On the Use of Kappa Coefficients to Measure the Reliability of the Annotation of Non-acted Emotions

Zoraida Callejas and Ramón López-Cózar

Dept. of Languages and Computer Systems, 18071 Granada, Spain
{zoraida,rlopezc}@ugr.es

Abstract. In this paper we study the impact of three main factors on measuring the reliability of the annotation of non-acted emotions: the annotator biases, the similarity between the classified emotions, and the usage of contextual information during the annotation. We employed a corpus collected from real interactions between users and a spoken dialogue system. The user utterances were classified by nine non-expert annotators into four categories. We discuss the problems that the nature of non-acted emotional corpora impose in evaluating the reliability of the annotations using Kappa coefficients. Although deeply affected by the so-called paradoxes of Kappa coefficients, our study shows how taking into account context information and similarity between emotions helps to obtain values closer to the maximum agreement rates attainable, and allow the detection of emotions which are expressed more subtly by the users.

1 Introduction

One of the difficulties of non-acted emotion recognition is that in most application domains the corpora obtained are very unbalanced, because there is usually a higher proportion of neutral than emotional utterances [Morrison et al., 2007]. Thus, the Kappa coefficients indicate very low inter-annotator agreement even when the actual observed agreement between the annotators is high. This is called the *prevalence* phenomena, which is caused by the high probability of agreeing by chance in the neutral category. Hence, interpretation approaches based uniquely on already established values of acceptability such as the ones proposed by [Landis and Koch, 1977] and [Krippendorff, 2003] are not suitable for this application domain, as they would consider most of the annotation results not reliable.

As prevalence appears as an unavoidable consequence of the natural skewness of non-acted emotional corpora, some authors report additional measures to complement the information provided with the Kappa coefficients. For example, [Forbes-Riley and Litman, 2004] report on both observed agreement and Kappa, whereas [Lee and Narayanan, 2005] report on Kappa along with an hypothesis test. Although reported Kappa values in emotion recognition employing unbalanced corpora are usually low, e.g. from 0.32 to 0.42 in [Shafran et al., 2003]

E. André et al. (Eds.): PIT 2008, LNAI 5078, pp. 221–232, 2008.

and below 0.48 in [Lee and Narayanan, 2005] and [Ang et al., 2002], there is not a deep discussion about the problematic of Kappa values in the area, not even in papers explicitly devoted to challenges in emotion annotation (for instance, [Devillers et al., 2005]). Furthermore, even when other agreement measures are reported along with Kappa, e.g. [Forbes-Riley and Litman, 2004] and [Lee and Narayanan, 2005], there is only one Kappa coefficient calculated (usually multi-π) and no discussion about why there is such a big difference between the Kappa values and the other measures reported.

In this paper, we report experimental results on the annotation of the recordings of real interactions of users with a spoken dialogue system. The procedure was carried out by nine non-expert annotators following two strategies: in the former the annotators had information about the dialogue context and the users' speaking style; in the latter, their decision was based only on the acoustics of the utterances. With the recorded emotional corpora, we address three main issues related to the use and interpretation of kappa coefficients in the annotation of real emotions: i) the impact of annotator bias, that is, given a fixed number of agreements, the effect that the distribution of disagreements between categories has in the Kappa value; ii) the level of importance of all possible disagreements in our task, i.e. disagreements between emotions which are easily distinguishable should have a more negative impact in the Kappa coefficient than disagreements in more similar categories; and iii) the benefits yielded by the use of contextual information on the obtained agreement values and the emotions annotated.

2 Experimental Set-Up

The UAH (Universidad al Habla - University On the Line) dialogue system was developed in our laboratory to provide telephone-based spoken access to the information in our Department web page [Callejas and López-Cózar, 2005]. The corpus used for the experiments described in this paper is comprised of 85 dialogues of 60 different users interacting with the system. The corpus contains 422 user turns, with an average of 5 user turns per dialogue. The recorded material has a duration of 150 minutes. The users were mainly students and professors at the University of Granada, which is in South Eastern Spain. The way the users expressed themselves was influenced by the Eastern Andalusian accent, which although similar to Spanish Castilian has several differences, for example a faster rhythm and lower expiratory strength.

