process.

[1] http://www.english.bham.ac.uk/staff/omason/software/qtag.html

- Once all the words of the sentences have been evaluated, we add up the value of each dimension of the different words and assign to the sentence the average value of valence, arousal and dominance, that is, we divide the total value of each dimension by the number of words which have taken part in the process

A sample part of a marked tale:

<valence=5.43 arousal=5.34 dominance=4.87> *A Fox once saw a Crow fly off with a piece of cheese in its beak and settle on a branch of a tree.* </valence=5.43 arousal=5.34 dominance=4.87>
<valence=5.31 arousal=5.37 dominance=4.90> *That's for me, as I am a Fox, said Master Reynard, and he walked up to the foot of the tree.* </valence=5.31 arousal=5.37 dominance=4.90>
...
<valence=4.29 arousal=5.54 dominance=5.56> *That will do, said he.* </valence=4.29 arousal=5.54 dominance=5.56>
<valence=4.45 arousal=5.62 dominance=4.79> *That was all I wanted.* </valence=4.45 arousal=5.62 dominance=4.79>
<valence=4.68 arousal=5.95 dominance=5.28> *In exchange for your cheese I will give you a piece of advice for the future:* </valence=4.68 arousal=5.95 dominance=5.28>
<valence=5.94 arousal=5.25 dominance=5.98> *Do not trust flatterers.* </valence=5.94 arousal=5.25 dominance=5.98>

Evaluation

In order to evaluate our work we carried out some tests, in these tests four tales are going to take part, two of them have been in our original corpus, the corpus we have used to obtain our LEW list and the other two are new tales, which did not take part in our extraction method. This way we will measure on the one hand how well our process marks the tales from which we have obtained our LEW list and on the other hand how well our approach works with tales that have not been involved in our extraction process. The results obtained are explained in the following section.

The data on emotional dimensions we have available for each tale are the values that each dimension takes for each sentence. To evaluate our tagger we have divided the evaluation according to the different dimensions: valence, arousal and dominance. In order to get a measure of our tagger we have take measures first from the evaluators' tales and then from our tagger's tales.

- Evaluators' tales: As reference data, we have the values assigned for each dimension and each sentence by the human evaluators. An average emotional score for each dimension of a sentence is calculated as the average value of those assigned to the corresponding dimension by the human evaluators. The deviation among these values is calculated to act as an additional reference, indicating the possible range of variation due to human subjectivity. The average deviation between evaluators is 1.5. Figure 2 shows the average deviation of evaluators in each of the tales mark up by them.

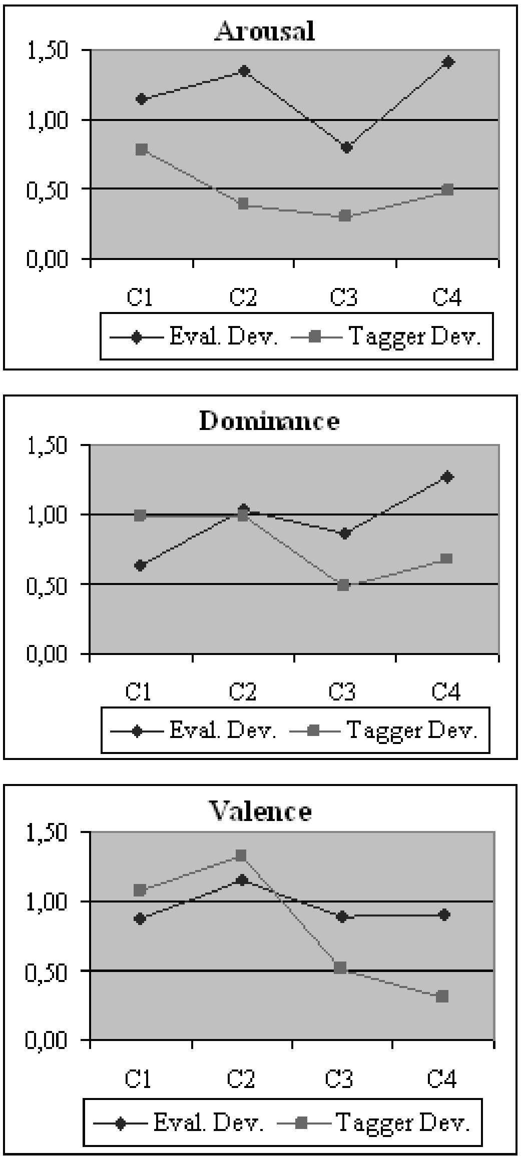

Figure 2: Evaluator deviation and tagger deviation for different emotional dimensions

