Determining Reliability of Subjective and Multi-label Emotion Annotation through Novel Fuzzy Agreement Measure

Plaban Kr. Bhowmick, Anupam Basu, Pabitra Mitra

Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur, India - 721302
plaban@gmail.com, anupambas@gmail.com, pabitra@gmail.com

Abstract

The paper presents a new fuzzy agreement measure γ_f for determining the agreement in multi-label and subjective annotation task. In this annotation framework, one data item may belong to a category or a class with a belief value denoting the degree of confidence of an annotator in assigning the data item to that category. We have provided a notion of disagreement based on the belief values provided by the annotators with respect to a category. The fuzzy agreement measure γ_f has been proposed by defining different fuzzy agreement sets based on the distribution of difference of belief values provided by the annotators. The fuzzy agreement has been computed by studying the average agreement over all the data items and annotators. Finally, we elaborate on the computation γ_f measure with a case study on emotion text data where a data item (sentence) may belong to more than one emotion category with varying belief values.

1. Introduction

Several coefficients of reliability (Artstein and Poesio., 2008) for measuring agreement among the annotators are available and they have widely been used in measuring reliability of different annotation tasks. These reliability measures consider judgements that classify a data item into a single category out of a set of predefined categories. For example, the parts of speech (POS) of a word in a sentence may be one of the discrete categories like *noun, pronoun, verb, adverb, adjective, etc*. There is no uncertainty involved in this kind of judgement process. But there are a number of domains where the judgement processes are ambiguous and one annotator may assign a data item to different categories with different degrees of belief or certainty. For example, one text segment may evoke multiple emotions in a reader's mind. As the emotion is subjective entity, different emotions may be evoked with different levels of intensity.

This kind of subjective and multi-label annotation task can be formally stated as follows:

Definition 1 *Let X be the domain of data instances and let Y be the set of discrete classes considered in an annotation task. The annotation or judgment task $\mathcal{J}$ is defined as $\mathcal{J} : X \longrightarrow \overrightarrow{Y}$ where $\overrightarrow{Y}$ is a vector and the k^{th} element y_k ($k = 1, 2, \ldots, |Y|$) of $\overrightarrow{Y}$ can take a value from the range [0,1]. The value of element y_k refers to the strength of belief or confidence of the annotator in labeling the data item with k^{th} label.*

In this work, we aim at measuring reliability of emotion annotation with a proposed agreement measure considering the classification process to be multi-label and subjective in nature. The agreement measure function is represented as follows:

$$\gamma = f(J_1, J_2, \ldots, J_N), \gamma \in [0, 1]$$

where J_i is the annotation provided by the i^{th} annotator and N is the number of annotators.

2. Related Works

Different coefficients of agreement have been proposed and widely been used in reliability assessment in different domains. One of the most popular among these is perhaps the Kappa coefficient introduced by Cohen (Cohen, 1960) for measurement of agreement in nominal scale. 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 software quality control (Park and 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 sample survey research. Kappa and π measures differ in the way they determine the chance related agreements. Scott's π assumes the distribution of proportions over categories to be same for all the coders. But Cohen's Kappa treats the individual coder distributions separately.

One of the drawbacks of π and Kappa like coefficients is that they do not consider the fact that inter-class ambiguity may widely vary over different class pairs. 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 to measure agreement in nominal, interval, ordinal and ratio scale data..

The above mentioned reliability coefficients cannot be applied to measure agreement in multi-label and subjective annotation tasks mentioned before. To deal with agreement in subjective annotation, Dou et. al. (Dou et al., 2007) proposed an agreement measure between two fuzzy classifiers, where the fuzzy agreement function for two classifications

considers the fuzzy *min* composition (Ross, 1997) of the membership values of a data item. The observed agreement is the average fuzzy agreement over all the data items. The expected agreement is computed by using the probability distributions of the membership values and the fuzzy agreement function. Although, this work is relevant in comparing the fuzzy classification process, it has some limitations as listed below.

- Agreement measurement is limited to only two annotators.
- *Min* composition has been used in the agreement function which may be applicable to fuzzy classifiers. But, in judgements provided by the human judges, the membership assignment process is subjective in nature. Using *Max* or *Min* composition in agreement function introduces a bias towards an annotator in an annotator pair. The annotator with smaller membership value will be rewarded for *Max* composition and the annotator with higher membership value will be penalized for *Min* composition.

