Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

An Annotation Scheme for Sentiment in Behaviour Reports

Norton Trevisan Roman · Paul Piwek · Ariadne Maria Brito Rizzoni Carvalho

Received: date / Accepted: date

Abstract This paper proposes an annotation scheme for the notoriously difficult task of sentiment annotation. We show how the task can be made tractable by focusing on one of the many aspects of sentiment: sentiment as it is recorded in behaviour reports of people and their interactions. Together with a number of measures for supporting the reliable application of the scheme, this allows us to obtain sufficient to good agreement scores (in terms of Krippendorf's alpha) on three key dimensions: polarity, evaluated party and type of clause. Evaluation of the scheme is carried out through the annotation of a novel corpus of dialogue summaries by nine annotators. Results show that three out of five proposed dimensions are reliable. Our contribution to the field is threefold: (i) We present a reliable multi-dimensional annotation scheme for sentiment in behaviour reports; (ii) we describe an annotated corpus that was used for testing the reliability of the scheme and which will be available to the research

This research was sponsored by CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico – and CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Part of it was also supported by the EC Project NECA IST-2000-28580.

Norton Trevisan Roman
Schools of Arts, Sciences and Humanities – University of São Paulo. São Paulo, Brazil
Tel.: +55-11-3091-1008
E-mail: norton@usp.br

Paul Piwek
Centre for Research in Computing – The Open University. Milton Keynes, UK
Tel.: +44-1908-85-89-14
Fax: +44-1908-65-21-40
E-mail: p.piwek@open.ac.uk

Ariadne Maria Brito Rizzoni Carvalho
Institute of Computing – University of Campinas. Campinas, Brazil
Tel.: +55-19-3521-5864
Fax: +55-19-3521-5847
E-mail: ariadne@ic.unicamp.br

community; and (iii) our study provides further strong empirical evidence for the use of more than two annotators.

Keywords Corpus Annotation · Automatic Dialogue Summarization · Natural Language Processing

1 Introduction

Recent years have witnessed a growing interest in emotion and sentiment analysis by the computational linguistics community. Current research topics include identification and classification of sentiment in reviews of products (*e.g.* [22,21,5]), people (*e.g.* [27,49,18]) and open domains (*e.g.* [7,33,4]), determination of semantic orientation of words, whether within (*e.g.* [52]) or without context (*e.g.* [22,21]), and flame detection in e-mails (*e.g.* [48]). As in other areas of computational linguistics, much of the extant work relies on corpora with gold standard annotations for training and evaluation. Unfortunately, marking up texts for sentiment has proven to be an extremely difficult task. For example, [11] report inter-annotator reliability of sentiment classification of around 0.6 in terms of Krippendorf's α. With $\alpha \geq 0.8$ as good reliability and $0.67 \leq \alpha < 0.8$ allowing tentative conclusions [30,3], such levels of agreement are not very satisfactory.

One possible reason for such disappointing results is that sentiment in text is in fact a many-faceted phenomenon (as illustrated by the many different sources of sentiment, ranging from properties of objects to the appearance and behaviour of people). To address this issue, we have identified one, in our view, important class of sentiment reports and developed and evaluated an annotation scheme for this particular class of reports. The class of sentiment reports that we focus on concerns human behaviour reports. In other words, we deal with sentiment classification for utterances where a behaviour of or interaction between human agents is reported. Our aim is to show that given this specific notion of sentiment and a particular design of the annotation process, reasonable scores for reliability can be achieved.

To our knowledge, our focus on sentiment identification and classification in behaviour reports is novel. There are, however, good reasons for investigating behaviour reports. Within the area of automatic summarization there is an emerging area of dialogue summarization (*e.g.* [53,44,39]). Whereas most research in this area has focused on the content of the dialogue and, for example, what the interlocutors agreed to (*e.g.* [24,1]), there is a range of applications which would benefit from also being able to report the sentiment evaluations of conversational participants' behaviour and interaction. Such applications include summaries of conversational exchanges of customer services representatives to monitor call quality, but could also, for example, be used to provide better meeting summaries (conveying not only what has been agreed, but also any animosity or positive atmosphere). To build such summarization systems knowledge of how dialogue behaviours are reported in a dialogue summary is required. With no theoretical account available, a corpus-based approach

seems appropriate. This, however, presupposes annotated dialogue summaries and source dialogues.

In this paper, we aim to help fill in this gap. We proceed as follows. Firstly, in Section 2 we describe our method for collecting a corpus of human-authored dialogue summaries which are based on dialogues that systematically vary in terms of the behaviour of their participants. Next, Section 3 presents our multi-dimensional annotation scheme for sentiment in behaviour reports. Section 4 describes how the scheme has been used. This includes an account of the training that the annotators received and the procedure that they followed when applying the scheme. Section 5 assesses the scheme's reliability by measuring the amount of agreement between annotators. We present conclusions and avenues for further research in Section 6.

