Emotions from text: machine learning for text-based emotion prediction

Cecilia Ovesdotter Alm*
Dept. of Linguistics
UIUC
Illinois, USA
ebbaalm@uiuc.edu

Dan Roth
Dept. of Computer Science
UIUC
Illinois, USA
danr@uiuc.edu

Richard Sproat
Dept. of Linguistics
Dept. of Electrical Eng.
UIUC
Illinois, USA
rws@uiuc.edu

Abstract

In addition to information, text contains attitudinal, and more specifically, emotional content. This paper explores the *text-based emotion prediction problem* empirically, using supervised machine learning with the SNoW learning architecture. The goal is to classify the emotional affinity of sentences in the narrative domain of children's fairy tales, for subsequent usage in appropriate expressive rendering of text-to-speech synthesis. Initial experiments on a preliminary data set of 22 fairy tales show encouraging results over a naïve baseline and BOW approach for classification of emotional versus non-emotional contents, with some dependency on parameter tuning. We also discuss results for a tripartite model which covers emotional valence, as well as feature set alternations. In addition, we present plans for a more cognitively sound sequential model, taking into consideration a larger set of basic emotions.

1 Introduction

Text does not only communicate informative contents, but also attitudinal information, including emotional states. The following reports on an empirical study of *text-based emotion prediction*.

Section 2 gives a brief overview of the intended application area, whereas section 3 summarizes related work. Next, section 4 explains the empirical study, including the machine learning model, the corpus, the feature set, parameter tuning, etc. Section 5 presents experimental results from two classification tasks and feature set modifications. Section 6 describes the agenda for refining the model, before presenting concluding remarks in 7.

2 Application area: Text-to-speech

Narrative text is often especially prone to having emotional contents. In the literary genre of fairy tales, emotions such as HAPPINESS and ANGER and related cognitive states, e.g. LOVE or HATE, become integral parts of the story plot, and thus are of particular importance. Moreover, the story teller reading the story interprets emotions in order to orally convey the story in a fashion which makes the story come alive and catches the listeners' attention.

In speech, speakers effectively express emotions by modifying prosody, including pitch, intensity, and durational cues in the speech signal. Thus, in order to make text-to-speech synthesis sound as natural and engaging as possible, it is important to convey the emotional stance in the text. However, this implies first having identified the appropriate emotional meaning of the corresponding text passage.

Thus, an application for emotional text-to-speech synthesis has to solve two basic problems. First, what emotion or emotions most appropriately describe a certain text passage, and second, given a text passage and a specified emotional mark-up, how to render the prosodic contour in order to convey the emotional content, (Cahn, 1990). The *text-based emotion prediction* task (TEP) addresses the first of these two problems.

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 579–586, Vancouver, October 2005.

3 Previous work

For a complete general overview of the field of *affective computing*, see (Picard, 1997). (Liu, Lieberman and Selker, 2003) is a rare study in text-based inference of sentence-level emotional affinity. The authors adopt the notion of *basic emotions*, cf. (Ekman, 1993), and use six emotion categories: ANGER, DISGUST, FEAR, HAPPINESS, SADNESS, SURPRISE. They critique statistical NLP for being unsuccessful at the small sentence level, and instead use a database of common-sense knowledge and create affect models which are combined to form a representation of the emotional affinity of a sentence. At its core, the approach remains dependent on an emotion lexicon and hand-crafted rules for conceptual polarity. In order to be effective, emotion recognition must go beyond such resources; the authors note themselves that lexical affinity is fragile. The method was tested on 20 users' preferences for an email-client, based on user-composed text emails describing short but colorful events. While the users preferred the emotional client, this evaluation does not reveal emotion classification accuracy, nor how well the model generalizes on a large data set.