3. Emotion Text Corpus Annotation

The emotion text corpus collected by us consists of 1000 sentences extracted from *Times of India* news paper archive. The sentences were collected from headlines as well as articles belonging to political, social, sports and entertainment domain.
The annotation scheme considers the following points:

- Our annotation scheme considers six basic emotions, namely, *Anger, Disgust, Fear, Happiness, Sadness, and Surprise* as specified by Ekman (Ekman et al., 1982).
- A sentence may trigger multiple emotions simultaneously. So, one annotator may classify a sentence to more than one emotion category.
- An annotator may assign belief value while assigning a sentence into an emotion category. This belief value reflects the extent to which the concerned sentence triggers a particular emotion. The range of the belief value is in [0, 1] with intervals of 0.1.

Five human judges with the same social background participated in the study, assigning emotion 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. An example annotation is provided in Table 1.
Distribution of the sentences across the emotion categories for the five judges is given in Figure 1.

4. Fuzzy Agreement Measure (γ_f)

The proposed fuzzy agreement measure γ_f is defined with notion of disagreement. As discussed earlier, the annotators not only categorizes a data item into a number of classes but also provide a belief or confidence value against each class. The annotators are said to perfectly agree if the difference between the belief values is zero.

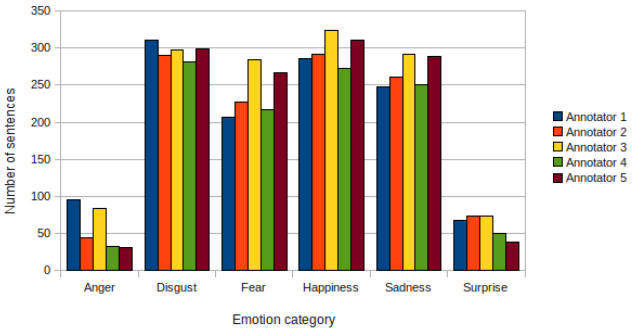


Figure 1: Distribution of sentences for five judges.

4.1. Notion of Disagreement

For a single category and annotator pair, we define a disagreement function (d), which yields the disagreement between an annotator pair in providing subjective values against a particular class for a data item. Following points have been considered while defining the function.

- The disagreement value is proportional to the absolute difference between the values provided by the annotators.
- High belief value against one class for a data item by an annotator indicates that the annotator assigns the data item into the class with high confidence, whereas, the assignment of low belief value signify that the data item belongs to the class with low confidence. Disagreement at higher belief values thus should contribute more to the disagreement function as compared to the disagreement at lower values.

Accordingly, the normalized disagreement function, d, for a <*data item, annotator pair, class*> triplet is given by

$$\alpha = d(x, y) = \mid x - y \mid (1 + e^{max(x,y)}) \qquad (1)$$

Here x and y are the belief values of a data item provided against a class by two annotators. Lower the α value, better is the agreement. Notice that, α value ranges from 0 to 3.72. The value of α becomes 0 when both the annotators provide the same belief value. The value of α is maximum (3.72) when one annotator provides 0 and the other assigns 1.

4.2. Determining Fuzzy Agreement Functions

After obtaining the disagreement values, they are partitioned into several clusters using k-means clustering algorithm such that each partition contains disagreement values which are close. To obtain optimal number of clusters, we adopt a *silhouette* based cluster validity measure (Rousseeuw, 1987).
Let M be the set of cluster centers after optimal clustering of the α values. Each data point in these clusters is an entry for the triplet <*data item, annotator pair, class*>. The clusters can be ordered according to the values of their respective cluster centers. The triplets belonging to the cluster with lower cluster center value are better agreed upon

Sentence	Anger	Disgust	Fear	Happy	Sad	Surprise
The four terrorists in the Taj Mahal hotel have killed virtually anyone and everyone they saw.	0	0	0.9	0	0.6	0

Table 1: An example annotation.

than the other clusters. To incorporate this ordering, relative weights are assigned to each cluster. The weight for i^{th} cluster C_i is assigned based on its relative ordering with respect to the lowest disagreement value ($\alpha_0 = 0$). The weight for cluster C_i is given by

$$w_i = \frac{1 - |\alpha_0 - M_i|}{\sum_{j=1}^{I} 1 - |\alpha_0 - M_j|} \quad (2)$$

where I is the number of clusters and M_i is the center of cluster C_i.

