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 employs 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 and against new text samples to test its coverage. The results and additional techniques and solutions that may be employed to improve the results are discussed.

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 be 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), the most important for our study are: *Emotional categories* - which involve a description of emotions by the use of emotion-denoting words, or category labels -, and *Emotional dimensions* - specific dimensions are used to represent the essential aspects of emotion concepts: evaluation (positive/negative), activation (active/passive) and sometimes power (dominant/submissive). Other methods are: *descriptions based on psychology*, *descriptions based on evaluation*, *circumflex models* ...

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 emotional dimensions

Existing approaches can be grouped in five main categories (H.Liu, Lieberman, & Selker 2003): keyword spotting, lexical affinity, statistical natural language processing, approaches based on large-scale real-world knowledge and hand-crafted methods. Our proposed method is based mainly in two of these categories: *Keyword spotting* which marks up text with emotions based on the presence of emotional words like "angry" or "sad". The disadvantages of this approach are: errors in the mark up when negation is involved and reliance on obvious emotional words. An example of this approach is the ANEW word list (Bradley & Lang 1999). *Lexical affinity* not only detects affective words but also assigns arbitrary words a probability of indicating different emotions. These probabilities are usually obtained from a corpus. This approach can easily have problems with negation and it is difficult to develop a reusable model because words and affinity are obtained from a corpus.

On deciding the parts of the text which are going to be marked with emotions there are different options: word, phrase, paragraph, chapter ... The simplest approach is classified sentences into one of the emotions. Another more sophisticated approach is to combine into large units the affectively marked sentences using an algorithm. Boundaries between larger regions of text can be determined using lay-

out structure (paragraph, scene, chapter breaks ...) or discourse cues (keywords and phrases which denote a break in the discourse).

EmoTag

EmoTag, which is the system described in this paper mark up texts with the three emotional dimensions: valence, arousal and dominance. EmoTag relies on a dictionary of word to emotion assignments - the LEW list of words. 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.

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.

The Corpus and the LEW Dictionary

A corpus of texts is marked-up with emotions by human evaluators and then analyzed in order to obtain a set of key words which we will use in the mark up process. 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). 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". 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 from our corpus, then from the mark-up texts of the previous phase we obtain emotional words.

In order to process computationally the information at word level - both when extracting emotional words from text and when tagging text with emotions - sentences must be submitted to a basic procedure of lexical tagging, stemming, and filtering by means of a stop list. Lexical tagging is carried out by means of the qtag [1] tagger. Filtering is done using list of stop POS tags. Our stop POS tags are composed of tags such as: verbs "to be", "to do", "to have" and all their conjugations, conjunctions, numbers... If the label is not in the stop POS tags we proceed to extract the stem of the word using a slightly modified version of the Porter stemming algorithm (Porter).

Based on the tales marked up by different persons we obtain a data base of words and their relation to emotional dimensions. When extracting emotional words from the corpus, each word - already labelled, filtered, and stemmed - is inserted into our word data base with the values for emotional dimensions assigned during mark up to the sentence it comes from. 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. For the normalization 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. Then 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. 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 (value antonym = 9 - original value).

A Method for Automated Mark Up of Emotions

Our process assigns emotions to sentences based on the relation between the words in the sentence and different emotions. Sentences are first labelled, filtered and stemmed as described above. Emotions are then assigned to each resulting word as follows: first each word is looked up in the lists of emotional words(LEW). If the word is present we get the value of each of the three dimensions. If this lookup fails, the ANEW list (Bradley & Lang 1999) is consulted, and the corresponding values used in case of success. The second step is, if the word is not in any of the lists available, to obtain the hypernyms of the word from WordNet, and to 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. Once all the words of the sentences have been evaluated, the third step is to 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 of a marked tale is given in Table 1.

Evaluation

In order to evaluate our work we carried tests over four tales: two that been in the corpus used to obtain the LEW dictio-

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

<emo 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. </emo>
<emo 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. </emo>
...
<emo 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: </emo>
<emo valence=5.94 arousal=5.25 dominance=5.98> Do not trust flatterers. </emo>

Table 1: Marked Up Portion of a Tale

nary and two new tales which did not take part in our extraction method. This allows us to measure how well our process marks the tales from which we have obtained our LEW list and how well our approach works with tales that have not been involved in our extraction process.

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 taken measures first from the tales tagged manually by evaluators and then from the tales tagged automatically by the tagger.

For tales tagged manually by evaluators we have used as reference data 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 1 shows the average deviation of evaluators in each of the tales mark up by them.

For tales tagged automatically by the tagger, in order to determine if each dimension of 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 than 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 1, where the average values of the various deviations are plotted against the four tales that have been evaluated.

The graph in Figure 2 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.

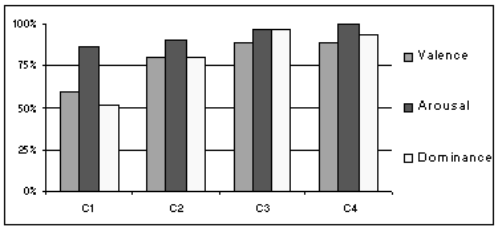

Figure 2: Percentage of success in automated tagging for the different dimensions of the evaluated tales

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

We have developed an emotional synthesizer which take into account five different 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. 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 an 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 . . .).

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 to simple keyword spotting, because the same word may have different meanings in different con-

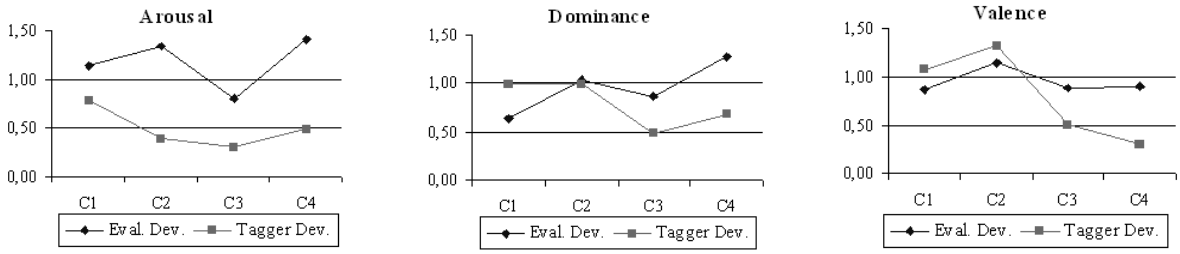

Figure 1: Evaluator deviation and tagger deviation for different emotional dimensions

texts. 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.

Some issues 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 on including 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. Additionaly, we have observed that very long sentences lead to confusion when assigning emotions. The use of MINIPAR will allow us to 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.

A good improvement will be consider only the words of the sentence which really have an emotional intention. EmoTag currently considers all the words of the sentence as affective words (excluding only words with a tag present in our stop POS tags list). A resource such as General Inquirer lexicon (Stone *et al.* 1966) may be used to identify emotional words. Another important improvement may be to replace WordNet with WordNet Affect (Strapparava & Valitutti May 2004), which is an extension of WordNet with "affective domains labels".

Acknowledgements

This work has been supported by the Spanish Committee of Science & Technology, Acción Integrada Hispano-Portuguesa (HP2003-0068), and partially supported by the Spanish Ministerio de Educación y Ciencia, project TIN2005-09382-C02-01.

References

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*.

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.

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

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

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

Stone, P. J.; Dunphy, D. C.; Smith, M. S.; and Ogilvie, D. M. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press.

Strapparava, C., and Valitutti, A. May 2004. WordNet-Affect: an affective extension of WordNet. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.