Computational Intelligence, Volume 000, Number 000, 0000

EmoTag: An Approach to Automated Mark-Up of Emotions in Texts

VIRGINIA FRANCISCO
Departamento de Ingeniería del Software e Inteligencia Artificial.
Universidad Complutense de Madrid, Spain,
C/ Prof. José García Santesmases, s/n. 28040 Madrid (Spain)
E-mail: virginia@fdi.ucm.es

PABLO GERVÁS
Instituto de Tecnología del Conocimiento.
Universidad Complutense de Madrid, Spain,
C/ Prof. José García Santesmases, s/n. 28040 Madrid (Spain)
E-mail: pgervas@sip.ucm.es.

This paper presents an approach to the automated mark-up of texts with emotional labels. The approach considers two possible representations of emotions in parallel: emotional categories (emotional tags used to refer to emotions) and emotional dimensions (measures that try to model the essential aspects of emotions numerically). 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 algorithm proposed for the automated mark-up of text closely mirrors the steps taken during feature extraction, employing a combination of the LEW resource and the ANEW word list for the actual assignment of emotional features, and WordNet for knowledge-based expansion of words not occurring in either and an ontology of emotional categories. The algorithm for automated mark-up is tested and the results are discussed with respect to three main issues: the relative adequacy of each of the representations used, correctness and coverage of the proposed algorithm, and additional techniques and solutions that may be employed to improve the results. The average precentage of success obtained by our approach when it marks up with emotional dimensions is around 80% and when it marks up with emotional categories is around 50%. The main contribution of the approach presented in this paper is that it allows dimensions and categories at different levels of abstraction to operate simultaneously during mark-up.

Key words: Emotional Mark-up, Emotional Dimensions, Emotional Categories

1. INTRODUCTION

This work deals with the computational mark-up of the emotions that are present in a text, and it is included in what has come to be known as the field of *Sentiment Analysis*. Sentiment Analysis is part of the broader area of *Affective Computing* which aims to enable computers to recognize and express emotions (Picard, 1997). Initial work on Sentiment Analysis focused on the specific application of classifying reviews according to their polarity (positive or negative). However, nowadays the term Sentiment Analysis refers generally to the computational treatment of opinion, emotion and subjectivity in any kind of document.

Sentiment Analysis has advanced enormously in recent years. The first projects in the area were centered on beliefs (Carbonell, 1979). Later work has focused mostly on interpretation of metaphor, narrative, affect, point of view (Sack, 1994; Wiebe, 1994), and other areas related to these topics (Wiebe et al., 1999; Wiebe and Rapaport, 1988). The year 2001 marked the beginning of widespread awareness of the research opportunities that this new field provides (Liu et al., 2003; Pang et al., 2002; Turney, 2002). Factors behind this success

include the following three (Pang and Lee, 2008): the rise of machine learning methods in Natural Language Processing and Information Retrieval, the availability of datasets to be trained on and to be analyzed, and the creation of competitions and commercial applications. As an example of the current interest in the research problem this paper **engages** with is the workshop on "Computational Approaches to Analysis and Generation of Emotion in Text" at NAACL 2008 [1].

This general progress in the field notwithstanding, research in general on Sentiment Analysis has so far been restricted to the most basic case studies that can be considered: analysis applied only to texts that have a single overarching emotional bias (news items, blog articles, opinion pieces, customer complaints...) and analysis aimed at identifying emotional connotations of a very particular kind (positive or negative attitude, for instance), singled out from a larger range of emotional connotations that are then ignored. Whereas this undoubtedly leads to more impressive qualitative results at the experimental level, the field has yet to expand its coverage over other kinds of texts or other kinds of emotion are not open to similar analysis.

Emotions are inherent to any human activity, including our interactions with computers. Processing emotions expressed in natural language within a speech or text document is becoming a requirement demanded of any computational system that aims to offer a natural interface to its users. For example, recognizing certain emotions in a human speaker would permit a computer to react to the commands according to the emotional situation, instead of giving a neutral response; and that response from the machine can also be generated after considering which emotion it should express to the user (Evens, 2002; Krenn et al., 2002). Synthesized speech would also be significantly improved by reproducing different emotional connotations when modulating the synthesized voice.

The recognition of emotions expressed in natural language is not only important for classic interfaces but also for on-line advice and recommendation systems (Pang and Lee, 2008). The interest that users show in on-line opinions and the potential influence of such opinions is something that vendors are paying more and more attention to (Hoffman, 2008), which makes it very important to identify the emotions behind them automatically. Emotional analysis algorithms have been recently applied to the creation of computational models of human opinion from customers' on-line reviews (Wright, 2009).

In addition, the automatic generation of text and speech have been widely developed over the last two decades, and have often given rise to technological solutions for restricted domains. Affective Computing aims for more natural interactions, particularly in the areas of the recognition and generation of emotions (Pang et al., 2002; Turney and Littman, 2003; Merola, 2007; Busso et al., 2008).

A fundamental technique in Sentiment Analysis (Pang and Lee, 2008) is the classification of emotions. An example of classification is making a decision for a particular sentence ("What emotion is evoked by this sentence?"). For the automatic mark-up of emotions in texts it is clear that some guidelines about how people express their feelings are required, and a text corpus with emotional annotations might be a reasonable first step towards that goal. This paper presents the creation of a corpus of texts marked up with emotions and an approach, based on this corpus, which automatically annotates texts with emotions. Annotating text with emotional content is a difficult task. As the identification and assignment of emotions are subjective decisions, it is common to find that different human annotators assign different emotional tags to the same sentence or piece of text. Therefore, it is very important to study how the emotional annotation process of a corpus is performed in order

[1] http://www.site.uottawa.ca/ diana/naacl2010_EmotionWorkshop.html

to define that process properly, while reducing its dependence on subjective criteria as much as possible.

The role of narrative as a vehicle for exercising and communicating emotions, and in the process helping people to learn about them and come to terms with them, has been well documented since the Poetics of Aristotle. From this perspective, it makes sense within the domain of our research to consider narrative texts as a valuable source for exploring what the full range of emotional connotations might be. The choice to study narrative texts would also make it possible to link the results of this research to the greater effort currently being undertaken by the entertainment industry to explore further uses of information technology in providing new experiences for gamers and consumers of other interactive media. Where such efforts involve the identification, representation, reproduction or induction of emotion in the user, a rich computational representation of emotion, a procedure for attributing emotion to text, and a corpus of material annotated with such a representation, would be very valuable resources.

In such a context, it would also be extremely useful if the representation chosen for annotation were devised in such a way as to provide flexible transitions between different degrees of granularity in the annotation. Because some research efforts may wish to concentrate on a small set of generic emotions, and others may want to consider a broader range, a resource that allows easy conversion from annotations in terms of generic emotions to annotations in terms of larger sets of emotional labels, or conversion across different methods of representing emotion, would be very useful.

Our research goal is therefore to create an annotated corpus for narrative applications and an approach to automatically mark up texts with emotional content by using the two most relevant methods to represent emotional states: emotional categories (emotional tags used to refer to emotions) and emotional dimensions (measures that try to model the essential aspects of emotions numerically). Previous attempts have been carried out in recent years in order to obtain a text corpus marked up with emotions and systems that automatically annotate texts with emotions. However, most of these attempts were oriented towards the identification of the positive or negative polarity of the texts (Bestgen, 1993; Pang et al., 2002; Read, 2005; Popescu and Etzioni, 2005) or their classification within a small set of preset emotions (Zhe and Boucouvalas, 2002; Alm and Sproat, 2005; Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2008). In contrast, our approach uses a broader spectrum of emotional concepts for the annotation.

The rest of the paper is structured as follows. Section 2 presents the essential aspects of the emotions. Section 3 presents a brief outline of the previous work on emotional mark-up. Section 4 provides a detailed explanation of how EmoTag works. Section 5 presents the results obtained by EmoTag. Finally, Section 6 discusses the results of our work and Section 7 shows the main conclusions of our work and presents some ideas for future work which will improve the results obtained by this approach.

2. EMOTIONS

Emotions are not an easy phenomenon; there is a large number of factors that contribute to the generation of emotions. Izard (1971) suggested that a good definition of emotion must take into account: the conscious feeling of the emotion, the processes that appear in the nervous system and in the brain, and the expressive models.

Appraisal Theory (Scherer et al., 2001; Read et al., 2007) is the idea that emotions are extracted from our evaluations (appraisals) of events that cause specific reactions in different people. Essentially, our appraisal of a situation can cause three different attitudes: affect (personal emotion), judgement (appraisal of other's behaviour) or appreciation (evaluation

of phenomena). All three ways of feeling can be either positive or negative. In this paper we are going to focus on appraisals that cause an emotional, or affective, response.

There is a number of theories about how to represent emotions, which are explained in Section 2.1, and different theories about how these emotions could be structured, which are explained in Section 2.2.

2.1. Representation of Emotions

There is a number of theories about how many emotions there actually are. In these theories the number of emotions varies from two up to hundreds (Ortony and Turner, 1990). There are different methods to represent emotions (Cowie and Cornelius, 2003) but two are the most important and the most often used in existing approaches in Sentiment Analysis: *emotional categories* and *emotional dimensions*.

Emotional categories: This analytical perspective deals with verbal tags as commonly understood by speakers to refer to emotions. Different languages provide assorted words with varying degrees of expressiveness for the description of emotional states. That is why several approaches have been proposed to reduce the number of words used to identify emotions, for example with the use of *basic emotions*, *super-ordinate emotional categories* or *essential everyday emotion terms*. *Basic emotions* refer to those that are more well-known and understandable to everybody than others (Cowie and Cornelius, 2003). In the *super-ordinate emotional categories* approach some emotional categories are proposed as more fundamental, with the argument that they subsume the others (Scherer, 1984). Finally, the *essential everyday emotion terms* approach focuses on emotional words that play an important role in everyday life (Cowie et al., 1999).

Emotional dimensions: Emotional dimensions are measures that try to model the essential aspects of emotions numerically. Emotional dimensions deal with scale perceptions and placements on artificially imposed scales of identified characteristics of emotions. Although there are different dimensional models with different dimensions and numerical scales (Fontaine et al., 2007), most of them agree on three basic dimensions called *evaluation*, *activation* and *power* (Osgood et al., 1957). *Evaluation* represents how positive or negative an emotion is. At one extreme we have emotions such as *happiness*, *satisfaction* and *hope* while at the other we find emotions such as *unhappiness*, *dissatisfaction* and *despair*. *Activation* represents an activity versus passivity scale of emotions, with emotions such as *excitation* at one extreme, and at the other emotions such as *calmness* and *relaxation*. *Power* represents the sense of control which the emotion exerts on the subject. At one end of the scale we have emotions characterized as completely controlled, such as *fear* and *submission* and at the other end we find emotions such as *dominance* and *contempt*. To assess the three dimensions, there is an affective rating system originally devised by Lang (1980) called SAM. The graphic SAM figures comprise bipolar scales that depict different values along each emotional dimension. Figure 1 illustrates a version of SAM. For the evaluation dimension SAM ranges from a smiling, happy figure to a frowning, unhappy figure; to represent the activation dimension SAM ranges from a excited, wide-eyed figure to a relaxed, sleepy figure. For the power dimension, SAM ranges from a small figure (dominated) to a large figure (in control). The subject can select any of the 5 figures on each scale or the space between two figures, which results in a 9-point rating scale for each dimension.

The clearest distinction between the two methods is that emotional dimensions allow the representation of any point in the space of emotional values captured by their three axes of representation, irrespective of whether there exists a specific lexical label for the emotional category that would correspond to that point. In this sense, they can be considered to represent a continuous space, even if the actual assignment of values to each dimension is done in terms of discrete numbers. There may well be large volumes of the possible

emotional space for which there are no lexical labels available, or volumes that are covered differently in terms of lexical labels across different languages. The representation in terms of emotional dimensions is therefore considered to be more generic.

An emotional markup language that allows to mark up emotions both as emotional categories and as emotional dimensions was proposed by the W3C Emotion Markup Language Incubator group (Schröder et al., 2008).

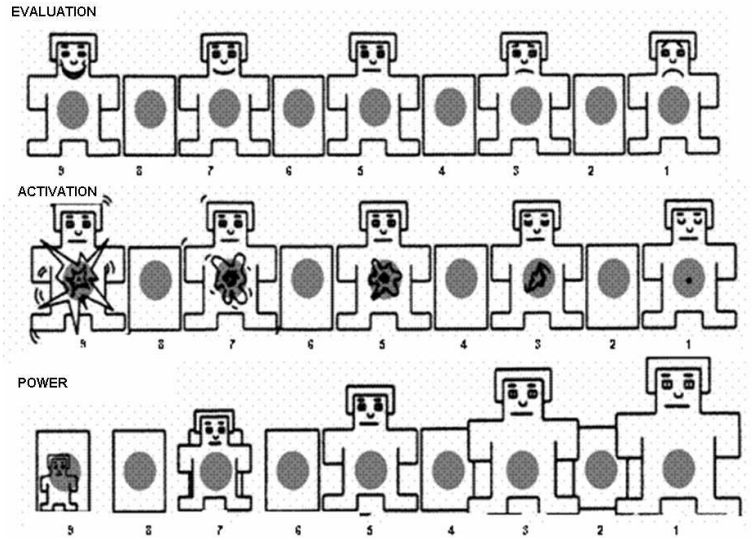


FIGURE 1. Dimensional scales according to SAM: evaluation, activation and power

2.2. Structure of Emotions

Psychologists have been searching for a suitable way to structure our emotional repertoire. Several methods have been proposed, each with its own advantages and disadvantages.

Methods based on emotional dimensions aim to capture the similarities and differences among emotions. Some researchers propose a two-dimensional space that exclusively considers the emotions of evaluation and activation. This is called the *circumflex model* where the points that correspond to all possible emotions form a circle (Russell, 1980; Watson and Tellegen, 1985). Viewing the multitude of emotions as points in a two-dimensional space can be useful in understanding the most generic emotions but not the most specific ones. This model reduces the variety of emotional states, and does not capture the slight differences found beyond the most generic sensations.

As an alternative to dimensional spaces some researchers have used cluster analysis (Storm and Storm, 1987; Shaver et al., 1987; Parrott, 2001; Arnold, 1960). These approaches group emotions into clusters, with the number of clusters depending on each specific approach. Storm and Storm (1987) **propose** the use of 12 clusters: *love*, *happiness*, *sadness*, *anger*, *fear*, *anxiety*, *contentment*, *disgust*, *hostility*, *liking*, *pride* and *shame*. Shaver et al. (1987) **propose** the use of 5 clusters called *affection*, *happiness*, *sadness*, *anger* and *fear*. Parrott (2001) presents a more detailed list of emotions categorized in a short tree structure. This structure has three levels for primary, secondary and tertiary emotions. As primary emotions, Parrot presents *love*, *joy*, *surprise*, *anger*, *sadness* and *fear*. Secondary emotions give nuance to primary emotions, e.g. *love* has *affection*, *lust* and *longing* as secondary emotions. Finally, tertiary emotions give further nuance to secondary emotions, e.g. *lust* is a secondary emotion with *arousal*, *desire*, *passion* and *infatuation* as tertiary emotions.

Arnold (1960) uses 11 basic emotions: *anger*, *aversion*, *courage*, *dejection*, *desire*, *despair*, *fear*, *hate*, *hope*, *love* and *sadness*.

Instead of grouping emotions according to their global similarity, other researchers prefer to group emotions based on different criteria such as the components of their appraisals (Scherer, 1984) or the events that give rise to them (Ortony et al., 1988).

To summarize, there are many different ways to structure emotions and each approach may be useful for a different purpose. Any approach that aims to be useful in a great variety of applications should take advantage of all these different representations of the world of emotions.