Whereas work on emotion classification from the point of view of natural speech and human-computer dialogues is fairly extensive, e.g. (Scherer, 2003), (Litman and Forbes-Riley, 2004), this appears not to be the case for text-to-speech synthesis (TTS). A short study by (Sugimoto et al., 2004) addresses sentence-level emotion recognition for Japanese TTS. Their model uses a composition assumption: the emotion of a sentence is a function of the emotional affinity of the words in the sentence. They obtain emotional judgements of 73 adjectives and a set of sentences from 15 human subjects and compute words' emotional strength based on the ratio of times a word or a sentence was judged to fall into a particular emotion bucket, given the number of human subjects. Additionally, they conducted an interactive experiment concerning the acoustic rendering of emotion, using manual tuning of prosodic parameters for Japanese sentences. While the authors actually address the two fundamental problems of emotional TTS, their approach is impractical and most likely cannot scale up for a real corpus. Again, while lexical items with clear emotional meaning, such as *happy* or *sad*, matter, emotion classification probably needs to consider additional inference mechanisms. Moreover, a naïve compositional approach to emotion recognition is risky due to simple linguistic facts, such as context-dependent semantics, domination of words with multiple meanings, and emotional negation.

Many NLP problems address attitudinal meaning distinctions in text, e.g. detecting *subjective* opinion documents or expressions, e.g. (Wiebe et al, 2004), measuring *strength* of subjective clauses (Wilson, Wiebe and Hwa, 2004), determining word *polarity* (Hatzivassiloglou and McKeown, 1997) or texts' attitudinal valence, e.g. (Turney, 2002), (Bai, Padman and Airoldi, 2004), (Beineke, Hastie and Vaithyanathan, 2003), (Mullen and Collier, 2003), (Pang and Lee, 2003). Here, it suffices to say that the targets, the domain, and the intended application differ; our goal is to classify emotional text passages in children's stories, and eventually use this information for rendering expressive child-directed storytelling in a text-to-speech application. This can be useful, e.g. in therapeutic education of children with communication disorders (van Santen et al., 2003).

4 Empirical study

This part covers the experimental study with a formal problem definition, computational implementation, data, features, and a note on parameter tuning.

4.1 Machine learning model

Determining emotion of a linguistic unit can be cast as a multi-class classification problem. For the flat case, let T denote the text, and s an embedded linguistic unit, such as a sentence, where $s \in T$. Let k be the number of emotion classes $E = \{em_1, em_2, .., em_k\}$, where em_1 denotes the special case of *neutrality*, or absence of emotion. The goal is to determine a mapping function $f : s \rightarrow em_i$, such that we obtain an ordered labeled pair (s, em_i). The mapping is based on $F = \{f_1, f_2, .., f_n\}$, where F contains the features derived from the text.

Furthermore, if multiple emotion classes can characterize s, then given $E' \subset E$, the target of the mapping function becomes the ordered pair (s, E'). Finally, as further discussed in section 6, the hierarchical case of label assignment requires a sequen-

tial model that further defines levels of coarse versus fine-grained classifiers, as done by (Li and Roth, 2002) for the *question classification* problem.

4.2 Implementation

Whereas our goal is to predict finer emotional meaning distinctions according to emotional categories in speech; in this study, we focus on the basic task of recognizing emotional passages and on determining their valence (i.e. positive versus negative) because we currently do not have enough training data to explore finer-grained distinctions. The goal here is to get a good understanding of the nature of the TEP problem and explore features which may be useful.

We explore two cases of flat classification, using a variation of the Winnow update rule implemented in the SNoW learning architecture (Carlson et al., 1999),[1] which learns a linear classifier in feature space, and has been successful in several NLP applications, e.g. semantic role labeling (Koomen, Punyakanok, Roth and Yih, 2005). In the first case, the set of emotion classes E consists of EMOTIONAL versus non-emotional or NEUTRAL, i.e. $E = \{N, E\}$. In the second case, E has been incremented with emotional distinctions according to the valence, i.e. $E = \{N, PE, NE\}$. Experiments used 10-fold cross-validation, with 90% train and 10% test data.[2]

4.3 Data

The goal of our current data annotation project is to annotate a corpus of approximately 185 children stories, including Grimms', H.C. Andersen's and B. Potter's stories. So far, the annotation process proceeds as follows: annotators work in pairs on the same stories. They have been trained separately and work independently in order to avoid any annotation bias and get a true understanding of the task difficulty. Each annotator marks the sentence level with one of eight *primary emotions*, see table 1, reflecting an extended set of *basic emotions* (Ekman, 1993). In order to make the annotation process more focused, emotion is annotated from the point of view of the text, i.e. the *feeler* in the sentence. While the primary emotions are targets, the sentences are also marked for other affective contents, i.e. background *mood*, *secondary* emotions via *intensity*, *feeler*, and *textual* cues. Disagreements in annotations are resolved by a second pass of tie-breaking by the first author, who chooses one of the competing labels. Eventually, the completed annotations will be made available.

