On the Contextual Analysis of Agreement Scores

Dennis Reidsma, Dirk Heylen, Rieks op den Akker
Human Media Interaction
University of Twente
E-mail: {dennisr,infrieks,heylen}@ewi.utwente.nl

Abstract

Annotators of multimodal corpora rely on a combination of audio and video features to assign labels to the events observed. The reliability of annotations may be influenced by the presences or absence of certain key features. For practical applications it can be useful to know what circumstances determined fluctuations in the interannotator agreement. In this paper we consider the case of annotations of addressing on the AMI corpus.

1. Introduction

To a large extent multimodal behaviour is a holistic phenomenon in the sense that the contribution of a specific behaviour to the meaning of an utterance needs to be decided upon in the context of other behaviours that coincide, precede or follow. A nod, for instance, may contribute in different ways when it is performed by someone speaking or listening, when it is accompanied by a smile, when it is a nod in a series of 3 or 5, etcetera. When we judge what is happening in conversational scenes, our judgements become more accurate when we know more about the context in which the actions have taken place. The record of gaze, eye-contact, speech, facial expressions, gestures, and the setting determine our interpretation of events and help to disambiguate otherwise ambiguous activities.

Annotators, who are requested to label certain communicative events, be it topic, focus of attention, addressing information or dialogue act get cues from both the audio and the video stream. Some cues are more important than others, some may be crucial for correct interpretation whereas others may become important only in particular cases. The reliability of annotations may crucially depend on the presence or absence of certain features. Also one annotator may be more sensitive to one cue rather than another. This means that the agreement between annotators may vary with particular variations in the input. Rather than relying simply on a single overall reliability score, it can be informative to know whether there are particular features that account for some of the disagreements. This may influence the choice of features to use for training machine learning algorithms.

The rest of the paper is organised as follows. First we introduce the AMI[1] project and corpus (Carletta, 2007). Then we summarize the role of addressee in interaction and its place in the AMI corpus. For the case of determining who is being addressed in the AMI data, we have looked at the reliability scores of the annotations under different circumstances. The rest of the paper discusses the results and some implications of that analysis.

2. The AMI Corpus

The AMI corpus that was used in this study consists of more than 100 hours of audio and video data of non-scripted, role played meetings (Carletta (2007). In a series of four meetings, a group of designers, marketing experts and project leaders (4 people each time) go through different phases of discussing the design of a new type of remote control. The data has been annotated on many levels. The addressee labels that are the subject of this paper are part of the dialogue act annotations (Jovanovič, 2007). For, more or less each dialogue act, annotators were instructed to indicate whether it was addressed to the group, to a specific individual. Annotators could also use the label unknown.

3. Addressee in Interaction

Addressing occurs in a variety of flavors, more explicitly or less so, verbally or non-verbally. Thus, deciding whether or not the speaker addresses one individual partner in particular can be far from trivial an exercise. In small group discussions, like those in the AMI meetings with 4 participants, most contributions are addressed to the whole group. But sometimes speakers direct themselves to one listener in particular. Group members bring in different expert knowledge and have different tasks in the design process. If someone says to a previous speaker *"can you clarify what you just said about ..."* it is clearly addressed to that previous speaker. This doesn't rule out that a non-addressed participant takes the next turn. But generally this will not happen in an unmarked way.

The basis of our concept of addressing originates from Goffman (1981). The addressee is the participant "oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants". Thus, according to Goffman, the addressee is the listener the speaker has selected because he expects a response from that listener. The addressee coincides with the one the speaker has selected to take the next turn. But addressing an individual does not always imply turn-giving. For example, a speaker can invite one of the listeners to give feedback (either verbally, or non-verbal by eye-gaze) when he thinks that is required, but continue speaking.

Lerner distinguished explicit addressing and tacit addressing. To characterize the latter he writes: "When the requirements for responding to a sequence-initiating action limit eligible responders to a single participant, then that participant has been tacitly selected as next speaker. Tacit addressing is dependent on the situation and content." (Lerner, 2003). An example from our corpus is when a presenter says *"Next slide please"* during his

[1] http://www.amiproject.org.

presentation, a request that is clearly addressed to the one who operates the laptop.
Explicit addressing is performed by the use of vocatives (*"John, what do you think?"*) or, when the addressee's attention need not be called, by a deictic personal pronoun: *"What do you think?"*. There is one form of address that always has the property of indicating addressing, but that does not itself uniquely specify who is being addressed: the recipient reference term "you" (Lerner, 2003). The use of "you" as a form of person reference separates the action of "addressing a recipient" from the designation of just who is being addressed. In interactional terms, then, "you" might be termed a recipient indicator, but not a recipient designator. As such, it might be thought of as an incomplete form of address (Lerner, 2003).