3. PREVIOUS WORK IN EMOTIONAL MARK-UP

This section presents a brief outline of the previous work related to the mark-up of texts with emotional content. First we present a brief outline of the affective dictionaries which have been developed in the recent past and second, a review of the current systems for tagging texts with emotional content.

3.1. Affective Dictionaries

There are different types of affective dictionaries, depending on the type of emotion taken into account. There are affective dictionaries which classify words into emotional dimensions, emotional categories, or both, and others that measure the subjectivity of the words. There are some dictionaries described hereafter that are not just about emotion per se, such as the work of Hatzivassiloglou and McKeown (1997), Turney and Littman (2003) or SentiWordNet (Esuli and Sebastiani, 2006) but we feel that they should be included here in order to show the current state of sentiment analysis.

First we have the affective dictionaries which use emotional dimensions for the classification of the words.

An initial group includes those research **projects** that classify words into those with positive or negative connotations. Hatzivassiloglou and McKeown (1997) elaborated a list of frequent adjectives with an associted orientation (*positive* or *negative*). The process involved an initial manual assignment followed by an extension to a broader list based on co-occurrence. The basic idea is that if an adjective with an unknown polarity appears together with an adjective with a known polarity**,** the first adjective takes the polarity of the second one. Turney and Littman (2003) **classify** words into positives and negatives. In order to do that, given a word, they obtained the frequency of **appearances of** that word along with positive words (*good*, *nice*, *excellent*, etc.) and the frequency of appearances of the word along with negative words (*bad*, *nasty*, *poor*, etc.). Based on these two frequencies they determined the valence of the word. Grefenstette et al. (2006) use patterns of co-occurrence with indicative words to identify **emotion-bearing** words over the web. Patterns consisted of pair of words likely to precede **emotion-bearing** words, and they consisted of a specific **word** from a list of 21 words (*appear*, *appears*, *appeared*, *appearing*, *feel*, *feels*, etc.) followed by a quantifying adverb (*almost*, *extremely*, *so*, *too* or *very*). This resulted in 105 patterns which were used as queries in `www.alltheweb.com`. The word which appeared in each case just after the pattern was taken and classified by one person into positive, negative or not affective words.

A second group involves research **systems** that assign to words values related to emotional dimensions. Whissell's Dictionary of Affect in Language (DAL) (Whissell, 1989) is a resource for measuring the evaluation and activation of words as well as the images associated with that word. It does this in terms of three dimensions: *evaluation*, *activation* and *imagery*. There are no scores for the *power* dimension. The Affective Norms for English

Words (ANEW) (Bradley and Lang, 1999) is a set of normative emotional ratings, in terms of *evaluation*, *activation* and *power*, for a large number of words in the English language. This database of emotional words is content-independent.

A third group involves research efforts that associate with each word values along a different set of dimensions. The Lasswell Value Dictionary (Lasswell and Namenwirth, 1969) marks up words with binary values which correspond with eight basic dimensions: *wealth*, *power*, *rectitude*, *respect*, *enlightenment*, *skill*, *affection* and *well-being*. Based on the Lasswell Value Dictionary, Stone et al. (1966) created the content dictionary for The General Inquirer. The General Inquirer's dictionary has a large number of labels such as *active*, *passive*, *strong*, *weak*, *pleasure*, *pain*, *feeling*, *arousal* or *virtue*. The words in this case either have an attribute or not; no intermediate degrees are considered.

There are other dictionaries that assign each word emotional categories, such as the Clairvoyance Affect Lexicon (Huettner and Subasic, 2000) which was developed by hand at the beginning of the nineties. The dictionary entries have an associated affective label, a weight which measures the position of the word in the affective label and an intensity.

There are dictionaries that combine affective labels with emotional dimensions such as WordNet Affect. WordNet Affect was developed by Strapparava and Valitutti (2004) semi-automatically. In WordNet Affect each word in WordNet (Fellbaum, 1998) has an affective label. The labels include: semantic labels based on the psychological and social theories of Ortony, Elliot and Ekman, labels for valence (positive or negative), for activation (active or passive), etc.

Finally, there are dictionaries that measure the subjectivity of words, such as SentiWordNet. SentiWordNet (Esuli and Sebastiani, 2006) is an affective dictionary that marked up the synsets of WordNet as WordNet Affect. Each WordNet synset is associated to three numerical scores *Obj(s)*, *Pos(s)* and *Neg(s)*. WordNet Affect describes how objective, positive and negative the terms contained in the synset are. To develop SentiWorNet, the glosses associated to the synsets are analyzed and then the resulting vectorial term representations are used for a semi-supervised synset classification.

These dictionaries constitute important resources to consider when working with emotions. However, it is clear that there is no consensus as to how emotion should be **represented, not only** in terms of whether lexical labels or numerical values along a set of chosen dimensios should be used, but also in terms of what sets of values, dimensions or labels to employ.

If the representation of choice is emotional dimensions, it is clear that the General Inquirer refines the Laswell Value Dictionary to a set of abstractions some of which correlate well with the power and the activity emotional dimensions. The DAL Dictionary and the ANEW list also constitute important resources when representing emotion in terms of emotional dimensions. Turney and Litman, and Grefenstette et al. only assign positive or negative values (with an optional not affective category in the case of the latter). Some of the labels in WordNet Affect can also provide information on emotional dimensions. If a representation in terms of emotional categories is chosen, The Clairvoyance Affect Lexicon, WordNet Affect and SentiWordNet all constitute valuable sources. However, their applicability is limited to the extent that the set of labels used in each case needs to match the set of labels used in the chosen representation. This is a general limitation on the reusability of this kind of resources. The ontology presented in section 4.3 was designed in part with the hope of interrelating a number of existing sets of labels in such a way that labels corresponding to one set might be interpreted in terms of labels corresponding to another set.

3.2. Existing Approaches for the Emotional Marked Up of Texts

In this section a brief outline of the most important approaches for the emotional mark-up of texts is presented. We are going to group these approaches according to the theory used to classify emotions: emotional categories, emotional dimensions or both.

3.2.1. *Approaches that use Emotional Categories.* Zhe and Boucouvalas (2002), Liu et al. (2003), Alm and Sproat (2005), Mihalcea and Liu (2006), Sugimoto and Yoneyama (2006), Aman and Szpakowicz (2007) or Strapparava and Mihalcea (2008) are examples of approaches that use emotional categories to classify emotions.

Zhe and Boucouvalas (2002) has developed an emotion extraction engine which can analyze sentences given by the user. This system is included in a human-machine communication domain so it only takes into account the emotions referring to the speaker. The system analyzes the sentences, detects the emotion (*happiness*, *sadness*, *fear*, *surprise*, *anger* or *disgust*) and shows the suitable facial expression. In order to obtain the emotion of the sentence, all the words that the sentence is composed of are looked up in a dictionary of 16,400 words. To test the system, a questionnaire was presented to 50 people, the same situations were presented to each of them and 450 sentences in total were submitted. Users were asked to write down their emotional responses to a set of prescribed situations. These sentences were then input into the engine. The evaluation measurement was the percentage of sentences correctly marked up by the system. The results showed that 90% of the sentences were correctly tagged.

Liu et al. (2003) created an approach based on large-scale real-world knowledge. This mark-up system uses the emotions defined by the OCC Model. The data used is the OMCS (Open Mind Common Sense) Corpus. Facts with affective relevance are extracted from the corpus. On this basis, a 'common sense affect model' is constructed. The model consists of a set of component models which compete with and complement each other. To construct the models, emotion keywords are propagated in three passes over the corpus. Emotion values initially are 1, then with each propagation are reduced by a factor *d*. To classify a text, it is first segmented into clauses, then linguistically processed, and finally evaluated by a 2-stage process using the models. To test their approach they incorporated their affect-sensing engine into an email browser and added the use of emotional faces. A 20-user study was conducted in order to see what was preferred by the users: a browser with neutral faces, a browser with randomized emotional faces or a browser with the faces generated by Liu's system. In this case the correctness of the emotional faces in each situation is not evaluated; the only aim of the evaluation was to decide if a browser with faces generated by the system was better than a browser with neutral or randomized emotional faces. User evaluations suggested that the system was **good** enough to bring measurable benefit to an affective user interface application.

Alm and Sproat (2005) have studied the emotional distributions in 22 fairy tales in terms of patterns of emotional sequencing and positioning, and also in terms of emotional development through the story temporally. The corpus analyzed has each sentence marked up with 8 basic emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness*, *positive surprise*, *negative surprise* and *neutral*). Stories are independently annotated by two people. The annotated texts are subsequently post-processed by one of the evaluators who tie-breaks disagreement by choosing the most appropriate of the conflicting labels. The study concludes that, first, *neutral* occurred more frequently in the first sentence and *happy* in the last sentence. Second, for all emotions except *disgust* and *negative surprise*, *neutral* was more frequently compared to other emotions. And finally, given consecutive sentences, *angry* and *sad* preceded and followed themselves significantly more often compared to other emotions. The study does not include a method for automatic mark-up based on the conclusions of the study. Alm

(2009) used classification methods to automatically infer affect in text and Alm (2010) discussed characteristics of a dataset with affect annotation.

Mihalcea and Liu (2006) employed 'linguistic ethnography' to seek out where happiness lies in our everyday lives by considering a corpus of blogposts annotated with happy and sad emotions. By analyzing the corpus they obtained a list of happy and sad words and phrases annotated by their 'happiness factor'. The list of happy and sad words **was** then used as a basis for studies centered around the topic of happiness.

Sugimoto and Yoneyama (2006) has a system that marks up text with emotions for Japanese in the narrative domain. The emotions used by this system are: *joy*, *sorrow*, *anger*, *surprise* and *neutral*. It decides on the emotion of the sentence from the emotion of the nouns, adjectives and verbs which compose the sentence and from the grammatical structure of the sentence. Japanese generally has three sentence types: adjective sentence (S+V+Adjective Complement), noun sentence (S+V+Noun Complement) or verb sentence (S+V, S+V+O). The rules in order to determine the emotion of the sentence are different depending on the type of the sentence. In the adjective and noun sentences the emotion with the greatest weight is the emotion assigned to the adjective or to the noun. In verbal sentences the emotion is determined by the combination of emotions assigned to the subject and the verb. The system was not evaluated.

Aman and Szpakowicz (2007) marks up blog posts with 6 basic emotions (*happiness*, *sadness*, *anger*, *disgust*, *surprise* and *fear*) and emotional intensity (*high*, *medium* and *low*). A corpus of 10,000-sentence blog posts was collected from the Web and was marked up by 4 evaluators. The words in the blog posts were looked up in The General Inquirer and WordNet Affect in order to extract tags such as *EMOT* (emotion), *Pos/Pstv* (positive), *happiness*, *fear*, etc. associated to each word. Then, Naive Bayes and Support Vector Machines (SVM) techniques were used to automatically mark up the blog posts. **In the evaluation, accuracy was calculated, obtaining an average accuracy of 73%.**

Strapparava and Mihalcea (2008) marks up texts with 6 basic emotions (*anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*). This work uses Latent Semantic Analysis (LSA) and Naive Bayes classifiers to automatically mark up text with basic emotions using the corpus of SEMEVAL 2007 (a corpus of 250 headlines annotated by 6 evaluators) and a collection of blog posts annotated with moods that were mapped to the six emotions. To evaluate the system, a gold-standard data set was provided with emotional annotations, and then both fine-grained and coarse-grained metrics were used for the evaluation. Fine-grained evaluations were conducted using the Pearson measure of correlation between the system scores and the gold standard scores, averaged over all the headlines in the data set. In the coarse-grained evaluations each emotion was mapped to a 0/1 classification. For the coarse-grained evaluations, precision, recall, and the F-measure were calculated. For fine-grained evaluation the average Pearson measure obtained was 28.38. For coarse-grained evaluation the average precision was 38.28, the average recall was 90.22 and the average F-measure was 17.57.

3.2.2. *Approaches that use Emotional Dimensions.* We have not found specific research efforts aimed at automatically assigning emotional dimensions to text in the broadest sense. Nevertheless, several research **systems** in sentiment analysis use approaches that distantly relate to emotional dimensions, in as much as they classify text in terms of negative or positive emotional polarity. For instance, the systems created by Bestgen (1993), Pang et al. (2002), Read (2005), Popescu and Etzioni (2005) and Snyder and Barzilay (2007).

These systems can be grouped into two different sets: those that consider sentences as basic units for opinion assignment and those that operate over larger segments of text.

Bestgen (1993) and Popescu and Etzioni (2005) develop systems that assign a negative or positive evaluation to sentences. The Bestgen system (Bestgen, 1993) focuses on

determining the evaluation of the sentences (positive or negative). The text is divided into segments and the words which compose the segment are looked up in a dictionary which contains words and their evaluation. In order to obtain the final value of the segment, the average evaluation value is calculated. In this system the words are taken out of context but negations which appear in the texts are taken into account. For the evaluation four tales were selected as a corpus; these tales were divided into sentences and each sentence was marked up by 15 subjects. As an evaluation measurement the correlation factor between the texts marked up by human annotators and the texts marked up by the system were calculated. The mean correlation obtained **was** 0.50. The Bestgen (1993)'s study highlights the necessity of using an adapted and exhaustive dictionary and the necessity of continually perfecting the proposed technique by taking other variables besides negation into account. Popescu and Etzioni (2005) created OPINE which is an unsupervised information extraction system that extracts the semantic orientation of opinions (positive or negative). OPINE uses a relaxation-labeling technique to determine the semantic orientation of potential opinion words and specific review sentences. The system was trained with a corpus of 200 tuples (word, feature, sentence) marked up by two annotators who assigned positive, negative and neutral labels to each tuple. To evaluate the system 800 tuples were annotated by human evaluators and these annotations were compared with the annotations made by OPINE; precision and recall measurements were obtained. OPINE obtained a precision of 86% and a recall of 89%.

Pang et al. (2002), Read (2005), and Snyder and Barzilay (2007) develop systems that assign negative or positive evaluation to larger fragments of text (film reviews, article extracts, restaurant reviews). Pang et al. (2002) created a system that classifies film reviews into positive or negative. A corpus of 700 negative and positive reviews were used. Two evaluators chose the indicator words for positive and negative sentiments in the reviews. A baseline was created by looking at the frequency of those words in the document. Three classifiers were tried: the Naive Bayes classification, maximum entropy classification, and support vector machines. No stemming or stop lists were used, punctuation was treated as words and negation tags were added to negated words. They did not create an a priori selection of keywords, but instead used all occurring words. The same word, whether negated or not, was actually counted as two distinct words. To evaluate the system, a corpus of 700 negative and positive reviews were used. The evaluation measurement used was the percentage of sentences correctly classified by the system with respect to the annotators' classifications. The best results **were** obtained by SVM which obtained 82.9% of sentences correctly classified. The results of the evaluation suggest that some form of discourse analysis is necessary and that the identification of features indicating whether sentences are on-topic is important. Read (2005) carried out several experiments which demonstrated the influence of domain, topic and time on machine learning based sentiment classification (determine if a text is generally positive or negative). Then, Read developed a sentiment classification approach based on machine learning. A corpus of 26,000 article extracts marked-up with emoticons was used in their evaluation. The mean accuracy obtained by this system was 61.5% for Naive Bayes Classifier and 70.1% for SVM classifier. Snyder and Barzilay (2007) marked up the evaluation of different aspects of reviews. Their system considered the problem of analyzing multiple related aspects of user reviews. The algorithm presented learns to rank models for individual aspects by modeling the dependencies among assigned ranks. The strength of the algorithm lies in its ability to guide the prediction of individual rankers using rhetorical relations among aspects such as agreement and contrast. To evaluate the algorithm a corpus of restaurant reviews was used. Each review is accompanied by a set of five scores, each on a scale of 1-5, covering food, ambience, service, value, and overall experience. These scores are provided by consumers who wrote original reviews. From this corpus 500 reviews were randomly selected for development and 500 for testing. The system was evaluated

by using ranking loss which measures the average distance between the true score and the one predicted; lower values of this measurement correspond to a better performance of the algorithm. The average ranking loss obtained by the system was 0.324.