- Tagger's tales: For each dimension, to determine if a sentence is tagged correctly we have compared the deviation of the tagger with respect to the average score against the average deviation among the human evaluators. If the deviation of the tagger is less or equal to the average deviation among evaluators, we consider that the sentence is tagged correctly. The average deviation of the tagger stands at 1.5 for the valence dimension, 0.75 for the arousal dimension, and 1 for the dominance dimension. This seems to indicate that the tagger is obtaining better results in terms of deviation from the average obtained by humans for the arousal and dominance dimensions, and comparable results in the case of valence. The actual values are shown in the graphs given in Figure 2, where the average values of the various deviations are plotted against the four tales that have been evaluated.

The graph in Figure 3 shows the percentage of success - the percentage of sentences in which the deviation of the automatically tagged dimensions from the human average is within the deviations observed between human evaluators. Looking at the graph we can see the greatest percentages of success are reached in the last two tales, which are the ones that were used to obtain the list of emotional words LEW.

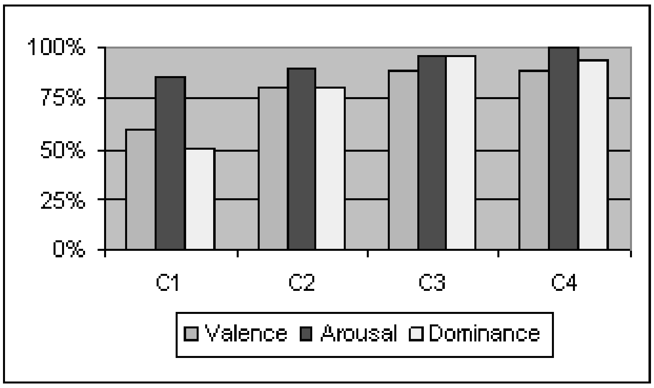

Figure 3: Percentage of success in automated tagging for the different dimensions of the evaluated tales

With respect to the percentage of success we can conclude that the best results are obtained with the tales which took part in our extraction method (C3 and C4). Analysis of the sentences that were tagged incorrectly indicates that most of them are either very long, include negations, or correspond to sentences with very high deviation between human evaluators - no consensus among human evaluators.

An Application of EmoTag: Emotional Story Tellers

Once we have our texts marked up with emotions, what can we do with it? In this section we are going to explain our goals and applications of EmoTag in different fields.

Voice generated by speech systems is not as natural as human voice and that it is the main cause of rejection by the public. An important challenge for text to speech systems is to generate voice more natural with different emotional states. We have developed an emotional synthesizer which take into account five differents emotions (sad, happy, fear, angry and surprise), in this synthesizer emotions are classified according to emotional categories it will be interesting modifiying it in order to use emotional dimensions, that way the texts marked up by EmoTag could be read by the synthesizer. Why emotional dimensions could be better than emotional categories? In a tale, an emotional state may build up rather gradually, and may change over time as the interaction moves on. Consequently, a speech synthesis system should be able to gradually modify the voice in a series of steps towards an emotional state. In addition, it seems reasonable to assume that most of the dialogues between characters will require the machine not to express only extreme emotional states. Emotional dimensions are a representation of emotional states which are naturally gradual, and are capable of representing low-intensity as well as high-intensity states. While they do not define the exact properties of an emotional state in the same amount of detail as a emotional category, they do capture the essential aspects of the emotional state. Another important issue to take into account is that the representation of emotions in terms of emotional dimensions is much better suited for translation into the kind of parameters required by a speech synthesizer (pitch, volumen, rate . . .).

