On the Contextual Analysis of Agreement Scores

Dennis Reidsma*, Dirk Heylen, and Rieks op den Akker

Human Media Interaction, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands
{dennisr,heylen,infrieks}@ewi.utwente.nl
http://hmi.ewi.utwente.nl/

Abstract. This paper explores the relation between agreement, data quality and machine learning, using the AMI corpus. The paper describes a novel approach that uses contextual information from other modalities to determine a *more reliable subset* of data, for annotations that have a low overall agreement.

Keywords: reliability, annotation, corpus, multimodal context.

1 Introduction

Researchers working with annotated multimodal corpora often find that annotations of many interesting phenomena can only be produced with a relatively low level of agreement. Sometimes this problem can be solved by spending more (time consuming) effort on defining the annotation scheme and training the annotators. Sometimes however, this is not possible, because of a lack of resources, or because the phenomenon is simply too difficult to annotate with a higher level of agreement. When one wants to use the annotated data for machine learning purposes, low agreement means lower quality training data, lower machine learning performance and less generalizable results.

Beigman Klebanov and Shamir [2006] argued that, if data has been annotated with a very low level of inter-annotator agreement, one could improve the quality of (parts of) the data by finding out whether one can pinpoint a *subset* of the data that has been annotated with a higher level of inter-annotator agreement. This more reliable subset can then be used for training and testing of machine learning, with a higher confidence in the validity of the results (see also the discussion in Reidsma and Carletta [2008]).

The approach of Beigman Klebanov and Shamir [2006] works as follows. In order to find the more reliable subset, they proposed an approach in which *all* data is annotated multiple times. They used annotations from 20 separate annotators

* 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 Jovanović for many discussions about addressing. This work is supported by the European IST Programme Project AMIDA (FP6-033812). 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.

M. Kipp et al. (Eds.): Multimodal Corpora, LNAI 5509, pp. 122–137, 2009.

on a data set annotated for lexical cohesion. Given these annotations they induced random *pseudo-annotators* from each annotator. Each pseudo-annotator marked up the data with the same distributions as the actual annotator, but chose the items at random. Given these pseudo-annotators, they calculated the probabilities that a certain item would be marked with a certain label by more than N of the random pseudo-annotators. They found that, for items that were marked with a specific label by at least 13 out of the 20 human annotators, the label could not have been the result of random annotation processes, with an overall confidence of 99%. For a different data set, concerning markup of metaphors in text annotated by only 9 annotators, they showed that an item needed to be marked by at least 4 out of the 9 annotators to make it sufficiently improbable that the assignment of that particular label to that particular item was the result of random coding behavior [Beigman Klebanov et al., 2008]. The particular proportion (here: 13 out of 20 or 4 out of 9) may depend on factors such as the number and distribution of class labels and the pairwise agreement between annotators. By taking the subset of only those items that were marked the same by at least that proportion of annotators, they obtain a subset of the data that has a higher reliability. Machine learning results obtained on this subset will potentially be more valid. Note, though, that the resulting classifiers are no longer qualified to render a judgement on all items: they have been trained only on the 'more reliable subset', and therefore are only qualified to render a judgement on items that belong in this same subset.

A major drawback of the method described above is, firstly, that it requires all training and test data to be multiply annotated — without exception. This requires an investment that otherwise might be spent on annotating more content, or different content, or on feature selection and classification experiments, and so forth. An important drawback to this approach appears when the classifier, trained and tested on such a subset, is applied to unseen data. This data has not been annotated by humans, so it is unknown *a priori* whether specific new instances would belong to the domain in which the classifier is qualified to render a judgement, that is, the reliable subset of data for which the classifier was trained and tested. The problem is, in other words, that the performance of the classifier as observed *on the reliable subset* in the testing phase is not necessarily a valid indicator of the performance of the classifier on the new, unseen data, as the classifier will assign a label to *all instances* in the new data. The problem would be solved if it were possible to deduce for new, unseen instances (from the same domain) whether they should belong to this more reliable subset.