There are some initiatives whose aim is to enhance research in Information Access (IA) technologies, including opinion tasks whose aim is to classify sentences into positive, negative or neutral; two of these initiatives are NTCIR [2] and TREC blog-track[3]. NTCIR is a series of evaluation workshops to enhance the research in information access technologies by providing the infrastructure for evaluation and research including large-scale re-usable test collections, evaluation metrics and methodologies, and a forum for researchers who are interested in exchanging research ideas and evaluation methodologies. The Sixth NTCIR Workshop selected Opinion Analysis as a pilot task. A test collection for 32, 30, and 28 topics (11,907, 15,279, and 8,379 sentences) in Chinese, Japanese and English was created. Using this test collection, the opinion extraction subtask was conducted. The subtask was defined from four perspectives: opinionated sentence judgment, opinion holder extraction, relevance sentence judgment, and polarity judgment. As evaluation measurements, NTCIR used precision, recall and f-measure using lenient gold standard and strict gold standard. TREC blog-track explores information-seeking behaviour in the blogosphere. In TREC 2006 the main task was opinion retrieval, which focused on the opinionated nature of many blogs. A test collection of blog data was created for the purposes of the TREC Blog track. The collection included a selection of "top blogs" covering topics such as news, sports, politics, health, etc. Then, a selection of blogs assumed to be spam was inserted to ensure that Blog track participants had a realistic research setting. The opinion retrieval task can be summarised as *What do people think about X*, *X* being a target. Each post can be assessed as -1 (Not judged), 0 (Not relevant) or 1 (Relevant). The metrics used for the opinion retrieval task are mean average precision (MAP), R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10).

Of the systems reviewed **in** this section, the Bestgen system (Bestgen, 1993) is much closer to EmoTag, in the sense that it works over individual sentences, it uses a similar bag-of-words approach enhanced with additional procedures to take negation into account. Popescu and Etzioni (2005) also works over sentences. Both of these **projects** rely on small data sets. In contrast, efforts that operate over larger fragments of text apply much larger sets of data. This is related to the comparative difficulty of annotating each sentence with a different emotion as compared to annotating a complete text with a single emotion. If a single emotion per text is enough, annotation is much faster and it becomes much easier to collect large sets of data. This issue must be taken into account when establishing comparisons in terms of data sets.

3.2.3. *Approaches that use Emotional Categories and Emotional Dimensions.* Finally, there are systems that combine both representation theories, that of emotional dimensions and that of emotional categories such as the system developed by O'Connor et al. (2007) and the systems presented in SemEval 2007.

O'Connor et al. (2007) has a system for emotional detection integrated in a system of textual improvisation. This system marks up narrative texts written in English with three labels: evaluation, emotional label and intensity. The system obtains the emotion based on the type of sentence, the metaphors or the similarity to some patterns previously generated. In this case the correctness of the emotional mark-up of the system is not evaluated; the only aim of the evaluation is to decide if the system of textual improvisation with the emotional

[2] http://research.nii.ac.jp/ntcir/index-en.html
[3] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

detection motor integrated improves the perceived quality of social interaction above and beyond the original system.

SemEval 2007 [4] had a task called Affective Task (Strapparava and Mihalcea, 2007). This task had the objective of classifying headlines with an appropriate emotion label and with a valence indication (positive / negative). Each headline was annotated with six emotions and valence. Each emotion was marked on an interval [0,100], where 0 meant the emotion was missing from the given headline and 100 represented maximum emotional load. The interval for the valence annotations was set to [-100,100], where 0 represents a neutral headline, -100 represents a highly negative headline, and 100 corresponds to a highly positive headline. Five teams participated in the task, with five systems for valence classification and three systems for emotional labeling: *UPAR7*, *SICS*, *CLaC*, *CLaC-NB*, *UA* and *SWAT*. *UPAR7* is a rule-based system using a linguistic approach. Each word was first rated separately for each emotion and for valence. Next, the main subject rating was boosted. Contrasts and accentuation between "good" or "bad" were detected. The system also took into account: human will, negation and mood, high-tech context and celebrities. *SICS* uses a very simple approach to valence annotation based on a word-space model and a set of seed words. *CLaC* assigns positive, negative and neutral valence to headlines. The system used three main kinds of knowledge: a list of sentiment-bearing words, a list of valence shifters and a set of rules that define the scope and the result of the combination of sentiment-bearing words and valence shifters. *CLaC-NB* uses a Naïve Bayes classifier in order to assign valence to headlines. *UA* determines the kind and the amount of emotions in a headline by gathering statistics from three different web Search Engines: MyWay, AlltheWeb and Yahoo. This information is used to observe the distribution of the nouns, verbs, adverbs and adjectives extracted from the headline and the different emotions. *SWAT* uses a unigram model trained to annotate emotional content. Synonym expansion on the emotion label words was also performed. To evaluate the system, the same measurements explained above for the (Strapparava and Mihalcea, 2008) system were used.

Finally, there are other systems that mark up the objectivity and subjectivity of text, such as the work of Wiebe et al. (2001). The marking of this type of content is completely outside of the scope of the work presented in this paper so we are not going to emphasize it. For more information about this work consult (Wiebe et al., 2001).

In Table 1 we show a summary of the methods presented above and provide information about the descriptors used by the system (dimensions, categories or others), the domain of the system, and the characteristics of the evaluation performed (material used, number of evaluators and measurements).

All the systems explained above are compared with EmoTag in Section 7.

4. EMOTAG

This section provides a detailed explanation of how EmoTag works and what resources it employs. First, the domain of the application is explained; then, the mark-up granularity selected and the representation of emotions used by EmoTag are discussed. Then we explain the ontology of emotional categories developed to provide a complex representation of emotional categories, the corpus of texts marked up by human evaluators created as a source from which to obtain basic assignations of emotions to words, and the List of Emotional Words (LEW) distilled from the corpus as a resource for emotion-to-word assignments. Finally the process which classifies English sentences into emotions is explained.

[4] http://nlp.cs.swarthmore.edu/semeval/

System	Descriptors			Domain	Evaluation		
	Dimensions	Categories	Others		Material	Evaluators	Measurements
Zhe & Boucouvalas		6 basic		Human-machine communication	9 sentences	50	% of sentences correctly marked up
Liu et al.		OCC Model basic emotions					
Alm & Sproat		8 basic		Fairy tales			
Mihalcea & Liu		2 basic		Blogposts			
Sugimoto & Yoneyama		5 basic		Narrative texts			
Aman & Szpakowicz		6 basic	Emotion intensity	Blog posts			
Strapparava & Mihalcea		6 basic		Headlines	250 headlines	6	Pearson measure of correlation, Precision, Recall and F-measure
Bestgen	Evaluation (+ or -)			Tales	4 tales	15	Correlation factor
Pang et al.	Evaluation (+ or -)			Film reviews	1,400 reviews	1	% of sentences correctly classified
Read	Evaluation (+ or -)			Article extracts with emoticons	26,000 article extracts		Precision
Popescu & Etzioni	Evaluation (+ or -)			Opinion texts	800 tuples	2	Precision and Recall
NTCIR	Evaluation (+ or -)			Opinion texts	90 topics	1	Precision, recall and F-measure
TREC	Evaluation (+ or -)			Blog data			MAP, R-Prec, bPref and P@10
Snyder & Barzilay	Evaluation (+ or -)			Reviews	500 restaurants reviews	1	Ranking loss
O'Connor et al.	Evaluation (+ or -)	Basic emotions	Intensity	Narrative texts	9 sentences	50	% of sentences correctly marked up
SEMEVAL 2007	Evaluation (+ or -)	6 basic		Headlines	250 headlines	6	Pearson measure of correlation, Precision, Recall and F-measure

TABLE 1. Approaches for the emotional mark-up of texts found in the scientific literature

4.1. Domain of Application

Due to our special interest in narrative applications and previous experiences in story generation, we decided to focus the effort of marking up texts with emotions on a very specific domain: fairy tales. Fairy tales are generally intended to help children to better understand their feelings, and they usually involve instances of the emotions that most children experience on their way to maturity (e.g. *happiness*, *sadness*, *anger* or *fear*). Furthermore, the domain selection of tales is not new in this field since this domain was also chosen in the work of (Alm and Sproat, 2005), (Sugimoto and Yoneyama, 2006) and (Bestgen, 1993).

Emotions in tales, considered from the point of view of a storyteller, have two main functions: to express the personality and internal feelings of a given character at a given moment in the tale, and to induce a certain emotional response in the audience (Kready, 1916; Alm et al., 2005). Moreover, tales are especially suitable for the identification and study of emotions because the emotions presented in them are more obvious and explicitly represented than those presented in more complex domains.

We have selected such a specific domain (fairy tales) due to three main factors:

- Narrative has great cultural importance as a means of communicating, exemplifying, transmitting and teaching complex abstract ideas about values and emotions.
- There is a considerable shortage of work in this domain from the point of view of the representation, identification and annotation of emotions.
- The complexity of the emotional information involved in narrative texts is much higher than in those domains that have so far been the focus of research in Sentiment Analysis domains (blogs, news items, opinion pieces, etc.).

The system is not domain-dependent; the process can be applied to any other domain. At most, transferring to a new domain might require a new corpus of texts from the new domain marked up with emotions, in order to provide coverage for any emotion-to-word assignments particular to the new domain.

4.2. Mark-up Granularity and Representation of Emotions

Our process assigns emotions to sentences; that is, the emotional unit of our system is the sentence. Clauses are a natural "unit" of linguistic communication; they contain a complete thought-package, so it seems very suitable to assign emotional content to the sentences in the text. However, it is also clear that subordinate clauses contribute a significant emotional value to their parent sentence. We have opted for a three-stage solution, where subordinate clauses are identified, the assignment of emotion to these individual clauses is calculated, and then the emotional value of sentences is calculated taking into account the emotional value of their constituent clauses. We consider subordinate clauses as emotional units inside a larger unit, the **main** sentence. In this way, we are going to obtain the emotion associated with each subordinate clause and then consider this emotion as a single emotion-bearing element within the encompassing sentence. Subordinate clauses play an important role in the emotional content associated with a sentence. For example, in the sentence *"The boy came into a castle which was magnificent and luminous, and found the wicked ogre"* if all the words are assigned the same weight, this results in two positive words (*magnificent* and *luminous*) and two negative words (*wicked* and *ogre*), and the final emotion will result from the combination of the four emotions with the same weight. If the emotion of the subordinate sentence (*which was magnificent and luminous*) is computed separately during a first step, and then the subordinate clause is considered to be a single emotion-bearing element within the main sentence, which contributes this previously computed value, the negative connotations of *wicked* and *ogre* contribute with greater weight to the emotion assigned to the complete sentence, resulting in an assignment that better matches the intuitive interpretation.

Since there is currently no agreement in the literature as to a preferred method for representing emotions (as explained in Section 2.1), we have chosen to develop a system capable of tagging text using both *emotional dimensions* and *emotional categories*. This ensures compatibility with a large number of the representations of emotions currently available, and makes the system applicable to the highest number of applications. These two methods (category-based mark-up and dimension-based mark-up) are used in parallel. When it is time to mark up a text, the user of EmoTag can select the most suitable representation for whatever purpose he has in mind. In this way, our system is more flexible and can be adapted to the different systems that may need a text marked up with emotions. If a synthesizer needs the text marked up with emotional dimensions, the user can select this representation as his preferred output format. If the user needs the text marked up with categories to provide input to a 3D system for rendering facial expressions, he can select categories as the corresponding output format. For each representation format, some decisions were

required in terms of choices among the various parameters that are handled by existing representations. To represent emotional dimensions in EmoTag we have selected the three main dimensions: *evaluation*, *activation* and *power*. To represent emotional categories we have selected 92 emotional categories which try to cover the possible emotional requirements of any text.

4.3. Ontology of Emotions

We have developed an ontology of emotional categories[5] as a fundamental resource for the management of emotional information represented as emotional categories. By using this ontology we can guarantee interoperability between different definitions of emotional values, whether expressed in terms of generic emotional categories, or more specific emotional categories. The ontology is intended as a source to document the relations between different emotional categories. We took emotional categories (i.e. emotion-denoting words such as *happiness*, *sadness* and *fear*) as "first class citizens" of our ontology. This is significantly different from WordNet Affect, which is an ontology of words and their relation with emotions. In our ontology, the emotional categories are structured in a taxonomy that ranges from generic emotions to the most specific emotional categories.

4.3.1. *Generic Emotions.* Based on cluster analysis theory we structured the emotions into clusters. The intention was to integrate the cluster approaches explained in Section 2.2. The first step in structuring the emotions was to decide what the generic emotions in our ontology (the different clusters in our approach) would be. As was concluded in (Ortony and Turner, 1990) researchers cannot identify the basic emotions and we did not even have a satisfactory criterion for identifying basic emotions that is generally acceptable to emotion theorists. However, we tried to find a set of basic emotions in order to create an ontology which allows for the comparison of certain emotions with others. To achieve this we asked ourselves the questions suggested by (Ortony and Turner, 1990): "What exactly do we mean with basic emotions? In what sense are we using the word 'basic'? What would we do with them if we had them?". The answers to those questions were the following:

- For us basic emotions are superordinate emotions such as *sadness* which subsume other more specific emotions such as *grief* or *despair*.
- The word "basic" is used in the sense of super-ordinate emotion, that is, an emotion that it is not subsumed by any other emotion.
- Once we had our set of generic emotions, our goal was to create a hierarchy of emotions whose roots were the generic emotions. This hierarchy would allow us to make comparisons between different emotions in order to determine whether two emotions are the same, similar or totally different. Two emotions are equal if they are different tags for naming the same abstract emotion. They are similar if they belong to the same cluster (i.e. to the same branch of generic emotions), and they are totally different if they belong to opposite abstract emotions or to two different clusters (i.e. two branches of different generic emotions).

We analyzed the cluster analysis approaches presented in Section 2.2, all these approaches include *sadness*, *anger* and *fear* as generic emotions so these three emotions were included in our list of generic emotions. When we compare the rest of basic emotions we can see that there are basic emotions that are also shared by all systems. As explained in (Ortony and Turner, 1990) sometimes the differences between collections of basic emotions are due only to the choice of the tag to refer to the emotion. This is true for *happiness*,

[5] The ontology can be downloaded from nil.fdi.ucm.es/index.php?q=node/201

(Parrott, 2001) refers to *happiness* as *joy*. We can conclude that *happiness* is a common emotion to all the cluster approaches and add **it** to our set of generic emotions with the tag *happiness*. *Surprise* is not included in any of the basic emotion sets, but it is included in most of the classic emotional theories of basic emotions (Frijda, 1986; Izard, 1971; Plutchik, 1980), and there is no emotion in our current list of generic emotions that subsumes the emotion *surprise*; therefore it must also be considered a generic emotion in our ontology.

We also added the term *neutral*, which refers to the lack of emotion, to our set of generic emotions. So far we had defined *sadness*, *anger*, *fear*, *happiness*, *surprise* and the additional term *neutral* as generic emotions.

4.3.2. *Specific Emotions.* Once the ontology was established with its generic emotions we placed each of the emotions from cluster analysis approaches presented in Section 2.2 in the clusters obtained from our set of generic emotions. The next step was to complete our ontology by adding those specific emotions that are found in existing emotion literature such as *disappointment*, *grief*, *intrigue* or *melancholy*.

4.3.3. *Structure of the Ontology.* In our ontology there are concepts that represent language-independent emotions corresponding to common experiences in life. The hypothesis is that we all have the same abstract conception of *Happiness*, for instance, while different words can be used to refer to it. There are also instances in the ontology that represent the words provided by specific languages (e.g. English) for referring to emotions. Therefore, a concept can have multiple instances as a language can give us multiple words to refer to the same emotion. Those instances that correspond to words in a specific language are the ones that were presented to the subjects during annotation.