2 A Corpus of Human-authored Dialogue Summaries

Our aim was to collect a corpus of dialogue summaries, where summaries include not only information on what was said, but also evaluations of the content of what was said and how it was said. For this purpose we required, as a basis for the human-authored summaries, dialogues which involve behaviours that are amenable to evaluative reporting. In particular, we focused on (im)politeness of interlocutors. Though we considered using naturally-occurring dialogues, we eventually decided to work with machine-generated dialogues. The principal reason for this decision was the ability to systematically control for the dialogue participants' behaviours, an almost impossible goal to achieve with naturally-occurring dialogues.

We used the eShowroom NECA system [13], a system for automatically generating scripted dialogues between a virtual sales person and buyer, to create a basis of four sales dialogues with (i) two dialogues in which both participants act politely, and (ii) two dialogues in which one of the participants is impolite. We had 30 independent summarizers produce summaries for each of the 4 dialogues. Summarizers were asked to produce both a summary where there was no limit on the number of words that could be used and one with a size restriction (10% of the words of original dialogue).[1] Additionally, our group of summarisers was split up into three groups of ten: we asked one group to simply summarise the dialogue as if they were a neutral observer of the dialogue, whereas each of the other two groups was asked to summarise specifically from the perspective of one of the interlocutors. They were asked to adopt a specific point of view by pretending to have been one of the participants in the dialogue.

As a result, we obtained a total of 240 different dialogue summaries. Since source dialogues were in English and summarizers were native speakers of

[1] The requests to summarise unrestricted and restricted size summaries were separated in time by a couple of months, in order to reduce any experimental bias that might occur as a result of repeated runs on the same task.

Portuguese[2], they were allowed to produce their summaries in either language, resulting in 36 (15%) summaries in English and 204 (85%) in Portuguese. We refer the interested reader to [46] for more detail on the construction of the dialogue summary corpus, including transcripts of the four source dialogues.

3 The Multi-dimensional Annotation Scheme

In our scheme, the clause functions as the basic unit of annotation, defined following [36] as a unit consisting, as a minimum, of a verb and its complements. As a unit, clauses have a number of desirable properties for annotation, notably their independence of a specific linguistic theory [34] and their purely syntactic nature [30].

3.1 Dimensions

The annotation scheme we propose classifies every clause in a summary according to five distinct dimensions, shown in Figure 1. At the root of the scheme is a distinction between clauses that do and those that do not report behaviour or interaction with sentiment. The other dimensions classify clauses according to whether the reported sentiment was positive or negative, its intensity, who or what was reported and any inferential relation of the current report with other statements in the summary. In what follows, we define each of the dimensions in detail.

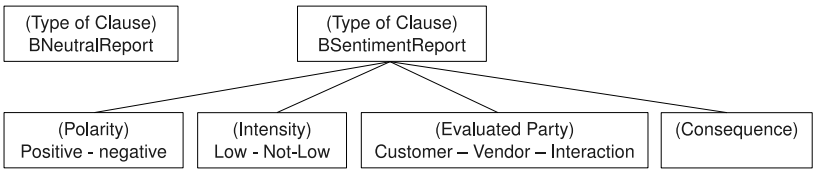

Fig. 1 Dimensions hierarchy in the annotation scheme [45].

3.1.1 Type of Clause

Following Batliner *et al.* [6], Pang [41] and Fisher [16], the basic distinction in this scheme rests on the dimension *Type of Clause*, between clauses classified as **BNeutralReport** and **BSentimentReport**, for Behaviour Neutral and Behaviour Sentiment report. In contrast with Batliner *et al.*, Pang and Fisher, the distinction underlying our notion of *Type of Clause* is more specific than the distinction between clauses that convey sentiment and those that do not.

[2] Although all of them were also fluent speakers of English.

Our distinction is grounded in what Mills [37] calls social politeness (or political behaviour), that is behaviour, humour, feelings and emotions concerning an interactional feature (akin to what Keenan, MacWhinney, and Mayhew [28] call the interactional content of the clause).

In particular, we propose to label a clause *BSentimentReport* only when it evaluates, either positively or negatively, the individual dialogue participants' behaviour or the interaction as a whole. This covers all the evaluation forms proposed by [35], to wit, affect (emotional responses), judgements (moral evaluations of behaviours) and appreciation (an aesthetic quality of semiotic processes and natural phenomena characterized as, for example, harmonious or elegant); applying them to the interaction and its participants. Within this context, *BNeutralReport* characterizes the complement of *BSentimentReport*. It applies to clauses with no evaluations of the interaction or its participants. Note that our notion of **BSentimentReport** is also more specific than emotions as classified by the appraisal theory of emotions by [40]. The latter covers reactions to events, agents and objects. In contrast, our scheme focuses on agents and events (in as far they are about the interaction between the agents). In particular, sentiment reports regarding objects (e.g. "The car was beautiful") are classified as *BNeutralReport*.

3.1.2 Polarity

For Polarity we follow [40] and [35] by taking it to range over two values only: **Positive** and **Negative**. These values must be applied only to *BSentimentReport* clauses (along the lines presented in [14,31,42,50], who apply such classification to clauses with some emotional/interactional content). According to this dimension, and following [32] and [35], a clause is classified as *Positive*, when the clause describes pleasant actions or feelings, or when it assesses them positively, like in "She was patient" and "The service was fine". A clause is classified as *Negative* when it describes actions or feelings that are unsatisfactory/inappropriate (and must be avoided) as in "He is rude" or "Tossed all her dissatisfaction on me". Our scheme differs from those proposed by [32] and [19,20], which allow for three polarity values (positive, negative and neutral). Neutral behaviour reports are assigned the label*BNeutralReport* (at the *Type of clause* level), where they are grouped together with sentiment reports that do not pertain to the interlocutors or their interaction.

