An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text

Plaban Kr. Bhowmick, Pabitra Mitra, Anupam Basu
Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India – 721302
{plaban,pabitra,anupam}@cse.iitkgp.ernet.in

Abstract

An affective text may be judged to belong to multiple affect categories as it may evoke different affects with varying degree of intensity. For affect classification of text, it is often required to annotate text corpus with affect categories. This task is often performed by a number of human judges. This paper presents a new agreement measure inspired by Kappa coefficient to compute inter-annotator reliability when the annotators have freedom to categorize a text into more than one class. The extended reliability coefficient has been applied to measure the quality of an affective text corpus. An analysis of the factors that influence corpus quality has been provided.

1 Introduction

The accuracy of a supervised machine learning task primarily depends on the annotation quality of the data, that is used for training and cross validation. Reliability of annotation is a key requirement for the usability of an annotated corpus. Inconsistency or noisy annotation may lead to the degradation of performances of supervised learning algorithms. The data annotated by a single annotator may be prone to error and hence an unreliable one. This also holds for annotating an affective corpus, which is highly dependent on the mental state of the subject. The recent trend in corpus development in NLP is to annotate corpus by more than one annotators independently. In corpus statistics, the corpus reliability is measured by coefficient of agreement. The coefficients of agreement are applied to corpus for various goals like measuring *reliability, validity* and *stability* of corpus (Artstein and Poesio, 2008).

Jacob Cohen (Cohen, 1960) introduced Kappa statistics as a coefficient of agreement for nominal scales. The Kappa coefficient measures the proportion of observed agreement over the agreement by chance and the maximum agreement attainable over chance agreement considering pairwise agreement. Later Fleiss (Fleiss, 1981) proposed an extension to measure agreement in ordinal scale data.

Cohen's Kappa has been widely used in various research areas. Because of its simplicity and robustness, it has become a popular approach for agreement measurement in the area of electronics (Jung, 2003), geographical informatics (Hagen, 2003), medical (Hripcsak and Heitjan, 2002), and many more domains.

There are other variants of Kappa like agreement measures (Carletta, 1996). Scott's π (Scott, 1955) was introduced to measure agreement in survey research. Kappa and π measures differ in the way they determine the chance related agreements. π-like coefficients determine the chance agreement among arbitrary coders, while κ-like coefficients treats the chance of agreement among the coders who produced the reliability data (Artstein and Poesio, 2008).

One of the drawbacks of π and Kappa like coefficients except Fleiss' Kappa (Fleiss, 1981) is that they treat all kinds of disagreements in the same manner. Krippendorff's α (Krippendorff, 1980) is a reliability measure which treats different kind of disagreements separately by introducing a notion of distance between two categories. It offers a way

Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, pages 58–65
Manchester, August 2008

to measure agreement in nominal, interval, ordinal and ratio scale data.

Reliability assessment of corpus is an important issue in corpus driven natural language processing and the existing reliability measures have been used in various corpus development tasks. For example, Kappa coefficient has been used in developing parts of speech corpus (Mieskes and Strube, 2006), dialogue act tagging efforts like MapTask (Carletta et al., 1997) and Switchboard (Stolke et al., 1997), subjectivity tagging task (Bruce and Wiebe, 1999) and many more.

The π and κ coefficients measure the reliability of the annotation task where a data item can be annotated with one category. (Rosenberg and Binkowski, 2004) puts an effort towards measuring corpus reliability for multiply labeled data points. In this measure, the annotators are allowed to mark one data point with at most two classes, one of which is primary and other is secondary. This measure was used to determine the reliability of a email corpus where emails are assigned with primary and secondary labels from a set of email types.

Affect recognition from text is a recent and promising subarea of natural language processing. The task is to classify text segments into appropriate affect categories. The supervised machine learning techniques, which requires a reliable annotated corpus, may be applied for solving the problem. In general, a blend of emotions is common in both verbal and non-verbal communication. Unlike conventional annotation tasks like POS corpus development, where one data item may belong to only one category, in affective text corpus, a data item may be fuzzy and may belong to multiple affect categories. For example, the following sentence may belong to *disgust* and *sad* category since it may evoke both the emotions to different degrees of intensity.

> *A young married woman was burnt to death allegedly by her in-laws for dowry.*

This property makes the existing agreement measures inapplicable for determining agreement in emotional corpus. Craggs and Wood (2004) adopted a categorical scheme for annotating emotion in affective text dialogue. They claimed to address the problem of agreement measurement for the data set where one data item may belong to more than one category using an extension of Krippendorff's α. But the details of the extension is yet to be disseminated.

