Subjective Machine Classifiers

Dennis Reidsma Rieks op den Akker

Human Media Interaction, University of Twente, PO Box 217
NL-7500 AE, Enschede, The Netherlands

Abstract

Many interesting phenomena in conversations require interpretative judgements by the annotators. This leads to data which is annotated with lower levels of agreement due to the differences in how annotators interpret conversations. Instead of throwing away this data we show how and when we can exploit it. 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.

Exploiting Subjective Annotated Data

Human classifications of events in terms of rather intuitive notions ask from the annotator to interpret social aspects of human behavior, such as the speaker's intention expressed in a conversation. We argue that dis-agreements between different observers is unavoidable and an intrinsic quality of the interpretation and classification process of such type of content. Any sub-division of these type of phenomena into a pre-defined 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. These difference may reveal some personal preference for one class over another. (For references see the full paper [4].)

Instead of throwing away the data as not being valuable at all for machine learning purposes, we show 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 is 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 illustrate this with several subsets of addressing events in the AMI meeting corpus and we show 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 is 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.

Some types of disagreement are more structural and other types are more noise like. We focus on a way of coping with disagreements resulting from a low level of intersubjectivity that actively exploits the structural differences in the annotations caused by this. From the patterns in the disagreements between annotators, we are able to formulate constraints and restrictions on the use of the data and on the reliability of the classifier's judgements. (see also Reidsma and Carletta [3]).

To illustrate how these ideas work out we used the hand annotated face-to-face conversations from the 100 hour multi-modal AMI meeting corpus [1]. A part of this corpus is annotated (by three annotators) with addressee information. Real dialogue acts were assigned a label indicating who the speaker 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. DAs are either addressed to the group (*G-addressed*) or to an individual (*I-addressed*). Another layer of the corpus contains *focus of attention*

information. so that for any moment it is known whether a person is looking at the table, white board, or some other participant.

The level of agreement with which an utterance is annotated with addressee is dependent on the FOA context of an utterance. 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 utterance based on lexical and multimodal features. 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). The outcome shows a performance gain in contexts with a distinctive addressee-directed focus of attention of the speaker.

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. 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 agreement of the voters, and the 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 *"Yeah-utterances"*, utterances that start with the word "yeah". They make up about eight percent of the dialogue acts in the AMI meeting conversations. 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 [2]).

The class variables for dialogue act types of Yeah utterances that are distinguished are: Assess (as), Backchannel (bc), Inform (in), and Other (ot). For each annotator, a disjunct train and test set have been defined. The inter-annotator agreement on the Yeah utterances is low (pair-wise alpha values are around 0.4).

We train three classifiers DH, S9 and VK, each trained on train data taken from one single annotator, and we build a Voting Classifier that outputs a class label when all three 'voters' give the same label; 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 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.

References

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

[2] D. Heylen and H.J.A. op den Akker. Computing backchannel distributions in multi-party conversations. In J. Cassell and D. Heylen, editors, *Proceedings of the ACL Workshop on Embodied Language Processing, Prague*, volume W07-19, pages 17–24, Prague, Czech Republic, June 2007. Association of Computational Linguistics.

[3] Dennis Reidsma and Jean C. Carletta. Reliability measurement without limits. *Computational Linguistics*, 2008. to appear.

[4] Dennis Reidsma and H.J.A. op den Akker. Exploiting 'subjective' annotations. In Ron Artstein, Gemma Boleda, Frank Keller, and Sabine Schulte im Walde, editors, *Proceedings of the Coling Workshop on Human Judgments in Computational Linguistics*, August 2008. to appear.