The root of all emotional concepts in the ontology is the concept *Emotion*. Each emotional concept is a subclass of this root. Emotions are structured in a taxonomy, with the number of levels under each generic emotion depending on the level of available specification for it. For example, *Sadness* has two sublevels of specification. The second level indicates different types of *Sadness*: *Despair*, *Disappointment*, *Grief* or *Shame*. Some of these emotions are specialized again in the third level. For example, *Shame* is divided into *Regret* and *Guilt*. On the other hand, *Surprise* only has one sublevel with two emotional concepts: *Amazement* and *Intrigue*.

Figure 2 shows a fragment of the ontology. It shows emotional concepts like *Happiness*, *Sadness*, *Fear* and *Surprise*. Under those emotional concepts there are instances of these emotional concepts (emotional words) such as *happiness*, *dismay*, *displeasure* and *depression*.

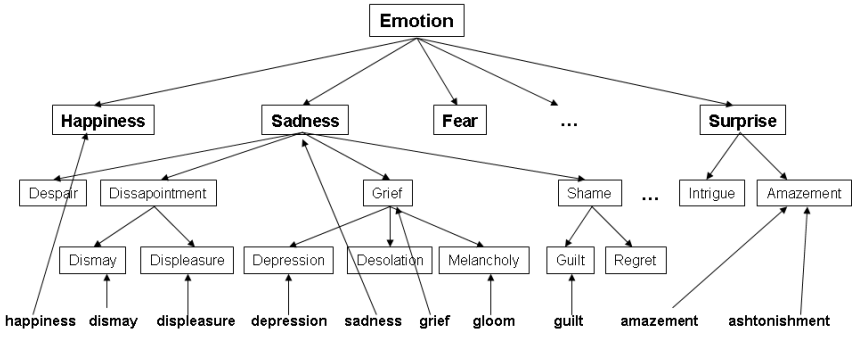


FIGURE 2. Fragment of the emotional ontology

According to the semantics we chose for our ontology, all the instances of the same

emotional concept are synonyms. For example, the words *astonishment* and *amazement* are considered synonyms because both are instances of the emotional concept *Amazement*. Using the ontology we can also obtain the emotion concept directly associated with an emotion word in the ontology, i.e. its parent, as well as with other, more general emotion concepts related to that word, according to the conceptual hierarchy. Finally we can also obtain the synonyms of an emotion word, by noting the siblings of a particular instance.

We have used Solomon (1980)'s theory of motivation/emotion as a theory of opposite emotions in the treatment of negation. This theory views emotions as pairs of opposites (for example, *happiness-sadness*). We have encoded the set of opposite relations between two emotion-denoting words as a table of pairs of opposing emotions (see Table 2). We have designed our table so that one emotional concept has only one opposite emotion. Any emotion in the ontology, regardless of its level of specification, may have an opposite emotion. That is, not only do generic emotions have opposite emotions in our system but so do specific emotions. For example, in the case of *happiness* its opposite emotion is *sadness*, for *euphoria* it is *depression*, for *boredom*, *fascination*, etc. Not all the emotions have an opposite; for example, the opposite of *surprise* is nothing but the neutral state.

Admiration	Contempt
Apathy	Enthusiasm
Boredom	Fascination
Compassion	Gloating
Contempt	Admiration
Depression	Euphoria
Despair	Hope
Disgust	Sympathy
Displeasure	Pleasure
Disappointment	Relief
Enthusiasm	Apathy
Euphoria	Depression
Fascination	Boredom
Gloating	Compassion
Gratification	Remorse
Grief	Pleasure
Happiness	Sadness
Hate	Love
Hope	Despair
Love	Hate
Pleasure	Displeasure
Pride	Shame
Relief	Disappointment
Remorse	Gratification
Sadness	Happiness
Shame	Pride
Sympathy	Disgust

TABLE 2. Table of Pairs of Opposing Emotions

4.4. Corpus

The basic assignment of emotions to words in our system is obtained from a corpus of sentences that has been hand-tagged by human evaluators. Part of this corpus was mined to obtain the set of word-to-emotion assignments used to drive the automatic tagging, and part of it was used to evaluate the tagger. The corpus and the method for constructing it are described in this section.

We have selected eight tales of different length, written in English, making a total of 10,331 words and 1,084 sentences. Tales were chosen according to the practical requirements of our applications, but one of our goals was to cover a broad spectrum of styles by having tales from different authors and time periods. The eight tales are marked up with emotional categories and emotional dimensions by human evaluators. Table 3 shows the author, number of sentences, words and average of words per sentence of each tale contained in the corpus.

Tale	**Author**	**Sentences**	**Words**	**W/S**
The Crystal Ball	Brothers Grimm	84	1,400	17
The Emperor's New Suit	H. C. Andersen	140	1,584	11
The Frog Prince	Brothers Grimm	86	1,205	14
The Image of the Lost Soul	Saki	48	891	19
The Little Match-Seller	H. C. Andersen	56	991	18
The Ox and the Frog	Aesop	16	164	10
The Princess and the Pea	H. C. Andersen	26	373	14
The Tortoise and the Hare	Aesop	15	153	10

TABLE 3. Distribution of sentences, words and words per sentence (W/S) in the tales chosen.

Fifteen evaluators were asked to tag the sentences in the corpus with both emotional categories and emotional dimensions. Evaluators were 53% male and 47% female, with an average age of 27. All of them have a high level of academic studies. Evaluators were provided with a list of different emotions as an aid for tagging sentences with emotional categories, and with the SAM standard (explained in Section 2.1) as an aid for tagging sentences with emotional dimensions.

The tagging results obtained for each sentence from the different evaluators were unified into a single value for the sentence (called the *reference value* for the sentence from now on) using a different method depending on the representation of emotion involved.

Emotional dimensions: In the case of emotional dimensions, the reference value for each sentence in the text was obtained by computing the average of the corresponding values assigned by the evaluators for each emotional dimension. This resulted in a value in the range of [1,9] for each dimension.

Emotional categories: In the case of emotional categories, the simplest possible approach would be to select the emotional category assigned to the sentence by the highest number of evaluators. However, this solution fails to take into account the relations between generic categories and more specific categories which are captured in the ontology. When different evaluators have tagged a sentence at different levels of abstraction, such a solution would arrive at incorrect overall tagging, by failing to identify more specific descriptions of an emotion with neighbouring descriptions at a more abstract level. To avoid this problem, we rely on the taxonomy of emotions provided by the ontology to identify the most specific abstraction of the emotional concepts used that subsumes at least half of the taggings provided by the evaluators. More precisely, we carry out the following process: if at least half of the evaluators agree on the assignment of one emotion, this emotion is taken as the reference value for the sentence. Otherwise, the process of considering interrelations between the taggings due to differences in degree of abstraction is set in motion. First we

group together those taggings that occur at the same level of abstraction, which represent subsets of the taggings that occur at comparable levels of abstraction. Starting at the lowest level, the abstract emotional concepts corresponding to the taggings given are identified, and they are added to the representation of the sentence at the appropriate level of abstraction. If after this operation we have any emotion supported by at least half of the evaluators, we take it as the reference value. Otherwise the process is iterated upwards to the next level of abstraction until we have an emotion supported by at least half of the evaluators.

Figure 3 shows an example of this process. In this example we have a sentence marked up by six evaluators, and there is no emotion supported by at least half of the evaluators. Therefore we follow the process explained above. First, we group the emotions in levels as seen in the Table 3.a. Then we obtain the related concepts for *Grief* (*Distress* and *Sadness*), *Helplessness* (*Powerlessness* and *Sadness*) and *Remorse* (*Regret* and *Sadness*) which are the emotions at the lowest level of abstraction. Third, we add these new concepts to the previous ones at the level of abstraction immediately above (Table 3.b.) In this case we add *Sadness* which is now supported by 4 evaluators (Eval 1, Eval 3, Eval 4 and Eval 5), *Distress* which are supported by 2 evaluators (Eval2 and Eval3), *Powerlessness* supported by Eval4 and *Regret* supported by Eval5. Finally, we take *Sadness* as the reference value because it is supported by 4 evaluators, which is more than half of the evaluators.

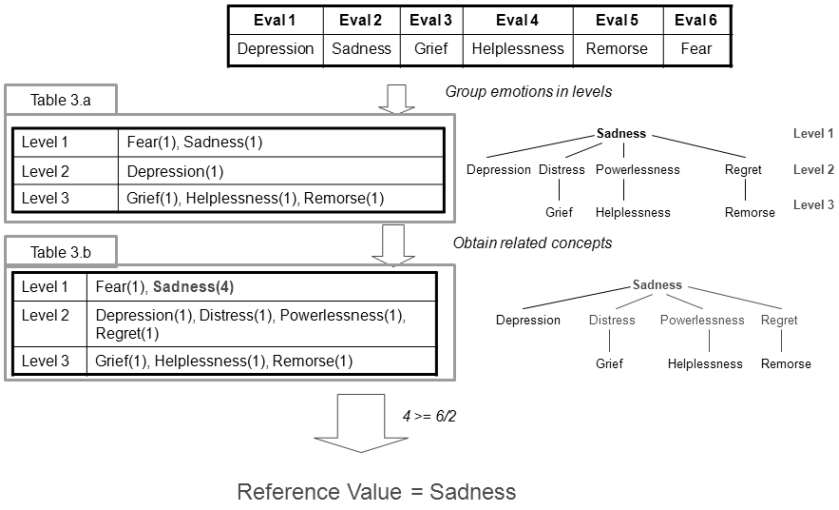


FIGURE 3. Example of how the reference value is obtained from the assignment of emotional categories to a sentence by the evaluators.

In our corpus, the emotions most supported by the majority of the evaluators are *sadness*, *happiness* and *anger*, followed by *fear*, *surprise*, *hope*, *arrogance* and *grief*.

4.5. List of Emotional Words (LEW Resource)

EmoTag marks up text with emotional categories – not just with generic emotions but with specific emotions as well – and emotional dimensions – with numerical values between 1 and 9 for each emotional dimension. To support this process we required a resource capable of associating each word with values in the corresponding ranges: generic and specific emotions when assigning categories, and values for all three dimensions when assigning emotional dimensions. Existing dictionaries based on emotional categories (as explained in Section 3.1) do not mark-up words with the whole set of emotional categories that we are

considering, but rather rely on specific subsets of the available categories, usually restricted to generic emotions. For our purposes, it was imperative that we develop an extended resource capable of meeting the project's needs. In the case of emotional dimensions, if we look at the dictionaries mentioned in Section 3.1 we can conclude that there are only two of them which mark up words with the three emotional dimensions: the General Inquirer's dictionary and ANEW. The General Inquirer´s dictionary does not assign numerical values to each dimension, but instead relies on a binary value. Therefore it would be insufficient for our purpose. ANEW marks up words with a numerical value for each dimension but it is composed of relatively rare words, so its coverage would be inadequate for the purpose under consideration. Nevertheless, both the General Inquirer and ANEW are useful resources and they are taken into account in the process presented in the paper. After considering these aspects we have decided to obtain our own affective dictionary, the List of Emotional Words (LEW), and complement this dictionary with ANEW and the General Inquirer´s Dictionary.

Our List of Emotional Words (LEW)[6] is a resource that associates each word with a set of emotional categories and a tuple of emotional dimensions. Both types of assignments contained in LEW (emotional categories and emotional dimensions) are extracted from the analysis of the assignments given by human evaluators to the sentences in the corpus described in section 4.4. That is, a corpus of eight tales which results in 1,084 sentences. A different extraction process is employed for each method of representing emotional information.

Extraction of Relevant Words. The process of obtaining a set of words with an associated emotional content from a sentence tagged with emotional information involves two basic processes: identifying the set of words to be considered, and computing the emotional content that should be attributed to each of those words.

To achieve the task of identifying the set of words to be considered, the sentences are processed to identify their lexical and syntactic information. The part-of-speech (e.g. noun, verb, etc.) for each word in a sentence is obtained by using the Qtag[7] tagger. The MINIPAR (Lin, 1998) dependency parser is used to obtain the dependency tree for the sentence and the stem of each word. Not all words are considered capable of bearing emotional content. Our system represents this by using a stop list of lexical categories that we consider neutral in terms of emotional content. These lexical categories are represented as part-of-speech tags in the format returned by the Qtag tagger. Our stop list is composed of the following labels: verbs "to be", "to do", "to have" and all their conjugations, conjunctions, cardinal numbers, ordinal numbers, determiners, existential **"there"**, prepositions, modal auxiliaries (might, should), determiners (a, the, this, that), indefinite pronouns (anyone, nothing), possessive particles, personal pronouns, possessive pronouns, reflexive pronouns, symbols (US$500), interjections and adverbs. The set of words in a sentence is filtered to eliminate all words whose lexical category is present in the stop list.

Additionally, we use a second filter based on the binary assignments of emotional value given in the General Inquirer (Stone et al., 1966). A word is considered to bear emotion if it is assigned one of the following labels in The General Inquirer: Positive, Negative, Pstv, Ngtv, Active, Passive, Power, Strong, Submit or Weak. All other words are also filtered out.

To compute the emotional content that should be attributed to each word, a different process is followed for each mode of representing emotional information. In both modes, the issue is whether or not to associate the word with the emotional information assigned to the sentence as the overall result obtained from all the human evaluators (as described in section 4.4).

In the case of *Emotional dimensions*, the decision to associate a given word with the

[6] The LEW resource can be downloaded from nil.fdi.ucm.es/index.php?q=node/186
[7] http://www.english.bham.ac.uk/staff/omason/software/qtag.html

values for specific dimensions is taken based on the information available for that word in The General Inquirer. As described above, the General Inquirer provides binary decisions on each dimension, but no numerical values. The construction process presented here relies on the binary decisions from the General Inquirer to decide when to assign the numerical values provided by the human evaluators to a word. This process, in detail is as follows. For the activation dimension the value associated to the sentence by the evaluators is assigned to the word if the word is labeled as *Active* or *Passive* in the General Inquirer; otherwise a neutral value (5.0) is taken as activation value. For the evaluation dimension the value associated to the sentence is taken if the word is labeled as *Positive*, *Negative*, *Pstv* or *Ngtv* in the General Inquirer; otherwise the neutral value (5.0) is taken. Finally, for the power dimension the value associated for this dimension to the sentence is taken if the word is labeled as *Power*, *Strong*, *Submit* or *Weak* in the General Inquirer; otherwise the neutral value (5.0) is taken.

In the case of *Emotional categories*, the emotional category associated to the sentence is taken directly as the emotion category associated to all the words identified as relevant in the sentence.

Finally, if the evaluation shows that the use of the General Inquirer gives better results than only using the stopwords list, we will use only the General Inquirer as a resource to remove words because the stopwords from our list of stopwords appears in it.

Negation. An important aspect to be considered in the mark-up process is the effect of negation on the emotional connotations of fragments of text. In transferring the emotional connotations of a sentence to the words that appear in it, those words affected by negation require special treatment. Because negation inverts the polarity of the linguistic units under its scope, the construction process takes this feature into account by inverting wherever possible the emotional content attributed to words in this situation. We use a dependency analyzer, MINIPAR, in order to find the negations present in the sentence and the words in the sentence under their scope. For example, for the sentence "I am not happy" we must identify the negation and the word *happy* as the word under its scope. For words under the scope of negation, the complement of the emotional connotation assigned to the sentence must be computed. This is done differently, depending on the particular mode of representation of emotion involved.

In the case of *Emotional dimensions*, the complement of an emotional connotation is defined as the point in the range of emotional dimensions that is symmetrically opposite the given emotional connotation with respect to the point of reference in the range. This symmetric point is obtained by getting the complementary value for each dimensional value (*activation*, *evaluation* and *power*). For example, if the sentence *"I am not happy"* is marked up as *eval=1*, *act=4* and *pow=2* by evaluators, the opposite emotional value for *happy* (which is the word under the negation scope) is *eval=9*, *act=6* and *pow=8*.