The narration of the tale could be livening up with some music, of course emotional music, that is music depends on the marking up of the tale, if the value of the activation dimension is high (around 9) the music in this region will be very lively and on the other hand if the value of activation is low (around 1) the musice will be very slow.

Once we have the music and the emotional voice of the narrator and characters it leaves only the aparence of the screen during the story telling. Based on marked up emotions the screen could change his colour from black (the tale's region is very sad) to white (the tale's region is neutral) and the text could be animated with some emotional icons.

Conclusions

Our method for marking emotions uses ideas from two of the main existing methods for marking texts with emotions: keyword spotting and lexical affinity. Our aim was to combine the advantages of each method in a way that avoided their disadvantages. The fact that we have considered words in a context instead of individually reduces some of the disadvantages associated with simple keyword spotting, because the same word may have different meanings in different contexts. Some issues related to context still need further work. Negation, for instance, may have the effect of inverting the polarity of the emotional content of words under its scope. We are working in include in EmoTag the use of Minipar (Lin May 1998), which is a dependency-based method for parser evaluation, to determine the scope of negations appearing in the sentences, in order to take their effect into account, both when computing word emotion from sentence emotion and viceversa.

With respect to methods based on lexical affinity we have reduced the dependency on a given corpus by resorting to two different data bases: LEW (corpus dependent) and ANEW (corpus independent). We have also complemented our data base of emotional words with synonyms, antonyms, and hypernyms. Nonetheless, we still get better results for the tales used to obtain the LEW corpus than for new tales, so we consider necessary to continue exploring better solutions for this problem.

Aside from these issues requiring improvement, we have observed that very long sentences lead to confusion when assigning emotions. In future versions we will consider a finer granularity for representing sentences. Another problem was the large observable disagreement between human evaluators. This may be reduced by carrying out experiments with a larger number of evaluators, and by introducing metrics to keep track of it.

References

Alter, K.; Rank, E.; Kotz, S.; Toepel, U.; Besson, M.; Schirmer, A.; and Friederici, A. 2000. Accentuation and emotions - two different systems? In *Proceedings of the ISCA Workshop on Speech and Emotion*, 138–142.

Bradley, M., and Lang, P. 1999. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. technical report c-1. Technical report, The Center for Research in Psychophysiology, University of Florida.

Cowie, R., and Cornelius, R. 2003. Describing the emotional states that are expressed in speech. In *Speech Communication Special Issue on Speech and Emotion*.

Dyer, M. 1987. Emotions and their computations: Three computer models. *Cognition and Emotion* 1:323–347.

Goertzel, B.; Silverman, K.; Hartley, C.; Bugaj, S.; and Ross, M. 2000. The baby webmind project. In *Proceedings of AISB*.

H.Liu; Lieberman, H.; and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of IUI*.

Izard, C. 1971. *The face of emotion*. New York: Appleton-Century-Crofts.

Lang, P. 1980. Behavioural treatment and bio-behavioural assessment: Computer applications. In Sidowski, J.; Johnson, J.; and Williams, T., eds., *Technology in mental health care delivery systems*, 119–137. Norwood, NJ: Ablex Publishin.

Lenat, D. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11):33–38.

Lin, D. May 1998. Dependency-based evaluation of minipar. In *Proc. of Workshop on the Evaluation of Parsing Systems*.

Liu, H.; Lieberman, H.; and Selker, T. November, 2002. Automatic affective feedback in an email browser. Technical report, MIT Media Lab Software Agents Group Technical Report.

Miller, G. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38:39–41.

Mueller, E. 1998. Thoughttreasure: A natural language/commonsense platform.

Ortony, A.; Clore, G.; and Collins, A. 1988. *The cognitive structure of emotions*. New York: Cambridge University Press.

Porter, M. An algorithm for suffix stripping. In *Readings in information retrieval*, 313–316. San Francisco, CA, USA.: Morgan Kaufmann Publishers Inc. A.

Russell, J. 1980. A circumflex model of affect. *Journal of Personality and Social Psychology* 39:1161–1178.

Singh, P. 2002. The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium*. Palo Alto, CA: AAAI.