Journal of Systems and Information Technology

Predicting meeting participants' note-taking from previously uttered dialogue acts

Antje Bothin Paul Clough

Received 21 July 2015
Revised 20 January 2016
Accepted 22 January 2016

Predicting meeting participants' note-taking from previously uttered dialogue acts

Antje Bothin and Paul Clough
Information School, University of Sheffield, Sheffield, UK

Abstract

Purpose – The purpose of this paper is to describe a new supervised machine learning study on the prediction of meeting participant's personal note-taking from spoken dialogue acts uttered shortly before writing.

Design/methodology/approach – This novel approach of providing cues for finding important meeting events that would be worth recording in a meeting summary looks at temporal overlaps of multiple people's note-taking. This research uses data of 124 meetings taken from the AMI meeting corpus.

Findings – The results show that several machine learning methods that the authors compared were able to classify the data significantly better than a random approach. The best model, decision trees with feature selection, achieved 70 per cent accuracy for the binary distinction writing for any number of participants simultaneously or no writing, whereas the performance for a more fine-grained distinction of the number of participants taking notes showed only about 30 per cent accuracy.

Research limitations/implications – The findings suggest that meeting participants take personal notes in accordance with the utterance of previously uttered speech acts, particularly dialogue acts about disfluencies and assessments appear to influence the note-taking activities. However, further research is necessary to examine other domains and to determine in what way this behaviour is helpful as a feature source for automatic meeting summarisation, which is useful for more efficiently satisfying people's information needs about meeting contents.

Practical implications – The reader of an Information Systems (IS) journal would be interested in this paper because the work described and the findings gained could lead to the development of novel information systems that facilitate the work for businesses and individuals. Innovative meeting capture and retrieval applications, satisfying automatic summaries of important meeting points and sophisticated note-taking tools that suggest content automatically could make people's daily lives more convenient in the future.

Social implications – There are wider implications in terms of productivity and efficiency. Business value is increased for the organisation, as human knowledge is built more or less automatically. There are also cognitive and social implications for individuals and possibly an impact on the society as a whole. It is also important for globalisation, social media and mobile devices.

Originality/value – The topic is new and original, as there has not been much research on it yet. Similar work was carried out recently (Murray, 2015; Bothin and Clough 2014). This is why it is relevant to an IS journal and interesting for the reader. In particular, dialogue acts about disfluencies and assessments appear to influence the note-taking activities. This behaviour is helpful as a feature source

Journal of Systems and Information Technology
Vol. 18 No. 2, 2016
pp. 170-185

1328-7265
DOI 10.1108/JSIT-07-2015-0064

Our thanks go to the creators of the AMI meeting corpus for providing us with large amounts of annotated meeting data. We are also grateful to the anonymous reviewers for their valuable comments.

for automatic meeting summarisation, which is useful for more efficiently satisfying people's information needs about meeting contents.

Keywords Knowledge management, Information management, Data analysis

Paper type Research paper

1. Introduction

Meetings are a vital part of many professional organizations and academic institutions today. They are the events where information exchange and distribution as well as knowledge generation and sharing occur. People usually attend many meetings in the workplace (3M Online Survey, 1998). However, over time, they tend to forget what has happened in the conversations. To satisfy the information needs of meeting participants, who cannot remember what happened in a particular meeting that took place a while ago, or for people who are unable to attend such a gathering at all, a concise meeting summary is necessary. Traditional minutes are sometimes not sufficient because they do not record every detail; above all, they cannot capture emotions and certain discussion elements such as what led to a decision (Whittaker *et al.*, 2008; Renals, 2010). In the business world, it is time-consuming to interrupt co-workers on missed or forgotten meeting content; thus, innovative automatic meeting information capture and retrieval systems are likely to improve employees' productivity (Benson and Standing, 2008).

It is now easily possible to record meetings at low cost and store them online or in a corporate network, but there is usually too much information available to quickly search for what users require to know. To overcome this problem of information overload, meeting browsers (Wellner *et al.*, 2004; Castronovo *et al.*, 2008; ICT Results, 2010) have been developed to display a better overview of a recorded meeting, such as audio, video, speech transcripts, presentation slides, summaries, keywords or personal notes. This facilitates the storage and retrieval of important meeting information and improves corporate memory by providing a better record of such multi-party conversations. The automatic creation of meeting summaries particularly enhances the performance of meeting browser environments because this approach saves time, effort and money, as opposed to producing handcrafted documents about the most informative meeting events. This also increases the productivity, as manual minutes are expensive to create and sometimes incomplete.