In the case of *Emotional categories*, the opposite emotional value is obtained by looking up the given emotional category in our table of pairs of opposing emotions. For example, if the sentence *"I am not happy"* is marked up by evaluators as *sadness*, EmoTag obtains *happiness* from the opposite emotions table as the opposite emotion and the word *happy* is inserted in the LEW resource with the emotion *happiness* associated to it.

Insertion of Words in LEW Resource. As a given word may occur more than once in the corpus, and each occurrence will be part of a different sentence, different emotional values for the same word may result from occurrences in different contexts. The construction method for LEW takes this into account by building the entries for particular words incrementally, updating them progressively whenever they occur more than once, and normalising the final values once the entire corpus has been processed. This incremental process is defined as follows. Each particular occurrence of a word in the corpus provides a candidate entry containing the word, its part-of-speech, its stem and the emotion associated to it at

that occurrence. This candidate entry is incorporated into LEW differently for each mode of representation.

In the case of *Emotional dimensions*, every new word is inserted into the word database with the values for dimensions corresponding to that particular occurrence; if the word is already in the list, the values for each dimension corresponding to that particular occurrence are added to the values for the corresponding dimensions stored in the list (which represent the accumulated values for all occurrences so far).

In the case of *Emotional categories*, every new word is inserted into the word data base and associated with a vector corresponding to the complete set of available categories, with the value 1 in entry of the vector corresponding to the emotion assigned to the particular occurrence, and the rest of the entries in the vector set to 0; if the word was already in the list, the system adds 1 to the field of the vector corresponding to the emotion assigned to the particular occurrence.

Normalization. Once the whole corpus has been processed, the accumulated values in the list of entries must be normalised to convert them into the correct range for representing emotion in the chosen representation methods.

In the case of *Emotional dimensions*, after the incremental construction process each word is associated with values for the three dimensions which represent a description of the emotional connotation of the word over a range with dimensions described over a scale of N times the original range of values, with N being the number of occurrences of the word in the corpus. These values can be converted to the expected range of values by dividing the value for each dimension by the number of occurrences of the word in the corpus.

In the case of *Emotional categories*, following the incremental construction process, each word in the list has become associated with a set of non-zero numerical values corresponding to those emotions assigned to the word by the construction process in at least one occurrence. Based on these values we obtain the final weighting for each word-to-emotion assignment. We obtain the collocation factor, a variation of pointwise mutual information (Manning and Schütze, 1999; Yang et al., 2007) for each pair *word-emotion* in our LEW resource. The collocation factor measures the collocation strength between an emotion *e* and a word *w*:

$$co(e,w) = c(e,w) * log\left(\frac{P(e,w)}{P(e)P(w)}\right)$$

$$P(e,w) = \frac{c(e,w)}{N}, P(e) = \frac{c(e)}{N}, P(w) = \frac{c(w)}{N}$$

c(e,w) = Co-occurrences of the word w and the emotion e in the corpus.

c(e) = Occurrences of the emotion e in the corpus.

c(w) = Occurrences of the word w in the corpus.

N = Total word occurrences in the corpus.

Expansion: Since the process of automated tagging described in the paper relies on the LEW list to provide basic emotional connotations to lexical items, from which estimations of the emotional connotation of sentences are built, the tagger can only be expected to perform well when the lexicon employed in the input texts it is processing is reasonably covered by the LEW list. To maximise coverage, the LEW list is extended by means of the relations of synonymy given in WordNet (Fellbaum, 1998). The assumption underlying this operation is that words that are synonyms are likely to have similar emotional connotations. This obviously constitutes an approximation, but the possible loss of precision introduced is compensated by the extension in coverage. The set of synonyms appearing in WordNet for each word resulting from the initial construction process of the LEW list is added to the LEW

list (associated with the same emotional connotation as the word already appearing in the list). This process is not applied recursively (the additional words obtained from WordNet are not themselves looked up for synonyms, as this would either point back to synonyms already found, or contaminate the list with alternative meanings not so directly related to the initial concept). In case of polysemy the first synset (the most frequent) is used by EmoTag. The expansion method results in a list of 3,027 emotional words.

Examples. For emotional dimensions the LEW resource stores a value for *activation*, *evaluation* and *power* for each pair (word's stem, label). The stored values for the dimensions correspond to the average value of *activation*, *evaluation* and *power* of that pair in all the texts analyzed. Table 4 shows the entries for 4 different words in the List of Emotional Words (LEW) resource for emotional dimensions. For each word, each entry includes the stem, the label corresponding to the POS tag, and the values for the dimensions of activation, evaluation and power as obtained from the extraction process.

TABLE 4. Fragment of the LEW resource for emotional dimensions

Word's Stem	Label	Act.	Eval.	Power
death	Noun	6.5	3	3.5
dark	Noun	5.2	4.8	5
grateful	Adj.	5	8	6
happy	Adj.	4.8	5.8	4.1

For emotional categories the LEW resource stores the collocation factor for each pair (word's stem, label). Table 5 shows the entries for 4 different words in the List of Emotional Words (LEW) resource for emotional categories. For each word, each entry includes the stem, the label corresponding to the POS tag, and the values obtained from the extraction process for each of the emotional categories under consideration.

TABLE 5. Fragment of the LEW resource for emotional categories

Word's Stem	Label	Grief	Sad	Happy	Neutral	...
death	Noun	22.234	6.590	0	2.317	...
dark	Noun	0	6.844	6.704	2.945	...
grateful	Adj.	0	0	1.376	0	...
happy	Adj.	0	0	19.011	2.772	...

4.6. Automatic Mark-up of Emotions in Texts

Our process assigns emotional connotations (represented in two different frames of reference) to English sentences. An input text is split into sentences and tokenised. Then assignment of emotional connotations to the sentence is carried out based on the set of emotion words that appear in the sentence and the relative positions they occupy in its syntactic structure.

Obtaining the Words: For each word in the input sentence we obtain its part-of-speech (using Qtag) and its stem (as given by MINIPAR), and we identify whether it occurs under the scope of a negation (based on the dependency tree built by MINIPAR). Words unlikely to bear emotional connotations are filtered out (based on the General Inquirer and the same stoplist of POS tags used during the construction of the LEW list).

Obtaining the Emotional Value Associated to the Words: Once we have the stem and the part-of-speech of the word, we look up this pair in the available data bases depending on the representation of the emotions selected.

In the case of *emotional dimensions* there are two available data bases. First EmoTag

looks up the words in the LEW resource. If the word is not in the LEW resource, EmoTag looks it up in the ANEW list.

In the case of *emotional categories* there is only one available database, the LEW resource.

If the word is not in the available data bases, EmoTag looks for its hypernym in WordNet, and looks up this hypernym in the available data bases. This is done recursively until a matching entry in the database is found. If none of the hypernyms appear in the available data bases, the word is left out of the mark-up process.

Negation: EmoTag identifies the sentences with a negation and reverses the emotional content of the words under the scope of the negation.

In the case of *emotional dimensions*, in order to obtain the reverse emotional content of an emotion which is identified by the values of the three emotional dimensions, EmoTag obtains the complementary value in the scale of each dimensional value. For example, the opposite value for the emotional value *act=8*, *eval=5* and *pow=3* is *act=2*, *eval=5* and *pow=7*.

In the case of the *emotional categories* the reverse emotional value is obtained by means of our table of pairs of opposing emotions.

Obtaining the Final Value of the Sentence: Once we have the emotional value associated to each of the emotion-bearing words of the sentence, we have to determine the final emotional value of the sentence based on the emotions associated to each of the words.

In the case of *emotional dimensions*, a representation of the emotional connotation of the sentence is obtained by taking the average values of each dimension for all the words in the set of emotion-bearing words in the sentence.

In the case of *emotional categories*, a first approximation for a representation of the emotional connotation of the sentence is obtained by merging the vectors of weights for emotional categories of all the emotion-bearing words from the sentence (adding together the values when different words provide different values for the same category). This provides a single vector for the sentence. However, this approximation suffers from the problem of emotional descriptions from different origins operating at different levels of abstraction (already described in section 4.4 when discussing the task of post-processing the results of human evaluators). As in that case, we can use the taxonomy implicit in the ontology to reduce the impact of this problem. If we have two concepts at different levels of generalization, it will be better to consider their collocation factors together under the more general concept using the emotional ontology. For all emotional categories with non-zero values (collocation factors) in the vector representing the emotional connotation of the sentence, each one is recursively replaced by its immediate parent in the ontology. If that emotional concept is already present, the two corresponding collocation factors are added together. Finally, the most specific emotion (the emotion with the most abstract level in the ontology) with the biggest collocation factor is assigned to the sentence. Figure 4 shows an example of how the final value of a sentence containing three emotional words is obtained. In this example we have a sentence with three emotional words. Table 4.a. shows the collocation factors for each word. The first step is to group the emotions according to their level in the ontology. Table 4.b. shows the result of this step. The second step is to obtain the related emotional concepts for each emotional category. For *Indignation*, *Sulking* and *Displeasure* we have *Anger* as a related concept and for *Amazement* we have *Surprise*. The related emotional concepts are added to the previous ones as can be seen in Table 4.c. If we look at the last table, it can be seen that the emotional category with the biggest collocation factor is *Anger*, and this is the emotion selected by EmoTag as final emotional value for the sentence.

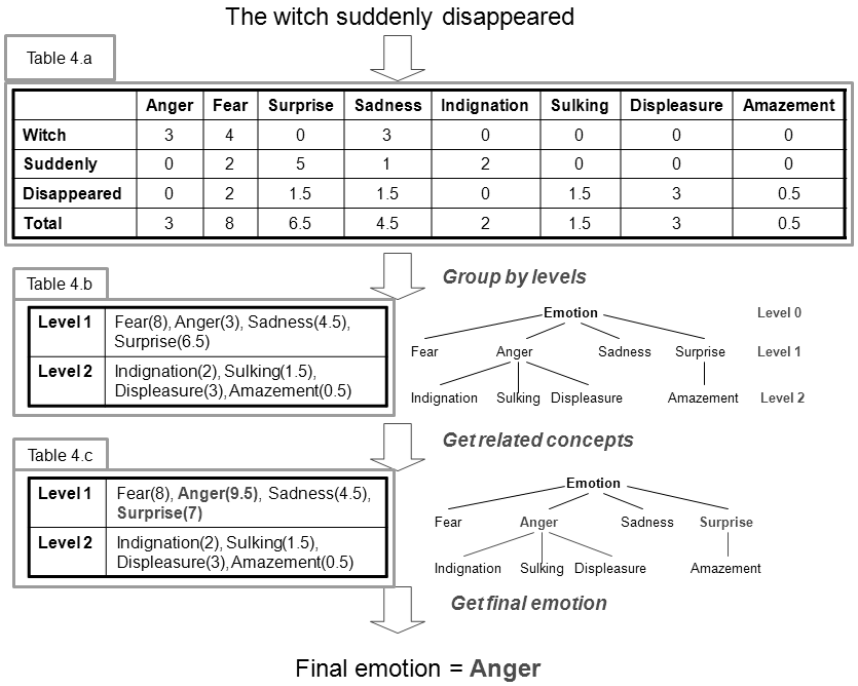


FIGURE 4. Example of how EmoTag obtains the final emotion for a sentence based on the emotional values of the words of which it is composed.

5. EVALUATION

This section presents the evaluation of EmoTag and the results obtained. As explained in section 3.2 there is no system which does the mark-up of text using the emotional descriptors used by EmoTag (the three main emotional dimensions, generic emotional categories and specific emotional categories) so there are no systems that can be taken as a baseline to see how well our system's method performs. Moreover, the accuracy measures that have been seen, also in Section 3.2, are not standard measures because they have not been used in most of the systems; each system has used its own standard measure. We are dealing with a spectrum of emotional descriptors that is much larger than any of the systems presented in section 3.2 and there is an absence of any standard measure, so we created our own evaluation measures, which will be explained next in section 5.2.

In order to evaluate the system explained in this paper we must take into account that we are evaluating a system which aims to mark up text with emotions as a human would. As we have explained in Section 4.4 the mark-up done by the evaluators is far from unanimous. The assignment of emotions to sentences is a subjective task and this subjectivity must be taken into account when it comes time to evaluate the system. This idea was suggested by some of the works presented in the AAAI Spring Symposium on Exploring Attitude and Affect in Text in 2004. To achieve this, we consider that the assignments provided by the automatic tagger are correct whenever they match those originally provided by the evaluators, accepting a margin of error equivalent to the disagreement between different human evaluators. For example, suppose we have the sentence *Cinderella had a wonderful time at the ball* marked up by six evaluators. Three of them selected 8 as the value for evaluation and the other three selected 7. In this case the reference value that would be compared with the value returned by our application would be 7.5. If we did not take into account a margin of error equivalent to the disagreement between evaluators, we would consider the sentence correctly marked up

only if our system marks the sentence with 7.5. But we consider that **it** is better to take into account the margin of error due to disagreement and consider the sentence correctly marked up if our system marks up the sentence with a value between 7 an 8.

5.1. Material

In order to evaluate our work we carried out two different tests.

(1) We measured the results of the different individual steps we have adopted during the mark-up process (handling negation, looking up hypernyms on WordNet, etc.). We obtain the results for the basic version of our system; that is, the system splits the sentence into words and discards the words which appear in our POS stop list. It then looks for the remaining words in LEW. If a word is not in LEW, EmoTag discards it and does not take part in the process. In the basic version the final value of the sentence is based on the emotions associated to the words which compose the sentence, selecting the emotion associated to the majority of the words as the final emotion, instead of using the ontology of emotions. Once we have the results for the basic version of our mark-up process we apply the modifications (treatment of negation and subordinates, use of ANEW and ontology, etc.) one by one to see their individual effect. The results obtained with **these** modifications are shown in Figures 6, 7, 8 and 10. We selected the best modifications after **analysing** the results obtained.
(2) We obtain the results for the definitive version, the one which contains only the best modifications detected in the previous test. These results are shown in Figures 9 and 10.

Two tales formed part of these tests which are English folk tales with different numbers of words. Table 6 shows the author, number of sentences, words and average of words per sentence of each tale that formed part of the evaluation.

Tale	Author	Sentences	Words	W/S
Rapunzel	Brothers Grimm	104	1,400	14
The Fox and the Crow	Aesop	10	164	16

TABLE 6. Distribution of sentences, words and words per sentence (W/S) in the tales chosen for the evaluation.

5.2. Metrics

Emotional Dimensions: To evaluate our tagger we have divided the evaluation according to the different dimensions. In order to get a measure of the precision of our tagger we have taken metrics first from the tales marked up by the evaluators and then from the tales marked up by the tagger.

For tales marked up by the evaluators we have, as reference data, the values assigned to each dimension and each sentence by the human evaluators. An average emotional score for each dimension in a sentence is calculated as the average value of those assigned to the corresponding dimension by the human evaluators, as it was explained in Section 4.4. For each sentence, the deviation between the mark-up of each evaluator and the average emotional score is obtained. The average value of the individual deviations of each evaluator with respect to the average emotional score of each sentence is called the *margin of subjectivity*. This value is taken to indicate the possible range of variation due to human subjectivity.

In the case of tales marked up by the tagger for each dimension, if the deviation of the tagger with respect to the average of the values provided by the evaluators is less than or equal to the margin of subjectivity among evaluators, we consider that the sentence is tagged

correctly. The margin of subjectivity for the entire corpus in the case of the evaluation and power dimension is 1 and for activation dimension is 1.2. The scale of each dimension is a 9-point scale, so the margin can be considered low.

Emotional Categories: For each tale we have the emotional label assigned to each sentence available. As we have done for emotional dimensions we have taken metrics first from the tales marked up by the evaluators and then from the tales marked up by the tagger.