3.1.3 Intensity

This dimension is intended to capture the degree of intensity or strength of the reported sentiment. There are many ways to construct an intensity scale. For example, [14] and [50] only allow for the values "low" (or "<NORM") and "high" (or ">NORM"). In contrast, [31], [42] and [35] allow for a continuum of values, ranging from "calm" to "excited". Our scale only has two values: low intensity and non-low intensity (i.e., normal to high). This approach was motivated by the aim to use the scheme to determine whether sentiment reports

are biased. A sign of this would be situations where reporters tone down the intensity of, for example, a negative report, if this suits their purposes (*cf.* [29, 17, 8, 26]).

3.1.4 Evaluated Party

Following [23], this dimension records the dialogue participant whose behaviour or emotion is reported. The identity of the participant may be explicit in the report, as in "the vendor treated me very well", or implicit, as in "the service was very good", in which case a service necessarily implies a server. In the context of our sales dialogues it takes one of three values: (a) **Vendor**, meaning that the evaluation concerned the vendor's emotions and behaviour; (b) **Client**, when it concerns the client's behaviour; and (c) **Interaction**, which must be used whenever the clause does not evaluate either of the dialogue participants individually but, instead, their interaction with each other.

The *Interaction* value means that the clause in question presents a summary of what happened, as in "what a horrid day". Clauses should only be classified as *Interaction* if they refer to the interaction between the participants. For example, if the fact that the day was horrid had something to do with the weather (as may be inferrable from the context), then "what a horrid day" should not be labelled *BSentimentReport* (and consequently evaluated party would not apply).

3.1.5 Consequence

Consequence should be applied to *BSentimentReport* clauses describing situations or feelings that were caused by something reported in another clause, as in "I was so badly served *that I lost my nerve*", where "I lost my nerve" describes a consequence of "I was so badly served". The motivation for this feature, beyond determining a cause, is to establish whether blame may have been transferred. It labels those clauses in which one party tries to justify his/her actions in terms of something else that happened. This dimension holds a link to the clause identified as describing the cause for the situation reported in the current clause, or "-" in case there is no such clause.

3.1.6 Multiple classifications

Since clauses can convey more than one evaluation, from the five dimensions described above, three can take multiple classifications. These are *Polarity*, *Intensity*, and *Evaluated Party*. Thus, in our scheme, clauses such as "I gently served the rude client" are classified as *BSentimentReport*, holding different evaluations about two participants (*Vendor* and *Client*), with opposite polarities (*Positive* and *Negative*, respectively) and with the same *Intensity*. The scheme was designed to independently keep track of the dialogue participants. As such, it is possible to determine which party was evaluated negatively and which positively (see Section 4 for more details on how this is done).

4 Testing the Scheme: Corpus Annotation

The test bed for our annotation scheme comprises 1773 clauses, from the 240 human generated summaries. The summaries were independently annotated with our scheme by nine volunteers (seven men and two women, all graduate students from the Institute of Computing at the University of Campinas, Brazil). As an additional measure to increase reliability [30, 3], and to make annotators more familiar both with the annotation scheme and with the computer program developed for the annotation task, annotators had to independently go through a training stage of about one and half hour, before doing the annotation.

During the training stage, they were given a description of the annotation scheme, along with a set of guidelines to help them understand what each category meant (*cf.* [10]). They were then asked to annotate a set of 18 summaries, with 128 clauses in total, constructed specifically for the training phase. These training summaries were written to provide a number of specially designed examples, making it easier for annotators to get to grips with the task. To further facilitate performing the task, we developed an annotation tool (Figure 2), specifically designed for this annotation effort. When using this tool, annotators would have a view that included the clause under consideration, its identification number, entire source summary (so they could put the clause into context), and the viewpoint under which it was written. They could also browse through the dataset, going back and forth, and revisiting already classified clauses, in case they were in doubt about the current annotation.

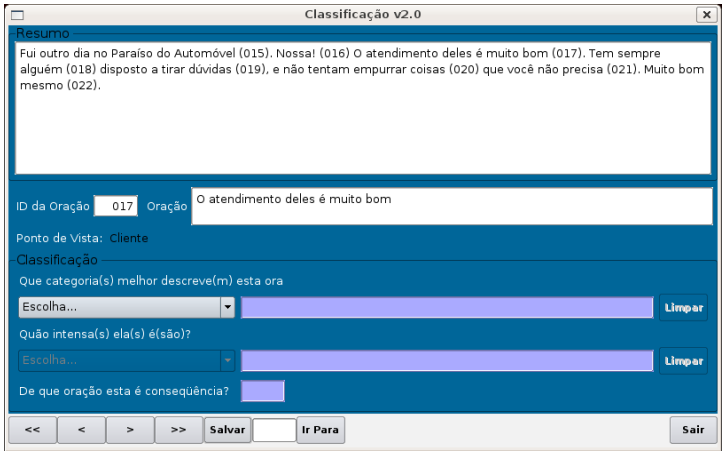

Fig. 2 Annotation program. "Resumo" shows the source summary, while "ID da oração" and "Oração show the clause's id and text, respectively. The point of view ("Ponto de vista") is then shown to the annotator (in this case, *client*), who must choose the proper labels to apply to the clause (all below "Classificação"). Buttons at the bottom are used for navigation purposes only.

