Exploiting 'Subjective' Annotations

Dennis Reidsma
Human Media Interaction
University of Twente, PO Box 217
NL-7500 AE, Enschede, The Netherlands
dennisr@ewi.utwente.nl

Rieks op den Akker
Human Media Interaction
University of Twente, PO Box 217
NL-7500 AE, Enschede, The Netherlands
infrieks@ewi.utwente.nl

Abstract

Many interesting phenomena in conversation can only be annotated as a subjective task, requiring interpretative judgements from annotators. This leads to data which is annotated with lower levels of agreement not only due to errors in the annotation, but also due to the differences in how annotators interpret conversations. This paper constitutes an attempt to find out how subjective annotations with a low level of agreement can profitably be used for machine learning purposes. We analyse the (dis)agreements between annotators for two different cases in a multimodal annotated corpus and explicitly relate the results to the way machine-learning algorithms perform on the annotated data. Finally we present two new concepts, namely 'subjective entity' classifiers resp. 'consensus objective' classifiers, and give recommendations for using subjective data in machine-learning applications.

1 Introduction

Research that makes use of multimodal annotated corpora is always presented with something of a dilemma. One would prefer to have results which are reproducible and independent of the particular annotators that produced the corpus. One needs data which is annotated with as few disagreements between annotators as possible. But labeling a corpus is a task which involves a judgement by the annotator and is therefore, in a sense, always a subjective task. Of course, for some phenomena those judgements can be expected to come out mostly the same for different annotators. For other phenomena the judgements can be more dependent on the annotator *interpreting* the behavior being annotated, leading to annotations which are more subjective in nature. The amount of overlap or agreement between annotations is then also influenced by the amount of *intersubjectivity* in the judgements of annotators.

This relates to the spectrum of content types discussed extensively by Potter and Levine-Donnerstein (1999). One of the major distinctions that they make is a distinction in annotation of *manifest content* (directly observable events), *pattern latent content* (events that need to be inferred indirectly from the observations), and *projective latent content* (loosely said, events that require a subjective interpretation from the annotator).

Manifest content is what is directly observable. Some examples are annotation of instances where somebody raises his hand or raises an eyebrow, annotation of the words being said and indicating whether there is a person in view of the camera. Annotating manifest content can be a relatively easy task. Although the annotation task involves a judgement by the annotator, those judgements should not diverge a lot for different annotators.

At the other end of the spectrum we find *projective latent content*. This is a type of content for which the annotation schema does not specify in extreme detail the rules and surface forms that determine the applicability of classes, but in which the coding relies on the annotators' existing mental conception[1] of the classes. Such an ap-

[1] Potter and Levine-Donnerstein use the word "mental scheme" for this. We will use "mental conceptions" in this

Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, pages 8–16
Manchester, August 2008

proach is useful for everyday concepts that most people understand and to a certain extent share a common meaning for, but for which it is almost impossible to provide adequately complete definitions. Potter and Levine-Donnerstein use the example 'chair' for everyday concepts that are difficult to define exhaustively. But this concept is also especially relevant in an application context that *requires the end user of the data to agree with the distinctions being made*. This is very important when machine learning classifiers are developed to be used in everyday applications. For example, one can make a highly circumscribed, ethologically founded definition of the class 'dominant' to guide annotation. This is good for, e.g., research into social processes in multiparty conversations. However, in a scenario where an automatic classifier, trained to recognize this class, is to be used in an application that gives a participant in a meeting a quiet warning when he is being too dominant (Rienks, 2007) one would instead prefer the class rather to fit the mental conceptions of dominance that a 'naive' user may have. When one designs an annotation scheme for projective latent content, the focus of the annotation guidelines is on instructions that trigger the appropriate existing mental conceptions of the annotators rather than on writing exhaustive descriptions of how classes can be distinguished from each other (Potter and Levine-Donnerstein, 1999).

Interannotator agreement takes on different roles for the two ends of the spectrum. For manifest content the level of agreement tells you something about how accurate the measurement instrument (schema plus coders) is. Bakeman and Gottman, in their text book *observing interaction: introduction to sequential analysis* (1986, p 57), say about this type of reliability measurement that it is a matter of "calibrating your observers". For projective content, we have additional problems; the level of agreement may be influenced by the level of intersubjectivity, too. Where Krippendorff (1980) describes that annotators should be interchangeable, annotations of projective latent content can sometimes say as much about the mental conceptions of the particular annotator as about the person whose interactions are being annotated. The personal interpretations of the data by the annotator should not necessarily be seen as 'errors', though, even if those interpretations lead to low interannotator agreement: they may simply be an unavoidable aspect of the interesting type of data one works with.