In this paper, we propose a new agreement measure for multiclass annotation which we denote by A_m. The new measure is then applied to an affective text corpus to

- *Assess Reliability:* To test whether the corpus can be used for developing computational affect recognizer.
- *Determine Gold Standard:* To define a gold standard that will be used to test the accuracy of the affect recognizer.

In section 2, we describe the affective text corpus and the annotation scheme. In section 3, we propose a new reliability measure (A_m) for multiclass annotated data. In section 4, we provide an algorithm to determine *gold standard* data from the annotation and in section 5, we discuss about applying A_m measure to the corpus developed by us and some observations related to the annotation.

2 Affective Text Corpus and Annotation Scheme

The affective text corpus collected by us consists of 1000 sentences extracted from *Times of India* news archive[1]. The sentences were collected from headlines as well as articles belonging to political, social, sports and entertainment domain.

Selection of affect categories is a very crucial and important decision problem due to the following reasons.

- The affect categories should be applicable to the considered genre.
- The affect categories should be identifiable from language.
- The categories should be unambiguous.

We shall try to validate these points based on the results obtained, after applying the our extended measure on the text corpus with respect to a set of selected basic emotional categories.

Basic emotions are those for which the respective expressions across culture, ethnicity, age, sex, social structure are invariant (Ortony and Turner, 1990). But unfortunately, there is a long persistent debate among the psychologists regarding

[1] http://timesofindia.indiatimes.com/archive.cms

the number of basic emotional categories (Ortony and Turner, 1990). One of the theories behind the basic emotions is that they are biologically primitive because they possess evolutionary significance related to the basic needs for the survival of the species (Plutchik, 1980). The universality of recognition of emotions from distinctive facial expressions is an indirect technique to establish the basic emotions (Darwin, 1965).

Six basic affect categories (Ekman, Friesen and Ellsworth, 1982) have been considered in emotion recognition from speech (Song et al., 2004), facial expression (Pantic and Rothkrantz, 2000). Our annotation scheme considers six basic emotions, namely, *Anger, Disgust, Fear, Happiness, Sadness, Surprise* as specified by Ekman for affect recognition in text corpus.

The annotation scheme considers the following points:

- Two types are sentences are collected for annotation.
  - *Direct Affective Sentence:* Here, the agent present in the sentence is experiencing a set of emotions, which are explicit in the sentence. For example, in the following sentence *Indian supporters* are the agents experiencing a disgust emotion.

    Indian supporters are disgusted about players' performances in the World Cup.
  - *Indirect Affective Sentence:* Here, the reader of the sentence is experiencing a set of emotions. In the following sentence, the reader is experiencing a *disgust* emotion because the event of *accepting bribe*, is an indecent act carried out by responsible agents like *Top officials*.

    Top officials are held for accepting bribe from a poor villager.
- A sentence may trigger multiple emotions simultaneously. So, one annotator may classify a sentence to more than one affective categories.
- For each emotion, the keywords that trigger the particular emotion are marked.
- For each emotion, the events or objects that trigger the concerned emotion are marked.

Here, we aim at measuring the agreement in annotation. The focus is to measure the agreement in annotation pattern rather than the agreement in individual emotional classes.

3 Proposed Agreement Measure

To overcome the shortcomings of existing reliability measures mentioned earlier, we propose A_m measure, which is an agreement measure for corpus annotation task considering multiclass classification. We present the notion of agreement below.

3.1 Notion of Paired Agreement

In order to allow for multiple labels, we calculate agreement between all the pairs of possible labels. Let *C1* and *C2* be two affect categories, e.g., *anger* and *disgust*. Let <*C1, C2*> denote the category pair. An annotator's assignment of labels can be represented as a pair of binary choices for each category pair <*C1, C2*>, namely, $< 0, 0 >$, $< 0, 1 >$, $< 1, 0 >$, and $< 1, 1 >$. It should be noted that the proposed metric considers the non-inclusion in a category by an annotator pair as an agreement.

For an item, two annotators *U1* and *U2* are said to agree on <*C1, C2*> if the following conditions hold.

$$U1.C1 = U2.C1$$
$$U1.C2 = U2.C2$$

where $U_i.C_j$ signifies that the value for C_j for annotator U_i and the value may either be 1 or 0. For example, if one coder marks an item with *anger* and another with *disgust*, they would disagree on the pairs that include these labels, but still agree that the item does not express *happiness* and *sadness*.