To classify some clause with multiple labels, all annotators have to do is to select them, in order, in the program interface. The label will be shown in the box by the combo box holding the annotation categories. Whenever necessary, annotators can go back in their choice by pressing the "clean" ("limpar" at the interface) button.

Instead of separately presenting annotators with all possible values for *Type of Clause*, *Polarity* and *Evaluated Party*, we show them grouped together into a single overall category. Thus, instead of classifying some clause as, for example, <*Type of Clause*: BSentimentReport; *Polarity*: Positive; *Evaluated Party*: Vendor>, annotators are offered the "Positive Report about the Vendor" alternative (Figure 3). This procedure leaves annotators with a single choice to make, as opposed to three separate choices, without loss of generality. The aim behind this approach was to achieve better inter-annotator agreement. Evidence has emerged (subsequent to our study being carried out) that this approach may indeed improve agreement. In particular, [52] noticed that executing annotation tasks in a single step performed about as well or better than doing it in separate steps.[3]

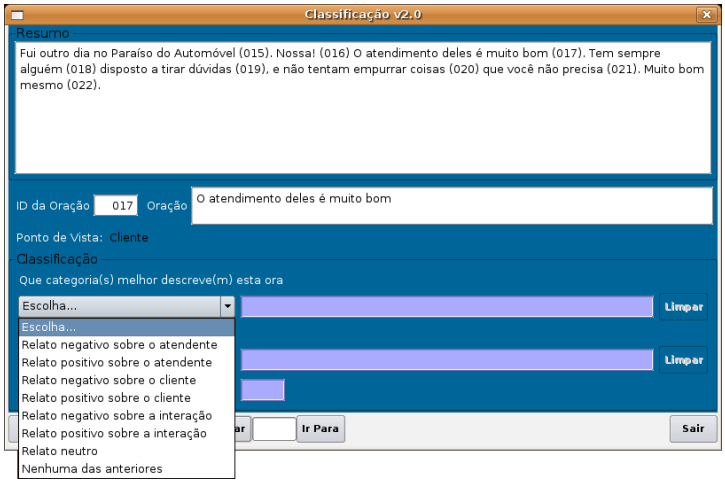

Fig. 3 The first three dimensions grouped together into a single one. At first, a neutral value ("Escolha...", *i.e.* "Choose...") is presented to the user. Remaining choices range from "Relato negativo sobre o atendente" ("negative report about the vendor") to "Relato positivo sobre a interação" ("Positive report about the interaction"). The last two values represent the *Neutral report* category ("Relato neutro") and *None of the above* ("Nenhuma das anteriores"), which is used to avoid biasing the results.

Regarding the program's interface, although there are seven possible combinations for the unified dimensions, nine are actually presented to users. The

[3] In this case, they compared a two-step approach, where annotators first decide on whether some word is neutral, and then classify it further in case it is not, to a one-step approach, where annotators do the entire classification in a single go.

two extra categories – "Choose ..." and "None of the above" – were added as a way to avoid biasing the results [43]. While "Choose ..." represents the preselected neutral alternative, indicating that no category was assigned to the clause, "None of the above" is meant to be used when the annotator does not agree with any of the existing categories, preventing him/her from picking a category s/he does not really agree with. As such, data would not suffer from the ambiguity brought by an empty classification, whereby one cannot tell whether the annotator simply skipped the clause, or did not agree with any of the proposed labels (a position that, if shared by the majority of them, would indicate an insufficiency in the annotation scheme). Note, however, that dimensions were unified at the interface level only. For example, when the user chooses "Negative Report about the Customer", s/he is actually choosing *Type of Clause = BSentimentReport*, *Polarity = Negative* and *Evaluated Party = Customer*, separately. Similarly, by choosing "Neutral Report", the user actually selects *Type of Clause = BNeutralReport*.

Upon finishing the training stage, annotators were given the annotation program and the set of instructions, and asked to complete the annotation whenever they found it most convenient (*cf.* Birnbaum 2004). Also, before giving the program away, we shuffled the summaries in the database, ensuring that each annotator had a different summary order to avoid any bias as a result of the order of data presentation. Resulting annotations, along with the annotators' characteristics (such as study level, gender, knowledge area where s/he took his/her degree etc), will be made available together with the corpus itself.[4]

4.1 Results

Figure 4 shows the distribution of labels for *Type of Clause*, *Polarity* and *Evaluated Party*. In this figure, we have followed [38] and used data from all nine annotators. The total number of data points is 1,773 clauses × 9 annotators = 15,957. However, with some clauses involved in multiple classifications, we ended up with the slightly higher number of 16,011 items. The dimensions *Intensity* and *Consequence* were left out of this and the next figure, given their low reliability (we will discuss this further in Section 5.2).