This paper investigates whether this solution can be achieved, for the case of addressee detection on dialog acts in the AMI corpus, by taking the *multimodal context* of utterances into account. Naturally, the approach still relies on a certain amount of multiply annotated data. However, in contrast to the method described above, only a limited part of the data needs to be annotated more than once, and it is possible for new, unseen data to determine whether it falls in the subset of data for which the classifier was trained, without requiring a large number of human judgements on this new data first. The approach set out

in this paper might be used for other data sets as well, using in-depth analyses of the contextual agreement and disagreement patterns in annotations to gain more insight in the quality and usability of (subsets of) the data.

The paper is structured as follows. Section 2 introduces the AMI corpus, of which the addressee annotations were used for this paper. Section 3 concerns the basic inter-annotator agreement for the addressee annotations. Section 4 considers the relevance of the multimodal context of utterances to the level of inter-annotator agreement with which they are annotated. In Section 5 it is shown that the multimodal context of utterances can indeed be used to determine a more reliable subset of the annotations. Finally, the paper ends with a discussion and conclusions.

2 The AMI Corpus

The data used for this study was taken from the hand annotated face-to-face conversations from the 100 hour AMI meeting corpus. This corpus has been described before in other publications [Carletta, 2007; Carletta et al., 2006]. In this section a brief overview of the relevant annotations is given.

The corpus consists of 100 hours of recorded meetings. Of these recordings, 65 hours are of meetings that follow a guided scenario [Post et al., 2004]. In the scenario-based meetings, design project groups of four players have the task to design a new TV remote control. Group members have roles: project manager (PM), industrial designer (ID), user interface design (UD), and marketing expert (ME). Every group has four meetings (20-40 minutes each), dedicated to a subtask. Most of the time the participants sit around a table. During the meetings, as well as between the meetings, participants will get new information about things such as market trends or changed design requirements, via mail. This process is coordinated by a scenario controller program. The whole scenario setup was designed to provide an optimal balance between control over the meeting variables and the freedom to have natural meetings with realistic behavior from the participants [Post et al., 2004].

All meetings were recorded in meeting rooms full of audio and video recording devices (see Figure 1) so that close facial views and overview video, as well as high quality audio, is available. Speech was transcribed manually, and words were time-aligned. The corpus has several layers of annotation for a large number of modalities, among which dialog acts, topics, hand gestures, head gestures, subjectivity, visual focus of attention (FOA), decision points, and summaries. The corpus uses the Nite XML Toolkit (NXT) data format as reference storage format, making it very easy to extend the corpus with new annotations either by importing data created in other formats or by using one of the many flexible annotation tools that it comes with [Carletta et al., 2005, 2003; Reidsma et al., 2005a,b]. Of these annotations, the dialog act, addressee and Focus of Attention annotations are presented in more detail in the rest of this section.

Fig. 1. A still image of the meeting recording room in Edinburgh

2.1 The AMI Dialog Act Annotations

The AMI dialog act annotation schema concerns the segmenting and labeling of the transcripts into dialog acts. The *segmentation* guidelines are centered around the speaker's intention, with a few rules that describe how the annotators should deal with the different situations they are likely to encounter. The rules are summarized below; more details can be found in the annotation manual [AMI Consortium, 2005].

- The *first* rule is: *each segment should contain a single speaker intention.*
- The *second* rule is that *all segments only contain transcription from a single speaker.* This rule allows dialog act segmentation to be carried out on the speech of one speaker.
- The *third* rule is that *everything in the transcription is covered in a dialog act segment, with nothing left over.*
- Finally, in case of doubt, annotators were instructed *to use two segments, instead of one.*

The guidelines for *labeling* dialog acts again center around the speaker's intention — as expressed in an utterance — to, for example, exchange information, contribute to the social structure of the group, carry out an action, get something clarified, or express an attitude towards something or someone. The schema contains fifteen types of dialog acts: eleven proper dialog acts, three 'Quasi-acts' (Backchannel, Stall, and Fragment) and the 'bucket' class Other. The 'proper dialog acts' represent certain speaker's intentions. The 'Quasi-acts' are

not proper dialog acts at all, but are present in the annotation schema to account for something in the transcript that does not really convey a speaker's intention. Furthermore, although the class OTHER *does* actually represent a speaker intention ('any intention not covered by the rest of the label set'), it is present as a 'bucket' class rather than a real part of the label set, and therefore it has also been included in the group 'quasi acts' for all analyses presented in this paper. The term 'proper dialog act' will apply to the labels not taken as 'quasi-acts'.