To get the best possible annotation employing non-expert annotators, the labelling process must be rigorously designed. We have followed some of the ideas suggested by [Vidrascu and Devillers, 2005] to decide the list of labels and annotation scheme. The first step is to decide the labels to be used for annotation. Our goal was to annotate negative emotional states of the user during the interaction with the UAH system in order to obtain an emotional corpus to train an emotion recognizer for the system. We have used four categories in the annotation of the corpus: *angry*, *bored*, *doubtful* and *neutral*. The first three categories represented

the major negative emotions encountered in the UAH corpus; whereas *neutral* represented a non-negative state[1].

We decided to use an odd, high number of annotators (nine) which is more than is typically reported in previous studies [Forbes-Riley and Litman, 2004] [Lee and Narayanan, 2005]. In our group of annotators, six were used to the Andalusian accent and three were not. Regarding the "segment length", in our study this is the whole utterance because our goal was to analyze the emotion as a whole response to a system prompt, without considering the possible emotional changes within the response. The utterances were annotated twice by every annotator following both annotation schemes, the annotations were carried out in different sessions separated by a long period of time to avoid obtaining a biased second annotation. In the first case the annotators had information about the dialogue context and the users' speaking style. In the second case, the annotators did not have this information, so their annotations were based only on acoustic information.

The final emotion category assigned to each utterance in the ordered and unordered schemes was the one annotated by the majority of annotators. Global emotions for the whole corpus were then computed from the results of each of the schemes. In situations where there was no majority emotion (e.g. 4 *neutral*, 4 *bored* and 1 *doubtful*), priority was given to non-neutrals (*bored* in the example). If this conflict was between two non-neutral emotions (e.g. 4 *doubtful*, 4 *bored* and 1 *neutral*), the results were compared between both annotation schemes to choose the emotion annotated by majority among the 18 annotations (the 9 of the ordered and the 9 of the unordered schemes).

On average, among the nine annotators, more than 85% of the utterances were annotated as *neutral*. We have also observed that this proportion is affected in 3.4% of the cases by the annotation style. Concretely, for the ordered annotation, 87.28% were tagged as *neutral*, whereas for the unordered annotation the corpus was even more unbalanced: 90.68% of the utterances were annotated as *neutral*.

3 Calculation of the Agreement between Annotators

Several Kappa coefficients were used to study the degree of inter-annotator agreement for both annotation styles (ordered and unordered). Kappa coefficients are based on the idea of rating the proportion of pairs of annotators in agreement (P_o) with the expected proportion of pairs of annotators that agree by chance (P_c). The result is a proportion between the agreement actually achieved beyond chance ($P_o - P_c$) and all the possible agreements that are not by chance ($1 - P_c$):

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{1}$$

For our study we used five different Kappa coefficients with which we studied two main issues: i) the impact of annotator bias, i.e. given a fixed number of

[1] Positive emotions were treated as neutral because our interest was only on those emotions that could lead to user frustration and interaction failure.

agreements, the effect that the distribution of disagreements between categories has in the Kappa value; and ii) the level of importance of all the possible disagreements, i.e. disagreement between emotions which are easily distinguishable should have a more negative impact on the Kappa coefficient than disagreements in very different categories.

[Artstein and Poesio, 2005] made a considerable effort to clarify the definitions of the different Kappa coefficients. In order to avoid inconsistencies, we follow their notation for all the Kappa coefficients employed in this paper. The simplest Kappa coefficient used was proposed by [Fleiss, 1971], which we have noted as **multi-π**. The calculation of multi-π is based on Equation 1, where the observed agreement (P_o) is computed as the number of cases in which two different annotators agreed to annotate a particular utterance with the same emotion category:

$$P_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{e=1}^{E} n_{ue}(n_{ue} - 1) \tag{2}$$

In Equation 2, U is the number of utterances to be annotated, A the number of annotators, E the number of emotions, and n_{ue} the number of times the utterance 'u' was annotated with the emotion category 'e'.