Inherent Ambiguity in Addressing

At a party the host asks Ben - Ben's wife at his side - whether he wants another drink. Ben answers "No, thanks, it was an enjoyable evening, but we should go now," gazing at his wife while uttering the final excuse to his host. What is an excuse for the host is an urgent request addressed to his wife. The example shows that the same words can simultaneously express different speaker intentions directed to different addressees. The AMI annotation scheme was not devised to handle these cases. In the corpus we hardly see cases where addressing is a problem for the participants themselves. Only in a few instances, for example when the speaker uses a wrong address term, or when his utterances containing "you" is not supported by eye gaze to his intended addressee, confusion with respect to the intended addressee occurs for the participants involved in the interaction (see Op den Akker and Theune 2008 for more examples).

4. Addressee Annotation in AMI

Addressing information is part of the Dialogue Act Annotations in the AMI meeting corpus. The AMI dialogue act scheme distinguishes between 16 labels. Some of these labels are not really referring to a speech act as such but mark the special status of the utterance as a *stall*, *fragment*, *backchannel*, or *other*. Excepting these 'non-real' dialog act types, the annotators have indicated for all dialog acts of the remaining 11 types who was being addressed by the speaker: the group or a particular individual. We used meeting IS1003d of the AMI corpus which was annotated by four different annotators. Table 1 shows the confusion matrix for the full set of addressee labels for all 'agreed segments' of two of the annotators (i.e. segments where the annotators agreed on the start and end boundaries).
We ran a series of pairwise agreement analyses for each pair of annotators on the addressee labels assigned to dialogue act segments (i.e. excluding the 'non-real' dialog act types). The agreement is expressed using Krippendorff's alpha (1980). In the following sections we discuss several cases comparing scores for different label sets or class maps and different conditions (contexts).

	A	B	C	D	G	U	Σ
A	46				26	2	74
B	1	25			12	1	39
C			38	1	10	1	50
D				63	16	4	83
G	7	5	9	10	155	5	191
U	16	1	4	4	15	2	42
Σ	70	31	51	78	234	15	479

Table 1: Confusion matrix for two annotators for addressee labels of agreed segments.

Besides annotation of Goffman's notion of addressing, the meetings in the AMI corpus were also annotated with visual Focus Of Attention (FOA), an important cue for addressing behavior (see Section 3). This annotation marks for every participant in the meeting at all times throughout the meeting whom or what he is looking at. The FOA annotation was done with a very high level of agreement at a very precision: changes are marked in the middle of eye movement between old and new target (Jovanovič, 2007).

5. 'Unknown Addressee'

Annotators indicated whether an utterance was addressed to a particular person or to the whole group (note that the AMI meetings are multi-party meetings involving four participants). The annotators also had the choice to use the label *unknown addressee* in case they could not decide who was being addressed.
One can imagine two possibilities for the subset of dialog acts annotated with the *unknown* addressee label. Firstly, annotation of addressee may be a task containing inherently ambiguous instances as discussed by Poesio and Artstein (2005), with the intended addressee of some utterances being ambiguous by design. Secondly, the use of the *unknown* addressee label may reflect more the attentiveness of the annotator or his certainty in his own judgement rather than inherent properties of certain dialog acts.
The difference between the two has clear consequences for machine learning applications of the addressee annotations. It might make sense to try and learn to classify dialog act instances that are *inherently ambiguous with respect to addressing* as such, but less so to train a classifier to emulate the uncertainty of the annotators.
It is not possible to determine solely from the instruction manual which of the two interpretations most accurately reflects the meaning of the *unknown* addressee label as it was applied in the AMI corpus. Inspection of the confusion matrices however suggests that the *unknown* label is about randomly confused with every other possible addressee label. This strongly hints at the second interpretation. This conclusion is also borne out by the alpha agreement score for addressee computed on all dialog act segments vs the alpha agreement on only those dialog act segments not annotated with this *unknown* label. Leaving out the unknown addressee cases shows consistent improvements on the alpha scores, not only for the overall data set reported in Table 2 but also for each and every contextual selection of the data set reported later in this paper.