Recently, there has been a growing interest in automatic meeting summarization (Buist *et al.*, 2004; Yu and Nakamura, 2010; Wang and Cardie, 2013; Murray, 2015a). Dialogue act (DA) types play a vital role in the meeting discussion, as they are usually meaningful, longer connected parts of the language that express speech acts. Weigand (2016) describes speech acts in more detail. Such utterances usually contain the speaker's intentions and communicative goals. Generally, they can be seen as a suitable information source for summary-worthy utterances in meetings (Wrede and Shriberg, 2003; Hsueh and Moore, 2007). In particular, important decisions and action items occur when certain DAs are uttered in meeting discussions. Thus, it is encouraging to examine whether the occurrence of such DA types is linked to other meeting activities, for example, note-taking.

This paper presents a supervised machine learning study that explores the relationship between DAs and multiple meeting participants' note-taking shortly after

their utterance. As people usually take written notes on salient meeting information, this investigation aims to determine whether the note-taking behavior can be predicted by the preceding DA utterances. To the best of our knowledge, this is the first work that examines this issue.

The remainder of this paper is divided as follows. Section 2 describes previous work on DA and note-taking research. In Section 3, we introduce the hypothesis for this work. Section 4 outlines the data sources we have used in this study – note-taking and DA information extracted from the AMI meeting corpus. In Section 5, we briefly depict the machine learning approaches we have used. Section 6 presents the results of our experiments and discusses their limitations and implications. Finally, Section 7 concludes this paper and provides recommendations for future work in this research area.

2. Previous work

Previous research in the field of meeting understanding and summarization has been carried out by Murray and Renals (2007). They used speech-related cues such as the fundamental frequency and lexical features such as different measures of term frequency counts. Hsueh and Moore (2007) found that summary-worthy meeting elements, for example, decisions occur when DA types such as *information exchange, suggestion* or *assessment* are uttered. Furthermore, disfluencies such as *fragment* or *stall* often occur after the utterance of such decision-related DAs, i.e. very informative and less salient DA types appear to alternate during the course of a meeting discussion. Related work on finding hot spots in naturally occurring meetings, i.e. areas of high participant involvement in the meeting discussion, using the ICSI meeting corpus (Janin *et al.*, 2003) suggests that such important meeting parts co-occur with the utterance of certain DA types (Wrede and Shriberg, 2003). Moreover, their research implies that meeting participants' individual attitudes toward the meeting events and not necessarily the semantic information itself may also be an influencing factor that distinguishes regions containing important meeting discussions from less involved meeting parts.

One way to express meeting participants' thoughts is through private note-taking. However, people need to think about what they consider relevant and subsequently write it down (Gimenez and Pinel, 2013). Recent work on personal note-taking in meetings has shown that people usually record the most important meeting content, particularly decisions and personal action items (Whittaker *et al.*, 2005; Bothin and Clough, 2014). They do this quite frequently, and typically their notes are short points, as opposed to long, grammatically correct sentences.

Nowadays, social summarization has become more and more common, as people like to share information on the Internet. With the arrival of social networking sites, people like to share music, pictures or videos. Here it is difficult to determine the content of the material automatically. In the semantic web philosophy, the idea of tagging is described, i.e. the picture or video obtains a caption as a description of its content. These captions can be seen as an index for the content of the documents. This makes it possible to exploit the user behavior, e.g. the number of clicks on a certain item, especially regarding what multiple people find important and do not want to forget. Research on social summarization demonstrated that multiple people find certain parts of conversations important (Kalnikaite and Whittaker, 2008). They reviewed the same (most informative)

parts of the speech when notes were used as an index into an audio recording of a lecture. For this reason, notes appear to point to the salient content. As human expertise is valuable for businesses, the thought of many people accessing salient information can be applied to recording the salient parts of meetings. As meeting participants are likely to take their personal notes at slightly different times during the meeting, i.e. not exactly at the same moment, group behavior such as note-taking of many people seems to be promising to investigate. Several people may note the same item, but over a larger time-span. Therefore, we decided to investigate multiple people's temporal note-taking overlaps and their dependence on certain DA utterances shortly before the writing.