Many different sources of low agreement levels, and many different solutions, are discussed in the literature. It is important to note that some types of disagreement are more systematic and other types are more noise like. For projective latent content one would expect more consistent *structure* in the disagreements between annotators as they are caused by the differences in the personal ways of interpreting multimodal interaction. Such systematic disagreements are particularly problematic for subsequent use of the data, more so than noise-like disagreements. Therefore, an analysis of the quality of an annotated corpus should not stop at presenting the value of a reliability metric; instead one should investigate the patterns in the disagreements and discuss the possible impact they have on the envisioned uses of the data (Reidsma and Carletta, 2008). Some sources of disagreements are the following.

(1) *'Clerical errors'* caused by a limited view of the interactions being annotated (low quality video, no audio, occlusions, etc) or by slipshod work of the annotator or the annotator misunderstanding the instructions. Some solutions are to provide better instructions and training, using only good annotators, and using high quality recordings of the interaction being annotated.

(2) *'Invalid or imprecise annotation schemas'* that contain classes that are not relevant or do not contain classes that are relevant, or force the annotator to make choices that are not appropriate to the data (e.g. to choose one label for a unit where more labels are applicable). Solutions concern redesigning the annotation schema, for example by merging difference classes, allowing annotators to use multiple labels, removing classes, or adding new classes.

(3) *'Genuinely ambiguous expressions'* as described by Poesio and Artstein (2005). They discuss that disagreements caused by ambiguity are not so easily solved.

(4) *'A low level of intersubjectivity'* for the interpretative judgements of the annotators, caused by the fact that there is less than perfect overlap between the mental conceptions of the annotators. The solutions mentioned above for issue (2) partly also apply here. However, in this article we focus on an additional, entirely different, way of coping

paper to avoid confusion with the term "annotation scheme".

with disagreements resulting from a low level of intersubjectivity that actively exploits the systematic differences in the annotations caused by this.

1.1 Useful results from data with low agreement

Data with a low interannotator agreement may be difficult to use, but there are other fields where partial solutions have been found to the problem, such as the information retrieval evaluation conferences (TREC). Relevance judgements in TREC assessments (and document relevance in general) are quite subjective and it is well known that agreement for relevance judgements is not very high (Voorhees and Harman report 70% three-way percent agreement on 15,000 documents for three assessors (1997)). Quite early in the history of the TREC, Voorhees investigated what the consequences of this low level of agreement are for the usefulness of results obtained on the TREC collection. It turns out that specifying a few constraints[2] is enough to be able to use the TREC assessments to obtain meaningful evaluation results (Voorhees, 2000). Inspired by this we try to find ways of looking at subjective data that tells us what constraints and restrictions on the use of it follow from the patterns in the disagreements between annotators, as also advised by Reidsma and Carletta (2008).

1.2 Related Work

In corpus research there is much work with annotations that need subjective judgements of a more subjective nature from an annotator about the behavior being annotated. This holds for Human Computer Interaction topics such as affective computing or the development of Embodied Conversational Agents with a personality, but also for work in computational linguistics on topics such as emotion (Craggs and McGee Wood, 2005), subjectivity (Wiebe et al., 1999; Wilson, 2008) and agreement and disagreement (Galley et al., 2004).

If we want to interpret the results of classifiers in terms of the patterns of (dis)agreement found between annotators, we need to subject the classifiers with respect to each other and to the 'ground truth data' to the same analyses used to evaluate and compare annotators to each other. Vieira (2002) and Steidl et al. (2005) similarily remark that it is not 'fair' to penalize machine learning performance for errors made in situations where humans would not agree either. Vieira however only looks at the *amount* of disagreement and does not explicitly relate the classes where the system and coders disagree to the classes where the coders disagree with each other. Steidl et al.'s approach is geared to data which is multiply coded for the whole corpus (very expensive) and for annotations that can be seen as 'additive', i.e., where judgements are not mutually exclusive.