The figure clearly shows that *BNeutralReport* clauses outnumber *BSentimentReport* clauses. Also, there is a predominance of negative labels (The set of Negative Reports, represented in the second, fourth and sixth column), even though two of the source dialogues were more balanced. This is somewhat in line with psychological findings about a natural tendency, both in humans and animals, to give more importance to negative data [47].

When analysing the figures for the majority of annotators[5] (*cf.* Vieira and Poesio 2000), we come to a similar distribution, as shown in Figure 5. The

[4] http://www.each.usp.br/norton/corpora

[5] *I.e.*, when taking, for each clause, the label given by the majority of annotators, as opposed to summing up all the labels separately.

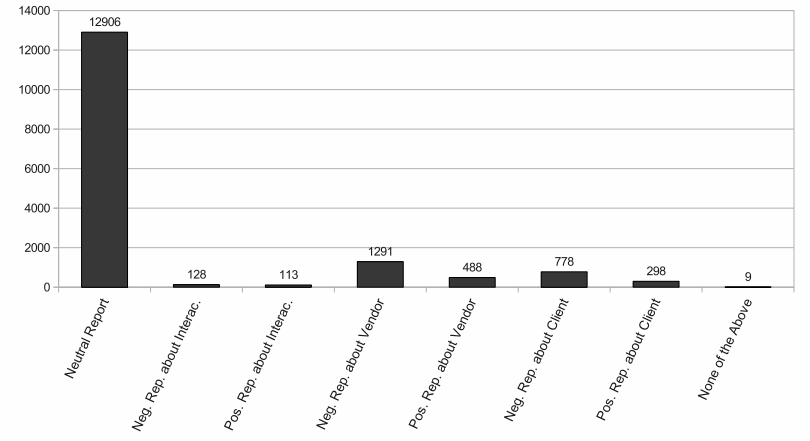

Fig. 4 Overall distribution of category labels, amongst 9 annotators.

only difference is that "None of the above" has dropped from nine to naught, meaning that, even though some annotators might have found the alternatives for classification inappropriate for the task, that view was not shared by the majority of them.

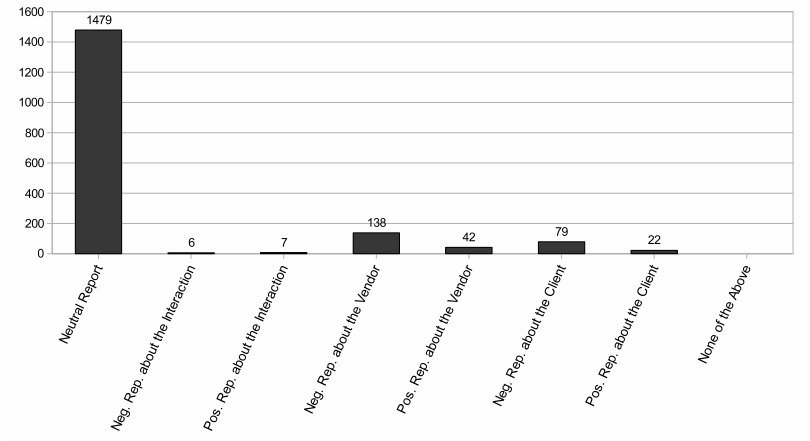

Fig. 5 Distribution of category labels, for the majority of annotators.

5 Evaluating the Scheme: Inter-Annotator Agreement

To test the reliability of our scheme, we measured its reproducibility, that is the extent to which annotators will produce the same classifications, working under varying conditions, at different times and locations [30]. As one of the

main goals of an annotation scheme is to allow other people to apply it to different data sets, the least we must guarantee is that its concepts are consistently understood by annotators [9,38]. This is the reason why we have chosen reproducibility to assess our scheme, by measuring the amount of agreement amongst annotators. Agreement and reliability[6], however, are not the same matter, since the former represents what is measured, whereas the later determines what we wish to infer from this measurement [30]. Still, if we consider the whole process of data codification as a mapping between units of analysis and a set of categories, the higher the inter-annotator agreement, when applying this mapping, the more we can trust it is correct [12].

In this research, we have elected Krippendorff's α as our coefficient of agreement. This choice was guided by two important features presented by α, to wit, (i) it accounts for the amount of agreement that is expected by chance; (ii) it calculates expected agreement by looking at the overall distribution of annotations, as opposed to inspecting distribution individually by annotator [3]; and (iii) it is not influenced by systematically biased distributions (*cf.* [15,30]), *i.e.*, distributions slanted towards some of the categories.

With α, agreement is measured by taking the proportion of the difference between the observed disagreement and the disagreement expected by chance, related to the disagreement expected by chance, as follows [30]

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

where D_o stands for the number of observed disagreements and D_e represents the number of expected disagreements by chance.