Most of the scenario data in the AMI corpus has been annotated for dialog acts, resulting in over 100,000 utterances. Details on the distribution of class labels, and the level of inter-annotator agreement obtained on meeting IS1003d, annotated by four annotators, can be found elsewhere [Reidsma, 2008, page 44].

2.2 The AMI Addressee Annotations

A part of the AMI corpus is also annotated with addressee information [Jovanović et al., 2006; Jovanović, 2007]. In that subset, all proper dialog acts were assigned a label indicating who the speaker addressed his speech to (was talking to). In the type of meetings considered in the AMI project, most of the time the speaker addresses the whole group, but sometimes his dialog act is addressed to some particular individual. This can be, for example, because he wants to know that individual's opinion, or is presenting information that is particularly relevant for that individual. The basis of the concept of addressing underlying the AMI addressee annotation schema originates from Goffman [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"*. Sub-group addressing hardly occurs, at least in the meetings that make up the AMI corpus, and was not included in the schema. Thus, dialog acts are either addressed to the group (*G-addressed*) or to an individual (*I-addressed*). Annotators could also use the label UNKNOWN when they were unsure about the intended addressee of an utterance.

The AMI addressee annotation schema was applied to a subset of 14 meetings from the corpus[1], containing 9987 dialog acts in total. In total, three annotators contributed to the addressee/dialog act annotations of those 14 meetings. For every one of those 14 meetings, the addressee annotation was performed by the annotator who also performed the dialog annotation of that particular meeting. Table 1 shows the label distribution in the 14 meetings annotated with addressee information. In addition, one meeting (meeting IS1003d) was annotated with dialog act and addressee labels four times, by four independent annotators (DHA, S95, VKAR, and MA). The resulting annotations on this meeting were used as reliability data, to determine the level of inter-annotator agreement. Table 2 presents Krippendorff's α for multiple annotators for the dialog acts annotated with addressee, for all annotators and once for each of the single annotators left

[1] This concerns the meetings ES2008a, TS3500a, IS1000a, IS1001a, IS1001b, IS1001c, IS1003b, IS1003d, IS1006b, IS1006d, IS1008a, IS1008b, IS1008c, and IS1008d.

Table 1. The distribution of labels in the part of the AMI corpus annotated with the addressee annotation schema

Type	Number of utterances	Frequency
Quasi-acts (no addressee)	3397	34.0%
I-addressed		
A	804	8.1
B	598	6.0
C	638	6.4
D	703	7.0
Total	2743	27.5%
G-addressed	3104	31.1%
Unknown	743	7.4%
Total	9987	100.0%

Table 2. Overview of the multi-annotator α values for addressee annotation, for the group of all four annotators and for each of the single annotators left out of the group once. The number of agreed segments for each group is given as N.

Group	N	α
All	120	0.38
Without VKAR	213	0.36
Without MA	157	0.39
Without S95	162	0.37
Without DHA	198	0.53

out of the calculation. A more detailed analysis of these annotations is presented in later sections.

2.3 The AMI Focus of Attention Annotations

A subset of meetings in the AMI corpus were also annotated with visual Focus of Attention (FOA) information, which annotators had to derive by watching the head, body and gaze of the participant [Ba and Odobez, 2006]. FOA forms an important cue for, among other things, addressing behavior. The FOA annotation contains, for every participant in the meeting, at all times throughout the meeting, whom or what he is looking at. This annotation schema was applied to the same subset of 14 meetings that was used for addressee annotation (but by other annotators). The FOA annotation was done with a very high level of agreement and with very high precision: changes are marked in the middle of eye movement between old and new target with α agreement between annotators ranging from 0.84 to 0.95 [Jovanović, 2007, page 80][2].

[2] These results were obtained on the AMI corpus, with a label set of 8 possible targets. Voit and Stiefelhagen [2008a] report a reliability of $\kappa = 0.70$ for a label set of 36 possible FOA targets.