	Inc. unknown	Excl. unknown
1 vs 2	0.57	0.67
3 vs 4	0.31	0.47
4 vs 2	0.50	0.63
3 vs 2	0.36	0.47
1 vs 4	0.46	0.59
3 vs 1	0.32	0.43

Table 2: Alpha agreement for all segments vs only segments not annotated with the *unknown* addressee label.

For machine learning this suggests that it is better not to try to learn this label. For training and testing the addressee one should probably ignore the unknown addressee segments. The rest of this paper therefore reports only on segments *not* annotated with the unknown label.

6. Group/Single vs Group/A/B/C/D

The second aspect of the annotated data that we investigated in more depth was the difference between dialog act segments annotated as being *group addressed* and segments annotated as being *single addressed*, i.e. addressed to one of the individual meeting participants *A*, *B*, *C*, or *D*. Informal inspection of the confusion matrices suggests that making the global distinction between *group* and *single* addressed utterances is a difficult task: there is a lot of confusion between the label *G* on one hand and *A*, *B*, *C* and *D* on the other hand. However, if annotators see an utterance as *single* addressed they subsequently do not have much trouble determining *who* of the single participants was addressed: there is much less confusion between the 'single' labels *A*, *B*, *C* and *D*.

This is made more concrete by calculating alpha agreement for a class mapping of the addressee annotation in which the 'single' labels *A*, *B*, *C* and *D* are all mapped onto the label *S*. Table 3 shows pairwise alpha agreement for this class mapping, beside the values for the normal label set (excluding all segments annotated with the *unknown* addressee label, as described in Section 5). The consistent differences between the two columns make it is clear that agreement on *who* of the participants was addressed individually is a major factor in the overall agreement.

	Normal label set	Class map (A,B,C,D) => S
1 vs 2	0.67	0.55
3 vs 4	0.47	0.37
4 vs 2	0.63	0.52
3 vs 2	0.47	0.37
1 vs 4	0.59	0.46
3 vs 1	0.43	0.32

Table 3: Pairwise alpha agreement for full label set (left) and for class mapping (A, B, C, D) => S (right), both excluding the segments labelled *unknown*.

Agreement between annotators as to whether an utterance is addressed to the group or to an individual participant is low, but if two annotators agree that a segment is addressed to a single individual instead of the group they also agree on who this individual is.

7. Context: Focus of Attention

The visual focus of attention (FOA) of speakers and listeners is an important cue in multimodal addressing behaviour. To what extent is this cue important for annotators who observe the conversational scene and have to judge who was addressing whom?

We can start answering this question when we compare cases where the gaze is directed towards any person versus those cases where the gaze is directed to objects (laptop, whiteboard, or some other artefact), or nowhere in particular. One might expect that in the second case the annotation is harder and the agreement between annotators lower. When, during an utterance, a speaker looks at only one participant, the agreement may also be higher than when the speaker looks at more (different) persons during the utterance.

To investigate this difference we compare pairwise alpha agreement for four cross sections of the data:

1. all segments irrespective of FOA
2. only those segments during which the speaker does not look at another participant at all (he may look at objects, though)
3. only those segments during which the speaker does look at one other participant, but not more than one (he may also intermittently look at objects)
4. only those segments during which the speaker does look at one or more other participants (he may also intermittently look at objects)

In all four cross sections, only those segments were considered that were *not* annotated with the 'unknown' addressee label. Table 4 presents the pairwise alpha scores for the four conditions. Agreement is consistently lowest for condition 2 whereas conditions 3 and 4 consistently score highest.

	(1)	(2)	(3)	(4)
1 vs 2	0.67	0.60	0.78	0.77
3 vs 4	0.47	0.41	0.57	0.57
4 vs 2	0.63	0.59	0.69	0.66
3 vs 2	0.47	0.42	0.48	0.51
1 vs 4	0.59	0.57	0.63	0.62
3 vs 1	0.43	0.32	0.53	0.56

Table 4: Pairwise alpha agreement for the four contextual FOA conditions, all excluding the segments labelled *unknown*.

This shows that focus of attention is being used as an important cue for the annotators. When a speaker looks at one or more participants, the agreement between annotators on addressing consistently becomes higher. Contrary to our expectations there is no marked difference, however, between the cases where, during a segment, a speaker only looks at one participant or at more of them (cases (3) versus (4)).

