Automated Mark Up of Affective Information in English Texts

Virginia Francisco and Pablo Gervás

Departamento de Sistemas Informáticos y Programación
Universidad Complutense de 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 in parallel two possible representations of emotions: as emotional categories and emotional dimensions. For each representation, 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. The proposed algorithm for automated mark up of text 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 and the results are discussed with respect to three main issues: relative adequacy of each one of the representations used, correctness and coverage of the proposed algorithm, and additional techniques and solutions that may be employed to improve the results.

1 Introduction

The present paper proposes a method for automated tagging of texts with emotions. Before marking text with emotions we have to decide which emotions we are going to deal with. Then we need a corpus of emotional marked text in order to analyse how people mark text and how our program will have to do it. Based on this corpus we obtain a list of "emotional words" and the relation between these words and the emotions we are modelling. Finally, with this relation between words and emotions we model a method for marking up every text.

For the study of emotional texts we need to decide which emotions we are going to model, and how we are going to represent them. There are different methods in order to research emotions [1]: *emotional categories* - based on the use of emotion-denoting words -, *descriptions based on psychology* [2] and *evaluation* [1], *circumflex models* - emotional concepts are represented by means of a circular structure [3], so that two emotional categories close in the circle are conceptually similar of them. - and *emotional dimensions* which represent the essential aspects of emotion concepts: evaluation (positive/negative) and activation (active/passive) are the main dimensions, sometimes they are augmented with the power dimension (dominant/submissive).

2 Labelling Text with Emotions

There are several issues that need to be taken into account when attempting to mark up a document with emotions: the granularity to be employed, and the particular approach to be used to relate emotions and textual elements.

On deciding the parts of the text which are going to be marked with emotions there are different options [4]: word, phrase, paragraph, chapter ... One of the simplest approaches is to use sentences as the emotional structures. Another solution is to combine the sentences into larger units using an algorithm to summarise the affect of text over multi-sentence regions (winner-take-all scheme, Bayesian networks ...).

Existing approaches to emotional mark up can be grouped in four main categories [4]: *keyword spotting* [5] - text is marked up with emotions based on the presence of affect words -, *lexical affinity* - not only detects affective words but also assigns arbitrary words a probability, obtained from a corpus, of indicating different emotions -, *statistical natural language processing* [6] - this method involves feeding a machine learning algorithm a large training corpus of text marked-up with emotions -, *an approach based on large-scale real-world knowledge* [4] - this method evaluates the affective qualities of the underlying semantic content of text -.

The method we are going to use mixes keyword spotting and lexical affinity in the hope that the weaknesses of each individual approach are reduced by their combination. The disadvantages of keyword spotting approach are two: poor recognition of emotion when negation is involved and reliance on surface features. On the other hand the weaknesses of lexical affinity are: it is based only in the word-level, so it can easily have problems with negation; and lexical affinity is obtained from a corpus, which makes it difficult to develop a reusable, domain-independent model.

3 Building Emotion-Annotated Resources

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 two methods: using the emotional dimensions (valence, arousal and dominance) and marking its with emotional categories (happy, sad, angry ...). In the following sections we describe in detail how we have obtained the list of emotional words (LEW) and how our approach works.

3.1 Corpus Annotation Method

If we want to obtain a program that marks up texts with emotions, as a human would, we first need 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". To be precise 15 evaluators have marked each of the tales. 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 people mark up texts of our corpus, then from the marked up texts of the previous phase we obtain emotional words. As a working corpus, we selected 8 popular tales, with different lengths (altogether they result in 10.331 words and 1.084 sentences), written in English. In order to evaluate the tales we have split the experiment in two phases:

– Phase 1: We selected four different tales for each evaluator in order to mark them up with emotional dimensions. 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 [7] which can be seen in the Figure 1.

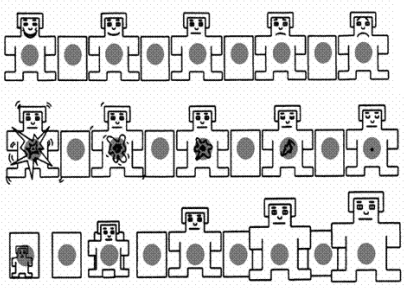

Fig. 1. Dimensional scales according to SAM: valence, arousal and dominance

– Phase 2: The four tales which did not take part in phase 1 are used in this second phase. At this point the mark up of the tales uses emotional categories: happy, sad, anger, surprise ... In order to help them in the assignment of emotional categories we provide a list of different emotional labels.