3 Basic Agreement and Class Maps for Addressee

The inter-annotator agreement for the AMI addressee annotations was given in Section 2.2. Recall that the value of Krippendorff's multi-annotator α was 0.44. This indicates a quite low level of agreement: it falls into the range usually reported on highly subjective annotation tasks. Before the contextual dependencies for the inter-annotator agreement are discussed in the next section, some more information about the basic agreement analysis is given here. Table 3 presents the pairwise agreement values expressed in Krippendorff's multi-annotator α [1980]. Table 4 shows an example of a confusion matrix for the addressee annotation, representative of the other confusion matrices. The values in the confusion matrix suggest that it is not so much problematic to decide *which* individual was addressed as it is to distinguish between I-addressed utterances versus utterances that are G-addressed or labeled UNKNOWN. In the remainder of this section, inter-annotator agreement is discussed for two derived versions of the label set, namely for the annotation without the label UNKNOWN and for the class map in which the annotation is reduced to the binary distinction I-addressed/G-addressed. Note that all results presented in the remainder of this paper concern only proper dialog acts and are based upon a pairwise comparison of agreed segments, as in the table below (the agreed segments of a pair of annotators are those segments for which the two annotators assigned exactly the same start and end boundaries).

Table 3. Pairwise agreement (Krippendorff's α) for addressee annotations by four annotators, on agreed segments annotated as proper dialog act

	MA	VKAR	DHA	S95
MA	.	0.57	0.32	0.46
VKAR		.	0.36	0.50
DHA			.	0.31
S95				.

Table 4. Confusion matrix for annotators VKAR and MA for the addressee labels of agreed segments in meeting IS1003d. Krippendorff's α is 0.57 for this matrix.

	A	B	C	D	G	U	$\sum$
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
$\sum$	70	31	51	78	234	15	479

3.1 Reliability for the Addressee Label Unknown

The annotators indicated whether an utterance was addressed to a particular person or to the whole group. They could also use the label UNKNOWN, if they could not decide who was being addressed. All four annotators used this label at some point in their annotation of meeting IS1003d. Given the annotation guidelines, there might have been two reasons why an annotator would use the label UNKNOWN. Firstly, the utterance may have been ambiguously or unclearly addressed, making it impossible to choose a single label like the annotation task requires. The reason for assigning the label UNKNOWN then lies within the content. A certain amount of inter-annotator agreement for this label could be expected, and the applicability of the label could be learnable and worth learning. Secondly, the utterance may have been unambiguously addressed, but nevertheless the annotator may have been uncertain about his own judgement, for example because he was tired, or did not understand what was being said. In that case, the reason for assigning the label UNKNOWN lies completely with the annotator, rather than with the content. This second type of uncertainty would *not* cause the label UNKNOWN to exhibit a large inter-annotator agreement, and would by far be less interesting to learn to classify.

The question to be answered here is then: does the uncertainty in the addressee annotation, expressed by the annotator assigning the label UNKNOWN, reflect an attribute of the content, or rather an attribute of the specific annotator who assigned the label at a certain point? Inspection of the confusion matrices shows a clear answer to this question. The matrix displayed in Table 4 is certainly representative in this respect. Inter-annotator agreement on the applicability of the label is virtually non-existent for each and every pair of annotators. This means that the occurrence of the label UNKNOWN in the corpus does not seem to give any useful information about the annotated content at all.

For this reason, it was decided to remove all UNKNOWN labels from the corpus before proceeding with further analysis. That is, for all segments that an annotator labeled UNKNOWN, the label was removed, and the segment was taken as if the annotator had not labeled it with addressee at all — reducing the number of segments available for the analyses presented later in this paper by one, for

Table 5. Inter-annotator agreement for all proper dialog acts versus only the dialog acts not annotated with the UNKNOWN addressee label

	Inc. UNKNOWN	Excl. UNKNOWN
MA vs VKAR	0.57	0.67
DHA vs S95	0.31	0.47
S95 vs VKAR	0.50	0.63
DHA vs VKAR	0.36	0.47
MA vs S95	0.46	0.59
DHA vs MA	0.32	0.43