Table 1: Basic emotions used in annotation

Abbreviation	Emotion class
A	ANGRY
D	DISGUSTED
F	FEARFUL
H	HAPPY
Sa	SAD
Su+	POSITIVELY SURPRISED
Su-	NEGATIVELY SURPRISED

Emotion annotation is hard; interannotator agreement currently range at $\kappa = .24 - .51$, with the ratio of observed annotation overlap ranging between 45-64%, depending on annotator pair and stories assigned. This is expected, given the subjective nature of the annotation task. The lack of a clear definition for emotion vs. non-emotion is acknowledged across the emotion literature, and contributes to dynamic and shifting annotation targets. Indeed, a common source of confusion is NEUTRAL, i.e. deciding whether or not a sentence is emotional or non-emotional. Emotion perception also depends on which character's point-of-view the annotator takes, and on extratextual factors such as annotator's personality or mood. It is possible that by focusing more on the training of annotator pairs, particularly on joint training, agreement might improve. However, that would also result in a bias, which is probably not preferable to actual perception. Moreover, what agreement levels are needed for successful expressive TTS remains an empirical question.

The current data set consisted of a preliminary annotated and tie-broken data set of 1580 sentence, or 22 Grimms' tales. The label distribution is in table 2. NEUTRAL was most frequent with 59.94%.

Table 2: Percent of annotated labels

A	D	F	H
12.34%	0.89%	7.03%	6.77%
N	SA	SU+	SU.-
59.94%	7.34%	2.59%	3.10%

[1] Available from http://l2r.cs.uiuc.edu/∼cogcomp/

[2] Experiments were also run for Perceptron, however the results are not included. Overall, Perceptron performed worse.

Table 3: % EMOTIONAL vs. NEUTRAL examples

E	N
40.06%	59.94%

Table 4: % POSITIVE vs. NEGATIVE vs. NEUTRAL

PE	NE	N
9.87%	30.19%	59.94%

Next, for the purpose of this study, all emotional classes, i.e. A, D, F, H, SA, SU+, SU-, were combined into one emotional superclass E for the first experiment, as shown in table 3. For the second experiment, we used two emotional classes, i.e. positive versus negative emotions; PE={H, SU+} and NE={A, D, F, SA, SU-}, as seen in table 4.

4.4 Feature set

The feature extraction was written in python. SNoW only requires active features as input, which resulted in a typical feature vector size of around 30 features. The features are listed below. They were implemented as boolean values, with continuous values represented by ranges. The ranges generally overlapped, in order to get more generalization coverage.

1. First sentence in story
2. Conjunctions of selected features (see below)
3. Direct speech (i.e. whole quote) in sentence
4. Thematic story type (3 top and 15 sub-types)
5. Special punctuation (! and ?)
6. Complete upper-case word
7. Sentence length in words (0-1, 2-3, 4-8, 9-15, 16-25, 26-35, >35)
8. Ranges of story progress (5-100%, 15-100%, 80-100%, 90-100%)
9. Percent of JJ, N, V, RB (0%, 1-100%, 50-100%, 80-100%)
10. V count in sentence, excluding participles (0-1, 0-3, 0-5, 0-7, 0-9, > 9)
11. Positive and negative word counts (≥ 1, ≥ 2, ≥ 3, ≥ 4, ≥ 5, ≥ 6)
12. WordNet emotion words
13. Interjections and affective words
14. *Content BOW*: N, V, JJ, RB words by POS

Feature conjunctions covered pairings of counts of positive and negative words with range of story progress or interjections, respectively.