Fleiss assumed that all the annotators share the same probability distribution. In our experiments, this means that the probability that an annotator classifies an utterance 'u' with a particular emotion category 'e', can be computed as the overall probability of annotating 'u' as 'e'. This global probability was computed as the total number of assignments to emotion category 'e' made by all annotators (n_e in Equation 3) divided by the total number of assignments ($U \cdot A$). Chance agreement (Equation 3) was then computed as the probability that any pair of labellers annotated the same utterance with the same category, which was assumed to be the joint probability of each of them making such assignment independently, as they judged all the utterances independently from each others.

$$P_c^{\pi} = \sum_{e=1}^{E} \left(\frac{1}{UA} n_e \right)^2 \tag{3}$$

The calculation of multi-π assumes that each annotator follows the same overall distribution of utterances into emotion categories. However, such a simplification may not be plausible in all domains due to the effect of th so-called *annotator bias* in the Kappa value. In our experiments, the annotator bias can be defined as the extent to which annotators disagree on the proportion of emotions, given a particular number of agreements. With the rest of the parameters fixed, the Kappa value increases as the bias value gets higher, that is, when disagreement proportions are not equal for all emotions and there is a high skew among them. This is the so-called *Kappa second paradox*. Different studies of the impact of this paradox can be found in the literature, e.g. [Feinstein and Cicchetti, 1990], [Lantz and Nebenzahl, 1996], and [Artstein and Poesio, 2005].

To study whether the inclusion of the different annotating behaviours could improve the Kappa values, we calculated the Kappa value that is proposed by [Davies and Fleiss, 1982], which we have noted as **multi-κ**. As happens with multi-π, the calculation of multi-κ also relies on Equation 1, and has the same observed agreement (Equation 2). However, for the chance agreement, it includes a separate distribution for each annotator. Thus, in this case the probability that an annotator 'a' classifies an utterance 'u' with an emotion category 'e' is computed with the observed number of utterances assigned to 'e' by that annotator (n_{ae}), divided by the total number of utterances (U). The probability that two annotators agree in annotating an utterance 'u' with the emotion category 'e' is again the joint probability of each annotator doing the annotation independently:

$$P_c^{\kappa} = \frac{1}{\binom{A}{2}} \sum_{e=1}^{E} \sum_{j=1}^{A-1} \sum_{k=j+1}^{A} \frac{n_{a_j e}}{U} \frac{n_{a_k e}}{U} \tag{4}$$

Despite of including differences between annotators, multi-κ gives all disagreements the same importance. In practice, all disagreements are not equally probable and do not have the same impact on the quality of the annotation results. For example, in our experiments, a disagreement between *neutral* and *angry* is stronger than between *neutral* and *doubtful*, because the first two categories are more easily distinguishable.

To take all this information into account we have used weighted Kappa coefficients, which put the emphasis on disagreements instead of agreements. The calculation of these coefficients is based on Equation 5 (equivalent to Equation 1):

$$\kappa_w = 1 - \frac{\overline{P}_o}{\overline{P}_c} \tag{5}$$

where $\overline{P}_o$ indicates observed disagreement, and $\overline{P}_c$ disagreement by chance. For all the coefficients used, the observed disagreement has been calculated as the number of times each utterance 'u' was annotated with two different emotion categories e_j and e_k by every pair of annotators, weighted by the distance between the categories:

$$\overline{P}_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{ue_j} n_{ue_k} distance(e_j, e_k) \tag{6}$$

Consequently, the computation of the weighted coefficients implies employing distance metrics between the four emotions used for annotation (*neutral*, *angry*, *bored* and *doubtful*). To do so, we have arranged our discrete list of emotions within a continuous space, using the bidimensional activation-evaluation space [Russell, 1980]. In the horizontal axis, evaluation deals with the "valence" of emotions, i.e. positive or negative evaluations of people, things or events. In the vertical axis, activation measures the user disposition to take some action rather than none. Emotions form a circular pattern in this space. This is why other

authors proposed a representation based on angles and distance to the centre. Taking advantage of this circular disposition, we have used angular distances between our emotions for the calculation of the weighted Kappa coefficients. Instead of establishing our own placement of the emotions in the space, we employed an already established angular disposition to avoid introducing measurement errors. We used the list of 40 emotions with their respective angles proposed by [Plutchik, 1980], which has been widely accepted and used by the scientific community. In this list, *bored* (136.0) and *angry* (212.0) were explicitly considered, but this was not the case for *doubtful*. The most similar emotions found were "uncertain", "bewildered" and "confused", which only differentiated in 2 in the circle. We chose "uncertain" (139.3) which was the one that better reflected the emotion we wanted to annotate. [Plutchik, 1980] did not reflect neutral in his list as it really is not an emotion but the absence of emotion. Instead, he used a state called "accepting" as the starting point of the circle (0), which we used as *neutral* in our experiments.