Table 6. Pairwise α agreement for the unmapped label set (left) and for the class mapping $(A, B, C, D) => S$ (right), both after removing the label UNKNOWN from the data set

	Normal label set (excl. UNKNOWN)	Class map $(A, B, C, D) => S$ (excl. UNKNOWN)
MA vs VKAR	0.67	0.55
DHA vs S95	0.47	0.37
S95 vs VKAR	0.63	0.52
DHA vs VKAR	0.47	0.37
MA vs S95	0.59	0.46
DHA vs MA	0.43	0.32

that annotator, but leaving the number of segments available from the other annotators unaffected.

The effect of this data set reduction on the inter-annotator agreement on the remaining segments is shown in Table 5. This table presents the α values for the addressee annotations computed on all proper dialog acts versus the α values calculated after removing all UNKNOWN labels from the corpus. The increase in level of inter-annotator agreement ranges from 0.10 to 0.16. This does not only hold for the overall data set reported in this table, but also for each and every contextual subset of the data set reported later in this paper.

3.2 Class Map: Group/A/B/C/D vs Group/Single

The second label that really contributed to the disagreement, according to the confusion matrix of Table 4, is the GROUP label. However, in large contrast to the label UNKNOWN discussed above, the majority of its occurrences are actually agreed upon by at least some of the annotators. From the confusion matrices it can nevertheless be seen that annotators cannot make the global distinction between G-addressed and I-addressed utterances with a high level of agreement: there is a lot of confusion between the label G on the one hand and A, B, C and D on the other hand. If annotators see an utterance as I-addressed they subsequently do not have much trouble determining who of the single participants was addressed: there is hardly any confusion between the individual addressee labels A, B, C and D.

This observation is quantified using a class mapping of the addressee annotation in which the individual addressee labels A, B, C and D are all mapped onto the label S. Table 6 shows pairwise α agreement for this class mapping, next to the values obtained for the full label set excluding only the label UNKNOWN (see also the previous section). Clearly, agreement on who of the participants was addressed individually is a major factor in the overall agreement.

4 The Multimodal Context of Utterances

The remainder of this paper concerns multimodal contextual agreement. To a large extent multimodal behavior is a holistic phenomenon, in the sense that the contribution of a specific behavior to the meaning of an utterance needs to be decided upon in the context of other behaviors that coincide, precede or follow. A nod, for instance, may contribute to a conversation in different ways when it is performed by someone speaking or listening, when it is accompanied by a smile, or when it is a nod in a series of more than one. 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 occurrences of gaze, eye-contact, speech, facial expressions, gestures, and the setting determine our interpretation of events and help us to disambiguate otherwise ambiguous activities.

Annotators, who are requested to label certain communicative events, be it topic, focus of attention, addressing information or dialog acts, 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, even if these features are not mentioned in the annotator instructions. Using or not using the video and audio while annotating may therefore have a large impact on the agreement achieved for certain annotations. Also, one annotator may be more sensitive to one cue rather than to another. This means that the agreement between annotators may depend on particular variations in the multimodal input.

Within the AMI corpus, one of the more obvious annotations to which this bears relevance is the combination of addressee and visual focus of attention (FOA) annotations. Visual focus of attention of speakers and listeners is an important cue in multimodal addressing behavior. The combination of these two layers will therefore be used in an attempt to determine a more reliable subset of the corpus.

5 Finding More Reliable Subsets

This section describes two 'more reliable subsets' within the AMI addressee annotations (with the UNKNOWN label removed as discussed in Section 3). The first is centered around the multimodal context of the utterance. The second uses the context determined by the type of dialog act for which the addressee was annotated. The aim of these contextual agreement analyses, as described in the introduction to this paper, is to be able to pinpoint a more reliable subset in the data without having all training and test data be annotated by multiple annotators.

