INTERSPEECH 2009 BRIGHTON

Improving acceptability assessment for the labelling of affective speech corpora

Zoraida Callejas, Ramón López-Cózar

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

Abstract

In this paper we study how to address the assessment of affective speech corpora. We propose the use of several coefficients and provide guidelines to obtain a more complete background about the quality of their annotation. This proposal has been evaluated employing a corpus of non-acted emotions gathered from spontaneous interactions of users with a spoken dialogue system. The results show that, due to the nature of non-acted emotional corpora, traditional interpretations would in most cases consider the annotation of these corpora unacceptable even with very high inter-annotator agreement. Our proposal provides a basis to argue their acceptability by supplying a more fine-grained vision of their quality.

Index Terms: affective corpora, non-acted emotions, inter-annotator agreement

10.21437/Interspeech.2009-731

1. Introduction

Non-acted emotional speech corpora gather the most realistic behaviour of speakers, but require an interpretation of the emotion conveyed in each recording. A lot of effort is necessary for the annotation of these corpora due to two main reasons. Firstly, because they are inherently skewed, being neutral states more frequent than emotional behaviours. Secondly, because emotions are more subtle than in the case of acted corpora, and thus it is more probable that two human labellers would not choose the same emotion category for the same utterance.

In this paper, we suggest to use different coefficients to measure the reliability of the annotations. Concretely, we describe how to calculate the complexity of the annotation task using entropy calculations. Also, we explain how to measure inter-annotator agreements by using several kappa coefficients.

High entropy values and low kappa coefficients are obtained in corpora in which approximately the same amount of utterances for each emotion (including neutral) are considered, as it is difficult to discern between spontaneous emotions. In corpora gathered from real interactions between users and a spoken dialogue system, as it is our case, the neutral category is highly predominant, which translates into low kappa and entropy values. Thus, traditional interpretations of these measures based on rules-of-thumb are not trustworthy for any of these corpora, as they would indicate that these are not reliable (high entropy and low Kappa), or even produce ambiguous assessments (e.g. low Kappa and low entropy).

To obtain a reliable interpretation of the agreement coefficients, we also suggest to report several sources of additional information. These are to study observed agreement and sources of disagreement both globally and pair-wise (for every two annotators), as well as to provide contextual values to interpret the results obtained, such as minimum, normal and maximum kappa values, and maximum entropy.

The rest of the paper is organized as follows. Section 2 presents the measures suggested for assessing the annotation. Section 3 proposes guidelines for creating a better background for interpreting the values of the described measures. Section 4 describes the corpus employed in our experiments along with the procedure followed to annotate it with emotional categories; and presents the results obtained when using our proposal to assess its reliability. Finally, Section 5 presents the conclusions.

2. Assessing emotional annotation

2.1. Task complexity

As stated in [1], annotators usually do not find a common label for a given utterance when tackling non-acted emotional corpora. This is not a problem of unreliability of the annotation, but is caused by the inherent difficulty of annotating these corpora. Thus, low agreement rates must not be the only source of information for assessing their reliability. This is the reason why [1] proposes to use entropy as an additional measure to compute how well human and automatic classifiers perform in the annotation task.

Entropy provides a quantitative measure of the complexity of the annotation. If annotators completely agree, there is a very low entropy, which might indicate that the emotion categories are clearly distinguishable. However, if they generally disagree, a high entropy is obtained, which gives a clue about the difficulty of the annotation (i.e. it is not straightforward for the labellers to decide the emotion category to assign to each utterance).

Following the proposal in [1], entropy can be calculated as shown in Equation 1:

$$H = \frac{1}{UA} \sum_{u=1}^{U} \sum_{a=1}^{A} H(a,u) \quad (1)$$

where U is the number of utterances to be annotated, A is the number of annotators, and $H(a,u)$ is the entropy for the utterance 'u' and annotator 'a'. $H(a,u)$ can be calculated as follows:

$$H(a,u) = -\sum_{e=1}^{E} l_e(\bar{a},u) Log_2(l_e(\bar{a},u)) \quad (2)$$

where E is the number of emotions, and $l_e(\bar{a},u)$, which is calculated following Equation 3, considers the new annotator differently from the reference annotators. To do this, it is possible to compute l_{ref} leaving annotator a out ($\bar{a}$), and compute l_{dec} considering all annotators [1]. Both measures represent the number of times the emotional category 'e' has been chosen for the utterance 'u' averaged by the number of annotators.

$$l_e(\bar{a}, u) = \frac{l_{ref_e}(\bar{a}, u) + l_{dec_e}(u)}{2} \quad (3)$$

2.2. Inter-annotator agreement

Several Kappa coefficients can be used to study the degree of inter-annotator agreement. They 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 ($1 - P_c$).