3.2 A_m Agreement Measure

With the notion of paired agreement discussed earlier, the *observed agreement*(P_o) is the proportion of items the annotators agreed on the category pairs and the *expected agreement*(P_e) is the proportion of items for which agreement is expected by chance when the items are randomly. Following the line of Cohen's Kappa (Cohen, 1960), A_m is defined as the proportion of agreement after expected or chance agreement is removed from consideration and is given by

$$A_m = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

When P_o equals P_e, A_m value is computed to be 0, which signifies no non-random agreement among the annotators. An A_m value of 1, the upper limit of A_m, indicates a perfect agreement among the annotators. We define P_o and P_e as follows.

Observed Agreement (P_o)**:**
Let **I** be the number of items, **C** is the number of categories and **U** is the number of annotators and **S** be the set of all category pairs with cardinality $\binom{\mathbf{C}}{2}$. The total agreement on a category pair p for an item i is n_{ip}, the number of annotator pairs who agree on p for i.

The average agreement on a category pair p for an item i is n_{ip} divided by the total number of annotator pairs and is given by

$$P_{ip} = \frac{1}{\binom{\mathbf{U}}{2}} n_{ip} \tag{2}$$

The average agreement for the item i is the mean of P_{ip} over all category pairs and is given by

$$P_i = \frac{1}{\binom{\mathbf{C}}{2}\binom{\mathbf{U}}{2}} \sum_{p \in \mathbf{S}} n_{ip} \tag{3}$$

The observed agreement is the average agreement over all the item and is given by

$$\begin{aligned} P_o &= \frac{1}{\mathbf{I}} \sum_{i=1}^{I} P_i \\ &= \frac{1}{\mathbf{I}\binom{\mathbf{C}}{2}\binom{\mathbf{U}}{2}} \sum_{i=1}^{I} \sum_{p \in \mathbf{S}} n_{ip} \\ &= \frac{4}{\mathbf{IC}(\mathbf{C}-1)\mathbf{U}(\mathbf{U}-1)} \sum_{i=1}^{I} \sum_{p \in \mathbf{S}} n_{ip} \end{aligned} \tag{4}$$

Expected Agreement (P_e**):**
The expected agreement is defined as the agreement among the annotators when they assign the items to a set of categories randomly. However, since we are considering the agreement on category pairs, we consider the expected agreement to be the expectation that the annotators agree on a category pair. For a category pair, four possible assignment combinations constitute a set which is given by

$$G = \{[0\ 0], [0\ 1], [1\ 1]\}.$$

It is to be noted that the combinations [0 1] and [1 0] are clubbed to one element as they are symmetric to each other. Let $\hat{P}(p_g|u)$ be the overall proportion of items assigned with assignment combination $g \in G$ to category pair $p \in S$ by annotator u and $n_{p_g u}$ be the total number of assignments of items by annotator u with assignment combination g to category pair p. Then $\hat{P}(p_g|u)$ is given by

$$\hat{P}(p_g|u) = \frac{n_{p_g u}}{\mathbf{I}} \tag{5}$$

For an item, the probability that two arbitrary coders agree with the same assignment combination in a category pair is the joint probability of individual coders making this assignments independently. For two annotators u_x and u_y the joint probability is given by $\hat{P}(p_g|u_x)\hat{P}(pg|u_y)$. The probability that two arbitrary annotators agree on a category pair p with assignment combination g is the average over all annotator pairs belonging to W, the set of annotator pairs and is given by

$$\hat{P}(p_g) = \frac{1}{\binom{\mathbf{U}}{2}} \sum_{(u_x,u_y) \in W} \hat{P}(p_g|u_x)\hat{P}(p_g|u_y) \tag{6}$$

The probability that two arbitrary annotators agree on a category pair for all assignment combinations is given by

$$\hat{P}(p) = \sum_{pg \in G} \hat{P}(p_g) \tag{7}$$

The chance agreement is calculated by taking average over all category pairs.

$$P_e = \frac{1}{\binom{\mathbf{C}}{2}} \sum_{p \in S} \hat{P}(p) \tag{8}$$

The A_m measure may be calculated based on the expressions of P_o and P_e as given in Equation 4 and Equation 8 to compute the reliability of annotation with respect to multiclass annotation.