3.2 Extraction Method for a List of Emotional Words

Based on the tales marked up by human evaluators we obtain a data base of words and their relation with emotional dimensions and categories. There are two different methods of extraction depending on the labeling method: emotional dimensions or emotional categories. For each extraction method, there is a generic part which is common to both, and a specific part which involves handling the particular representation of emotion in each case. The common part can be described as follows.

First we split the text into phrases and we obtain for every phrase the particular representation of the emotional content involved in each case. 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:

- If the label is in the list of stop POS tags we leave it out. Our stop list is composed of labels such as: conjunctions, numbers, determiners ...
- 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 [8].
- Once we have the stem of the word it is inserted into our data base with the particular representation of the emotional content involved in each case.
- After processing all the tales, 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 [9]. The extraction method results in a list of 3.027 emotional words.

Emotional dimensions In the case of emotional dimensions, the particular representation of the emotional content involved in a given phrase corresponds to the three emotional dimensions assigned to it. If the word was already in our list we add up the new values to the ones we had. In order to obtain the average value of the dimensions we divide the numeric value we have for each of the three dimensions, 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).

Emotional categories In the case of emotional categories, the particular representation of the emotional content involved in a given phrase corresponds to the emotion assigned to every phrase by most of the evaluators. When we have the stem of the word it is inserted into our word data base with the value 1 in the field of the emotion assigned to the phrase in which the word was. If the word was already in our list we add up 1 to the field of the phrase's emotion. In order to obtain the average value of the emotions we divide the numeric value of each of the emotions by the number of appearances of the word in the texts, to work out the probability of that word indicating the emotions we are studying. For inserting related words into the database, the same probabilities of the original word are used in the case of synonyms case and the opposite probability in the case of antonyms (1- original probability).

4 A Method for Automated Mark Up of Emotions

Our process classifies sentences into emotions. The first step is to perform sentence detection and tokenization in order to carry out our process based on the

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

relation between words and different emotions. We have created two mark-up processes, the first one marks up tales with emotional dimensions and the second one with emotional categories. This process also has a common part for each representation method. The common part of the mark up method is:

- By means of the tagger qtag, mentioned in the previous section, we obtain the tag for every word in the sentence; If the tag associated with the word is in our list of stop POS tags we leave it out.
- We get the stem of the word by means of the modified Porter stemming algorithm mentioned before.
- We look the pair stem-tag up in the lists of emotional words available for the specific representation method. If the word is present we get the particular representation of its emotional content.
- 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; the first appearance of a hypernym is taken and the emotional content associated to the hypernym 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.

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. For emotional dimensions, the words are looked up first in the list of emotional words (LEW) obtained from the annotation experiments. If the word is not in LEW we look up for it in the ANEW word list [10]. 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.

In order to obtain the emotion associated to the sentence in the case of emotional categories we look around every word of the sentence in the LEW list and assign to it the probability of carrying the emotions we are studying. Based on these probabilities of the words we obtain the final emotion of the sentence. Once all the words of the sentences have been evaluated, we add up the probability of each emotion of the different words and assign to the sentence the emotion which has a bigger probability.

5 Evaluation

In order to evaluate our work we carried out two different tests. In these tests four tales are going to take part, two of them have been in the corpus we have used to obtain our LEW list and the other two are new tales. 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 tales which

take part in these tests are English popular tales with different number of words (from 153 words and 20 lines to 1404 words and 136 lines). Each of our four tales will be tagged first with the emotional dimensions, and then with the categories.

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.

For evaluator's tales we have, 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. Figure 2 shows the average deviation of evaluators in each of the tales (C1, C2, C3 and C4).

In the case of tagger's tales for each dimension, if the deviation of the tagger is less or equal to the average deviation among evaluators, we consider that the sentence is tagged correctly. Figure 2 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 graph in Figure 3 shows the success percentage - the percentage of sentences in which the deviation of the automatically tagged dimensions from the human average is within the deviations between human evaluators.