With the angle that each of the four emotions forms in the space we calculated the distance between them in degrees. We chose always the smallest angle between the emotions being considered (x or 360-x). This way, the distance between every two angles was always between 0 and 180 degrees. For the calculation of the Kappa coefficients, distances were converted into weights with values between 0 and 1. A 0 weight (which corresponds to 0 distance in our approach) implies annotating the same emotion, and thus having no disagreement. On the contrary, $weight = 1$ (180 distance) corresponds to completely opposite annotations and thus maximum disagreement. The resulting distances and weights are listed in Table 1.

Table 1. Distance between emotions

Angle/ Weight	Neutral	Angry	Bored	Doubtful
Neutral	0.00 / 0.00	148.00 / 0.82	136.00 / 0.75	139.30 / 0.77
Angry	148.00 / 0.82	0.00 / 0.00	76.00 / 0.42	72.70 / 0.40
Bored	136.00 / 0.75	76.00 / 0.42	0.00 / 0.00	3.30 / 0.02
Doubtful	139.30 / 0.77	72.70 / 0.40	3.30 / 0.02	0 / 0.00

There is not a consensus in the scientific community about the properties of the distance measures. However, [Artstein and Poesio, 2005] have proposed some constraints: the distance between a category and itself should be minimal and the distance between two categories should not depend on the order (i.e. the distance from A to B should be equal to distance from B to A). As can be observed by the symmetry of the table, our distance measures and weights follow these restrictions as the angle an emotion forms with itself is 0 and, as we established to choose the minimal angle, the distance between two emotions is the same regardless of the order.

As can be observed in the table, the highest distances were between non-neutrals and neutral. Thus, when calculating weighted Kappa coefficients, disagreements in which an annotator judged an utterance as neutral and the other as non-neutral were given more importance than, for example, a disagreement between the *angry* and *bored* categories.

We calculated three weighted Kappa coefficients. The first one was α, proposed by [Krippendorff, 2003]. The second was a variant called α' proposed by Artstein and Poesio, and the third coefficient was the β that is proposed by [Artstein and Poesio, 2005]. All of them shared the same observed disagreement calculation (Equation 5). Disagreement by chance for α and α' was calculated as:

$$\overline{P}_c^{\alpha} = \frac{1}{UA(UA-1)} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{e_j} n_{e_k} distance(e_j, e_k) \tag{7}$$

$$\overline{P}_c^{\alpha'} = \frac{1}{(UA)^2} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{e_j} n_{e_k} distance(e_j, e_k) \tag{8}$$

As can be observed in Equations 7 and 8, these coefficients do not consider annotator bias. This was addressed by employing the β coefficient, with which we have measured also the observed behaviour of each annotator:

$$\overline{P}_c^{\beta} = \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} \left[\frac{1}{U^2 \binom{A}{2}} \sum_{m=1}^{A-1} \sum_{n=m+1}^{A} n_{a_m e_j} n_{a_n e_k} distance(e_j, e_k) \right] \tag{9}$$

4 Discussion of the Results

The results for each described coefficient are listed in Table 2. A plausible reason for these results is that the incorporation of context in the ordered case influences the annotators in assigning the utterances belonging to the same dialogues to the same emotional categories. This way, there were no very noticeable transitions between consecutive utterances. For example, if anger was detected in one utterance, then the next one was probably also annotated as *angry*. Besides, the context allowed the annotators to have information about the user's speaking style and the interaction history. In contrast, in the unordered case the annotators only had information about the current utterance. Hence, sometimes they could not decide whether the user was either angry or he normally spoke loudly and fast.

In addition, when listening to the corpus in the ordered scheme, the annotators had information about the position of the current user turn within the whole dialogue, which also gave a reliable clue to the user's state. For example, a user was more likely to get bored after a long dialogue, or to become angry after many confirmation prompts generated by the system.