4 Gold Standard Determination

Gold standard data is used as a reference data set for various goals like

- Building reliable classifier

- Determine the performance of a classifier

To attach a set of labels to a data item in the gold standard data, we assign the majority decision label to an item. Let n_O be the number of annotators, who have assigned an item i into category C and n_ϕ annotators have decided not to assign the same item into that category. Then i is assigned to C if $n_O > n_\phi$; otherwise it is not assigned to that category.

Algorithm 1: Algorithm for determining gold standard data

Input: Set of I items annotated into C categories by U annotators
Output: Gold standard data
foreach *annotator* $u \in U$ **do**
  $\xi_u \leftarrow 0$;
end
foreach *item* $i \in I$ **do**
  foreach *category* $c \in C$ **do**
    Θ = set of annotators who have assigned i in category c;
    ϕ = set of annotators who have not assigned i in category c;
    if *cardinality*(Θ)>*cardinality*(ϕ) **then**
      assign label c to i;
      $\xi_j \leftarrow \xi_j + 1$ where $j \in \Theta$;
    end
    else if *cardinality*(Θ)<*cardinality*(ϕ) **then**
      do not assign label c to i;
      $\xi_j \leftarrow \xi_j + 1$ where $j \in \phi$;
    end
    else if $\sum_\Theta \xi > \sum_\phi \xi$ **then**
      assign label c to i;
    end
  end
end

If $n_O = n_\phi$, then we resolve the tie based on the performances of the annotators in previous assignments. We assign an *expert coder index*(ξ) to each annotator and it is updated based on the agreement of their judgments over the corpus. There are two cases when the ξ values are incremented

- If the item is assigned to a category in the gold standard data, the ξ values are incremented for those annotators who have assigned the item into that category.
- If the item is not assigned to a category in the gold standard data, the ξ values are incremented for those annotators who have not assigned the item into that category.

If n_O and n_ϕ are equal for an item, we make use of the ξ values for deciding upon the assignment of the item to the category in concern. We assign the item into that category if the combined ξ values of the annotators who have assigned the item into that category is greater than the combined ξ values of the annotators who have not assigned the item into that category, i.e.,

$$\sum_{i=1}^{n_O} \xi_i > \sum_{j=1}^{n_\phi} \xi_j$$

The algorithm for determining gold standard data is given in Algorithm 1.

5 Experimental Results

We applied the proposed A_m measure to estimate the quality of the affective corpus described in section 2. Below we present the annotation experiment followed by some relevant analysis.

5.1 Annotation Experiment

Ten human judges with the same social background participated in the study, assigning affective categories to sentences independently of one another. The annotators were provided with the annotation instructions and they were trained with some sentences not belonging to the corpus. The annotation was performed with the help of a web based annotation interface[2]. The corpus consists of 1000 sentences. Three of judges were able to complete the task within 20 days. In this paper, we report the result of applying the measure with data provided by three annotators without considering the incomplete annotations. Distribution of the sentences across the affective categories for the three judges is given in Figure 1.

5.2 Analysis of Corpus Quality

The corpus was evaluated in terms of the proposed measure. Some of the relevant observations are presented below.

- **Agreement Value:** Different agreement values related to A_m measure are given in Table 1. We present A_m values for all the annotator pairs in Table 2.

[2]http://www.mla.iitkgp.ernet.in/Annotation/index.php

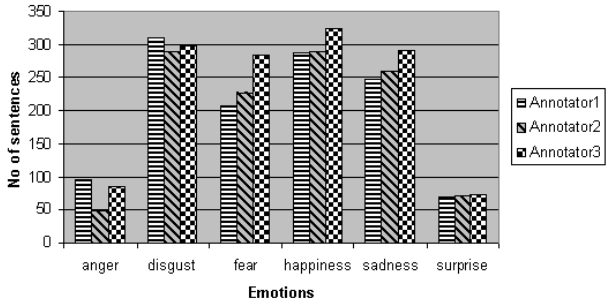

Figure 1: Distribution of sentences for three judges.

Agreement	A_m **Value**
Observed Agreement(P_o)	0.878
Chance Agreement(P_e)	0.534
A_m	**0.738**

Table 1: Agreement values for the affective text corpus.

Annotator Pair	P_o	P_e	A_m **Value**
1-2	0.858	0.526	0.702
1-3	0.868	0.54	0.713
2-3	0.884	0.531	0.752

Table 2: Annotator pairwise A_m values.

- **Agreement Study:** Table 3 provides the distribution of the sentences against individual observed agreement values. It is observed that 71.5% of the corpus belongs to [0.7 1.0] range of observed agreement and among this bulk portion of the corpus, the annotators assign 78.6% of the sentences into a single category. This is due to the existence of a dominant emotion in a sentence and in most of the cases, the sentence contains enough clues to decode it. For the non-dominant emotions in a sentence, ambiguity has been found while decoding.