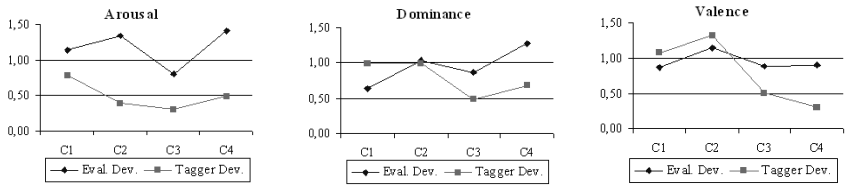

Fig. 2. Evaluator and tagger deviation for different emotional dimensions

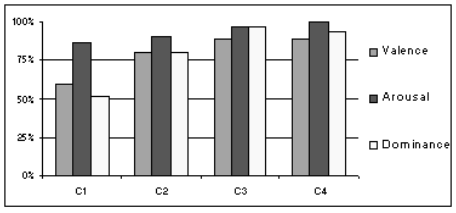

Fig. 3. Success percentage in automated tagging for the different dimensions

The data on emotional categories we have available for each tale are emotional label for each sentence. As we have done for emotional dimensions we have taken measures first from the evaluators' tales and then from our tagger's tales.

For evaluator's tales we have noticed that the percentage of sentences on which the majority of the human evaluators - half of their number plus one - agrees on the assignment of an emotion is very low, around 45%. This is an important data when it comes to interpreting the results obtained by our tagger. A reference value for the emotion of each phrase is obtained by choosing the emotion most often assigned to that sentence by the human evaluators.

In the case of tagger's tales the reference value obtained in the evaluator's tales is used to compare with the results generated by our tagger. The graph in Figure 4 shows the percentages of success obtained for each tale, each sentence has been considered successfully tagged if the emotion assigned by the tagger matched the reference value, and the relationship between the success percentage and the percentage of sentences whose reference value is supported by one more than half the number of evaluators.

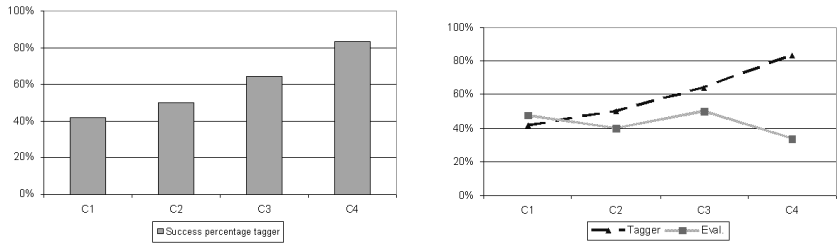

Fig. 4. Success percentage in automated tagging and relation between the success percentage of our tagger and majority-supported evaluators for emotional categories

Once human evaluators had finished with the mark up process they were asked to fill in a questionnaire about the evaluation process. Analysis of the results indicates that human evaluators find it easier to mark up tales with emotional categories than with emotional dimensions. However, if we look at the results we can see that values for different dimensions match reasonably well the categories assigned to sentences in the emotional categories approach.

With respect to the success percentage we can conclude that in both cases the best results are obtained with the tales which took part in our extraction method (C3 and C4). If we compare the results of the two approaches we can see that the best results are obtained with the emotional dimensions approach.

6 Conclusions

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 considering the use of shallow parsing techniques to determine the scope of negations appearing in the sentences, in order to take their effect into account.

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.

References

1. Cowie, R., Cornelius, R.: Describing the emotional states that are expressed in speech. In: Speech Communication Special Issue on Speech and Emotion. (2003)
2. Alter, K., Rank, E., Kotz, S., Toepel, U., Besson, M., Schirmer, A., Friederici, A.: Accentuation and emotions - two different systems? In: Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland (2000) 138–142
3. Russell, J.: A circumflex model of affect. Journal of Personality and Social Psychology **39** (1980) 1161–1178
4. H.Liu, Lieberman, H., Selker, T.: A model of textual affect sensing using real-world knowledge. In: Proceedings of IUI, Miami, Florida (2003)
5. Ortony, A., Clore, G., Collins, A.: The cognitive structure of emotions. Cambridge University Press, New York (1988)
6. Goertzel, B., Silverman, K., Hartley, C., Bugaj, S., Ross, M.: The baby webmind project. In: Proceedings of AISB. (2000)
7. Lang, P.: Behavioural treatment and bio-behavioural assessment: Computer applications. In Sidowski, J.B., Johnson, J.H., (Eds.), T.A.W., eds.: Technology in mental health care delivery systems, Norwood, NJ, Ablex Publishing (1980) 119–137
8. Porter, M.: An algorithm for suffix stripping. In: Readings in information retrieval, San Francisco, CA, USA., (Morgan Kaufmann Publishers Inc. A) 313–316
9. Miller, G.: Wordnet: a lexical database for english. Communications of the ACM **38** (1995) 39–41
10. Bradley, M., Lang, P.: 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 (1999)