5.1 Context: Focus of Attention

Visual Focus of Attention (FOA) of speakers and listeners is an important cue in multimodal addressing behavior. In this section it is investigated to what extent

Table 7. The three different contexts defined by different conditions on the FOA annotation that are used to find more reliable subsets of the addressee annotations

Context	Description
I	Only those utterances during which the speaker does not look at another participant at all (he may look at objects, though)
II	Only those utterances during which the speaker does look at one other participant, but not more than one (he may additionally look at objects)
III	Only those utterances during which the speaker does look at one or more other participants (he may additionally look at objects)

this cue impacts the task of annotators who observe the conversational scene and have to judge who was addressing whom. FOA annotations are a manifest type of content, do not need extensive discourse interpretation, and can be annotated with a very high level of inter-annotator agreement. This makes them especially useful when they can serve as multimodal context for finding a more reliable subset of the addressing data, because it is more likely that this context can be retrieved for new, unseen data, too[3].

Table 7 lists three different FOA contexts (I, II and III) that each define a different subset of all addressee annotations. The contexts are defined with respect to the Focus of Attention of the speaker during the utterance. Context I concerns utterances during which the speaker's gaze is directed only to objects (laptop, whiteboard, or some other artefact) or nowhere in particular. One might expect that in this context the annotation task is harder and the inter-annotator agreement lower. Contexts II and III concern the utterances during which the speaker's gaze is directed at least some of the time to other persons (only one person, for context II, or any number of persons for context III). The expectation was that utterances in contexts II and III respectively would also exhibit a difference in inter-annotator agreement. When a speaker looks at only one participant, agreement may be higher than when the speaker looks at several (different) persons during an utterance.

Table 8 presents α values for the pairwise inter-annotator agreement for the three subsets defined by the three FOA contexts from Table 7, compared to the α values for the whole data set that were presented in Section 3.1. Inter-annotator agreement for the addressee annotation is consistently lowest for context I whereas contexts II and III consistently score highest. When a speaker looks at one or more participants, the agreement between annotators on addressing consistently becomes higher. Contrary to expectations there is no marked difference, however, between the contexts where, during a segment, a speaker only looks at one participant or at several of them (context II versus III).

[3] Although it should be noted that state-of-the-art recognition rates are still too low for this, in the order of 60% frame recognition rate [Ba and Odobez, 2007; Voit and Stiefelhagen, 2008b].

Table 8. Pairwise α agreement for the subsets defined by the three contextual FOA conditions, compared to α agreement for the full data set (without the label UNKNOWN)

	All (excl. UNKNOWN)	I	II	III
MA vs VKAR	0.67	0.60	0.78	0.77
DHA vs S95	0.47	0.41	0.57	0.57
S95 vs VKAR	0.63	0.59	0.69	0.66
DHA vs VKAR	0.47	0.42	0.48	0.51
MA vs S95	0.59	0.57	0.63	0.62
DHA vs MA	0.43	0.32	0.53	0.56

In conclusion, it can be said that the subset of all utterances during which the speaker looks at some other participants at least some of the time, defined by context II or III, forms a more reliable subset of the addressee annotations as defined in the introduction to this paper. This subset contains two thirds of all utterances annotated with addressee.

5.2 Context: Elicit Dialog Acts

The second contextual agreement analysis presented here concerns a certain specific group of dialog acts. Op den Akker and Theune [2008] discussed that forward looking dialog acts, and more specifically, 'Elicit' types of dialog act (see Table 9), are more often I-addressed, and tend to be addressed more explicitly. If this were true, one would also expect elicit dialog acts to exhibit a higher inter-annotator agreement on addressing. This we can test on the data in the AMI corpus. Table 10 presents the pairwise α inter-annotator agreement values for all proper dialog acts, the 'elicit' dialog acts only, and the proper acts without the 'elicit' acts. Clearly, the agreement for only the 'elicit' acts is a lot higher. Apparently the intended addressee of elicits is relatively easy to determine for an outsider (annotator); this may support what Op den Akker and Theune [2008] say about the differences in how speakers express 'elicit' acts and other forward looking acts.