Observed Agreement	**No. of Sentences**
$0.0 < A_0 \leq 0.2$	14
$0.2 < A_0 \leq 0.4$	73
$0.4 < A_0 \leq 0.7$	198
$0.7 < A_0 \leq 1.0$	715

Table 3: Distribution of the sentences over observed agreement.

- **Disagreement Study:** In Table 4, we present the category wise disagreement for all the annotator pairs. From the disagreement table it is evident that the categories with maximum number of disagreements are *anger*, *disgust* and *fear*. The emotions which are close to each other in the evaluation-activation space are inherently ambiguous. For example, anger and disgust are close to each other in the evaluation-activation space. So, ambiguity between these categories will be higher compared to other pairs. If [a b] is the pair, we count the number of cases where one annotator categorized one item into [a -] pattern and other annotator classified the same item into [- b] pattern. In Table 5, we provide the confusion between two affective categories for all annotator pairs. This confusion matrix is a symmetric one. So, we have provided only the upper triangular matrix.

  In Figure 2, we provide ambiguity counts of the affective category pairs. It can be observed that *anger*, *disgust* and *fear* are associated with three topmost ambiguous pairs.

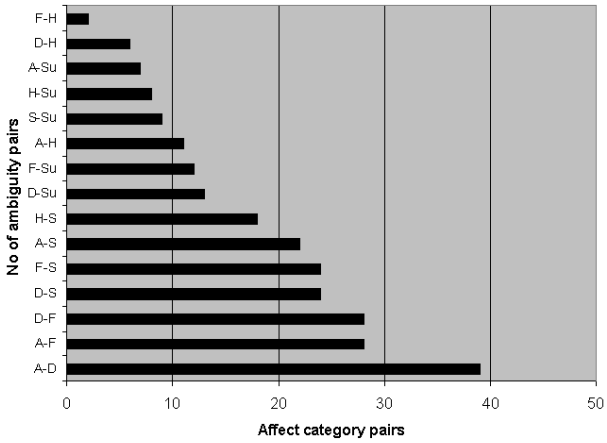

Figure 2: Category pair wise disagreement (A=Anger, D=Disgust, F=Fear, H=Happiness, S=Sadness and Su=Surprise).

5.3 Gold Standard for Affective Text Corpus

To determine the *gold standard* corpus, we have applied majority decision label based approach discussed in section 4 on the judgements provided by only three annotators. However, as the number of annotators is much less in the current study, the determined gold standard corpus may not have

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
1-2	68	94	74	64	74	45
1-3	74	86	105	57	54	45
2-3	65	49	58	22	50	20
Total	207	229	273	143	178	110

Table 4: Categorywise disagreement for the annotator pairs.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	-	39	28	11	22	7
Disgust	-	-	28	6	24	13
Fear	-	-	-	2	24	12
Happiness	-	-	-	-	18	8
Sadness	-	-	-	-	-	9
Surprise	-	-	-	-	-	-

Table 5: Confusion matrix for category pairs.

much significance. Here, we report the result of applying the gold standard determination algorithm on the data provided by three annotators. The distribution of sentences over the affective categories is depicted in Figure 3.

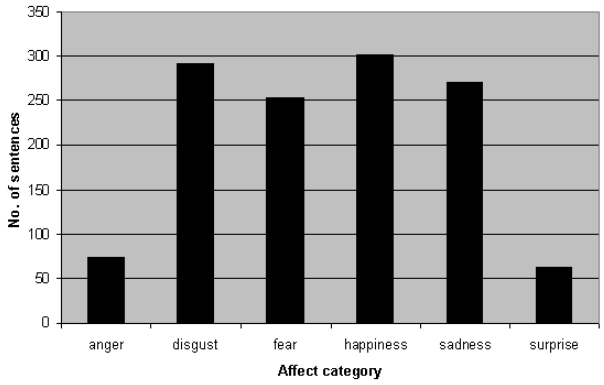

Figure 3: Distribution of sentences in gold standard corpus.

6 Conclusion and Future Work

Measuring the reliability of the affective text corpus where one single item may be classified into more than one single category is a complex task. In this paper, we have provided a new coefficient to measure reliability in multiclass annotation task by incorporating pairwise agreement in affective class pairs. The measure yields an agreement value 0.72, when applied to an annotated corpus provided by three users. This considerable agreement value indicates that the affect categories considered for annotation may be applicable to the news genre.