The simplest Kappa coefficient used was proposed by [2], which we have noted as multi-π following [3]. This notation will be employed for the remaining coefficients. The multi-π observed agreement (P_o) is computed as the number of cases in which two different annotators agree to label a particular utterance with the same emotion category:

$$P_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{e=1}^{E} n_{ue}(n_{ue} - 1) \quad (4)$$

In Equation 4, n_{ue} represents the number of times the utterance 'u' is annotated with the emotion category 'e'.

Fleiss [2] assumed that all the annotators share the same probability distribution. In our experiments, this means that the probability that an annotator labels an utterance 'u' with a particular emotion category 'e', can be computed as the overall probability of annotating 'u' as 'e'. This global probability is employed for computing agreement by chance as shown in Equation 5.

$$P_c^{\pi} = \sum_{e=1}^{E} \left(\frac{1}{UA} n_e \right)^2 \quad (5)$$

The calculation of multi-π assumes that each annotator follows the same overall distribution of utterances into emotion categories, and thus does not cope with *annotator bias*. To include different annotating behaviours we propose to use multi-κ, as done by [4]. The multi-κ coefficient has the same observed agreement (Equation 4), but it includes a separate distribution for each annotator in the calculation of chance agreement.

$$P_c^{\kappa} = \frac{1}{\binom{A}{2}} \sum_{e=1}^{E} \sum_{j=1}^{A-1} \sum_{k=j+1}^{A} \frac{n_{a_j e}}{U} \frac{n_{a_k e}}{U} \quad (6)$$

Despite of including differences between annotators, multi-κ assigns the same importance to all disagreements. In practice, all disagreements are not equally probable and do not have the same impact on the quality of the annotation results. To take this information into account we propose using weighted Kappa coefficients, which emphasize disagreements instead of agreements. Their calculation is based on Equation 7:

$$\kappa_w = 1 - \frac{\overline{P}_o}{\overline{P}_c} \quad (7)$$

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

$$\overline{P}_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{ue_j} n_{ue_k} distance(e_j, e_k) \quad (8)$$

To calculate the distance between the emotional categories, we propose to arrange these within the bidimensional activation-evaluation space, in which they form a circular pattern [5]. This can be done by employing already established angular dispositions, such as the list of 40 emotions with their respective angles proposed by [6], so that the angular distance between the emotions can be calculated in degrees.

In order to optimize the results, we propose to choose always the smallest angle between the emotions being considered (x or 360-x). This way, the distance between every two angles is always between 0 and 180 degrees. For the calculation of the Kappa coefficients, distances can be converted into weights with values between 0 (0^o distance and thus no disagreement) and 1 (180^o distance and maximum disagreement).

Three weighted Kappa coefficients have been usually reported in the literature. The first one is α, proposed by [7]. The second and third are α' and β respectively, both proposed by [3]. All of them share the same observed disagreement calculation (Equation 7). Disagreement by chance for α and α' is calculated as:

$$\overline{P}_c^{\alpha} = \frac{1}{UA(UA-1)} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{e_j} n_{e_k} distance(e_j, e_k) \quad (9)$$

$$\overline{P}_c^{\alpha'} = \frac{1}{(UA)^2} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{e_j} n_{e_k} distance(e_j, e_k) \quad (10)$$

As can be observed in Equations 9 and 10, these coefficients do not consider annotator bias. This can be addressed by employing the β coefficient, with which also the observed behaviour of each annotator is taken into account:

$$\overline{P}_c^{\beta} = \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} \left[\frac{1}{U^2 \binom{A}{2}} \sum_{m=1}^{A-1} \sum_{n=m+1}^{A} X \right]$$

$$X = n_{a_m e_j} n_{a_n e_k} distance(e_j, e_k) \quad (11)$$

3. Proposal for interpreting reliability

When emotional corpora are gathered from spontaneous spoken interactions, the majority of utterances correspond to a neutral user state. This causes that, even with very high inter-annotator agreement, the value of kappa coefficients is low. The situation in which although having an almost identical number of agreements, the distribution of these across the different annotation categories deeply affects Kappa, is typically known as *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 Kappa is considerably decremented [8].