We tested this difference between Elicits and other dialog acts again using a second set of annotations, that were not yet introduced in this paper. The dialog acts of AMI meeting IS1003d, segmented and labelled by annotator

Table 9. Types of 'Elicit' dialog acts

Description	Dialog act label
Acts about information exchange:	ELICIT-INFORM
Acts about possible actions:	ELICIT-OFFER-OR-SUGGESTION
Acts that comment on the previous discussion:	ELICIT-COMMENT-ABOUT-UNDERSTANDING and ELICIT-ASSESSMENT

Table 10. Pairwise α agreement for all proper dialog acts and for the elicit dialog acts only

	All proper acts	Elicits only	No elicits
MA vs VKAR	0.67	0.87	0.64
DHA vs S95	0.47	0.84	0.38
S95 vs VKAR	0.63	0.80	0.61
DHA vs VKAR	0.47	0.58	0.41
MA vs S95	0.59	0.76	0.57
DHA vs MA	0.43	0.57	0.40

VKAR, were annotated for addressee by another 10 annotators. In this case the DA-segments as well as their labels were already given, so annotators only had to label the addressee (of the proper acts, i.e. excluding BACKCHANNELS, STALLS and FRAGMENTS). This implies that we can study inter-annotator agreement on the whole set of all proper DActs in this meetings which is 454 out of a total of 693 acts. In this case, annotators were not allowed to use the Unknown label, so they were asked to decide if a DA is addressed to the Group of to some individual, and in the latter case who is addressed.

For the addressee labeling task, we computed Krippendorff's pairwise α and Krippendorf's group α for the whole group both for the whole set of proper dialogue acts and for the subset of elicit acts.

The results are as follows. The pairwise α is significantly higher for Elicit-acts (50 units) than for all proper acts (454 units). A paired t-test was performed to determine if there is a real difference in the α values for pairs of annotators of addressees for all acts and α values for the same pair of annotators for only addressees labels of elicit acts. The mean difference (M=0.1718, SD =0.01412, N= 45) was significantly greater than zero, $t(44) = 12.16$, two-tail $p < 0.001$, providing evidence that their is a real difference. A 95% C.I. about the mean difference is (0.1433, 0.2001). The group wise α for elicit acts is 0.80 which is much higher than the group wise α for the whole set of proper acts, 0.65.

Since not all of these 44 pairs are independent we also performed the same test on 9 pairs of annotators that are independent (one fixed annotator paired with all other 9). The mean difference (M=0.1667, SD =0.00915, N= 9) was significantly greater than zero, $t(8) = 5.46$, two-tail $p < 0.001$, providing evidence that their is a real difference. A 95% C.I. about the mean difference is (0.09, 0.23).

These findings again support the claim that it is easier for annotators to tell if the group is addressed or some individual, when the act is an elicit act, than in general for dialogue acts, and that this subset of the data in the corpus has a high reliability.

6 Discussion and Summary of Addressing Agreement

Throughout this paper pairwise α agreement scores have been presented for different class mappings and subsets of the addressee annotations in the AMI

corpus. The different effects noted about these scores were consistent. That is, although only a few combinations of scores are reported, all different combinations of mappings and subsets consistently show the same patterns. For example, all relative differences between the FOA contexts hold for the 'all agreed proper dialog acts' condition, the 'excluding UNKNOWN' condition, and for the $(A, B, C, D) => S$ class mapping.

The following conclusions can be summarized for the inter-annotator agreement of addressee annotations: (1) the label UNKNOWN does not give any information about the annotated content; (2) there is a large confusion between dialog acts being G-addressed or I-addressed, but if the annotators agree on an utterance being I-addressed they typically also agree on the particular individual being addressed; (3) utterances during which the speaker's focus of attention is directed to one or more other participants are consistently annotated with more agreement than those during which the speaker's FOA is not directed to any participant; and (4) 'elicit' dialog acts are easier to annotate with addressee than other types of dialog act.

The context defined by the different FOA conditions, and the context defined by the 'elicit' dialog acts, specify more reliable subsets of the annotated data. These subsets can be used in machine learning tasks in two ways. Firstly, classifiers can be trained on only the more reliable subset, in an effort to increase the relevance of the results. Secondly, classifiers can be built that restrict their judgements to those instances for which humans agree more easily (that is, the more reliable subset), yielding a 'no classification possible' judgement for instances that do not belong to the more reliable subset. This way, the consumer of the classifications can place more trust in the judgements returned by the classifier, even when the original annotations were produced with a low level of inter-annotator agreement. The approach set out in this paper might be used for other data sets as well, using in-depth analyses of the contextual agreement and disagreement patterns in annotations to gain more insight in the quality and usability of (subsets of) the data. It remains for future work to find out how much more this approach can make classifiers 'fit for purpose', but it is important to note that the FOA and the Elicit acts can be automatically detected with much better reliability than addressee.