3. Hypothesis

Until now, few investigations have focused on note-taking in regard to meeting summarization (Banerjee and Rudnicky, 2009; Bothin and Clough, 2012). This research, on the contrary, aims to analyze the importance of note-taking for this purpose. Our long-term goal is to find out whether personal notes predict the salient or informative parts of meetings. The purpose of this study is to show the role DAs and notes play in meeting summarization. As mentioned before, Hsueh and Moore (2007) argue that decisions in meetings are made at certain moments in time. Therefore, we assumed that we can predict when multiple people will all write by looking at the points in the meetings, where important information has been discussed. This study assumed that a segment would be important if multiple people took notes shortly after something informative was uttered. Based on this idea, the following hypothesis was developed:

H1. It is expected that the utterances of certain DA types such as information exchange, suggestion and assessment are more likely to cause note-taking; whereas, the utterances of other DA types, for example, disfluencies, such as stall or fragment, are considered to be unlikely correlated with writing activities.

4. Data

We used 124 meetings taken from the AMI meeting corpus (Carletta, 2006), as this is a large collection of meetings that was also utilized for related work in the field (Hsueh and Moore, 2007). These meetings were based on the scenario of the creation of an innovative TV remote control, but the participants were encouraged to behave as naturally as possible. There were no restrictions on the utterances or the note-taking, only the general agenda items to discuss and training for the functional roles to play were provided. Each meeting was 30 minutes long on average and had four participants with functional roles specified as project manager, user interface expert, industrial designer and marketing expert. The meetings were recorded in a smart meeting room and later features such as DAs and note-taking information were annotated by human experts.

4.1 Participant note-taking

The movement files provided in the AMI meeting corpus were examined for the *take notes* tag, and the start and end times of note-taking items for each meeting participant were extracted. Thereafter, temporal note-taking overlap windows of two-minute duration were created. The note-taking moments were individually different; however, within this time-span, we found several overlaps of all meeting participants. Figure 1 shows the approach. This example illustrates a "four-participant note-taking overlap" because this

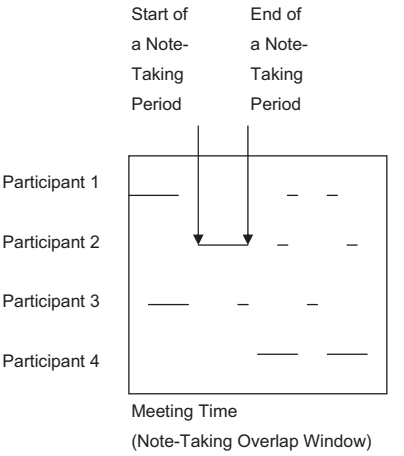

Figure 1.
Example of a note-taking overlap window

number of people is writing in the time-span of the temporal note-taking overlap window. In the figure, there are many short lines for each participant's note-taking. The beginning of such a line shows the start of the note-taking for this person, and when the line ends, note-taking is stopped. Each meeting participant took notes for several seconds and then did not take notes any more, which was followed by another time-span of note-taking and so on. Thus, we can see many short lines of note-taking for each participant at different moments in time compared to the other participants.

We calculated how many participants wrote simultaneously in such a temporal note-taking overlap window: 0, 1, 2, 3 or 4. We also applied a second method of counting note-taking overlap, where we distinguished nobody taking notes from any number of participants writing. The class distribution for the count of note-taking overlap was as follows. There were 1,866 note-taking overlap windows in total to examine. In case of the fine-grained distinction, i.e. 0, 1, 2, 3 or 4 participants writing simultaneously, no meeting participants were writing in 563 windows, which is 30 per cent of the total number of note-taking overlap windows. In all, 524 instances (28 per cent of the total) showed one participant taking notes, and two meeting participants recorded something in 335 note-taking overlap windows (18 per cent of the total number of windows). A three-people writing overlap occurred in 308 windows (17 per cent), and four participants wrote in 136 windows (7 per cent of all note-taking overlap windows). For the binary version, no or any number of people taking notes simultaneously, there were 563 instances of no writing and 1,303 cases of writing.

Furthermore, there was an issue of the processing time of a spoken utterance and the cognitive effort of note-taking (Piolat *et al.*, 2005); the participants cannot write before they have heard and understood the utterance. As we intended to compare the note-taking activities to the preceding DAs, a one-minute period of time appeared to be the best for the delay between the utterance of DAs and subsequent writing.