For tales marked up by the evaluators we have obtained a reference value as explained in Section 4.4.

In the case of the tales marked up by the tagger the reference value obtained in the evaluators' tales is used to compare with the results generated by our tagger. In order to determine how well a text is marked up, EmoTag assigns a score, which ranges from 0.0 to 1.0, to each sentence. The score is based on the number of levels in common between the two emotions and the level of the more specific emotion, as follows:

$$Correct = number\_common\_levels / level\_specific\_emotion$$

For example, in the case of a sentence marked by the evaluators as *excitement* and by EmoTag as *enthusiasm*, as can be seen in Figure 5, we see that *excitement* has *Excitement*, *Enthusiasm* and *Happiness* as related concepts, so it has the level 3, and *enthusiasm* has *Enthusiasm* and *Happiness* (level 2) as related concepts. *Excitement* and *enthusiasm* have two levels in common (*Enthusiasm* and *Happiness*) and the more specific concept is *Enthusiasm* (level 3) so the result is: correct = 2/3 = 0.67.

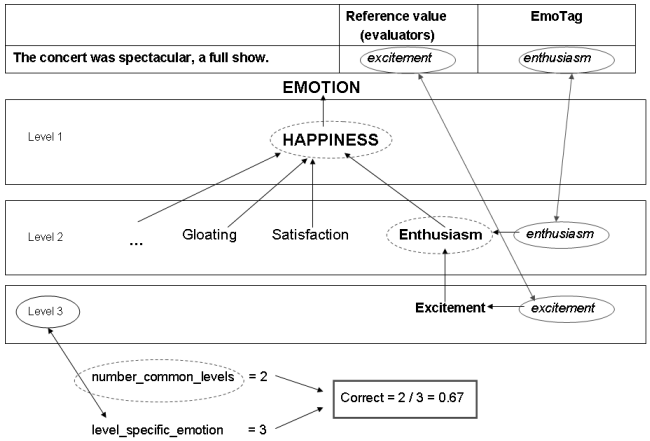


FIGURE 5. Example of the assignment of sentence scoring in the evaluation to texts marked up with emotional categories.

5.3. Results

First of all, we took metrics for the different important aspects that we added while we developed EmoTag. The first step was to obtain a simple baseline that shows the proportion correct if always assigning the affect label which covers the largest proportion of sentences. In the case of emotional categories the affect **label assigned** by the baseline is *neutral* which is the most frequent emotion assigned by the evaluators in the corpus. In the case of emotional dimensions the values assigned by the baseline to the sentences are the average value of each dimension in the whole corpus (evaluation=5.04, activation=5.18 and power=5.39). The second step was to obtain the results for the basic version of our system; that is, the system splits the sentence into words and discards the words which appear in our POS stop

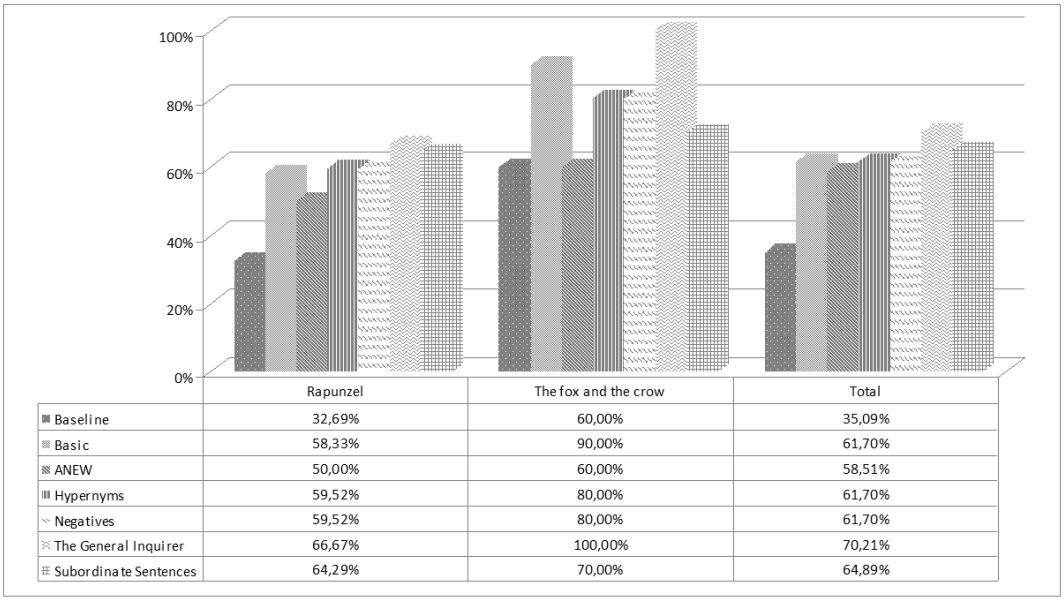
list. It then looks for the remaining words in LEW. If a word is not in LEW, EmoTag discards it and it does not take part in the process. In the basic version the final value of the sentence is based on the emotions **associated with** the words which compose the sentence, selecting the emotion **associated with** the majority of the words as the final emotion, instead of using the ontology of emotions. Once we have the results for the basic version of our mark-up process we apply the modifications. These modifications are:

- ANEW: ANEW list is used, in the mark-up process, as an additional list for obtaining the emotional value **associated with** the words that do not appear in the LEW resource (ANEW is only available for emotional dimensions).
- Hypernyms: In the mark-up process if a word is not in the available data bases, WordNet is used to look up hypernyms of those words.
- Emotional ontology: The emotional ontology is used in the mark-up with emotional categories to reduce the impact of the problem of emotional descriptions from different origins operating at different levels of abstraction. The emotional ontology is used in the last step of the mark-up process to obtain the final value of the sentence based on the emotional values **associated with** each of the emotions-bearing words of the sentence.
- Negation: The emotional content **associated with** words under the scope of negation **is** inverted.
- Subordinate clauses: Subordinate clauses are considered as emotional units inside a larger unit, the **main** sentence. The emotion associated with each subordinate clause is obtained and then this emotion is considered as a single emotion-bearing element within the encompassing sentence.
- The General Inquirer: Use of the General Inquirer in order to determine the "emotional" words. The General Inquirer is used, first, when we get the LEW resource (the General Inquirer is used to identify the set of words to be inserted in LEW) and second, in the mark-up process to filter out words unlikely to bear emotional connotations.

We established the success rate (percentage of sentences correctly marked-up by EmoTag, that is **the** number of sentences correctly marked-up by EmoTag with respect to the total number of sentences) obtained by EmoTag with each of the modifications and we determined which of these aspects really improved the results. The results based on these decisions were measured for the two tales that formed part of the evaluation and for both of them together (Total bars). The success rate shows the percentage of sentences that are correctly marked up by the system.

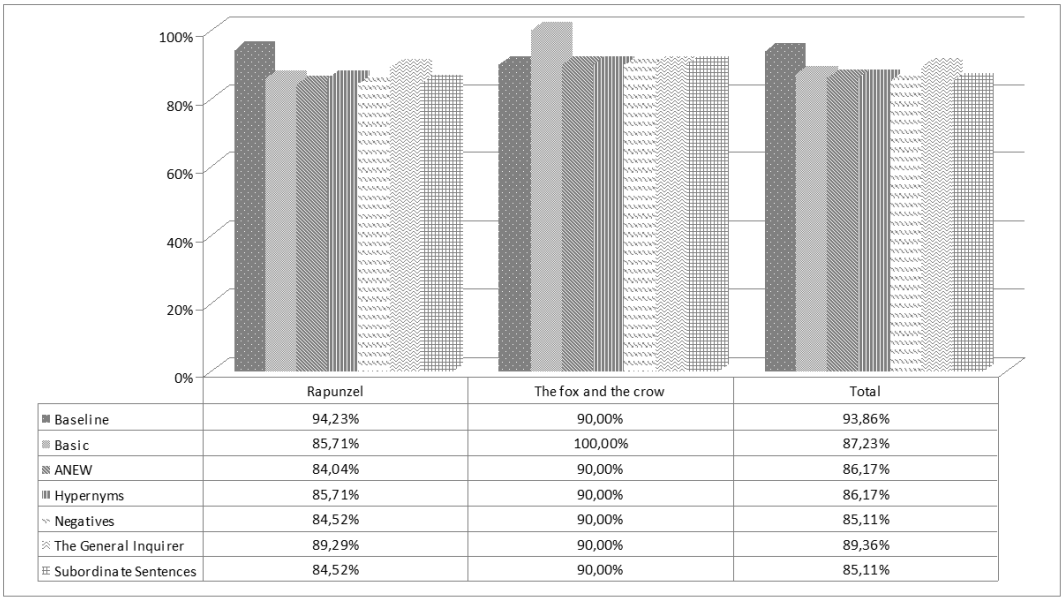
Emotional dimensions: Figures 6, 7 and 8 show the modifications of the different aspects that have been added to the basic prototype of EmoTag. The modifications added to the mark-up with the emotional dimensions are: the use of ANEW list, hypernyms, treatment of negation and subordinate clauses and the use of General Inquirer in order to determine the "emotional" words.

From results we can draw some important conclusions. First of all, we can conclude that **with** respect to the baseline the basic version of EmoTag increases the success rate for evaluation dimension in more than a 25% but decreases the success rate for activation and power dimensions. In the case of power dimension the version of EmoTag with The General Inquirer increases the success rate **with** respect to the baseline in more than 20%. With respect to modifications of the basic version of EmoTag we can conclude that the ANEW list decreases the success rate. This supports the hypothesis that the LEW resource (content dependent) is a better solution when used on its own than when combined with the ANEW list (**content-independent**). The correlation between Power and Evaluation in ANEW is 0.84, suggesting that they are not independent dimensions at all. This may be why power and evaluation lead to the most confusing tagging results. The use of hypernyms which are looked up in WordNet results in an improvement for evaluation dimension. In this



	Rapunzel	The fox and the crow	Total
Baseline	32,69%	60,00%	35,09%
Basic	58,33%	90,00%	61,70%
ANEW	50,00%	60,00%	58,51%
Hypernyms	59,52%	80,00%	61,70%
Negatives	59,52%	80,00%	61,70%
The General Inquirer	66,67%	100,00%	70,21%
Subordinate Sentences	64,29%	70,00%	64,89%

FIGURE 6. Success rate for the different modifications added to EmoTag for evaluation dimension.

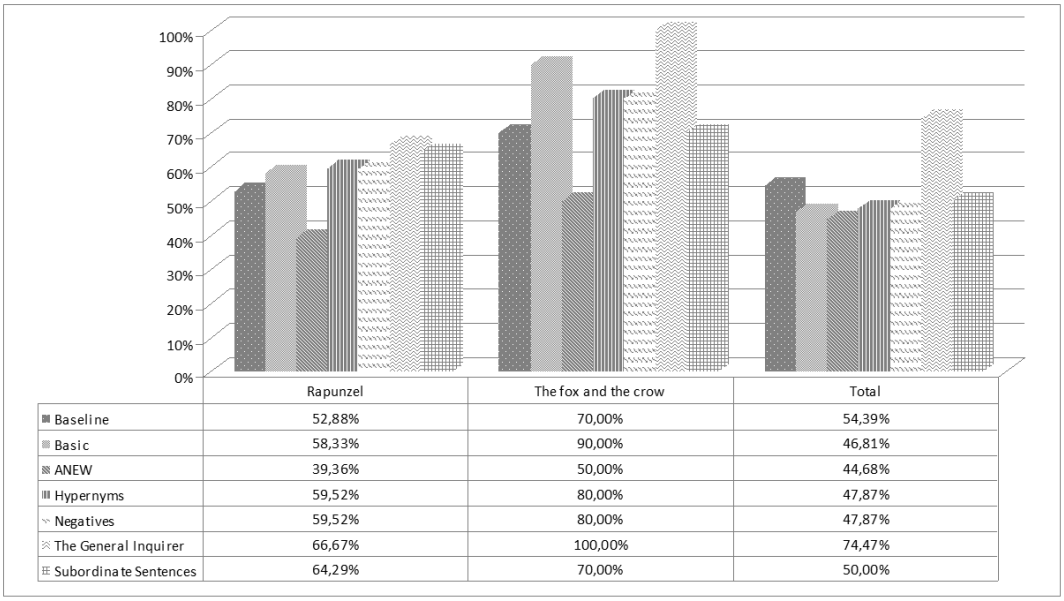


	Rapunzel	The fox and the crow	Total
Baseline	94,23%	90,00%	93,86%
Basic	85,71%	100,00%	87,23%
ANEW	84,04%	90,00%	86,17%
Hypernyms	85,71%	90,00%	86,17%
Negatives	84,52%	90,00%	85,11%
The General Inquirer	89,29%	90,00%	89,36%
Subordinate Sentences	84,52%	90,00%	85,11%

FIGURE 7. Success rate for the different modifications added to EmoTag for activation dimension.

way EmoTag partially rectifies the content dependency of the LEW resource. Treatment of negation does not result in any improvement. Treatment of subordinate sentences results in an improvement only for the *evaluation* dimension. The use of the General Inquirer results in an improvement for the *evaluation* and *power* dimensions.

Based on these results we have decided not to use the ANEW list in combination with the LEW resource and not to treat negation in the final version of EmoTag when it marks up



	Rapunzel	The fox and the crow	Total
Baseline	52,88%	70,00%	54,39%
Basic	58,33%	90,00%	46,81%
ANEW	39,36%	50,00%	44,68%
Hypernyms	59,52%	80,00%	47,87%
Negatives	59,52%	80,00%	47,87%
The General Inquirer	66,67%	100,00%	74,47%
Subordinate Sentences	64,29%	70,00%	50,00%

FIGURE 8. Success rate for the different modifications added to EmoTag for power dimension.

with emotional dimensions. We have decided to use WordNet in order to look up hypernyms and treat subordinates only for the *evaluation* dimension and to use The General Inquirer only for the *evaluation* and *power* dimensions while the list of stop POS tags will be used to determine the "emotional" words for the *activation* dimension.

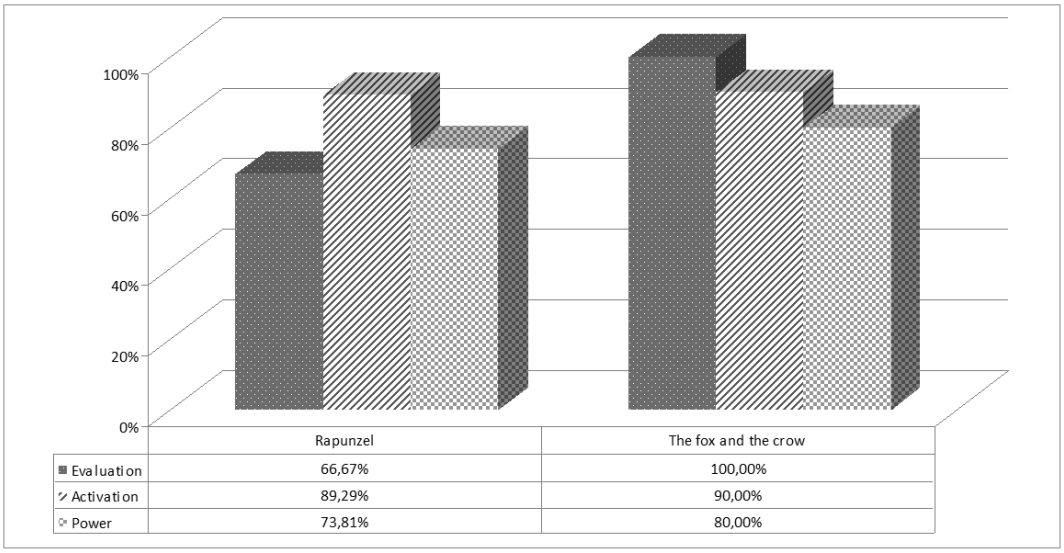
The deviation among the evaluators is around 1 for the *evaluation* and *power* dimensions and 1.2 for *activation*. The deviation between the values obtained by EmoTag and the reference values obtained from the corpus is around 0.9 for the *evaluation* dimension, and around 0.8 for the *activation* and *power* dimensions. If we look at these results, we can conclude that the deviation of EmoTag is lower than the deviation among the evaluators.