As can be observed in Table 2, the values of the different Kappa coefficients also vary slightly depending on the annotating scheme used. In the unordered

Table 2. Values of the Kappa coefficients for unordered and ordered annotation schemes

Coefficient	Unordered	Ordered
multi-π	0.3256	0.3241
multi-κ	0.3355	0.3256
α	0.3382	0.3220
α'	0.3381	0.3218
β	0.3393	0.3237

case, both taking into account annotator bias (multi-κ vs. multi-π, and β vs. α), and weighting disagreements (β and α vs. multi-κ) improves the agreement values. However, in the ordered case only taking into account annotator bias enhances the agreement values, whereas weighting the disagreements reduces Kappa. This is a consequence of the increment of non-neutral annotations in the ordered case. Taking into account that the great majority of agreements occur when annotators tag the same utterance as neutral, an increment in the number of emotions annotated as non-neutral provokes more discrepancies among the annotators and thus reduces the Kappa value. Furthermore, most of the disagreements occur between neutral and non-neutral categories, which are the emotions with higher distances according to our weighting scheme (Table 1), thus provoking weighted agreements to be lower in the case of the ordered scheme.

When we examined the annotation results, we found that there were remarkable differences between the annotators who were used to the Andalusian accent. From the non-neutral emotions encountered by the nine annotators, most of them were annotated by the ones that were not used to the Andalusian accent. This was probably caused by the confusion of characteristics of the accent with emotional cues, for example, confusing the Andalusian fast rhythm with an indication of anger. We studied the effect on the annotation schemes for both kinds of annotator and obtained the results shown in Table 3.

As can be observed in the table, the annotators used to the Andalusian accent obtained Kappa values for both annotation schemes which were more similar (ranging between 0.3234 to 0.3621). For these annotators, the Kappa values were smaller for the ordered scheme because there were fewer utterances annotated as neutral.

On the contrary, annotators not used to the Andalusian accent had very different Kappa values depending on the annotating scheme used: in the ordered case, values ranged from 0.5593 to 0.5697, whereas in the unordered the values ranged from 0.3639 to 0.3746. This is due to the big decrement of the chance agreement. The most likely reason for this is the lower number of neutrals annotated by annotators not used to Andalusian. This happens for both annotation schemes, but the number of neutrals annotated is higher in the unordered one, and this is why the results are more similar to those obtained by Andalusian annotators with the unordered annotation scheme. Even though the number of non-neutral annotations increased proportionally with the decrement of neutrals, the unbalancement of the corpus made the probability of agreeing

Table 3. Kappa values for the different annotator types

	Andalusian annotators		Non-andalusian annotators	
	Unordered	Ordered	Unordered	Ordered
multi-π	0.3608	0.3234	0.3734	0.5593
multi-κ	0.3621	0.3275	0.3746	0.5598
α	0.3595	0.3248	0.3644	0.5691
α	0.3592	0.3245	0.3639	0.5688
β	0.3607	0.3265	0.3703	0.5697

by chance in the neutral emotion more important in the computation of the overall agreement by chance. For example, in the case of multi-κ, the agreement by chance (P_c) was calculated as the sum of agreeing by chance in each emotion ($P_c = P_c^{neutral} + P_c^{bored} + P_c^{angry} + P_c^{doubtful}$). The values for agreeing by chance when annotators not used to Andalusian used the ordered scheme were $P_c^{neutral} = 0.6645$, $P_c^{bored} = 0.0052$, $P_c^{angry} = 0.0069$ and $P_c^{doubtful} = 0.0008$. For the rest of annotators these values were: $P_c^{neutral} = 0.8137$, $P_c^{bored} = 0.0010$, $P_c^{angry} = 0.0014$ and $P_c^{doubtful} = 0.0008$. Thus, $P_c^{neutral}$ was the determining factor in obtaining the global P_c.

The situation in which although having almost identical number of agreements, the distribution of these across the different annotation categories deeply affects Kappa, is typically known as the *first Kappa paradox*. This phenomenon establishes that other things being equal, Kappa increases with more symmetrical distributions of agreement. That is, if the prevalence of a category compared to the others is very high, then the agreement by chance (P_c) is also high and the Kappa is considerably decremented [Feinstein and Cicchetti, 1990].