Depending on the distribution of the α values different number of *fuzzy agreement sets (FAS)* can be defined. The left most interval contains the lower α values (i.e., triplets with higher agreement) and the other intervals with increasing cluster center values contains triplets with lower agreement. We designate the fuzzy agreement set residing at the left-most interval as the *High Agreement Set (HAS)* and those residing at the other remaining intervals are termed as *Low Agreement Set (LAS)*. The membership of a triplet in HAS and LASs are determined with a Z-function and π-functions respectively. We augment the list of cluster centers with minimum and maximum values of α to derive the parameters of the fuzzy functions. The augmented list (L) can be expressed as follows.

$$L = \{min(\alpha), M_1, M_2, ..., M_I, max(\alpha)\}$$

Next we fit membership functions to these intervals, as described below.

4.3. Assignment of Fuzzy Membership Functions

We assign membership function to each interval. The membership value of an α value within an interval is the degree of belongingness of that value to this interval. The membership value is highest (1.0) for the center of an interval. An example of membership functions for four intervals is shown in Figure 2. The left-most interval is assigned with a Z-function and the remaining intervals are assigned with π-functions. The assignment of fuzzy membership functions to the HAS and the LAS's are as follows.

- *HAS membership function:* First, we consider the case of left-most interval. In this case, the interval center M_1 is bounded by 0 and M_2. The points belonging to the range [0 M_1] possess the highest membership value (1.0) and the lowest membership point is located at M_2. The membership function for this interval is represented with a Z-function and is given by

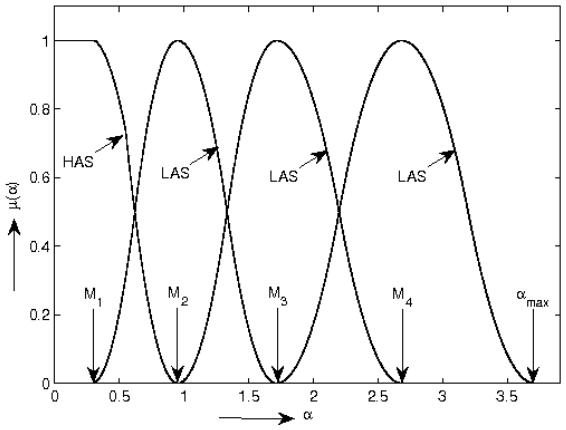


Figure 2: Example of the membership functions for fuzzy agreement sets. M_1, M_2, M_3 and M_4 are the cluster centers.

$$Z(u; a, b, c) = \begin{cases} 1 - 2\left(\frac{(u-a)}{(c-a)}\right)^2 & \text{if } a \leq u < b \\ 2\left(\frac{(u-c)}{c-a}\right)^2 & \text{if } b \leq u < c \\ 1 & \text{if } u < a \\ 0 & \text{if } u > c \end{cases} \quad (3)$$

The membership function μ^h for the left-most interval is given by

$$\mu^h(\alpha) = Z(\alpha; 0, M_1, M_2) \quad (4)$$

- *LAS membership function:* The LAS's are located in the remaining intervals. The membership functions for these intervals are represented using different π functions. An interval center M_i is bounded by its left interval center M_{i-1} and its right interval center M_{i+1}. The highest membership value occurs at M_i and lowest membership value is at M_{i-1} and M_{i+1}. The membership function μ^l is represented by π function and the π function is a combination of an S-Function bounded by M_{i-1} and M_i and a Z-function bounded by M_i and M_{i+1}. The S-function is given by

$$S(u; a, b, c) = \begin{cases} 2\left(\frac{(u-a)}{c-a}\right)^2 & \text{if } a \leq u < b \\ 1 - 2\left(\frac{(u-c)}{c-a}\right)^2 & \text{if } b \leq u < c \\ 0 & \text{if } u < a \\ 1 & \text{if } u > c \end{cases} \quad (5)$$

The membership function for interval centered at M_i is given by

$$\mu^l(\alpha) = \begin{cases} S(\alpha; M_{i-1}, \frac{(M_i+M_{i-1})}{2}, M_i) & \text{if } M_{i-1} \leq \alpha \leq M_i \\ Z(\alpha; M_i, \frac{(M_{i+1}+M_i)}{2}, M_{i+1}) & \text{if } M_i < \alpha \leq M_{i+1} \end{cases} \quad (6)$$