4.2 Dialogue act information

The start and end times of DA occurrences in the meeting speech and the DA types were extracted from the AMI corpus. To analyze the DA data, we divided each meeting into

one-minute windows, as this was a good basis for efficient computations. The total number of occurrences of each DA type in each window was calculated. There are 15 different types of DAs in the corpus. DA types with a speaker intention include: *information exchange*, *elicit information exchange*, *suggestion*, *offer*, *elicit offer or suggestion*, *assessment*, *elicit assessment*, *comment about understanding*, *elicit comment about understanding*, *social behavior be positive*, *social behavior be negative* and *other*. Contrary to that, DAs without a speaker intention are *backchannel*, *stall* and *fragment*.

We considered these DA types interesting for our experiment for the following reasons. *Information exchange* is an indicator of summary worthiness because facts are usually important for meeting minutes; therefore, decisions might have been made or action items might have been allocated at this time, and this is what people usually record in their personal notes (Khan, 1992; Whittaker *et al.*, 2008). These events should be part of the meeting summary. *Suggestion* and *offer* can be seen as actions or to do items that are personally important to the participants involved. Thus, such DA types look promising for assuming subsequent writing after their utterance. *Assessment* evaluates actions and might lead to decisions that should be included in a summary and in the notes of the meeting participants. *Comments about understanding* were seen as events to reassure that acoustic problems did not occur. This DA type seems unlikely to be of interest for the creation of automatic summaries, but might give participants time to think about the meeting discussion. This may then lead to note-taking activities. *Be positive* can be related to progress making or occur before criticizing, whereas *be negative* can mean the group has problems or does not make any progress. *Other*, e.g. talking to oneself, seems to be questionable again and not summary-worthy. This was the category for the rest of any DAs that did not fit into any other DA type in the annotation scheme.

Recent research on (dis-)agreement detection in meetings suggests that *backchannels*, for example, utterances, such as "yeah" as confirmations that one is listening, could be misinterpreted as signs of agreement (Germesin and Wilson, 2009); contrary to that, this DA type was found to be correlated with non-involvement (Wrede and Shriberg, 2003). Disfluencies such as *fragments* and *stalls* can be considered as floor holders; therefore, it is likely that something important is being discussed when people are trying to keep the attention of the group and have a desire to keep on speaking. However, it was found that floor holders occurred in less involved meeting regions because the speaker can afford to pause there without fear of losing the right to talk (Wrede and Shriberg, 2003). Also, these DA types may provide additional time for note-taking.

Information about *laughter* and *coughs* was also extracted from the AMI corpus. Our experimentation revealed that this generally improved the performance of the models. Laughter is an indicator of amusement or agreement that can lead to decisions; whereas, coughs usually occur without a speaker intention, but can be signs of disagreement (Niekrasz and Purver, 2005).

We used two approaches for our experiments:

(1) all 17 features (15 DA types + laughter and coughs); and

(2) feature selection.

In case of the latter, we determined the optimal subset of features by examining the predictive power of each individual feature using correlation-based attribute selection in Weka (Witten and Frank, 2005). This method selects features that are highly correlated with the class while having low correlation with other features. The selected features for

the second technique were *laughter*, *stall* and *fragment* for the five groups definition of note-taking overlap, i.e. 0, 1, 2, 3 or 4 meeting participants writing at the same time, and *laughter*, *fragment*, *assessment* and *social behavior be positive* for the binary case – no or any number of people writing simultaneously in a note-taking overlap window.

Additionally, we used a model that combined several DA types into six broad groups, i.e. *information exchange*, *disfluencies*, *evaluation and comments*, *actions*, *social behavior* and *other*. To obtain these DA feature groups, we added up the number of the single DA types in the AMI corpus belonging to each broad group. The DAs *information exchange* and *elicit information exchange* together form the first group. *Fragment*, *stall* and *backchannel* are disfluencies. The third group consists of *assessment*, *elicit assessment*, *comment about understanding* and *elicit comment about understanding*. *Suggestion*, *offer* and *elicit offer or suggestion* are considered as actions. Social behavior can be *positive* (amusement, politeness) or *negative* (bad comments). *Other* is again reserved for all DAs that remain and do not belong to any other group.

Again, we used two ways for counting the writing overlaps – all pen number values separately (0, 1, 2, 3 or 4 participants writing simultaneously in the overlap window) and only 0 or 1 for no pens down and 1-4 pens down as one single group, respectively. Weka's attribute selection chose only some of the six broad DA features to reduce complexity and minimize the risk of overfitting. The following features were selected: for five groups for number of people writing: *disfluencies;* for binary – no *or* any number of people writing: *evaluation and comments*, *social behavior* and *disfluencies*.