Passonneau et al. (2008) present an extensive analysis of the relation between per-class machine learning performance and interannotator agreement obtained on the task of labelling text fragments with their function in the larger text. They show that overall high agreement can indicate a high learnability of a class in a multiply annotated corpus, but that the interannotator agreement is not necessarily predictive of the learnability of a label from a single annotator's data, especially in the context of what we call projective latent content.

1.3 This Paper

This paper constitutes an attempt to find out how subjective annotations, annotated with a low level of agreement, can profitably be used for machine learning purposes. First we present the relevant parts of the corpus. Subsequently, we analyse the (dis)agreements between annotators, on more aspects than just the value of a reliability metric, and explicitly relate the results to the way machine-learning algorithms perform on the annotated data. Finally we present two new concepts that can be used to explain and exploit this relation ('subjective entity' classifiers resp. 'consensus objective' classifiers) and give some recommendations for using subjective data in machine-learning applications.

2 From Agreement to Machine Learning Performance

We used the hand annotated face-to-face conversations from the 100 hour AMI meeting corpus (Carletta, 2007). In the scenario-based AMI meetings, design project groups of four players have the task to design a new remote TV 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 min. each), dedicated to a subtask. Most of

[2] Only discuss *relative* performance differences on different (variations of) algorithms/systems run on *exactly the same set of assessments* using the *same set of topics*.

the time the participants sit at a square table.

The meetings were recorded in a meeting room stuffed with audio and video recording devices, 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 several modalities, such as dialogue acts, topics, hand gestures, head gestures, subjectivity, visual focus of attention (FOA), decision points, and summaries, and is easily extendible with new layers. The *dialogue act* (DA) layer segments speaker turns into dialogue act segments, on top of the word layer, and they are labeled with one of 15 dialogue act type labels, following an annotation procedure.

In this section we will inspect (dis)agreements and machine learning performance for two corpus annotation layers: the addressing annotations (Jovanović et al., 2006) and for a particular type of utterances in the corpus, the *"Yeah-utterances"* (Heylen and op den Akker, 2007).

2.1 Contextual Addressing

A part of the AMI corpus is also annotated with addressee information. Real dialogue acts (i.e. all dialogue acts but backchannels, stalls and fragments) were assigned a label indicating who the speaker addresses his speech to (is talking to). In these type of meetings most of the time the speaker addresses the whole group, but sometimes his dialogue act is particularly addressed to some individual (about 2743 of the 6590 annotated real dialogue acts); for example because he wants to know that individual's opinion. The basis of the concept of addressing underlying the addressee annotation in the AMI corpus 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 and was not annotated. Thus, DAs are either addressed to the group (*G-addressed*) or to an individual (*I-addressed*) (see Jovanovic et al. (2006)).

Another layer of the corpus contains *focus of attention* information derived from head, body and gaze observations (Ba and Odobez, 2006), so that for any moment it is known whether a person is looking at the table, white board, or some other participant. Gaze and focus of attention are important elements of addressing behavior, and therefore FOA is a strong cue for the annotator who needs to determine the addressee of an utterance. However, FOA is not the only cue. Other relevant cues are, for example, proper names and the use of addressing terms such as "you". Even when the gaze is drawn to a projection screen, or the meeting is held as a telephone conference without visuals, people are able to make the addressee of their utterances clear.

From an extensive (dis)agreement analysis of the addressing and FOA layers the following conclusions can be summarized: the visual focus of attention was annotated with a very high level of agreement (Jovanović, 2007); in the addressee annotation there is a large confusion between DAs being G-addressed or I-addressed; if the annotators agree on an utterance being I-addressed they typically also agree on the particular individual being addressed; 'elicit' DAs were easier to annotate with addressee than other types of dialog act; and reliability of addressee annotation is dependent on the FOA context (Reidsma et al., 2008). When the speaker's FOA is not directed to any participant the annotators must rely on other cues to determine the addressee and will disagree a lot more than when they are helped by FOA related cues. Some of these disagreements can be due to systematic subjective differences, e.g. an annotator being biased towards the 'Group' label for utterances that are answers to some question. Other disagreements may be caused by the annotator being forced to choose an addressee label for utterances that were not be clearly addressed in the first place.