4.4. Computation of Fuzzy Agreement Measure

The agreement function takes the judgement or annotation matrices (**J**) as parameter. The i^{th} annotation matrix is represented as follows.

$$\mathbf{J_i} = \begin{bmatrix} x^i_{11}\, x^i_{12} \ldots x^i_{1K} \\ x^i_{21}\, x^i_{22} \ldots x^i_{2K} \\ \ldots \\ \ldots \\ x^i_{D1}\, x^i_{D2} \ldots x^i_{DK} \end{bmatrix}_{D \times K}$$

where x^i_{dk} is the value provided by i^{th} annotator against class k for data item d, D and K be the number of data items and classes respectively. The determination of the fuzzy agreement over the data set involves the following steps.

i) Let $\mathcal{V}$ be the set of annotator pairs. For each $< v, d, k >$ triplet ($v \in \mathcal{V}$) we compute α value with belief values x^i_{dk} and x^j_{dk} (considering the annotator pair v consisting of i^{th} and j^{th} annotators) using Equation 1. In this step, we obtain $\binom{N}{2}$ matrices consisting of α values. The matrix for pair v is given below.

$$\mathbf{X_v} = \begin{bmatrix} \alpha^v_{11}\, \alpha^v_{12} \ldots \alpha^v_{1K} \\ \alpha^v_{21}\, \alpha^v_{22} \ldots \alpha^v_{2K} \\ \ldots \\ \ldots \\ \alpha^v_{D1}\, \alpha^v_{D2} \ldots \alpha^v_{DK} \end{bmatrix}_{D \times K}$$

where α^v_{dk} is the alpha value for $< v, d, k >$ triplet.

ii) The α values obtained for all $< v, d, k >$ triplets stored in the $\mathbf{X_v}$'s ($v \in \mathcal{V}$) are clustered using k-means clustering algorithm to obtain optimal number of intervals in the range of α.

iii) The membership values of α in the fuzzy agreement sets are computed for each $< v, d, k >$ triplet. In this step, for each annotator pair v, we obtain one membership matrix for HAS and one or more than one LAS which are of the following form.

$$\mathbf{H} = \begin{bmatrix} \mu^h_{11}\, \mu^h_{12} \ldots \mu^h_{1K} \\ \mu^h_{21}\, \mu^h_{22} \ldots \mu^h_{2K} \\ \ldots \\ \ldots \\ \mu^h_{D1}\, \mu^h_{D2} \ldots \mu^h_{DK} \end{bmatrix}_{D \times K}$$

$$\mathbf{L_i} = \begin{bmatrix} \mu^l_{11}\, \mu^l_{12} \ldots \mu^l_{1K} \\ \mu^l_{21}\, \mu^l_{22} \ldots \mu^l_{2K} \\ \ldots \\ \ldots \\ \mu^l_{D1}\, \mu^l_{D2} \ldots \mu^l_{DK} \end{bmatrix}_{D \times K}$$

iv) The agreement matrix ($\mathbf{A_v}$) for an annotator pair v is obtained by applying element wise weighted fuzzy aggregation on the obtained fuzzy agreement sets. This aggregation operation can be represented as presented below.

$$\mathbf{A_v} = \otimes(w^h, w^l_1, w^l_2, \ldots, w^l_q; \mathbf{H}, \mathbf{L_1}, \mathbf{L_2}, \ldots, \mathbf{L_q}) \quad (7)$$

where there are $q = I - 1$ number of LAS's. The weights are computed using Equation 2. The weighted fuzzy aggregation technique proposed by Cron and Dubuisson (Cron and Dubuisson, 1998) has been used for elementwise fuzzy aggregation computation. For any triplet, the aggregation operation is given by

$$g_{(w^h, w^l_1, w^l_2, \ldots, w^l_q)}(\mu^h, \mu^l_1, \mu^l_2, \ldots, \mu^l_q) \quad (8)$$

v) The average agreement value for each $< d, k >$ pair is obtained by aggregating the agreement values for all the annotator pairs for class k and data item d. The aggregation operation is performed on $\mathbf{A_v}$ matrices for all the annotator pairs by means of fuzzy conjunction operator ($\bigcap$) applied on every element as below.

$$\mathbf{A}^* = \bigcap_{v=1}^{|\mathcal{V}|} \mathbf{A_v} \quad (9)$$

This step produces a $D \times K$ matrix $\mathbf{A}^*$ of agreement values.

vi) Each row vector of an $\mathbf{A}^*$ matrix is a point in the K-dimensional space. The points are said to exhibit similar agreement pattern if they are close to each other in this K dimension space. The K dimensional row vectors of the $\mathbf{A}^*$ matrix are clustered into groups of data points with similar agreement patterns using k-means clustering algorithm.

vii) Distance from a cluster center to the lowest agreement point in the K dimension space signifies the average agreement value for the data points belonging to that cluster. The lowest agreement point (O) has the coordinate values as zero.