Traditionally, the interpretations of kappa are based on rules-of-thumb, which make a correspondence between intervals for Kappa values and interpretations of agreement [9]. Alternatively, other authors have established 0.65 as a threshold for acceptability of agreement results [7]. However, due to the effect of the first kappa paradox, using these alternatives for interpreting the reliability of non-acted emotional corpora, would lead in most cases to consider it as unreliable, even when there is a very high agreement between the labellers.

Thus, using a fixed benchmark of Kappa intervals does not provide enough information to make a justified interpretation of acceptability of the agreement results. In addition, the calculation of entropy with a skewed corpus would always lead to very positive interpretations as it is easy for the labellers to agree in the neutral category. Therefore, the interpretation of entropy would be just opposite to kappa interpretation, thus being either not informative or confusing.

In order to address these problems, we propose to provide additional information in two ways: studying closely the agreements between annotators and the sources of disagreement, and providing the Kappa and entropy coefficients context to achieve a better interpretation.

3.1. Disagreement between annotators

As kappa coefficients take into account agreement by chance, they have been preferred to calculating rates of observed agreement, information which is seldom reported as part of the reliability interpretation process. However, due to the difficulties of interpreting Kappa coefficients for non-acted emotional corpora, reporting observed agreement can be very valuable. Additionally, to obtain a more fine-grained basis for reliability interpretation, we propose to include not only general observed agreement, but also pair-wise agreement (agreement rates between every pair of annotators along all the utterances). To complete the information, also entropy values can be calculated for every annotator. To do this, we propose to use Equation 2 averaging over all the utterances and annotators considered without excluding the one under study when calculating l_{ref}.

3.2. Placing coefficients in context

In order to provide enough information to make a justified interpretation of acceptability, Kappa can be placed into context by computing *maximum*, *minimum* and *normal* values of Kappa, which can be done considering the observed agreement (P_o) as indicated in [10]. Given the same observed agreement, the possible values of Kappa can deeply vary from $Kappa_{min}$ to $Kappa_{max}$ depending on the balance 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 stated in [10], 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. Thus, reporting these values helps to better understand the complexity of the annotation task due to possible problems of inherent skewness, which is relevant in order to carry out correct interpretations of the Kappa and entropy values obtained.

The case of entropy is different in that it always has the same minimum value, which is 0 (when all annotators agree), and it is not so obvious how to calculate the maximum value if the annotation task does not have as many categories as labellers (in this case, the maximum entropy would indicate that each annotator assigned a different category for each utterance). In any case, an approximate computation of the maximum entropy can be supplied and brings important information to be taken into account for interpreting the entropy value.

4. Experiments

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. The corpus used for the experiments described in this paper is comprised of 85 dialogues of 60 different users interacting with the system [11]. The corpus contains 422 user turns, and has a duration of 150 minutes. Nine annotators labelled each utterance in the corpus with one of the following emotion categories: *angry*, *bored*, *doubtful* and *neutral*. The final emotion category assigned to the utterances was decided considering the majority opinion of the annotators. On average, more than 80% of the utterances were annotated as *neutral*.

To assess the reliability of the annotation of the corpus, we calculated the coefficients as described in Section 2 and obtained the results shown in Table 1. A traditional interpretation of these values would indicate high reliability if based on entropy and only fair agreement or even non-acceptable reliability with all the Kappa coefficients.

Table 1: Values of the coefficients

Entropy	0.318
multi-π	0.324
multi-κ	0.326
α	0.322
α'	0.322
β	0.324

In order to provide enough information to study the reliability of the annotation of the corpus and disambiguate the contradictory interpretations obtained, we followed the method proposed in Section 3.

4.1. Computation of disagreement between annotators

Firstly, we computed the observed agreement between the annotators and found that it was of 0.85. It is worth noting that the high difference between this value and the Kappa values (which are around 0.32), is due to the high probability of agreeing by chance. A high agreement by chance and low entropy indicates that it was easy for the annotators to agree in the same category. Thus, this result shows that there is a big imbalance of the corpus, which explains the low Kappas obtained. In other words, these low values are not due to a low reliability of the annotation.