References

AMI Consortium: Guidelines for dialogue act and addressee. Technical report (2005)

Ba, S.O., Odobez, J.-M.: A study on visual focus of attention recognition from head pose in a meeting room. In: Renals, S., Bengio, S., Fiscus, J.G. (eds.) MLMI 2006. LNCS, vol. 4299, pp. 75–87. Springer, Heidelberg (2006)

Ba, S.O., Odobez, J.M.: Head pose tracking and focus of attention recognition algorithms in meeting rooms. In: Stiefelhagen, R., Garofolo, J.S. (eds.) CLEAR 2006. LNCS, vol. 4122, pp. 345–357. Springer, Heidelberg (2007)

Beigman Klebanov, B., Beigman, E., Diermeier, D.: Analyzing disagreements. In: Artstein, R., Boleda, G., Keller, F., Schulte im Walde, S. (eds.) Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, Manchester, UK, Coling 2008 Organizing Committee, August 2008, pp. 2–7 (2008) ISBN 978-3-540-69567-7

Beigman Klebanov, B., Shamir, E.: Reader-based exploration of lexical cohesion. Language Resources and Evaluation 40(2), 109–126 (2006)

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

Carletta, J.C., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W.M., Reidsma, D., Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 28–39. Springer, Heidelberg (2006)

Carletta, J.C., Evert, S., Heid, U., Kilgour, J., Robertson, J., Voormann, H.: The NITE XML toolkit: flexible annotation for multi-modal language data. Behavior Research Methods, Instruments and Computers 35(3), 353–363 (2003)

Carletta, J.C., McKelvie, D., Isard, A., Mengel, A., Klein, M., Møller, M.B.: A generic approach to software support for linguistic annotation using xml. In: Sampson, G., McCarthy, D. (eds.) Corpus Linguistics: Readings in a Widening Discipline. Continuum International, London (2005)

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

Jovanović, N.: To Whom It May Concern - Addressee Identification in Face-to-Face Meetings. Phd thesis, University of Twente (2007)

Jovanović, N., op den Akker, H.J.A., Nijholt, A.: A corpus for studying addressing behaviour in multi-party dialogues. Language Resources and Evaluation 40(1), 5–23 (2006)

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

op den Akker, H.J.A., Theune, M.: 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 2008, April 2008, pp. 10–17 (2008) ISBN 1-902956-69-9

Post, W.M., Cremers, A.H.M., Blanson Henkemans, O.A.: A research environment for meeting behavior. In: Proceedings of the 3rd Workshop on Social Intelligence Design, pp. 159–165 (2004)

Reidsma, D.: Annotations and Subjective Machines — of annotators, embodied agents, users, and other humans. PhD thesis, University of Twente (October 2008)

Reidsma, D., Carletta, J.C.: Reliability measurement without limits. Computational Linguistics 34(3), 319–326 (2008)

Reidsma, D., Hofs, D.H.W., Jovanović, N.: Designing focused and efficient annotation tools. In: Noldus, L.P.J.J., Grieco, F., Loijens, L.W.S., Zimmerman, P.H. (eds.) Measuring Behaviour, Wageningen, NL, September 2005a, pp. 149–152 (2005a)

Reidsma, D., Hofs, D.H.W., Jovanović, N.: A presentation of a set of new annotation tools based on the NXT API. Poster at Measuring Behaviour 2005 (2005b)

Voit, M., Stiefelhagen, R.: Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In: IMCI 2008: Proceedings of the 10th international conference on Multimodal interfaces, pp. 173–180. ACM, New York (2008a)

Voit, M., Stiefelhagen, R.: Visual focus of attention in dynamic meeting scenarios. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 1–13. Springer, Heidelberg (2008)