We are in process of collecting annotated corpus from more annotators which will ensure a statistically significant result. According to the disagreement study presented in section 5.2, confusions between specific emotions is most likely between categories which are adjacent in the activation-evaluation space. The models of annotator agreement which use weights for different types of disagreement will be interesting for future study. The direct and indirect affective sentences have not been treated separately in this study. The algorithm for determination of gold standard requires more details investigation as simple majority voting may not be sufficient for highly subjective data like emotion.

Acknowledgement

Plaban Kr. Bhowmick is partially supported by Microsoft Corporation, USA and Media Lab Asia, India. The authors are thankful to the reviewers for their detailed suggestions regarding the work.

References

Artstein, Ron and Massimo Poesio. 2008. *Inter-coder Agreement for Computational Linguistics*. Computational Linguistics.

Bruce, Rebecca F. and Janyce M. Wiebe 1999. *Rec-*

ognizing Subjectivity: A Case Study of Manual Tagging. Natural Language Engineering. 1(1):1-16.

Carletta, Jean. 1996. *Assessing Agreement on Classification Tasks: The Kappa Statistic*. Computational Linguistics. 22(21):249-254.

Carletta, Jean, Isard .A, Isard S., Jacqueline C. Kowtko, Gwyneth D. Sneddon, and Anne H. Anderson. 1997. *The Reliability of a Dialogue Structure Coding Scheme*. Computational Linguistics. 23(1):13-32.

Cohen, Jacob. 1960. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement. 20(1):37-46.

Craggs Richard and Mary M. Wood. 2004. *A Categorical Annotation Scheme for Emotion in the Linguistic Content of Dialogue*. Tutorial and Research Workshop, Affective Dialogue Systems. Kloster Irsee, 89-100.

Darwin, Charles. 1965. *The Expression of Emotions in Man and Animals.*. Chicago: University of Chicago Press. (Original work published 1872)

Ekman, Paul., Friesen W. V., and Ellsworth P. 1982. *What Emotion Categories or Dimensions can Observers Judge from Facial Behavior?* Emotion in the human face, Cambridge University Press. pages 39-55, New York.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. Wiley. second ed., New York.

Hagen-Zanker, Alex. 2003. *Fuzzy Set Approach to Assessing Similarity of Categorical Maps*. International Journal for Geographical Information Science. 17(3):235-249.

Hripcsak, George and Daniel F. Heitjan. 2002. *Measuring Agreement in Medical Informatics Reliability Studies*. Journal of Biomedical Informatics. 35(2):99-110.

Jung, Ho-Won. 2003. *Evaluating Interrater Agreement in SPICE-based Assessments*. Computer Standards & Interfaces. 25(5):477-499.

Krippendorff, Klaus 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications. Beverley Hills, CA.

Mieskes, Margot and Michael Strube. 2006. *Part-of-Speech Tagging of Transcribed Speech*. Proceedings of International Conference on Language Resources and Evaluation. GENOA

Ortony, Andrew and Terence J. Turner. 1990. *What's Basic About Basic Emotions?*. Psychological Review. 97(3):315-331.

Pantic, Maja and Leon Rothkrantz. 2000. *Automatic Analysis of Facial Expressions: The State of the Art*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22(12):1424-1445.

Plutchik, Robert 1980. *A General Psychoevolutionary Theory of Emotion*. Emotion: Theory, research, and experience: Vol. 1. Theories of emotion. Academic Press, New York, 3-33.

Rosenberg, Andrew, and Ed Binkowski. 2004. *Augmenting the Kappa Statistic to Determine Interannotator Reliability for Multiply Labeled Data Points*. In Proceedings of North American Chapter of the Association for Computational Linguistics. Boston, 77-80.

Scott, William A. 1955. *Reliability of Content Analysis: The Case of Nominal Scale Coding*. Public Opinion Quarterly. 19(3):321-325.

Song, Mingli, Chun Chen, Jiajun Bu, and Mingyu You. 2004. *Speech Emotion Recognition and Intensity Estimation*. Internation Conference on Computational Science and its Applications. Perugia, 406-413.

Stolcke A., Ries K., Coccaro N., Shriberg E., Bates R., Jurafsky .D, Taylor P., Martin C. Van-Ess-Dykema, and Meteer .M. 1997. *Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech*. Computational Linguistics. 26(3):339-371.