In this section we will not so much focus on the *subjectivity* of the addressee annotation as on the *multimodal context* in which annotators agree more. Specifically, we will look further at the way the level of agreement with which addressee has been annotated is dependent on the FOA context of a set of utterances. We expect this will be reflected directly by the machine learning performance in these two contexts: the low agreement might indicate a context where addressee is inherently difficult to determine and furthermore the context with high agreement will result in annotations containing more consistent information that machine learning can model.

To verify this assumption we experimented with automatic detection of the addressee of an utter-

ance based on lexical and multimodal features. Compared to Jovanovič (2007), we use a limited set of features that does not contain local context features such as 'previous addressee' or 'previous dialogue act type'. Besides several lexical features we also used features for focus of attention of the speaker and listeners during the utterance. Below we describe two experiments with this task. Roughly 1 out of every 3 utterances is performed in a context where the speaker's FOA is not directed at any other participant. This gives us three contexts to train and to test on: all utterances, all utterances where the speaker's FOA is not directed at any other participant (1/3 of the data) and all utterances during which the speaker's FOA is directed at least once at another participant (2/3 of the data).

First Experiment For the first experiment we trained a Bayesian Network adapted from Jovanovič (2007) on a mix of utterances from all contexts, and tested its performance on utterances from the three different contexts: (1) all data, (2) all data in the context 'at least some person in speaker's FOA' and (3) all data in the context 'no person in speaker's FOA during utterance'. As was to be expected, the performance in the second context showed a clear gain compared to the first context, and the performance in the third context was clearly worse. The performance differences, for different train/test splits, tend to be about five percent.

Second Experiment Because the second context showed such a better performance, we ran a second experiment where we trained the network on only data from the second context, to see if we could improve the performance in that context even more. In different train/test splits this gave us another small performance increase.

Conclusions for Contextual Addressing The performance increases can mostly be attributed to the distinction between different individual addressees for I-addressed utterances. Precision and recall for the G-addressed utterances does not change so much for the different contexts. This result is reminiscent of the fact that when the annotators agreed on an utterance being I-addressed they typically also agreed on the particular individual being addressed.

These results are particularly interesting in the light of the high accuracy with which FOA was annotated. If this accuracy points at the possibility to also achieve a high automatic recognition rate for FOA we can exploit these results in a practical application context by defining a addressee detection module which only assigns an addressee to an utterance in the second FOA context (FOA at some participants), and in all other cases labels an utterance as 'addressee cannot be determined'. Such a detection module achieves a much higher precision than a module that tries to assign an addressee label regardless; of course this happens at the cost of recall.

2.2 Interannotator Training and Testing

Classifiers behave as they are trained. When two annotators differ in the way they annotate, i.e. have different "mental conceptions" of the phenomenon being annotated, we can expect that a classifier trained on the data annotated by one annotator behaves different from a classifier trained on the other annotator's data. As Rienks describes, this property allows us to use all data in the corpus, instead of just the multiply annotated part of it, for analyzing differences between annotators (Rienks, 2007, page 105). We can expect that a classifier A trained on data annotated by A will perform better when tested on data annotated by A, than when tested on data annotated by B. In other words, classifier A is geared towards modelling the 'mental conception' of annotator A. In this section we will try to find out whether it is possible to explicitly tease apart the overlap and the differences in the mental conceptions of the annotators as mirrored in the behavior of classifiers, on a subjective annotation task. Suppose that we build a Voting Classifier, based on the votes of a number of classifiers each trained on a different annotator's data. The Voting Classifier only makes a decision when all voters agree on the class label. How good will the Voting Classifier perform? Is there any relation between the (dis)agreement of the voters, and the (dis)agreement of the annotators? Will the resulting Voting Classifier in some way embody the overlap between the 'mental conceptions' of the different annotators?

As an illustration and a test case for such a Voting Classifier, we consider the human annotations and automatic classification of a particular type of utterances in the AMI corpus, the *"Yeah-utterances"*, utterances that start with the word "yeah".

class	train-tot	test-tot	DH-train/test	S9-train/test	VK-train/test
bc	3043	1347	1393/747	670/241	980/359
as	3724	1859	1536/1104	689/189	1499/566
in	782	377	340/229	207/60	235/88
ot	1289	596	316/209	187/38	786/349

Table 1: Sizes of train and test data sets used and the distribution of class labels over these data sets for the different annotators.