Feature groups 1, 3, 5, 6, 7, 8, 9, 10 and 14 are extracted automatically from the sentences in the stories; with the SNoW POS-tagger used for features 9, 10, and 14. Group 10 reflects how many verbs are active in a sentence. Together with the quotation and punctuation, verb domination intends to capture the assumption that emotion is often accompanied by increased action and interaction. Feature group 4 is based on Finish scholar Antti Aarne's classes of folk-tale types according to their informative thematic contents (Aarne, 1964). The current tales have 3 top story types (ANIMAL TALES, ORDINARY FOLK-TALES, and JOKES AND ANECDOTES), and 15 subtypes (e.g. *supernatural helpers* is a subtype of the ORDINARY FOLK-TALE). This feature intends to provide an idea about the story's general affective *personality* (Picard, 1997), whereas the feature reflecting the story progress is hoped to capture that some emotions may be more prevalent in certain sections of the story (e.g. the happy end).

For semantic tasks, words are obviously important. In addition to considering 'content words', we also explored specific word lists. Group 11 uses 2 lists of 1636 positive and 2008 negative words, obtained from (Di Cicco et al., online). Group 12 uses lexical lists extracted from WordNet (Fellbaum, 1998), on the basis of the primary emotion words in their adjectival and nominal forms. For the adjectives, Py-WordNet's (Steele et al., 2004) SIMILAR feature was used to retrieve similar items of the primary emotion adjectives, exploring one additional level in the hierarchy (i.e. similar items of all senses of all words in the synset). For the nouns and any identical verbal homonyms, synonyms and hyponyms were extracted manually.[3] Feature group 13 used a short list of 22 interjections collected manually by browsing educational ESL sites, whereas the affective word list of 771 words consisted of a combination of the non-neutral words from (Johnson-Laird and Oatley, 1989) and (Siegle, online). Only a subset of these lexical lists actually occurred.[4]

[3] Multi-words were transformed to hyphenated form.

[4] At this point, neither stems and bigrams nor a list of onomatopoeic words contribute to accuracy. Intermediate resource processing inserted some feature noise.

The above feature set is henceforth referred to as *all features*, whereas *content BOW* is just group 14. The *content BOW* is a more interesting baseline than the naïve one, *P(Neutral)*, i.e. always assigning the most likely NEUTRAL category. Lastly, emotions blend and transform (Liu, Lieberman and Selker, 2003). Thus, emotion and background mood of immediately adjacent sentences, i.e. the *sequencing*, seems important. At this point, it is not implemented automatically. Instead, it was extracted from the manual emotion and mood annotations. If *sequencing* seemed important, an automatic method using sequential target activation could be added next.

4.5 Parameter tuning

The Winnow parameters that were tuned included promotional α, demotional β, activation threshold θ, initial weights ω, and the regularization parameter, S, which implements a margin between positive and negative examples. Given the currently fairly limited data, results from 2 alternative tuning methods, applied to *all features*, are reported.

- For the condition called *sep-tune-eval*, 50% of the sentences were randomly selected and set aside to be used for the parameter tuning process only. Of this subset, 10% were subsequently randomly chosen as test set with the remaining 90% used for training during the automatic tuning process, which covered 4356 different parameter combinations. Resulting parameters were: $\alpha = 1.1$, $\beta = 0.5$, $\theta = 5$, $\omega = 1.0$, $S = 0.5$. The remaining half of the data was used for training and testing in the 10-fold cross-validation evaluation. (Also, note the slight change for *P(Neutral)* in table 5, due to randomly splitting the data.)
- Given that the data set is currently small, for the condition named *same-tune-eval*, tuning was performed automatically on all data using a slightly smaller set of combinations, and then manually adjusted against the 10-fold cross-validation process. Resulting parameters were: $\alpha = 1.2$, $\beta = 0.9$, $\theta = 4$, $\omega = 1$, $S = 0.5$. All data was used for evaluation.

Emotion classification was sensitive to the selected tuning data. Generally, a smaller tuning set resulted in pejorative parameter settings. The random selection could make a difference, but was not explored.

5 Results and discussion

This section first presents the results from experiments with the two different confusion sets described above, as well as feature experimentation.