Values for α range from -1 to 1, where 1 means total agreement and 0 implies that analysed data cannot be told apart from random events. Negative α values may arise from two sources [30]: sampling errors, as a consequence of using small datasets; and systematic disagreement, where observers "agree on disagreeing", that is when they systematically choose opposite labels, due to opposite interpretation of the annotation instructions, for instance.

5.1 Agreement and Sentiment Annotation

Although the major goal of a data annotation system is to obtain values for a coefficient of agreement lying near unity, that is, near unanimity, the annotation task results depend to a great extent both on the personality and mood of annotators [2]. Thus, even with totally unambiguous classification schemes, we can expect this ideal to be not always achieved. It becomes, however, even harder if we consider the natural subjectivity of classification schemes that deal with sentiment/emotion features [50]. The lack of clear definitions of emotion/sentiment terms can lead to low values of agreement [2]. Moreover,

[6] The characteristic whereby data "remain constant through variations in the measuring process"[25, pages 83–84] (as cited in Krippendorff 2004, page 211)

when, as in our case, sentiment is related with evaluative judgement of appropriate behaviour, there is the problem that people do not always agree on what counts as appropriate behaviour in a given situation [51].

In order to ameliorate the effects of subjectivity in data annotation, we have reduced the number of categories, whenever possible, to a minimum (see, for example, dimension *Intensity*, which allows for two categories only). This approach was inspired by an experiment carried out by [11], in which it was observed that adding a single category to a subjective dimension dropped agreement from $\alpha = 0.52$ to $\alpha = 0.37$. Also, as we mentioned in Section 4, and as an attempt to reduce the cognitive load on annotators, we have combined the dimensions *Type of Clause*, *Polarity* and *Evaluated Party*. With this additional measure, we aimed to make the annotation process as intuitive and natural for annotators as possible. We reasoned that giving an overall description of a clause would be more intuitive and natural than considering each of the dimensions and categories in isolation.

5.2 Results

Table 1 shows the results for Krippendorff's α, in its version for nominal categories, many observers and allowing for missing values (see [30, pp. 230]), for all nine annotators. Assuming $\alpha \geq 0.8$ as good reliability, with $0.67 \leq \alpha < 0.8$ allowing tentative conclusions [30,3], only *Polarity* turns out to be a reliable dimension, with *Type of Clause* and *Evaluated Party* permitting only tentative conclusions. Although these values are modest, they are very encouraging when compared to existing results on emotion classification, e.g., [11] report $\alpha = 0.6$ as their highest value. Also, the existence of a "Neutral" category which, however necessary, has been described as a common source of confusion [2], just makes it harder to expect a high agreement on a dimension such as *Type of Clause*.

Table 1 Alpha values for the annotation [45].

Dimension	α
Polarity	0.843
Evaluated Party	0.783
Type of Clause	0.674
Intensity	0.212
Consequence	0.085
Consequence$_g$	0.175

With $\alpha < 0.67$, *Intensity* and *Consequence* turned out to be unreliable. Regarding *Consequence*, the results show that there is very low agreement on whether a clause signifies the consequence of an event reported in another clause ($\alpha = 0.175$, at Consequence$_g$), and even less whether specific clause represents a cause ($\alpha = 0.085$). Actually, the very notion of cause/consequence

between clauses involving sentiment in behaviour reports seems to be highly subjective.

As for *Intensity*, its low score may be connected to the observation of [40, Page 34] that "[emotions] vary a great deal in intensity both within and between people. [...] that the intensity of emotions is influenced by a number of variables", which makes it improbable to achieve high agreement on this dimension. Finally, another possible reason for *Intensity* and *Consequence* scoring so low[7] might be that most of the annotated items fall under one single category (in this case, Neutral Report, as shown on Figures 4 and 5). This prevalence problem (see [15]) increases the expected agreement by chance, making it harder to get a high figure for this coefficient of agreement [3].

One further question raised in the related literature deals with the number of annotators necessary to achieve good confidence that agreement does imply reliability. Although it is suggested that one should have more than two annotators, in order to reduce the odds of deviant readings of the annotation guidelines [3], there is, to our knowledge, no empirical evidence to support this statement. To help answer this question, we have analysed agreement scores for each pair of annotators separately, so as to determine the best and worst agreement when taking annotators pairwise. Table 2 shows the results.

Table 2 Agreement values for pairs of annotators.

Dimension	*Highest* α	*Lowest* α	*Highest* − *Lowest*
Polarity	0.957	0.726	0.231
Evaluated Party	0.922	0.697	0.225
Type of Clause	0.758	0.578	0.180
Intensity	0.528	-0.428	0.953
Consequence	0.270	-0.032	0.302

In this table, the difference between the highest and lowest scores varies considerably, ranging from 0.180, for *Type of clause*, to 0.953 for *Intensity*. It seems that the lower the overall agreement (see Table 1), the higher the difference between pairs. Interestingly, two annotators, within the *Intensity* dimension, delivered $\alpha = -0.428$, that is they not only disagreed, but appear to have done so systematically.

Regardless of the dimension, the overall α value is substantially higher than that of the *lowest* pairwise agreement score. Similarly, the overall α value is substantially lower than that of the *highest* pairwise agreement score. This suggests that a good number of annotators is required to make sure that no undue weight is given to annotators that have slightly deviant interpretations of the annotation guidelines.