Moreover, it can be observed in Table 1 that weighting disagreement (β and α vs. multi-κ) reduces Kappa, which means that the main sources of disagreements occur for the most distant categories. When adding information about pairwise agreement as suggested in Section 3.1, we corroborated that there were not many disagreements (indicated by the high

agreement rates), and that they did occur in most cases between neutral and non-neutral categories (highest distance), whereas few disagreements occurred between non-neutral categories. This is again a consequence of employing spontaneous emotions which are very subtle and thus difficult to distinguish from neutral states.

Additionally, the entropy values calculated for each labeller also indicate good agreement rates for every annotator as shown in Table 2.

Table 2: Entropy measures per annotator

Annotator	Entropy
0	0.317
1	0.317
2	0.318
3	0.318
4	0.318
5	0.317
6	0.320
7	0.318
8	0.317

4.2. Taking context into account

Secondly, as proposed in Section 3.2, we calculated the context for all the coefficients. In the case of Kappa coefficients (Table 3), our results corroborate that reporting Kappa values is more informative when they are put into context, as we obtain a valuable indication of possible imbalance that must be considered to come to 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 is a big asymmetry in the categories. This is again due to the prevalence phenomena discussed before (first Kappa paradox).

Table 3: Kappa minimal, observed, normal and maximal values

	multi-π	multi-κ	α	α'	β
κ_{min}	-0.086	-0.085	-0.064	-0.064	-0.064
κ_o	0.324	0.326	0.322	0.329	0.324
κ_{nor}	0.686	0.686	0.759	0.759	0.759
κ_{max}	0.693	0.693	0.763	0.763	0.763

In the case of entropy, as we considered nine annotators and four emotions, we established the worst case scenario depicted in Table 4. As can be observed, the worst case takes place when each category is chosen a minimum number of times in every utterance (i.e. there is maximum disagreement). In our case, the minimum number of repetitions is 2 for 3 of the categories (the emotional category is chosen by two annotators), and 3 for one category (as there is an odd number of annotators). The entropy value obtained was 0.988, which was far from the 0.318 observed in our corpus. When we computed the entropies per annotator, values ranged from 0.987 to 0.990, which was also much higher than the observed entropy values. These results were also due to the inherent predominance of the *neutral* category, and not due to an unacceptable annotation.

5. Conclusions

In this paper we have proposed the use of several coefficients and information sources to enhance the evaluation of the human labelling of affective speech corpora. The proposal has been empirically evaluated with a non-acted emotional corpus gathered from spontaneous phone calls to a spoken dialogue system. Experimental results highlight the difficulty of assessing reliability of such corpora. Traditional interpretations of the value of inter-agreement and complexity coefficients would consider the labelling of this corpus as unreliable. However, this conclusion would not be trustworthy as these interpretations do not take into account the effects of the inherent skewness and subtlety of spontaneous emotional corpora. Our method enhances the interpretation of the traditional results and provides a solid basis to better assess the quality of the annotations.

Table 4: Worst scenario for entropy calculation

Annotator	0	1	2	3	4	5	6	7	8
Emotion	A	B	C	D	A	B	C	D	A

6. Acknowledgements

This research has been funded by the project HADA TIN2007-64718 of the Spanish Ministry for Education and Science.

7. References

[1] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of all things the measure is man. automatic classification of emotions and inter-labeler consistency," in *Proceedings of ICASSP 2005*, Philadelphia, USA, 2005, pp. 317–320.

[2] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[3] R. Artstein and M. Poesio, "$kappa_3$ = alpha (or beta)," University of Essex, Tech. Rep., 2005.

[4] M. Davies and J. L. Fleiss, "Measuring agreement for multinomial data," *Biometrics*, vol. 38, no. 4, pp. 1047–1051, 1982.

[5] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

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

[7] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*. Sage Publications, Inc, 2003.

[8] A. R. Feinstein and D. V. Cicchetti, "High agreement but low Kappa: I. The problems of two paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 543–549, 1990.

[9] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[10] C. A. Lantz and E. Nebenzahl, "Behavior and interpretation of the κ statistic: Resolution of the two paradoxes," *Journal of Clinical Epidemiology*, vol. 49, no. 4, pp. 431–434, 1996.

[11] Z. Callejas and R. López-Cózar, "Relations between de-facto criteria in the evaluation of a spoken dialogue system," *Speech Communication*, vol. 50, no. 8-9, pp. 646–665, 2008.