5.1 Classification results

Average accuracy from 10-fold cross validation for the first experiment, i.e. classifying sentences as either NEUTRAL or EMOTIONAL, are included in table 5 and figure 1 for the two tuning conditions on the main feature sets and baselines. As expected,

Table 5: Mean classification accuracy: N vs. E, 2 conditions

	same-tune-eval	sep-tune-eval
P(Neutral)	59.94	60.05
Content BOW	61.01	58.30
All features except BOW	64.68	63.45
All features	68.99	63.31
All features + sequencing	69.37	62.94

degree of success reflects parameter settings, both for *content BOW* and *all features*. Nevertheless, under these circumstances, performance above a naïve baseline and a BOW approach is obtained. Moreover, *sequencing* shows potential for contributing in one case. However, observations also point to three issues: first, the current data set appears to be too small. Second, the data is not easily separable. This comes as no surprise, given the subjective nature of the task, and the rather low interannotator agreement, reported above. Moreover, despite the schematic narrative plots of children's stories, tales still differ in their overall affective orientation, which increases data complexity. Third and finally, the EMOTION class is combined by basic emotion labels, rather than an original annotated label.

More detailed averaged results from 10-fold cross-validation are included in table 6 using *all features* and the separated tuning and evaluation data condition *sep-tune-eval*. With these parameters, approximately 3% improvement in accuracy over the naïve baseline *P(Neutral)* was recorded, and 5% over the *content BOW*, which obviously did poorly with these parameters. Moreover, precision is

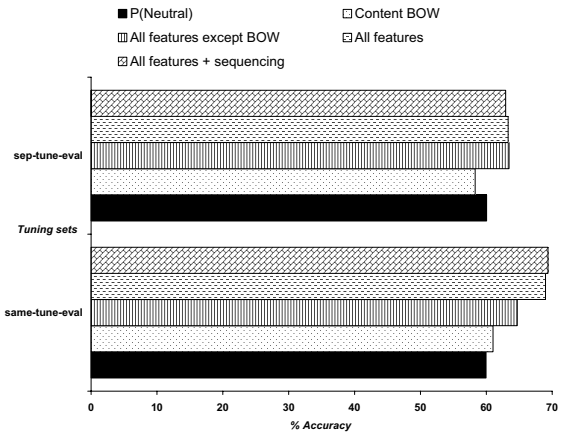

Figure 1: Accuracy under different conditions (in %)

Table 6: Classifying N vs. E (*all features, sep-tune-eval*)

Measure	N	E
Averaged accuracy	0.63	0.63
Averaged error	0.37	0.37
Averaged precision	0.66	0.56
Averaged recall	0.75	0.42
Averaged F-score	0.70	0.47

higher than recall for the combined EMOTION class. In comparison, with the *same-tune-eval* procedure, the accuracy improved by approximately 9% over *P(Neutral)* and by 8% over *content BOW*.

In the second experiment, the emotion category was split into two classes: emotions with positive versus negative valence. The results in terms of precision, recall, and F-score are included in table 7, using *all features* and the *sep-tune-eval* condition. The decrease in performance for the emotion classes mirrors the smaller amounts of data available for each class. As noted in section 4.3, only 9.87% of the sentences were annotated with a positive emotion, and the results for this class are worse. Thus, performance seems likely to improve as more annotated story data becomes available; at this point, we are experimenting with merely around 12% of the total texts targeted by the data annotation project.

5.2 Feature experiments

Emotions are poorly understood, and it is especially unclear which features may be important for their recognition from text. Thus, we experimented

Table 7: N, PE, and NE (*all features, sep-tune-eval*)

	N	NE	PE
Averaged precision	0.64	0.45	0.13
Averaged recall	0.75	0.27	0.19
Averaged F-score	0.69	0.32	0.13

Table 8: Feature group members

Word lists	interj., WordNet, affective lists, pos/neg
Syntactic	length ranges, % POS, V-count ranges
Story-related	% story-progress, 1st sent., story type
Orthographic	punctuation, upper-case words, quote
Conjunctions	Conjunctions with pos/neg
Content BOW	Words (N,V,Adj, Adv)

with different feature configurations. Starting with all features, again using 10-fold cross-validation for the separated tuning-evaluation condition *sep-tune-eval*, one additional feature group was removed until none remained. The feature groups are listed in table 8. Figure 2 on the next page shows the accuracy at each step of the cumulative subtraction process. While some feature groups, e.g. syntactic, appeared less important, the removal order mattered; e.g. if syntactic features were removed first, accuracy decreased. This fact also illustrated that features work together; removing any group degraded performance because features interact and there is no true independence. It was observed that features' contributions were sensitive to parameter tuning. Clearly, further work on developing features which fit the TEP problem is needed.