5. Machine learning methodology

In all, four different machine learning algorithms were applied and compared using the Weka toolkit (Witten and Frank, 2005): decision tree learning, the Naïve Bayes classifier, Bayesian networks and a random tree approach. We used the DA features as attributes or independent variables and the note-taking information (number of pens down per temporal note-taking overlap window) as class or dependent variable in the supervised learning approaches.

Ten-fold cross-validation was carried out, i.e. the data set was divided into ten equal parts or folds. Each fold in turn was taken as the test set and the remaining folds formed the training set; this procedure was executed ten times in a row. To increase the reliability of the results, we conducted the ten-fold cross-validation runs ten times, i.e. we carried out a total of 100 cycles for each machine learning algorithm we utilized. In the following, we briefly describe the machine learning techniques.

5.1 Decision tree learning

We used the C4.5 algorithm (Quinlan, 1993). This method uses decision trees and selects features based on the information gain, i.e. the expected reduction in entropy, to determine the outcome. We utilized this approach because it has been successfully used in prior work, for example, on topic detection (Banerjee and Rudnicky, 2006) and agreement detection (Hillard *et al.*, 2003; Wrede and Shriberg, 2003; Germesin and Wilson, 2009) in meetings. Decision tree learning is robust to noisy data; however, this machine learning method tends to favor a long tree that overfits the training data and does not generalize well (Witten and Frank, 2005); thus, tree pruning needs to be applied.

5.2 The Naïve Bayes Classifier

The Naïve Bayes classifier is a probabilistic learner that is frequently used for text classification in natural language processing (Sebastiani, 2002). It is based on the Bayes theorem of conditional probabilities, which is used by the classifier to compute the predicted class probability given the values of the other features for each instance. The method assumes that all features are equally important for the classification task and statistically independent (Mitchell, 1997; Sebastiani, 2002).

5.3 The Bayesian networks

This technique extends the Naïve Bayes method, i.e. it allows the predictor variables to be dependent from each other (Sebastiani, 2002). We used this approach as we expected dependencies among the DA features, and this method is suitable for accounting for them. Thus, it is possibly a good choice for examining whether it provides a better model.

5.4 The random tree approach

This method builds a tree that selects attributes randomly. In our experiments, it was used as a baseline to compare the performance of the other machine learning algorithms.

6. Results and discussion

6.1 Single dialogue act types

6.1.1 Algorithm comparison for all dialogue act features. Table I illustrates that the best model, the Bayesian network approach, achieved 30.02 per cent accuracy in the all the 17 DA features case for five note-taking overlap classes and 69.42 per cent in the binary case – no or any number of meeting participants taking notes in a writing overlap window. This is significantly better than a random tree method in both cases. In the binary case, the Naïve Bayes classifier (65.19 per cent accuracy) and the decision tree learning (63.69 per cent) also performed significantly better than the random technique (59.52 per cent).

6.1.2 Algorithm comparison for feature selection. Using feature selection, we observed that for the best models, Bayesian network and decision tree learning, 30.02 per cent and 70.37 per cent of the data were correctly classified for the five classes and the binary approach respectively. Both were significantly different from the random method (25.46 per cent and 60.79 per cent accuracy, respectively). This is given in Table II. The Naïve Bayes algorithm and decision tree learning improved for feature selection, but the Bayesian network did not. This is because, usually, feature selection

Algorithm/class description	Random	C4.5 decision tree	Naïve Bayes	Bayesian network
Five groups for number of participants writing in note-taking overlap window	25.40% (2.97)	26.48% (3.07)	26.15% (3.24)	30.02% (1.94)**
Binary – no or any number of participants writing in note-taking overlap window	59.52% (3.58)	63.69% (2.80)**	65.19% (2.58)**	69.42% (1.19)**

Notes: **Means significantly better than random; $p < 0.05$; standard deviation in brackets

Table I. Accuracy of machine learning approaches for all 17 dialogue act features

works well with the Naïve Bayes, as it removes redundant, i.e. dependent, features. Table II shows that the Naïve Bayes (29.78 per cent and 67.18 per cent accuracy, respectively) was significantly better than the random for both methods of counting note-taking overlap. The confusion matrices for the best models for both cases of note-taking overlap definitions are given in Tables III and IV.