As already reported by other authors, e.g. [Feinstein and Cicchetti, 1990], the first Kappa paradox can drastically affect Kappa values and thus must be considered in its interpretation. There is not an unique and generally accepted interpretation of the Kappa values. One of the most widely used is the one presented by [Landis and Koch, 1977], which makes a correspondence between intervals for Kappa values and interpretations of agreement. Following this approach, our experimental results indicate fair agreement for both annotating schemes and with the four different Kappa coefficients. Alternatively, [Krippendorff, 2003] established 0.65 as a threshold for acceptability of agreement results. Hence, considering this value, our 0.3393 highest Kappa would not be acceptable. However, most authors seem to agree in that using a fixed benchmark of Kappa intervals does not provide enough information to make a justified interpretation of acceptability of the agreement results. In order to provide a more complete framework, a number of authors, e.g. [Dunn, 1989], propose to place Kappa into perspective by reporting *maximum*, *minimum* and *normal* values of Kappa, which can be calculated from the observed agreement (P_o) as follows [Lantz and Nebenzahl, 1996]:

$$kappa_{max} = \frac{P_o^2}{(1-P_o)^2+1};\ kappa_{min} = \frac{P_o - 1}{P_o + 1};\ kappa_{nor} = 2P_o - 1 \qquad (10)$$

For the same observed agreement, the possible values of Kappa can deeply vary from $kappa_{min}$ to $kappa_{max}$ depending on the balancement of the corpus. $Kappa_{max}$ is obtained when maximally skewing disagreements while maintaining balanced agreements, whereas $kappa_{min}$ is obtained when agreements are skewed and disagreements balanced. $Kappa_{nor}$ does not correspond to an ideal value of Kappa, but rather to symmetrical distributions of both agreements and disagreements. As observed in Table 4, the displacement between actual and normal values was smaller in the ordered scheme. Thus, contextual information does not only allow recognizing more non-neutral emotions, but also obtaining Kappa values which, although smaller than in the unordered scheme in absolute value, are much closer to the *normal* and *maximum* agreement values attainable and further from the *minimum*.

Table 4. Kappa minimal, observed, normal and maximal values in the ordered and unordered schemes

	multi-π		multi-κ		α		α'		β	
	Unord.	Ord.	Unord.	Ord.	Unord.	Ord.	Unord.	Ord.	Unord.	Ord.
κ_{min}	-0.062	-0.086	-0.069	-0.085	-0.046	-0.064	-0.046	-0.064	-0.046	-0.064
κ_o	0.326	0.324	0.335	0.326	0.338	0.322	0.338	0.329	0.339	0.324
κ_{nor}	0.767	0.686	0.767	0.686	0.823	0.759	0.823	0.759	0.823	0.759
κ_{max}	0.770	0.693	0.770	0.693	0.825	0.763	0.825	0.763	0.825	0.763

As stated in [Lantz and Nebenzahl, 1996], departures from the $kappa_{nor}$ value indicate asymmetry in agreements or disagreements depending on whether they are closer to the minimum or maximum value respectively. Our results corroborate that reporting Kappa values is more informative when they are put into context, as we obtain a valuable indicative of possible unbalancements that has to be considered to reach appropriate conclusions about reliability of the annotations. For example, in our case there were significant departures from $kappa_{nor}$ in all cases, which corroborates that there was a big asymmetry in the categories. This is due to the prevalence phenomena discussed in Section 1 (first Kappa paradox).

Finally, to obtain a more approximate idea about the real level of agreement reached by the nine annotators, we report the values of the observed agreement

Table 5. Observed agreement for all annotation schemes and annotator types

		Observed agreement	Weighted observed agreement
Unordered	Total	0.8836	0.9117
	Andalusian	0.8950	0.9197
	Non-andalusian	0.8767	0.9050
Ordered	Total	0.8429	0.8800
	Andalusian	0.8761	0.9049
	Non-andalusian	0.8578	0.895

in Table 5, which has been used along with Kappa by other authors in different areas of study, e.g. [Ang et al., 2002] [Forbes-Riley and Litman, 2004]. As can be observed in the table, in all cases the observed agreement was above 0.85. This measure does not take into account the high probability of agreeing by chance in the neutral category, and thus values were not higher for the annotators not used to the Andalusian accent in the ordered case.