The average agreement is computed by calculating the average of the Euclidean distances from the lowest agreement point to the cluster centers. So the fuzzy agreement (γ_f) is given by

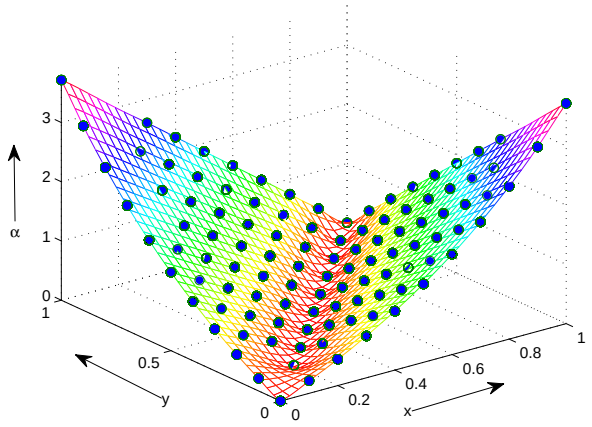
$$\gamma_f = \frac{1}{I}\Big(\sum_{i=1}^{I} e(C_i, O)\Big) \quad (10)$$

where C_i is the center of the i^{th} cluster and e is the Euclidean distance between two points.

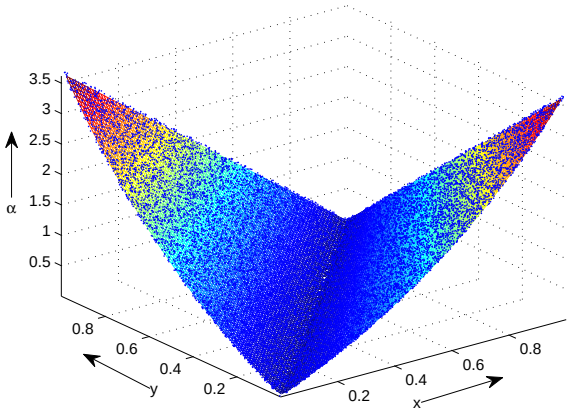
5. Measuring Reliability of Emotion Annotation through γ_f

We followed the steps outlined in Section 4.4. to compute the fuzzy agreement for the emotion text data. Following observations are made during the computation of the fuzzy agreement.

- *α Values*: The α values are computed using Equation 1. The α value ranges from 0 to 3.72 which is the maximum range possible. The distribution of α values are depicted in Figure 3.



(a) Distribution of α for emotion data



(b) Distribution of α for random data

Figure 3: Distribution of α values.

- *Interval Selection*: The α values are clustered using k-means algorithm and the optimal number of clusters is two. The centers of the clusters are given by

$$M = \{0.04, 1.14\} \tag{11}$$

- *Membership Function Assignment*: As there are two intervals, the number of fuzzy functions are two. The left-most membership function which is the HAS, is a Z-function with parameters a = 0, b = 0.04 and c = 1.14. The second function is the LAS represented by a π-Function with parameters a = 0.04, b = 1.14 and c = 3.72. The membership functions are depicted in Figure 4.

- *Weights of Fuzzy Sets*: The weights for HAS and LAS are computed to be 0.70 and 0.30 respectively.

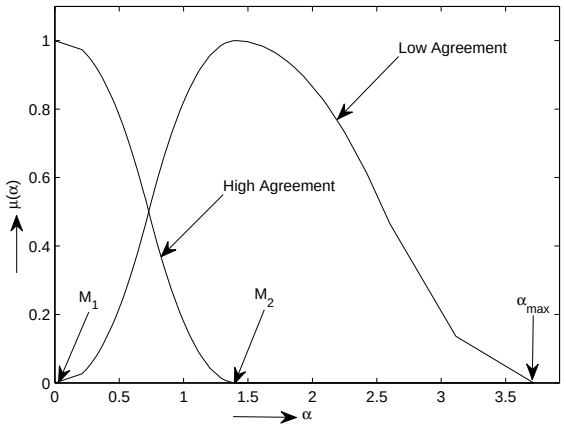


Figure 4: Membership functions for emotion data.