The graph in Figure 9 shows the percentages of success obtained for each tale in the final version of EmoTag when it marks up with emotional dimensions.

Emotional Categories: Figure 10 shows the results of the baseline, the basic version of EmoTag and in the different aspects that have been added to the basic prototype of EmoTag when it marks up with emotional categories. The basic decisions added to the mark-up with emotional categories are: the use of hypernyms, the treatment of negation, the use of the emotional ontology, the treatment of subordinate sentences and the use of the General Inquirer in order to determine "emotional" words.

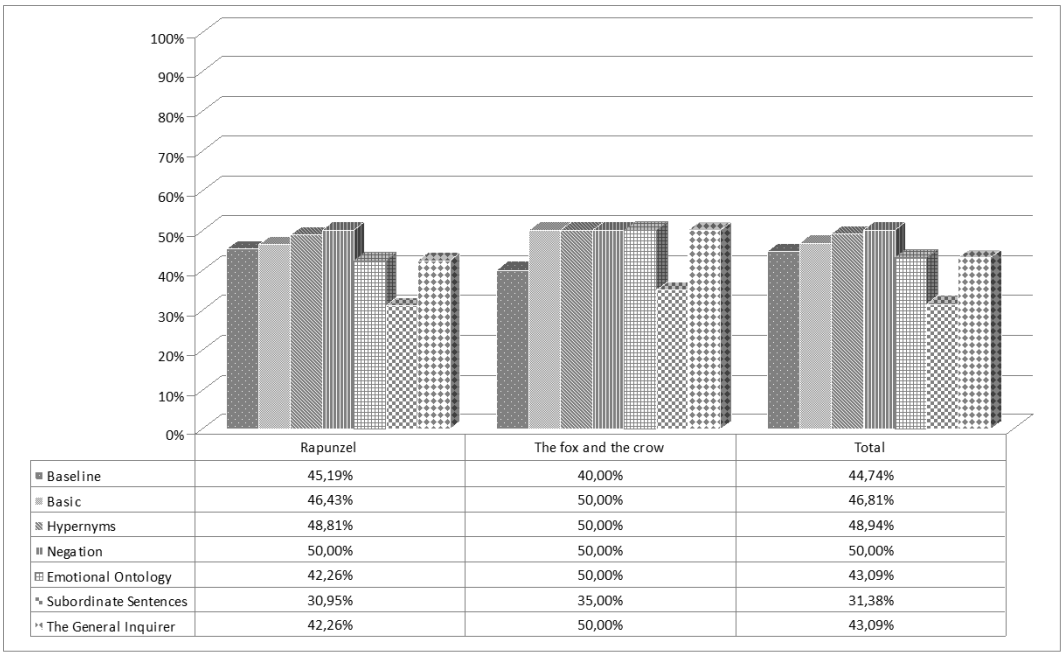
If we look at the results, we find some important conclusions. First of all, we can conclude that the basic version of EmoTag increases the success rate **with** respect to the baseline in more than a 2%. With respect to modifications of the basic version of EmoTag we can conclude that the use of hypernyms resulted in an improvement of 1.5% in the success rate. The treatment of negation resulted in an improvement of 1.5% in the success rate. The use of emotional ontology resulted in an improvement of 2%. Treatment of subordinate sentences resulted in an important reduction in the success rate. The use of the General Inquirer did not result in a reduction or an improvement.

After analyzing these results we have decided to use only the hypernyms, the emotional ontology and treatment of negation, and discard the treatment of subordinate sentences and



	Rapunzel	The fox and the crow
Evaluation	66,67%	100,00%
Activation	89,29%	90,00%
Power	73,81%	80,00%

FIGURE 9. Success rate of EmoTag for the emotional dimensions in the final version of EmoTag.



	Rapunzel	The fox and the crow	Total
Baseline	45,19%	40,00%	44,74%
Basic	46,43%	50,00%	46,81%
Hypernyms	48,81%	50,00%	48,94%
Negation	50,00%	50,00%	50,00%
Emotional Ontology	42,26%	50,00%	43,09%
Subordinate Sentences	30,95%	35,00%	31,38%
The General Inquirer	42,26%	50,00%	43,09%

FIGURE 10. Success rate for the different modifications added to EmoTag for emotional categories.

the use of the General Inquirer in the final version of EmoTag when it marks up with emotional categories instead of using the list of stop POS tags.

To obtain the reference value for each sentence in the corpus based on the emotions selected for each evaluator we had two options: either to consider that the evaluators agree only if they marked up the sentence with the same emotion or to use the ontology in order

to find out if the evaluators agree on the emotion selected. The percentage of sentences in which the majority of the human evaluators - half of their number plus one - agreed on the assignment of an emotion during the construction of the corpus is around 85% when the emotional ontology is used to handle different levels of abstraction (as explained in Section 4.4). In the previous version of EmoTag (without emotional ontology), where the evaluators were considered to agree only if the selected emotion was exactly the same one (without taking into account the relation of emotional categories in the ontology), agreement was around 75%. This shows that the use of the emotional ontology resulted in an improvement of 10%, regardless of the automatic mark-up process.

The graph in Figure 11 shows the percentages of success obtained for each tale and the percentage of sentences incorrectly annotated that correspond to a sentence in which the majority of the evaluators did not agree.

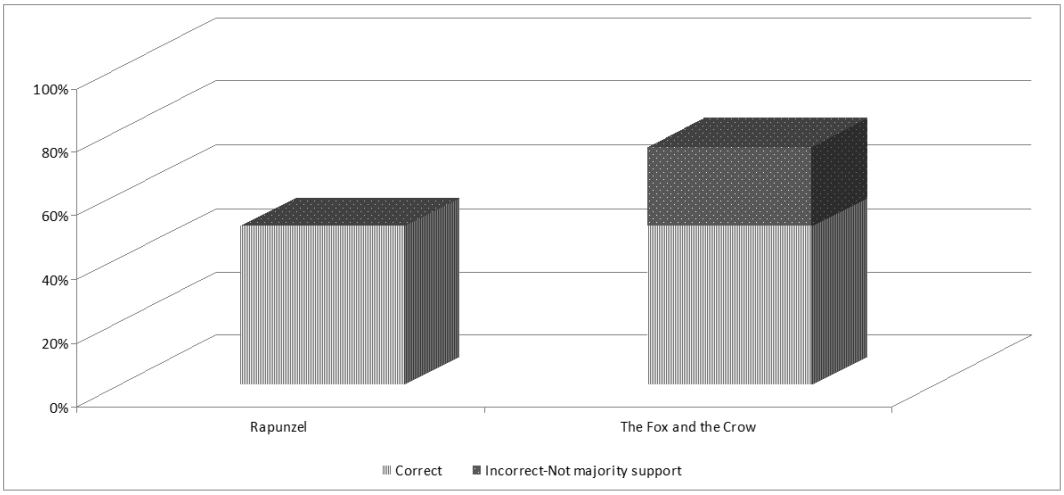


FIGURE 11. Success rate of EmoTag for the emotional categories in the final version of EmoTag.

The emotions with the highest percentage of correct sentences marked up are *bravery*, *optimism*, *relief*, *happiness* and *sadness*, followed by *anger* and *affection* and with a lower, but important, percentage of correct sentences we have *care for*, *fear*, *surprise*, *alarm* and *hope*.

The use of ANEW, a content-independent emotion list, does not improve upon the use of the LEW resource, which is content-dependent. A better way of complementing our LEW resource in order to reduce the weaknesses of a content-dependent approach is the use of hypernyms. In this way if a word is not in our LEW resource, EmoTag obtains its hypernyms from WordNet and then looks them up in the LEW resource. The use of hypernyms has resulted in an increment in the success rate for EmoTag with both emotional dimensions and emotional categories. ANEW is composed of relatively rare words and probably had a relatively low hit rate for the fairy tale corpus. As future work we would research the use of other emotional dictionaries such as the Dictionary of Affect in Language which has a 90% hit rate for natural English (Whissell, 2009). The treatment of negation resulted in an important improvement only for emotional categories.

The use of an emotional ontology in the mark-up with emotional categories resulted in an significant improvement in both the success rate and the agreement of evaluators.

Finally, the use of an approach to content analysis of textual data such as the General Inquirer in order to determine the "emotional" words of the sentence only resulted an improvement in the case of the *evaluation* and *power* dimensions.

If we compare the results of the two approaches (emotional categories and emotional dimensions), taking into account that the metrics used to evaluate each approach are different, we can see that the best results are obtained with the emotional dimensions approach, and more specifically with the *activation* dimension, which obtains **an** 89% success rate.

6. DISCUSSION

EmoTag is a system that marks up texts with emotional content in the domain of fairy tales. The emotional unit of our marker is the sentence; each sentence which makes up a text is assigned an emotion. The emotion associated to each sentence can be represented by means of emotional dimensions or emotional categories which makes EmoTag into a very flexible resource that could be useful both to research **systems** based on emotional categories (specific or generic) and to **systems** based on emotional dimensions. EmoTag marks up English texts. The assignment of emotions is based on the relation between the words and the emotions. These relations are stored in an affective dictionary (LEW), which is content-dependent and covers both representations of emotion (categories and dimensions). EmoTag uses WordNet in order to look for hypernyms of words that do not appear in our affective dictionary in an attempt to reduce the content-dependency of the LEW resource; that is, with WordNet we try to obtain the emotional value of words which do not appear in our LEW resource. If a word does not appear in our corpus, it is not in the LEW resource so by means of hypernym relations between words, we obtain a related word which appears in the LEW resource.

The approach applied to handle the compositionality of emotion across text units (that is, establishing a relation between the emotion of a sentence with the emotion of its words) corresponds in essence to a bag-of-words approach. This type of approach has proven to be extremely successful in information retrieval. It also presents the advantage of allowing **symmetrical** treatment of the processes of extracting emotional words from annotated corpora on one hand, and automated tagging of unseen text based on the list of emotional word on the other. More elaborate solutions may be considered that take into account the syntactic and semantic context of occurrence of each emotional word. However, such solutions are not popular in language processing, posssibly due to their high complexity and a lack of evidence of performance improvement.

EmoTag considers negation, so that all the words that are in the scope of the negation invert their original emotion (the one inherited from the sentence's emotion). The solution adopted for identifying the scope of negation is comparable to standard practices in treatment of polarity. The solution adopted for opposite emotions is based on Solomon (1980)'s theory of motivation/emotion (as detailed in the paper). The solution adopted for assigning complementary values in the emotional dimension space corresponds to the symmetrically opposite point in the space being considered. Once a spatial representation based on dimensions is accepted, this seemed to be the obvious solution for dealing with inversions of polarity. However, the poor results obtained by the modification for treatment of negation when applied over emotional dimensions suggest negation may affect some dimensions more than others.

Another important aspect of our system is the use of an emotional ontology, an ontology of emotional categories which provides the following advantages: it improves the comparison between descriptions of emotion based on categories (by allowing emotions expressed

at different levels of abstraction to be compared with one another) and it provides the means for obtaining more general categories from more specific categories.

These advantages improve the results obtained by EmoTag. Another improvement in our system is the use of the dictionary used by the General Inquirer as a filter for isolating the emotion-bearing words in a sentence. In this way, EmoTag does not take into account all the words that compose a sentence but only those words that are consider "emotional" by the General Inquirer.

If we compare our system to the ones mentioned in Section 3.2, we can see that EmoTag differs from these approaches in different ways. EmoTag marks up the 3 emotional dimensions, whereas (Bestgen, 1993), (Pang et al., 2002), (Read, 2005), (Popescu and Etzioni, 2005), NTCIR, TREC, (O'Connor et al., 2007), (Snyder and Barzilay, 2007) or the systems from SemEval (Strapparava and Mihalcea, 2007) only mark up the evaluation dimension. EmoTag can mark up text not only with a reduced number of emotions, as done by (Zhe and Boucouvalas, 2002), (Liu et al., 2003), (Alm, 2009), (Mihalcea and Liu, 2006), (Mihalcea and Liu, 2006), (Sugimoto and Yoneyama, 2006), (O'Connor et al., 2007), SemEval (Strapparava and Mihalcea, 2007), (Aman and Szpakowicz, 2007) or (Strapparava and Mihalcea, 2008) but also with labels of different granularity (from generic emotions to the more specific ones) by means of our emotional ontology. EmoTag is more flexible and more adaptable to the different environments of application which can benefit from our marker. It can mark up text with three different emotional dimensions, with generic emotional categories or with more specific emotional categories. EmoTag uses not only the emotional dictionary as done by (Bestgen, 1993) but also uses WordNet for knowledge expansion in order to improve the content-dependent dictionary. Finally, EmoTag takes negation into account, unlike (Sugimoto and Yoneyama, 2006) or (O'Connor et al., 2007).

Looking at the results of our approach we can conclude that the use of categories at different levels of abstraction is a very important contribution. If we analyze the role of specific emotions versus the role of generic emotions in the whole process presented, we can conclude the following:

- Human evaluators prefer specific emotions to refer to emotional states. In every tale of the corpus specific emotions are preferred by evaluators.
- Human evaluators do not always agree on the assignment of specific emotions to a sentence. This sometimes makes it better to select as a consensus a more generic emotion which subsumes specific emotions selected by different evaluators.

If we compare our corpus to the corpora used in the systems explained in Section 3.2, we can see that the number of evaluators we used (15 evaluators) is very similar to the number of evaluators used in the systems reviewed (from 2 evaluators to 50 evaluators). However, if we compare our corpus size (around 500 sentences, 8 tales) with the size of the corpus used in those systems, we can see that our corpus is bigger than the corpora used in (Bestgen, 1993), (Zhe and Boucouvalas, 2002), (Popescu and Etzioni, 2005) or (Strapparava and Mihalcea, 2008) but it is much smaller than the corpora used in (Pang et al., 2002), (Alm, 2009) or (Aman and Szpakowicz, 2007). We are working on the extension of the current corpus in order to validate the conclusions presented in this paper over a bigger corpus like the one used in (Pang et al., 2002), (Alm, 2009) or (Aman and Szpakowicz, 2007). As future work, when the corpus extension is ready, we will use classification accuracy measures at each level of the emotion taxonomy to compare the accuracy of classification for different categories. Another issue be to resolved with the extension of the corpus is the level of lexical similarity between the various fairy tales.

7. CONCLUSIONS AND FUTURE WORK

One very important contribution of the system presented in this paper is that it allows categories at different levels of abstraction to operate simultaneously during mark-up. This extension shifts the range of possible taggings from a limited set of five categories to a potential set of 92 categories. We believe this to be a considerable improvement, which provides an automatic emotional tagger much closer to human evaluators in qualitative performance. This is achieved with no loss of applicability in the tagger by relying on the ontology of emotions to provide the means for comparing them during extraction of information from the corpus, during mark-up and during evaluation.

The automatic mark-up of texts with emotions as EmoTag does is an important step in various areas:

- Human-Machine communication: EmoTag can process responses from users and identify their emotions in order to help them.
- Education: EmoTag can detect the status of students and respond differently depending on this. The emotions in the stories may be important for therapeutical education for children with communication problems. The success of computer tutoring systems could be higher if they are able to predict and adapt to the mood of the students by reinforcing positive statements and rectifying the negative (Evens, 2002).
- Entertainment: EmoTag could be beneficial in role-playing settings, in both therapy groups and role-playing games. For example, EmoTag permits the detection of the emotion from the texts introduced by players, and based on this it could change the way it presents the character, or the music of the game can be generated in accordance with this emotion.
- Personalized Systems: EmoTag allows the analysis of user responses, thereby obtaining their feelings and allowing a more natural customization than the existing ones, which need the user to explicitly provide the system with his emotions at each moment.
- Cinematography: EmoTag allows the recovery of dialogues from films which recreate a particular emotion. That allows students of audiovisual communication to make decisions that have to do with the creation of scenes that generate that emotion.