The Data Response tokens like "yeah", "okay", "right" and "no" have the interest of linguists because they may give a clue about the stance that the listener takes towards what is said by the speaker (Gardner, 2004). Jefferson described the difference between "yeah" and other backchannels in terms of speaker recipiency, the willingness of the speaker to take the floor (Jefferson, 1984). Yeah utterances make up a substantial part of the dialogue acts in the AMI meeting conversations (about eight percent). "Yeah" is the most ambiguous utterance that occurs in discussion segments in AMI meetings. In order to get information about the stance that participants take with respect towards the issue discussed it is important to be able to tell utterances of "Yeah" as a mere backchannel, from Yeah utterances that express agreement with the opinion of the speaker (see the work of Heylen and Op den Akker (2007)).

The class variables for dialogue act types of Yeah utterances that are distinguished are: Assess (as), Backchannel (bc), Inform (in), and Other (ot). Table 1 gives a distribution of the labels in our train and test data sets. Note that for each annotator, a disjunct train and test set have been defined. The inter-annotator agreement on the Yeah utterances is low. The pairwise alpha values for meeting IS1003d, which was annotated by all three annotators, are (in brackets the number of agreed DA segments that start with "Yeah"): alpha(VK,DH) = 0.36 (111), alpha (VK,S9) = 0.36 (132), alpha(DH,S9) = 0.45 (160).

Testing for Systematic Differences When one suspects the annotations to have originated from different mental conceptions of annotators, the first step is to test whether these differences are systematic. Table 2 presents the intra and inter annotator classification accuracy. There is a clear performance drop between using the test data from the same annotator from which the training data was taken and using the test data of other annotators or the mixed test data of all annotators. This suggest that some of the disagreements in the annotation stem from systematic differences in the mental conceptions of the annotators.

	TEST			
TRAIN	DH	S9	VK	Mixed
DH	**69**	64	52	63
S9	59	**68**	48	57
VK	63	57	**66**	63

Table 2: Performance of classifiers (in terms of accuracy values – i.e. percentage correct predictions) trained and tested on various data sets. Results were obtained with a decision tree classifier, J48 in the Weka toolkit.

Building the Voting Classifier Given the three classifiers DH, S9 and VK, each trained on the train data taken from one single annotator, we have build a Voting Classifier that outputs a class label when all three 'voters' (the classifiers DH, S9 and VK) give the same label, and the label 'unknown' otherwise. As was to be expected, the *accuracy* for this Voting Classifier is much lower than the accuracy of each of the single voters and than the accuracy of a classifier trained on a mix of data from all annotators (see Table 3), due to the many times the Voting Classifier assigns the label 'unknown' which is not present in the test data and is always false. The precision of the Voting Classifier however is higher than that of any of the other classifiers, for each of the classes (see Table 4).

Conclusions for the Voting Classifier For the data that we used in this experiment, building a Voting Classifier as described above gave us a high precision classifier. Based on our starting point, this would relate to the classifier in some way embodying the overlap in the mental conceptions of each of the annotators. If that were true, the cases in which the Voting Classifier returns an unanimous vote would be mostly those cases in which the different annotators would also have agreed.

TRAIN	Accuracy
train_MIX(8838)	67
DH(3585)	63
S9(1753)	57
VK(3500)	63
VotingClassifier(8838)	43

Table 3: Performance of the MaxEnt classifiers (in terms of accuracy values – i.e. percentage correct predictions) tested on the whole test set, a mix of three annotators data (4179 "Yeah" utterances). The first column between brackets the size of the train sets.

	Classifier				
Class	Voting	DH	S9	VK	train_MIX
BC	71	65	63	71	69
AS	73	62	64	61	66
IN	60	58	34	52	50
OT	86	59	32	57	80

Table 4: Precision values per class label for the classifiers.

This can be tested quite simply using multiply annotated data. Note that not *all* data needs to be annotated by more annotators: just enough to test this hypothesis. Otherwise, it will suffice to have enough data for each single annotator, be it overlapping or not. This is especially advantageous when the corpus is really large, such as the 100h AMI corpus. Another way to test the hypothesis that the voting behavior relates to intersubjectivity is to look at the type and context of the agreements between annotators, found in the reliability analysis, and see if that relates to the type and context of the cases where the Voting Classifier renders an unanimous judgement. That would be strong circumstantial evidence in support of the hypothesis.