Take, for example, *Evaluated Party*. If we were to rely only on two annotators, and these happened to be the ones with the highest agreement,

[7] And which may also have reduced the other dimensions' reliability to the same extent, but not enough to render them unreliable.

this dimension would come out as reliable, whereas our overall score for nine annotators indicates that we should go no further than making tentative conclusions. At the other end, *Type of Clause* could come out as unreliable, if we were to use only the pair with the lowest agreement; similarly *Polarity* – the only reliable dimension – could come out as only allowing tentative conclusions if we depend on only two annotators.

6 Conclusion

In this article we introduced an annotated corpus of 240 human produced summaries, along with its corresponding multidimensional annotation scheme. Primarily designed for emotional and behavioural assessment in dialogue summaries, this scheme can serve a variety of purposes, from identifying polite/impolite behaviour to detecting bias in sentiment behaviour reports (*cf.* [46]).

To test the soundness both of the developed scheme and its corresponding annotation guidelines, we carried out reliability studies with nine independent annotators. Results show that out of the five original dimensions, three were sufficiently reliable (*i.e.* $\alpha \geq 0.67$). These were *Type of Clause*, *Evaluated Party* and *Polarity*. For the two remaining dimensions (*Intensity* and *Consequence*) reliability could not be established. By determining the highest and lowest reliability amongst subsets of two to eight annotators, we presented new empirical evidence for the use of more than two annotators.

Finally, all nine versions of the annotated corpus will be made available to the community, so that other researchers can use the data. As for avenues for further research, it would be useful to identify a way to reliably measure intensity amongst reports, since this is a feature that plays a prominent role in many theories of emotion and evaluation (*e.g.* [40]).

References

1. Alexandersson, J., Poller, P., Kipp, M., Engel, R.: Multilingual summary generation in a speech-to-speech translation system for multilingual dialogues. In: Proceedings of the First International Conference on Natural Language Generation, pp. 148–155. Mitzpe Ramon, Israel (2000)
2. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: Machine learning for text-based emotion prediction. In: Proceedings of HLT/EMNLP 2005. Vancouver, Canada (2005)
3. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008)
4. Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., Martínez-Barco, P.: Summarizing threads in blogs using opinion polarity. In: Proceedings of the Events in Emerging Text Types Workshop of the RANLP. Borovets, Bulgaria (2009)
5. Balahur, A., Montoyo, A.: Determining the semantic orientation of opinions on products - a comparative analysis. Procesamiento del lenguaje natural (41), 201–208 (2008). ISSN 1135-5948
6. Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K.: Foundations of Speech-to-Speech Translation, chap. The Recognition of Emotion, pp. 122–130. Springer (2000)

7. Beineke, P., Hastie, T., Manning, C., Vaithyanathan, S.: An exploration of sentiment summarization. In: AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications. Stanford, USA (2004). Technical Report SS-04-07
8. Blackwood, N., Bentall, R., Ffytche, D., Simmons, A., Murray, R., Howard, R.: Self-responsibility and the self-serving bias: an FMRI investigation of causal attributions. Neuroimage **20**(2), 1076–1085 (2003)
9. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics **22**(2), 249–254 (1996)
10. Cohn, T., Callison-Burch, C., Lapata, M.: Constructing corpora for the development and evaluation of paraphrase systems. Computational Linguistics **34**(4), 597–914 (2008)
11. Craggs, R., Wood, M.M.: A two dimensional annotation scheme for emotion in dialogue. In: AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications. Stanford, USA (2004). Technical Report SS-04-07
12. Craggs, R., Wood, M.M.: Evaluating discourse and dialogue coding schemes. Computational Linguistics **31**(3), 189–295 (2005)
13. van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schröder, M., Baumann, S.: Fully generated scripted dialogue for embodied agents. Artificial Intelligence **172**(10), Elsevier (2008)
14. Dyer, M.: The role of affect in narratives. Cognitive Science **7**(3), 211–242 (1983)
15. Eugenio, B.D., Glass, M.: The kappa statistic: A second look. Computational Linguistics **30**(1), 95–101 (2004)
16. Fischer, K.: Annotating emotional language data. Tech. Rep. 236, Verbmobil Project (1999)
17. Higgins, N.C., Bhatt, G.: Culture moderates the self-serving bias: Etic and emic features of causal attributions in India and in Canada. Social Behavior and Personality **29**(1), 49–62 (2001)
18. Hijikata, Y., Ohno, H., Kusumura, Y., Nishida, S.: Social summarization of text feedback for online auctions and interactive presentation of the summary. Knowledge-Based Systems **20**(6), 527–541 (2007)
19. Hovy, E.: Generating Natural Language Under Pragmatic Constraints. Lawrence Erlbaum Associates (1988)
20. Hovy, E.: Pragmatics and natural language generation. artificial intelligence. Artificial Intelligence **43**(2), 153–198 (1990)
21. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. Seattle USA (2004)
22. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004), pp. 755–760. San Jose, USA (2004)
23. Hunston, S.: Evaluation in Text: Authorial Stance and the Construction of Discourse, chap. Evaluation and the Planes of Discourse: Status and Value in Persuasive Texts, pp. 176–207. Oxford University Press (1999)
24. Kameyama, M., Kawai, G., Arima, I.: A real-time system for summarizing human-human spontaneous spoken dialogues. In: Proceedings of the 4th International Conference on Spoken Language (ICSLP 96), vol. 2, pp. 681–684. Philadelphia, USA (1996)
25. Kaplan, A., Goldsen, J.M.: The reliability of content analysis categories. In: H.D. Lasswell, N. Leites (eds.) Language of politics: Studies in quantitative semantics, pp. 83–112. Cambridge: MIT Press (1965)
26. Kaplan, T., Ruffle, B.: The self-serving bias and beliefs about rationality. Economic Inquiry **42**(2), 237–246 (2004)
27. Katagiri, Y., Takahashi, T.: Social summarization for semantic society. In: JSAI2003 Workshop "From Semantic Web to Semantic World" (2003)
28. Keenan, J., MacWhinney, B., Mayhew, D.: Pragmatics in memory: A study of natural conversation. Journal of Verbal Learning and Verbal Behavior **16**(5), 549–560 (1977)
29. Knee, C.R., Zuckerman, M.: Causality orientations and the disappearance of the self-serving bias. Journal of Research in Personality **30**(1), 76–87 (1996)
30. Krippendorff, K.: Content Analysis: An Introduction to its Methodology, 2nd edn. SAGE (2004)