6.1.3 Detailed results for the best models. In Table III, we observe that most instances were classified as class 0 (no writing) and 441 cases were correctly predicted, whereas 1,019 were misclassified as class 0. There was a trend to predicting no writing in the data set regardless of the real class values. Class 1 was also predicted often. Classes 2 and 4 were not predicted at all, although there were 335 and 136 instances in these classes in reality. It appears that there was not enough data to classify these examples correctly. We found many misclassified instances for the 0 and 1 classes, where the majority of the data were situated. There were obviously no decisive DA features in the data set that could perform a more accurate separation of the classes.

Table IV shows the confusion matrix for the binary case, where we observed 550 misclassified instances in total. Almost all of the class 1 instances were correctly

Table II.
Accuracy of machine learning approaches for dialogue act feature selection

Algorithm/class description	Random	C4.5 decision tree	Naïve Bayes	Bayesian network
Five groups for number of participants writing in note-taking overlap window	25.46% (2.78)	27.23% (2.73)	29.78% (3.16)**	30.02% (1.94)**
Binary – no or any number of participants writing in note-taking overlap window	60.79% (3.34)	70.37% (1.48)**	67.18% (2.53)**	69.42% (1.19)**

Notes: **Means significantly better than random; $p < 0.05$; standard deviation in brackets

Table III.
Confusion matrix for feature selection for five groups of meeting participants (0, 1, 2, 3 or 4 people) writing simultaneously (Bayesian network)

	Classified as				
Class	0	1	2	3	4
0	441	101	0	21	0
1	399	108	0	17	0
2	257	69	0	9	0
3	252	45	0	11	0
4	111	16	0	9	0

Table IV.
Confusion matrix for feature selection for the binary case

	Classified as	
Class	0	1
0	59	504
1	46	1,257

Note: No (0) or any number (1) of meeting participants writing simultaneously (decision tree learning)

predicted, 1,257 out of 1,303, but the class 0 cases were not. Only 59 instances out of 563 were identified as class 0. This is because of the uneven data distribution in the data set.

6.2 Six broad dialogue act groups

6.2.1 Algorithm comparison for all dialogue act features using six broad dialogue act groups. Table V shows the results of the machine learning approaches used in case of all six broad DA group features. We can see that the Bayesian network has the highest accuracy for five classes of note-taking behavior, 29.98 per cent; this is significantly better than the random approach. For the binary case, any number of people writing or no writing at all, we observed 69.83 per cent accuracy for the Bayesian network, which is also significantly better than the random.

6.2.2 Algorithm comparison for feature selection using six broad dialogue act groups. The results for the six broad DA groups with feature selection are shown in Table VI. The Naïve Bayes improved through feature selection and gained the highest accuracy, 30.75 per cent, in the five class case. This finding is significantly better than the random. For the two class case, we found 70.11 per cent accuracy for the decision trees, which is also significantly different from the random result of 58.27 per cent.

6.2.3 Detailed results for the best models (six broad dialogue act groups). The Naïve Bayes classifier and decision tree learning were best in this study. The confusion matrix

Algorithm/class description	Random	C4.5 decision tree	Naïve Bayes	Bayesian network
Six DA groups – five groups for number of participants writing in note-taking overlap window	23.91% (3.28)	25.46% (3.37)	29.04% (3.33)**	29.98% (1.54)**
Six DA groups – binary – no or any number of participants writing in note-taking overlap window	59.26% (3.26)	69.79% (1.66)**	67.67% (2.18)**	69.83% (0.21)**

Notes: **Means significantly better than random; $p < 0.05$; standard deviation in brackets

Table V. Accuracy of machine learning approaches for all six broad dialogue act group features

Algorithm/class description	Random	C4.5 decision tree	Naïve Bayes	Bayesian network
Six DA groups with feature selection – five groups for number of participants writing in a note-taking overlap window	28.36% (2.47)	28.75% (1.72)	30.75% (2.12)**	29.98% (1.54)
Six DA groups with feature selection – binary – no or any number of participants writing in a note-taking overlap window	58.27% (3.09)	70.11% (1.34)**	68.47% (2.14)**	69.83% (0.21)**

Notes: **Means significantly better than random; $p < 0.05$; standard deviation in brackets

Table VI. Accuracy of machine learning approaches for six broad dialogue act groups with feature selection

for the five class case is presented in Table VII. We see that again classes 2, 3 and 4 were not predicted at all. We observed all instances to be classified as class 0 or 1, which does not reflect the real class distribution. There is a bias in the broad DA feature data toward no writing or one person writing only. It appears that the feature combination that we realized through producing the six DA groups, as opposed to 17 separate DA features, influenced the outcome in the direction of a smaller proportion of note-taking. Table VIII shows that for the binary case, we misclassified 559 instances. We only classified 51 correctly for class 0, where actually 563 would have been expected. Thus, we can only conclude that this data set contains too much variation for a more accurate prediction.