8. Context: Elicit Dialog Acts

The last contextual agreement analysis that we present here concerns the different types of dialog acts. Goffman's notion of addressing that was used for the annotation of

the corpus seems to be more applicable to initiatives than to responsive acts, given that it is formulated in terms of "that some answer is therefore anticipated from [the addressee]" (Goffman, 1981). Table 5 presents the pairwise alpha agreement for only the 'elicit' dialog acts opposed to that for all dialog acts. Clearly, the agreement for 'elicit' acts is a lot higher. Apparently the intended addressee of elicits is relatively easy to determine for an outsider (annotator); a closer inspection of the instances concerned may reveal what exactly are the differences in how speakers express 'elicit' acts and other acts (see also op den Akker and Theune, 2008).

	All 'real' dialog acts	Elicits only
1 vs 2	0.67	0.87
3 vs 4	0.47	0.84
4 vs 2	0.63	0.80
3 vs 2	0.47	0.58
1 vs 4	0.59	0.76
3 vs 1	0.43	0.57

Table 5: Pairwise alpha agreement for all 'real' dialog acts (left) and for only the elicit dialog acts (right), both excluding the segments labelled *unknown*.

9. Interaction Between the Different Views

Throughout this paper we presented pairwise alpha agreement scores for different class mappings or cross sections of the addressee annotations in the AMI corpus. The different effects noted about those scores were consistent. That is, although we report only a few combinations of scores, different combinations of mappings and cross sections consistently show the same patterns. For example, all differences for the different FOA conditions hold both for the 'all segments' and the 'excluding unknown labels' condition, and for the (*A*, *B*, *C*, *D*) => *S* class mapping, etcetera. Only the 'elicit' scores were not calculated in combination with each and every other cross section.

10. Discussion

Determining who is being addressed as an outsider is not easy as the alpha scores demonstrate. The above analysis shows some of the factors that influence annotators in the choices they make by comparing alpha values for different conditions.

Reidsma and Carletta (to appear) point out that reliability measures should not be treated as simple one shot indicators of agreement between annotators. A more detailed analysis is required to judge the usability of annotations for further analysis or machine learning.

Vieira (2002) and Steidl (2005) claim that it is 'unfair' to blame machine learning algorithms for bad performance in case human annotators are equally bad or worse in reaching agreement. In general, we agree with this point of view, but we want to argue for a more fine-grained analysis that allows one to understand better the disagreements between annotators. It is very well possible that an algorithm performs badly because of completely different reasons, for which one could "blame" the algorithm. On the other hand, creating algorithms can be improved by knowing the situations in which humans disagree and the reasons that lie behind this.

Acknowledgements

The authors would like to thank the developers of the AMI annotation schemes and the AMI annotators for all their hard work, as well as Nataša Jovanovič for many discussions about addressing. This work is supported by the European IST Programme Project FP6-033812 (AMIDA, publication 99). This article only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

11. References

op den Akker, R. and Theune, M. (2008), How Do I Address You? Modelling addressing behavior based on an analysis of multi-modal corpora of conversational discourse. In: Proceedings of the AISB symposium on Multi-modal Output Generation, MOG'08, Aberdeen.

Carletta, J.C. (2007), Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus, in: Language Resources and Evaluation, 41:2(181-190)

Goffman, E. (1981), Footing. In: Forms of Talk, pages 124-159. Philadelphia: University of Pennsylvania Press.

Gupta, S., John N., Matthew P., and Jurafsky, D. (2007), Resolving "you" in multiparty dialog. In: Proceedings of 8th SigDial Workshop, pages 227-230.

Jovanovič, N. (2007), To Whom It May Concern - addressee identification in face-to-face meetings, PhD Thesis, University of Twente

Krippendorff, K. (1980), Content Analysis: An Introduction to its Methodology, Sage Publications, The Sage CommText Series, volume 5

Lerner, G.H. (2003), Selecting next speaker: The context-sensitive operation of a context-free organization. Language in Society, 32:177-201.

Poesio, M. and Artstein, R. (2005), The Reliability of Anaphoric Annotation Reconsidered: Taking Ambiguity into Account, in: Proceedings of the ACL Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, pages 76-83

Reidsma D. and Carletta. J. (to appear), Reliability measurement: there's no safe limit, to appear in Computational Linguistics

Steidl, S. and Levit, M. and Batliner, A. and Nöth, E. and Niemann, H. (2005), "Of all things the measure is man" Automatic classification of Emotion and Intra Labeler Consistency. ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing

Vieira, R. (2002), How to evaluate systems against human judgment in the presence of disagreement. Proceedings of the Workshop on Joint Evaluation of Computational Processing of Portuguese