Note that the gain in precision is obtained at the cost of recall, because the Voting Classifier approach explicitly restricts judgements to the cases where annotators would have agreed and, presumably, therefore to the cases in which users of the data are able to agree to the judgements as well. It is possible that you 'lose' a class label in the classifier by having a high precision but a recall of less than five percent, which in our example happened for the 'other' class.

3 The Classifier as Subjective Entity vs the Classifier as Embodiment of Consensus Objectivity

Many annotation tasks are subjective to a larger degree. When this is simply taken as a given, and the systematic disagreements resulting from the different mental conceptions of the annotators are not taken into account while training a machine classifier on the resulting data, there is no simple reason to assume that the resulting classifier is any less subjective in the judgements it makes. Without additional analyses one cannot suppose the classifier did not pick up idiosyncrasies from the annotators. We have seen that machine classifiers can indeed considered to be subjective in their judgements, a property they have inherited from the annotations they have been trained on. A judgement made by such a classifier should be approached in a similar manner as a judgement made by another person[3]. We will call the resulting classifier therefore a *'subjective entity'* classifier.

A careful analysis of the interannotator agreements and disagreements might make it possible to build classifiers that partly embody the intersubjective overlap between the mental conceptions of the annotators. Because the classifier only tries to give a judgement in situations where one can expect annotators or users to agree, one can approach the judgements made by the classifier as a "common sense" of judgements that people can agree on, despite the subjective quality of the annotation task. We will call the resulting classifier a *'consensus objective'* classifier.

4 Discussion

In the Introduction we distinguished several uses of data annotation using human annotators. The analyses and research in this paper mainly concerns the use of annotated data for the training and development of automatic machine classifiers. Ideally the annotation schema and the class labels that are distinguished reflect the use that is made of the output of the machine classifiers in some particular application in which the classifier operates as a module. Imagine for example a system that detects when meeting participants are too dominant and signals the chairman of the meet-

[3]On a side note, letting the machine classifiers judgments be presented through an embodied conversational agent can be a way to present this human-like subjectivity for the user (Reidsma et al., 2007).

ing to prevent some participants being dissatisfied with the decision making processes. Or, a classifier for addressee detection that signals remote participants that they are addressed by the speaker. The way that users of the system interpret the signals output by the classifier should correspond to the meanings that were used by the annotators and that were implemented in the classifier.

When there is a lot of disagreement in the annotations this should be taken into account for machine learning if one does not want to obtain a 'subjective entity' classifier, the judgements of which the user will often disagree with. In Section 2 we presented two ways to exploit such data for building machine classifiers. Here we elaborate a bit on a difference between the two cases relating to the different *causes* of the inter-annotator disagreement.

For the addressing annotations, the annotators sometimes had problems with choosing between G-addressed and I-addressed. The *participants* in the conversation usually did not seem to have any problem with that. There are only a few instances in the data where the participants explicitly requested clarification. It is reasonable to expect that in cases where it really matters – for the conversational partners – who is being addressed, outside observers will not have a problem to identify this. Thus, in those cases where the annotators had problems to decide upon the type of addressing there maybe was no reason for the participants in the conversation to make that clear because it simply was not an issue. The annotators were then tripped by the fact that they were *forced* by the annotation guidelines to choose one addressee label.

In the dialogue act classification task something additional is going on. Here we see that annotators also have problems because many utterances themselves are ambiguous or poly-interpretable. Some annotator may prefer to call this act an assess where an other prefers to call it an inform, and both may have good reason to back up their choice. A similar situation occurs in the case of the classification of Yeah utterances. The disagreements then seem to be caused more explicitly by differing judgements of a conversational situation.

5 Conclusions

We have argued that dis-agreements between different observers of 'subjective content' is unavoidable and an intrinsic quality of the interpretation and classification process of such type of content. Any subdivision of these type of phenomena into a predefined set of disjunct classes suffers from being arbitrary. There are always cases that can belong to this but also to that class. Analysis of annotations of the same data by different annotators may reveal that there are differences in the decisions they make, such as some personal preference for one class over another.