6 Refining the model

This was a "first pass" of addressing TEP for TTS. At this point, the annotation project is still on-going, and we only had a fairly small data set to draw on. Nevertheless, results indicate that our learning approach benefits emotion recognition. For example, the following instances, also labeled with the same valence by both annotators, were correctly classified both in the binary (N vs. E) and the tripartite polarity task (N, NE, PE), given the separated tuning and evaluation data condition, and using *all features*:

(1a) E/NE: Then he offered the dwarfs money, and prayed and besought them to let him take her away; but they said, "We will not part with her for all the gold in the world."

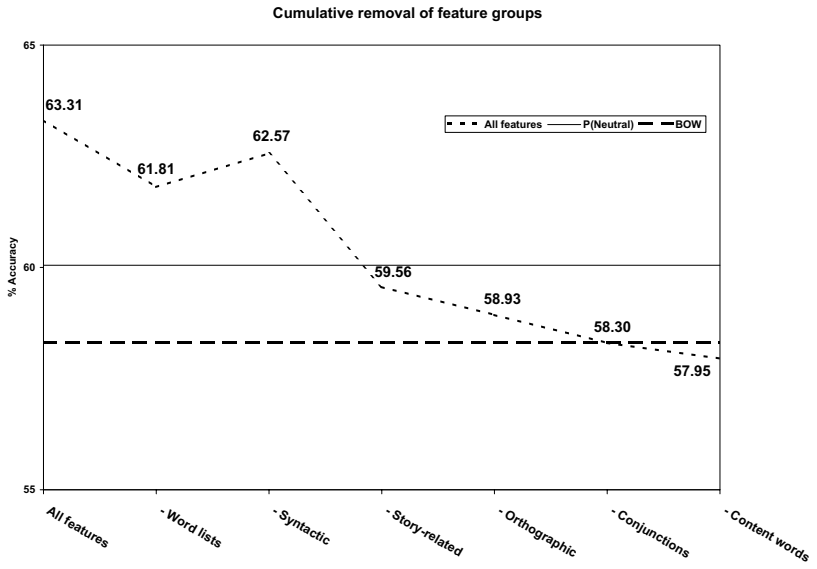

Figure 2: Averaged effect of feature group removal, using *sep-tune-eval*

(1b) N: And so the little girl really did grow up; her skin was as white as snow, her cheeks as rosy as the blood, and her hair as black as ebony; and she was called Snowdrop.

(2a) E/NE: "Ah," she answered, "have I not reason to weep?

(2b) N: Nevertheless, he wished to try him first, and took a stone in his hand and squeezed it together so that water dropped out of it.

Cases (1a) and (1b) are from the well-known FOLK TALE *Snowdrop*, also called *Snow White*. (1a) and (1b) are also correctly classified by the simple *content BOW* approach, although our approach has higher prediction confidence for E/NE (1a); it also considers, e.g. direct speech, a fairly high verb count, advanced story progress, connotative words and conjunctions thereof with story progress features, all of which the BOW misses. In addition, the simple *content BOW* approach makes incorrect predictions at both the bipartite and tripartite levels for examples (2a) and (2b) from the JOKES AND ANECDOTES stories *Clever Hans* and *The Valiant Little Tailor*, while our classifier captures the affective differences by considering, e.g. distinctions in verb count, interjection, POS, sentence length, connotations, story subtype, and conjunctions.

Next, we intend to use a larger data set to conduct a more complete study to establish mature findings. We also plan to explore finer emotional meaning distinctions, by using a hierarchical sequential model which better corresponds to different levels of cognitive difficulty in emotional categorization by humans, and to classify the full set of basic level emotional categories discussed in section 4.3. Sequential modeling of simple classifiers has been successfully employed to question classification, for example by (Li and Roth, 2002). In addition, we are working on refining and improving the feature set, and given more data, tuning can be improved on a sufficiently large development set. The three subcorpora in the annotation project can reveal how authorship affects emotion perception and classification.