- *Clustering of Aggregated Data*: Here, we also applied silhouette measure based optimal cluster selection approach to determine the number of optimal clusters where each cluster represents a pattern of similar agreement values across the emotion categories. We obtain three optimal clusters. The centers of the clusters are given in Table 2. The scatter plot of clusters in disgust-fear-sadness dimension is shown in Figure 5.

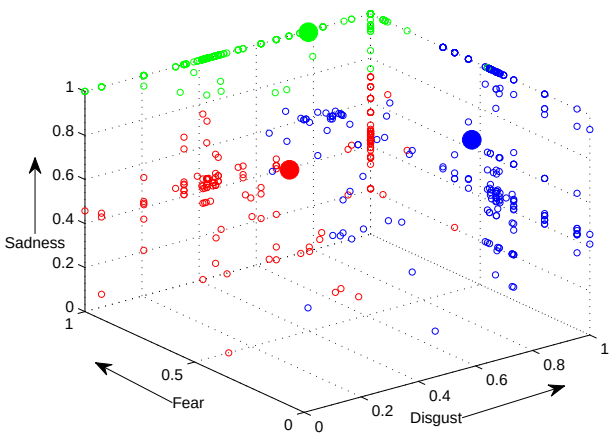


Figure 5: Scatter plot the clusters of data points in disgust-fear-sadness dimension

- *fuzzy Agreement Value*: The fuzzy agreement value (γ_f) is the average of the distances between the cluster centers and the lowest agreement point using Equation 10. γ_f is computed to be 0.86. The computed γ_f value signifies that the agreement is good for the emotion annotation task considered in this study.

6. Conclusions

In this paper, we have proposed a fuzzy measure for determining agreement in multi-label and subjective emotion data. The proposed measures are generalizations over π and κ like measures where the classification process considers

Cluster	Anger	Disgust	Fear	Happiness	Sadness	Surprise
C_1	0.97	0.98	0.99	0.65	0.90	0.93
C_2	0.92	0.94	0.41	0.96	0.69	0.93
C_3	0.85	0.41	0.93	0.95	0.83	0.94

Table 2: Centers of the clusters of aggregated data.

that a data item may either belong to a class or does not. So, the belief values provided for a data item in a class is either 0 or 1 which is specialization of the case where a data item may have any belief value within [0 1] range. The proposed agreement measure has been applied in emotion annotation task where one data item may be assigned to multiple categories with subjective belief value. The agreement value was computed to be 0.86.

Acknowledgment

Plaban Kumar Bhowmick is partially supported from projects sponsored by Microsoft Corporation, USA and Media Lab Asia, India.

7. References

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistics. *Computational Linguistics*, 22(2):249–254.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

G. Cron and B. Dubuisson. 1998. A weighted fuzzy aggregation method. In *Proceedings of IEEE International Conference on Fuzzy Systems*, pages 675–680, Anchorage, Alaka.

W. Dou, Y. Ren, Q. Wu, S. Ruan, Y. Chen, D. Bloyet, and J. M. Constans. 2007. Fuzzy kappa for the agreement measure of fuzzy classifications. *Neurocomputing*, 70(4-6):726–734.

P. Ekman, W. V. Friesen, and P. Ellsworth. 1982. What emotion categories or dimensions can observers judge from facial behavior? In P. Ekman, editor, *Emotion in the human face*, pages 39–55. Cambridge University Press, New York.

J. L. Fleiss. 1981. *Statistical methods for rates and proportions*. John Wiley & Sons, New York.

A. Hagen. 2003. Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, 17(3):235–249.

G. Hripcsak and D. F. Heitjan. 2002. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2):99–110.

K. Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage Publications, Newbury Park, CA.

H. M. Park and H.W. Jung. 2003. Evaluating interrater agreement with intraclass correlation coefficient in spice-based software process assessment. In *Proceedings of the Third International Conference on Quality Software*, page 308, Washington, DC, USA. IEEE Computer Society.

T. J. Ross. 1997. *Fuzzy logic with engineering applications*. McGraw-Hill Inc., New York.

Peter Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, November.

W. A. Scott. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.