Instead of throwing away the data as not being valuable at all for machine learning purposes, we have shown two ways to exploit such data, both leading to high precision / low recall classifiers that in some cases refuse to give a judgement. The first way was based on the identification of subsets of the data that show higher inter-annotator agreement. When the events in these subsets can be identified computationally the way is open to use classifiers trained on these subsets. We have illustrated this with several subsets of addressing events in the AMI meeting corpus and we have shown that this leads to an improvement in the accuracy of the classifiers. Precision is raised in case the classifier refrains from making a decision in those situation that fall outside the subsets. The second way is to train a number of classifiers, one for each of the annotators data part of the corpus, and build a Voting Classifier that only makes a decision in case all classifiers agree on the class label. This approach was illustrated by the problem of classification of the dialogue act type of Yeah-utterances in the AMI corpus. The results show that the approach indeed leads to the expected improvement in precision, at the cost of a lower recall, because of the cases in which the classifier doesn't make a decision.

Acknowledgements

The authors are in debt to many people for many fruitful discussions, most prominently Jean Carletta, Ron Artstein, Arthur van Bunningen, Henning Rode and Dirk Heylen. This work is supported by the European IST Programme Project FP6-033812 (AMIDA, publication 136). 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.

References

Ba, S. O. and J.-M. Odobez. 2006. A study on visual focus of attention recognition from head pose in a

meeting room. In Renals, S. and S. Bengio, editors, *Proc. of the MLMI 2006*, volume 4299 of *Lecture Notes in Computer Science*, pages 75–87. Springer.

Bakeman, R. and J. M. Gottman. 1986. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.

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

Craggs, R. and M. McGee Wood. 2005. Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3):289–296.

Galley, M., K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proc. of the 42nd Meeting of the ACL*, pages 669–676. ACL.

Gardner, R. 2004. Acknowledging strong ties between utterances in talk: Connections through right as a response token. In *Proceedings of the 2004 Conference of the Australian Linguistic Society*, pages 1–12.

Goffman, E. 1981. Footing. In *Forms of Talk*, pages 124–159. Philadelphia: University of Pennsylvania Press.

Heylen, D. and H. op den Akker. 2007. Computing backchannel distributions in multi-party conversations. In Cassell, J. and D. Heylen, editors, *Proc. of the ACL Workshop on Embodied Language Processing, Prague*, pages 17–24. ACL.

Jefferson, G. 1984. Notes on a systematic deployment of the acknowledgement tokens 'yeah' and 'mm hm'. *Papers in Linguistics*, 17:197–206.

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

Jovanović, 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*, volume 5 of *The Sage CommText Series*. Sage Publications, Beverly Hills, London.

Passonneau, R. J., T. Yano, T. Lippincott, and J. Klavans. 2008. Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. In *Proc. of the LREC 2008*.

Poesio, M. and R. Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. ACL.

Potter, J. W. and D. Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of applied communication research*, 27(3):258–284.

Reidsma, D. and J. C. Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3).

Reidsma, D., Z. M. Ruttkay, and A. Nijholt, 2007. *Challenges for Virtual Humans in Human Computing*, chapter 16, pages 316–338. Number 4451 in LNAI: State of the Art Surveys. Springer Verlag, Berlin/Heidelberg.

Reidsma, D., D. Heylen, and H. op den Akker. 2008. On the contextual analysis of agreement scores. In *Proc. of the LREC Workshop on Multimodal Corpora*.

Rienks, R. J. 2007. *Meetings in Smart Environments: Implications of progressing technology*. Phd thesis, SIKS Graduate School / University of Twente, Enschede, NL.

Steidl, S., M. Levit, A. Batliner, E. Nöth, and H. Niemann. 2005. "of all things the measure is man" automatic classification of emotion and intra labeler consistency. In *ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing*.

Vieira, R. 2002. How to evaluate systems against human judgment on the presense of disagreement? In *Proc. workshop on joint evaluation of computational processing of Portugese at PorTAL 2002*.

Voorhees, E. M. and D. Harman. 1997. Overview of the trec-5. In *Proc. of the Fifth Text REtrieval Conference (TREC-5)*, pages 1–28. NIST.

Voorhees, E. M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716.

Wiebe, J. M., R. F. Bruce, and T. P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proc. of the 37th Annual Meeting of the ACL*, pages 246–253. ACL.

Wilson, T. 2008. Annotating subjective content in meetings. In *Proc. of the Language Resources and Evaluation Conference (LREC-2008)*.