Examples of the application of the ideas presented here are those proposed by NECA (Krenn et al., 2002): the eShowroom which generates sales dialogues, and Socialite, in which the avatar tries to become integrated into society. A more abstract study is the combination of voice synthesis with the facial modeling of photography.

In order to improve the results we have obtained with EmoTag we will consider processing modifiers and modal verbs. When modifiers appear in a sentence, the emotion associated to that sentence should be increased or reduced, by using the ontology of emotions in the case of emotional categories or increasing or reducing the value of the emotional dimensions. When a modal verb appears in a sentence the words under their scope must be treated in a special way. For example, “I can sing” does not imply that the subject is singing so a possible solution could be to reduce the activation because the action is not actually taking place.

The choice of Minipar as parser to use in detecting syntactic structure may be contested, as there are other dependency parsers available (such as the Stanford parser (Marneffe et al., 2006) or RASP (Briscoe et al., 2006)). However, syntactic structure plays a very small role in the system as described in the paper. If the expansion to address modifiers and modal verbs is carried out, it may be worth exploring the use of alternative parsers.

At the moment EmoTag does not take into account the polysemy of words; when EmoTag look for hypernyms or synonyms in WordNet it recovers the first synset. It would be interesting to deal with the polysemy of words in order to recover the correct synset and not only the most likely. However, it is unlikely that very significant improvements will result. The method of assigning to each word the most frequent sense for the POS tag it has in the

context is used as baseline in SENSEVAL experiments, and many **sophisticated** WSD are unable to reach its effectiveness (Hidalgo et al., 2005).

The results obtained with the treatment of negation in emotional dimensions do not constitute any improvement in the results of EmoTag so we need to analyze the effect of negation over emotional connotations represented in terms of emotional dimensions in order to get better results.

The use of the General Inquirer does not result in an improvement for the activation dimension so in order to discard the use of General Inquirer definitively for this dimension we would like to extend the evaluation to other texts.

We would like to improve the results of this first approach with the use of statistical natural language or machine learning techniques as done in (Aman and Szpakowicz, 2007), (Strapparava and Mihalcea, 2008) or (Alm, 2009). If we apply these techniques to the corpus of previously annotated text, we can obtain some clues that may help us in the labeling of text with emotions.

SemEval provides a corpus of headlines for the systems that participated in it. All the results obtained by these systems were based on the headlines corpus. We would like to mark up this corpus with EmoTag in order to compare the results with the ones obtained by the systems that took part in the task. The systems of SemEval do not operate with the same emotional categories and emotional dimensions that EmoTag used, but this evaluation could be an extra measurement for the ones that we have.

We would like to add some of the conclusions obtained in (Alm and Sproat, 2005) to our mark-up process. We can add rules that help us to select a more appropriate emotion when two emotions are possible. For example, if we are marking up the last sentence of the tale and the final value of the sentence must be obtained from the emotions *happiness* and *surprise*. If we were to apply the conclusions in (Alm and Sproat, 2005) (*happiness* occurred more frequently in the last sentence), we would select *happiness* as final value.

Another important point that we would treat as future work is to obtain a final emotion for the whole text. For the moment, with EmoTag we know the emotional valence for each sentence, we could apply a similar process to the one we use to obtain the final emotion of a sentence based on the emotions **associated with** each word. We think it would be possible to adapt the process presented in Section 4.6 to multi-sentence deduction of the basic or overwhelming emotion of the whole text. However, in the context of narrative texts, the evolution of emotional connotations from the beginning to the end of a text is very likely to be strong competitor with the overall afffective value in terms of interpretative potential. This is supported by the evidence reported by Alm and Sproat (2005) on the relative frequency of particular emotions at the beginning and end of fairy tales.

The generalizability of this method´s success remains to be proved. Sentences in fairy tales have a very particular character and other communications may contain sentences with very unclear emotional tags. In principle, the only thing we would need to transfer our system to a new domain might be a new corpus of texts from the new domain marked up with emotions, in order to provide coverage for any emotion-to-word assignments particular to the new domain. As future work we would like try to generalize the method present in this paper to other domains.

Finally, another important improvement could be to convert EmoTag into a multilingual resource which not only marks up English text but also texts in other languages such as Spanish.

References

ALM, C.O. 2009. Affect in text and speech.

ALM, C.O. 2010. Characteristics of high agreement affect annotation in text. *In* Proceedings of the Fourth Linguistic Annotation Workshop, LAW IV '10, Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-932432-72-5. pp. 118–122.

ALM, C. O., D. ROTH, and R. SPROAT. 2005. Emotions from Text: Machine Learning for Text-based Emotion Prediction. *In* Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 579–586.

ALM, C. OVESDOTTER, and R. SPROAT. 2005. Emotional sequencing and development in fairy tales. *In* ACII, pp. 668–674.

AMAN, S., and S. SZPAKOWICZ. 2007. Identifying expressions of emotion in text. *In* TSD, Volume 4629 of *Lecture Notes in Computer Science*, Springer, pp. 196–205.

ARNOLD, M. B. 1960. Emotion and personality.

BESTGEN, Y. 1993. Can emotional valence be determined from words? Cognition and Emotion, **7**:21–36.

BRADLEY, M.M., and P.J. LANG. 1999. The international affective digitized sounds (IADS): stimuli, instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

BRISCOE, TED, JOHN CARROLL, and REBECCA WATSON. 2006. The second release of the [rasp] system. *In* Proceedings of the COLING/ACL on Interactive presentation sessions, COLING-ACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 77–80. . http://dx.doi.org/10.3115/1225403.1225423.

BUSSO, CARLOS, MURTAZA BULUT, CHI-CHUN LEE, ABE KAZEMZADEH, EMILY MOWER, SAMUEL KIM, JEANNETTE CHANG, SUNGBOK LEE, and SHRIKANTH S. NARAYANAN. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation, **42**(4):335–359.

CARBONELL, J. G. 1979. Subjective understanding: Computer models of belief systems. Ph. D. thesis, Yale.

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

COWIE, R., E. DOUGLAS-COWIE, and A. ROMANO. 1999. Changing emotional tone in dialogue and its prosodic correlates. *In* Proc ESCA International Workshop on Dialogue and prosody, Veldhoven, The Netherlands.

ESULI, A., and F. SEBASTIANI. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *In* In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pp. 417–422.

EVENS, M. 2002. New questions for circsim-tutor. *In* Symposium on Natural Language Tutoring, University of Pittsburgh.

FELLBAUM, CHRISTIANE editor. 1998. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press. ISBN 026206197X.

FONTAINE, J.R., K.R. SCHERER, E.B. ROESCH, and P. ELLSWORTH. 2007. The world of emotion is not two-dimensional. Psychological Science, **13**:1050–1057.

FRIJDA, N.H. 1986. The emotions. Studies in emotion and social interaction. Cambridge University Press.

GREFENSTETTE, G., Y. QU, D.A. EVANS, and J.G. SHANAHAN. 2006. Validating the Coverage of Lexical Resources for Affect Analyisi and Automatically Classifying New Words along Semantic Axes.

HATZIVASSILOGLOU, V., and K. R. MCKEOWN. 1997. Predicting the semantic orientation of adjectives. *In* Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp. 174–181.

HIDALGO, JOSÉ MARÍA GÓMEZ, MANUEL DE BUENAGA RODRÍGUEZ, and JOSÉ CARLOS CORTIZO. 2005. The role of word sense disambiguation in automated text categorization. *In* NLDB. *Edited by* A. Montoyo, R. Muñoz, and E. Métais, Volume 3513 of *Lecture Notes in Computer Science*. Springer. ISBN 3-540-26031-5. pp. 298–309.

HOFFMAN, T. 2008. Online reputation management is hot — but is it ethical? *In* Computer World.

HUETTNER, A., and P. SUBASIC. 2000. Fuzzy Typoing for Document Management. *In* ACL Software Demonstration.

IZARD, C.E. 1971. The face of emotion. Appleton-Century-Crofts, New York.

KREADY, L.F. 1916. A study of fairy tales. Houghton Mifflin Company.

KRENN, B., H. PIRKER, M. GRICE, P. PIWEK, K. VAN DEEMTER, M. SCHRÖDER, M. KLESEN, and E. GSTREIN. 2002. Generation of multimodal dialogue for net environments. *In* Proceedings of Konvens, Saarbrücken, Germany.

LANG, P.J. 1980. Behavioural treatment and bio-behavioural assessment: Computer applications. *In* Technology

in mental health care delivery systems. *Edited by* J. B. Sidowski, J. H. Johnson, and T. A. W. (Eds.). Ablex Publishing, Norwood, NJ, pp. 119–137.

LASSWELL, H.D., and J.Z. NAMENWIRTH. 1969. The Lasswell Value Dictionary. Yale University Press, New Haven.

LIN, D. 1998. Dependency-based evaluation of MINIPAR. *In* Proc. of Workshop on the Evaluation of Parsing Systems, Granada, Spain.

LIU, H., H. LIEBERMAN, and T. SELKER. 2003. A model of textual affect sensing using real-world knowledge. *In* Proceedings of IUI, Miami, Florida.

MANNING, C. D., and H. SCHÜTZE. 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts.

MARNEFFE, M., B. MACCARTNEY, and C. MANNING. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *In* Proceedings of LREC-06, pp. 449–454.

MEROLA, G. 2007. Emotional gestures in sport. Language Resources and Evaluation, **41**(3-4):233–254.

MIHALCEA, R., and H. LIU. 2006. A corpus-based approach to finding happiness. *In* In AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs, AAAI Press, pp. 139–144.

O'CONNOR, J., A. OLDROYD, D. ROBERTSON, L. ZHANG, K. DHALIWAL, and M. GILLIES. 2007. Edrama: Facilitating online role-play using emotionally expressive characters. *In* AISB'07, Artificial and Ambient Intelligence, Proceedings of the AISB Annual Convention, Culture Lab, Newcastle University, Newcastle upon Tyne, UK, pp. 179–186.

ORTONY, A., G.L. CLORE, and A. COLLINS. 1988. The cognitive structure of emotions. Cambridge University Press, New York.

ORTONY, A., and T.J. TURNER. 1990. What's basic about basic emotions? Psychological Review, **97**:315–331.

OSGOOD, C. E., G. SUCI, and P. TANNENBAUM. 1957. The measurement of meaning. University of Illinois Press.

PANG, B., and L. LEE. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., **2**(1-2):1–135.

PANG, B., L. LEE, and S. VAITHYANATHAN. 2002. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. *In* Conference on Emprirical Methods in Natural Language Processing.

PARROTT, W.G. 2001. Emotions in Social Psychology: Essential Readings. PA: Psychology Press, Philadelphia.

PICARD, R.W. 1997. Affective Computing. MIT Press.

PLUTCHIK, R. 1980. A general psychoevolutionary theory of emotion. *In* Emotion: Theory, research, and experience.

POPESCU, A.M., and O. ETZIONI. 2005. Extracting product features and opinions from reviews. *In* HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Morristown, NJ, USA, pp. 339–346.

READ, J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *In* Proceedings of the ACL Student Research Workshop, ACLstudent '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48.

READ, J., D. HOPE, and J. CARROLL. 2007. Annotating expressions of appraisal in english. *In* Linguistic Annotation Workshop (LAW '07), Association for Computational Linguistics, Morristown, NJ, USA, pp. 93–100.

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

SACK, W. 1994. On the computation of point of view. *In* 12th National Conference on Artificial Intelligence (AAAI'94), American Association for Artificial Intelligence, Menlo Park, CA, USA, p. 1488.

SCHERER, K.R., A. SCHORR, and T. JOHNSTONE. 2001. Appraisal processes in emotion: Theory, Methods, Research. Oxford University Press, New York and Oxford.

SCHERER, K. R. 1984. On the nature and function of emotion: A component process approach. *In* Approaches to emotion, pp. 293–317.

SCHRÖDER, M., I. WILSON, W. JARROLD, D. EVANS, C. PELACHAUD, E. ZOVATO, and K. KARPOUZIS. 2008. What is most important for an emotion markup language? *In* Proceedings of the 3rd Workshop on Emotion and Computing, o.A., pp. 9–16. ISSN 1865-6374.

SHAVER, P., J. SCHWARTZ, D. KIRSON, and C. O'CONNOR. 1987. Emotion knowledge: Further exploration of a prototype approach. Journal of Personality and Social Psychology, **52**(6):1061–1086.

SNYDER, B., and R. BARZILAY. 2007. Multiple aspect ranking using the good grief algorithm. *In* Human Language Technologies 2007: The Conference of the North American Chapter of the Association for

Computational Linguistics, Association for Computational Linguistics, Rochester, New York, pp. 300–307.

SOLOMON, R.L. 1980. The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain. American Psychologist, **35**:691–712.

STONE, P.J., D.C. DUNPHY, M.S. SMITH, and D.M. OGILVIE. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge.

STORM, C., and T. STORM. 1987. A taxonomic study of the vocabulary of emotions. Journal of Personality and Social Psychology, **53**.

STRAPPARAVA, C., and R. MIHALCEA. 2007. Semeval-2007 task 14: Affective text. *In* Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007).

STRAPPARAVA, C., and R. MIHALCEA. 2008. Learning to identify emotions in text. *In* Proceedings of the 2008 ACM symposium on Applied computing, ACM New York, NY, USA, pp. 1556–1560.

STRAPPARAVA, C., and A. VALITUTTI. 2004. Wordnet-affect: an affective extension of wordnet. *In* Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1083–1086.

SUGIMOTO, F., and M. YONEYAMA. 2006. A method for classifying emotion of text based on emotional dictionaries for emotional reading. *In* Proceedings of the 24th IASTED International Multi-Conference Artificial Intelligence and Applications, Innsbruck, Austria, pp. 91–96.

TURNEY, P., and M. LITTMAN. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, **21**:315–346.

TURNEY, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *In* 40th Annual Meeting on Association for Computational Linguistics (ACL '02), Association for Computational Linguistics, Morristown, NJ, USA, pp. 417–424.

WATSON, D., and A. TELLEGEN. 1985. Toward a consensual structure of mood. Psychological Bulletin, **98**(2):219–235.

WHISSELL, C.M. 1989. The dictionary of affect in language. Emotion: Theory, research and experience. The measurement emotions., **4**:113–131.

WHISSELL, C. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. Psychol Rep, **105**(2):509–21.

WIEBE, J. 1994. Tracking point of view in narrative. Computational Linguistics, **20**:233–287.

WIEBE, J., R. BRUCE, and T.P. O'HARA. 1999. Development and use of a gold-standard data set for subjectivity classifications. *In* 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99), Association for Computational Linguistics, Morristown, NJ, USA, pp. 246–253.

WIEBE, J., and W. J. RAPAPORT. 1988. A computational theory of perspective and reference in narrative. *In* 26th annual meeting on Association for Computational Linguistics (ACL'88), Association for Computational Linguistics, Morristown, NJ, USA, pp. 131–138.

WIEBE, J., T. WILSON, and M. BELL. 2001. Identifying collocations for recognizing opinions. *In* ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation, pp. 24–31.

WRIGHT, A. 2009. Our sentiments, exactly. Commun. ACM, **52**(4):14–15.

YANG, C., K. HSIN-YIH LIN, and H. CHEN. 2007. Building emotion lexicon from weblog corpora. *In* ACL 2007 Demo and Poster Sessions, pp. 133–136.

ZHE, X., and A. BOUCOUVALAS. 2002. Text-to-emotion engine for real time internet communication. *In* International Symposium on Communication Systems, Networks and DSPs, pp. 164–168.