31. Lang, P.: The emotion probe: Studies of motivation and attention. American Psychologist **50**(5), 372–385 (1995)
32. Lehnert, W.: Plot units: A narrative summarization strategy. In: W.G. Lehnert, M.H. Ringle (eds.) Strategies for Natural Language Processing, pp. 375–412. Erlbaum (1982)
33. Lloret, E., Balahur, A., Palomar, M., Montoyo, A.: Towards building a competitive opinion summarization system: challenges and keys. In: Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium, pp. 72–77. Boulder, USA (2009)
34. Mann, W., Thompson, S.: Rhetorical structure theory: Toward a functional theory of text organization. Text **8**(3), 243–281 (1988)
35. Martin, J.R.: Evaluation in Text: Authorial Stance and the Construction of Discourse, chap. Beyond Exchange: Appraisal Systems in English, pp. 142–175. Oxford University Press (1999)
36. Miller, J.: An Introduction to English Syntax. Edinburgh University Press Ltd, Edinburgh, Scotland (2002). ISBN 0 7486 1254 8
37. Mills, S.: Gender and Politeness. Cambridge University Press (2003)
38. Moens, S.T.M.: Summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics **28**(4), 409–445 (2002)
39. Murray, G., Renals, S., Carletta, J.: Extractive summarization of meeting recordings. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech'2005). Lisbon, Portugal (2005)
40. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press (1988)
41. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL, pp. 271–278 (2004)
42. Picard, R.: Affective computing. Tech. Rep. 321, MIT Media Laboratory, Perceptual Computing Section, Cambridge, USA (1995). URL citeseer.ist.psu.edu/picard95affective.html
43. Reips, U.D.: Internet-based psychological experimenting: Five dos and five don'ts. Social Science Computer Review **20**(3), 241–249 (2002)
44. Reithinger, N., Kipp, M., Engel, R., Alexandersson, J.: Summarizing multilingual spoken negotiation dialogues. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'2000), pp. 310–317. Hong Kong, China (2000)
45. Roman, N.T., Carvalho, A.M.B.R.: A multi-dimensional annotation scheme for behaviour in dialogues. In: A. Kuri-Morales, G.R. Simari (eds.) Proceedings of the 12th Ibero-American Conference on Artificial Intelligence (IBERAMIA 2010), *Advances in Artificial Intelligence*, vol. 6433, pp. 386–395. Springer, Baha Blanca, Argentina (2010). ISBN: 978-3-642-16951-9
46. Roman, N.T., Piwek, P., Carvalho, A.M.B.R.: Computing Attitude and Affect in Text: Theory and Applications, *The Information Retrieval Series*, vol. 20, 1st edn., chap. Politeness and Bias in Dialogue Summarization: Two Exploratory Studies, pp. 171–185. Springer, Dordrecht, The Netherlands (2006). ISBN: 978-1-4020-4026-9
47. Rozin, P., Royzman, E.: Negativity bias, negativity dominance and contagion. Personality and Social Psychology Review **5**(4), 296–320 (2001)
48. Spertus, E.: Smokey: Automatic recognition of hostile messages. In: Innovative Applications of Artificial Intelligence (IAAI 97), pp. 1058–1065 (1997)
49. Takahashi, T., Katagiri, Y.: Telmea2003: Social summarization in online communities. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI 03), pp. 928–929. Fort Lauderdale, USA (2003)
50. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems **21**(4), 315–346 (2003)
51. Watts, R.: Politeness. Cambridge University Press (2003)
52. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics **35**(3), 399–433 (2009)
53. Zechner, K., Waibel, A.: Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. In: Proceedings of COLING-2000, pp. 968–974. Saarbruecken, Germany (2000)