Moreover, arousal appears to be an important dimension for emotional prosody (Scherer, 2003), especially in storytelling (Alm and Sproat, 2005). Thus, we are planning on exploring degrees of emotional intensity in a learning scenario, i.e. a problem similar to measuring strength of opinion clauses (Wilson, Wiebe and Hwa, 2004).

Finally, emotions are not discrete objects; rather they have transitional nature, and blend and overlap along the temporal dimension. For example, (Liu, Lieberman and Selker, 2003) include parallel estimations of emotional activity, and include smooth-

ing techniques such as interpolation and decay to capture sequential and interactive emotional activity. Observations from tales indicate that some emotions are more likely to be prolonged than others.

7 Conclusion

This paper has discussed an empirical study of the *text-based emotion prediction* problem in the domain of children's fairy tales, with child-directed expressive text-to-speech synthesis as goal. Besides reporting on encouraging results in a first set of computational experiments using supervised machine learning, we have set forth a research agenda for tackling the TEP problem more comprehensively.

8 Acknowledgments

We are grateful to the annotators, in particular A. Rasmussen and S. Siddiqui. We also thank two anonymous reviewers for comments. This work was funded by NSF under award ITR-#0205731, and NS ITR IIS-0428472. The annotation is supported by UIUC's Research Board. The authors take sole responsibility for the work.

References

Antti Aarne. 1964. *The Types of the Folk-Tale: a Classification and Bibliography.* Helsinki: Suomalainen Tiedeakatemia.

Cecilia O. Alm, and Richard Sproat. 2005. Perceptions of emotions in expressive storytelling. *INTERSPEECH 2005*.

Xue Bai, Rema Padman, and Edoardo Airoldi. 2004. Sentiment extraction from unstructured text using tabu search-enhanced Markov blankets. In *MSW2004*, Seattle.

Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan. 2004. The sentimental factor: improving review classification via human-provided information. In *Proceedings of ACL*, 263–270.

Janet Cahn. 1990. The generation of affect in synthesized Speech. *Journal of the American Voice I/O Society*, 8:1–19.

Andrew Carlson, Chad Cumby, Nicholas Rizzolo, Jeff Rosen, and Dan Roth. 1999. *The SNoW Learning Architecture.* Technical Report UIUCDCS-R-99-2101, UIUC Comp. Sci.

Stacey Di Cicco et al. General Inquirer Pos./Neg. lists *http://www.webuse.umd.edu:9090/*

Paul Ekman. 1993. Facial expression and emotion. *American Psychologist*, 48(4), 384–392.

Christiane Fellbaum, Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Vasileios Hatzivassiloglou, and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL*, 174–181.

Philip Johnson-Laird, and Keith Oatley. 1989. The language of emotions: an analysis of a semantic field. *Cognition and Emotion*, 3:81–123.

Peter Koomen, Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2005. Generalized inference with multiple semantic role labeling systems. In *Proceedings of the Annual Conference on Computational Language Learning (CoNLL)*, 181–184.

Diane Litman, and Kate Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of ACL*, 351–358.

Xin Li, and Dan Roth. 2002. Learning question classifiers: the role of semantic information. In *Proc. International Conference on Computational Linguistics (COLING)*, 556–562.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *ACM Conference on Intelligent User Interfaces*, 125–132.

Tony Mullen, and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP*, 412–418.

Bo Pang, and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*, 271–278.

Rosalind Picard. 1997. *Affective computing*. MIT Press, Cambridge, Mass.

Dan Roth. 1998. Learning to resolve natural language ambiguities: a unified approach. In *AAAI*, 806–813.

Klaus Scherer. 2003. Vocal communication of emotion: a review of research paradigms. *Speech Commununication*, 40(1-2):227–256.

Greg Siegle. The Balanced Affective Word List *http://www.sci.sdsu.edu/CAL/wordlist/words.prn*

Oliver Steele et al. Py-WordNet *http://osteele.com/projects/pywordnet/*

Futoshi Sugimoto et al. 2004. A method to classify emotional expressions of text and synthesize speech. In *IEEE*, 611–614.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, 417–424.

Jan van Santen et al. 2003. Applications of computer generated expressive speech for communication disorders. In *EUROSPEECH 2003*, 1657–1660.

Janyce Wiebe et al. 2004. Learning subjective language. *Journal of Computational Linguistics*, 30(3):277–308.

Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, 761–769.