5 Conclusions

We have shown that when evaluating the reliability of the annotation of non-acted emotions corpora, very low Kappas can be obtained (Table 2) which are usually much lower than the agreement values observed (Table 5). This is due to the unavoidable natural skewness of such corpora, in which there is usually a noticeable prevalence of the neutral categories. We have discussed other coefficients that can be reported along with Kappa, such as observed agreement and minimal, maximal and normal Kappa values, in order to obtain meaningful interpretations about the reliability of the annotations.

Additionally, our experimental results show that employing contextual information about the users' speaking style and the history of the interaction allowed the annotation of more non-neutral emotions in our speech database. Unfortunately, this translates into lower Kappa coefficients as most of the agreements occur for neutrals. However, although the Kappa value and the observed agreement percentages were lower when using contextual information, we found that it can be useful to obtain results which are closer to the maximum Kappa values achievable. Besides, as shown in Table 5, giving a weight to the different disagreement types considerably incremented the observed agreement between annotators. We have presented a method to compute distances between such disagreements.

Our results indicate that multiple annotators should be used for annotating natural emotions to obtain reliable emotional corpora. One possible way to overcome the problem of high chance agreements, is maximizing the observed agreement. For example, [Litman and Forbes-Riley, 2006] propose the usage of "consensus labelling", i.e. to reach a consensus between annotators until a 100% observed agreement is obtained.

References

[Ang et al., 2002] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. of Interspeech'02 - ICSLP, Denver, USA, pp. 2037–2040 (2002)

[Artstein and Poesio, 2005] Artstein, R., Poesio, M.: $kappa_3$ = alpha (or beta). Technical report, University of Essex (2005)

[Callejas and López-Cózar, 2005] Callejas, Z., López-Cózar, R.: Implementing modular dialogue systems: a case study. In: Proc. of ASIDE 2005 (2005)

[Davies and Fleiss, 1982] Davies, M., Fleiss, J.L.: Measuring agreement for multinomial data. Biometrics 38(4), 1047–1051 (1982)

[Devillers et al., 2005] Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18(4), 407–422 (2005)

[Dunn, 1989] Dunn, G.: Design and analysis of reliability studies: the statistical evaluation of measurement errors. Edward Arnold (1989)

[Feinstein and Cicchetti, 1990] Feinstein, A.R., Cicchetti, D.V.: High agreement but low Kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 43(6), 543–549 (1990)

[Fleiss, 1971] Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)

[Forbes-Riley and Litman, 2004] Forbes-Riley, K., Litman, D.J.: Predicting emotion in spoken dialogue from multiple knowledge sources. In: Proc. of HLT-NAACL 2004, pp. 201–208 (2004)

[Krippendorff, 2003] Krippendorff, K.: Content Analysis: An Introduction to its Methodology. Sage Publications, Inc., Thousand Oaks (2003)

[Landis and Koch, 1977] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977)

[Lantz and Nebenzahl, 1996] Lantz, C.A., Nebenzahl, E.: Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. Journal of Clinical Epidemiology 49(4), 431–434 (1996)

[Lee and Narayanan, 2005] Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13(2), 293–303 (2005)

[Litman and Forbes-Riley, 2006] Litman, D.J., Forbes-Riley, K.: Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Communication 48(5), 559–590 (2006)

[Morrison et al., 2007] Morrison, D., Wang, R., Silva, L.C.D.: Ensemble methods for spoken emotion recognition in call-centers. Speech Communication 49, 98–112 (2007)

[Plutchik, 1980] Plutchik, R.: EMOTION: A psychoevolutionary synthesis. Harper and Row publishers (1980)

[Russell, 1980] Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)

[Shafran et al., 2003] Shafran, I., Riley, M., Mohri, M.: Voice signatures. In: Proc. of IEEE ASRU 2003 Workhop, pp. 31–36 (2003)

[Vidrascu and Devillers, 2005] Vidrascu, L., Devillers, L.: Real-Life Emotion Representation and Detection in Call Centers Data. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 739–746. Springer, Heidelberg (2005)