6.3 Limitations and implications

Our experiments showed that the machine learning approaches were fairly predictive. We achieved about 30 per cent and 70 per cent correctly classified for the five classes and the binary case, respectively. This was the case for the 17 single DA types as well as the six broad DA groups. All methods that had the highest accuracy for the case they examined were better than the random method. This suggests that there is a relationship between DAs and note-taking shortly afterwards. *H1* was partially supported. There was evidence in favor of the first part of *H1* (DA type *assessment*), but no evidence for the second part of *H1*, as *disfluencies* seemed to influence the note-taking behavior in the meetings we examined.

One limitation of this work is that we observed some misclassifications, which have been caused by a lack of training examples for every feature and each class. This is especially true for the five class case, where we conclude that our approach was too fine-grained compared to the binary discrimination of classes. We had an unequal class distribution in the original data set; thus, the number of instances per class biased the machine learning methods toward the majority class. To improve the performance, we can give a certain class a weight, i.e. make use of cost-sensitive learning (Witten and Frank, 2005). Cost-sensitive classifiers can be built. This allows us to control the

Table VII.
Confusion matrix for feature selection for five groups of meeting participants (0, 1, 2, 3 or 4 people) writing simultaneously for six broad dialogue act groups (Naïve Bayes)

	Classified as				
Class	0	1	2	3	4
0	463	100	0	0	0
1	413	111	0	0	0
2	269	66	0	0	0
3	249	59	0	0	0
4	114	22	0	0	0

Table VIII.
Confusion matrix for feature selection for the binary case

	Classified as	
Class	0	1
0	51	512
1	47	1,256

Note: No (0) or any number (1) of meeting participants writing simultaneously for six broad dialogue act groups (decision tree learning)

outcome regarding the number of false negatives and false positives. In this study, we thought these two types of errors should be treated as equally important, as we did not know what DAs encourage the meeting participants to take notes. Overall, it is clearly a research challenge how to alleviate the feature value scarceness and the class distribution imbalance we found in this data set.

We also need to account for the fact that 17 DA features are complicated to interpret and may have caused the algorithms to overfit. We used feature selection to provide a remedy for this problem and achieved a slightly better outcome for the decision trees and Naïve Bayes, but the Bayesian network remained unchanged. The latter is simply more robust against redundant features in the data set. The Naïve Bayes was strongly influenced by redundant attributes, as the improvement with feature selection shows, but the Bayesian network was not. The feature selection reduced the number of DA features, which made these approaches generally more interpretable. Our experiment suggests that disfluencies such as *fragment* appeared to influence the note-taking activities. This could be because of the short pause such an event causes in the meeting discussion that might provide time for note-taking. As s*ocial behavior be positive*, *laughter* and *assessment* were chosen by the feature selection approach, it appears that behavior that is related to amusement or agreement controlled the participants' note-taking activities. Research has found that *assessment* is a good predictor of agreement or disagreement (Germesin and Wilson, 2009); thus, areas of assessment are likely to lead to decision-making and, therefore, eventually to salient meeting events.

In the six broad DA groups study, it was also confirmed that certain DAs may cause writing of multiple participants in the meetings, whereas others may not. DAs about *disfluencies* that do not contain much meaningful information sometimes appear to make people write, which is because of the additional time for note-taking they provide during a meeting discussion. However, future work is required to improve our understanding of which DA types are especially interesting here.

Another limitation of this study is that we obtained our results using only one collection of meeting recordings. Thus, it is of interest to conduct further research on other types of meetings, for example, naturally occurring conversations and discussions in a real setting, e.g. in an academic environment, to examine whether the findings for this data set generalize well across different domains.

On the whole, the latest analyses merely provided moderate findings, but there seems to be at least some dependence between the utterance of DAs and subsequent note-taking in meetings. The machine learning experiments showed that approximately 30 per cent of the data can be correctly classified when we use five classes to describe note-taking; however, about 70 per cent were predicted correctly for a binary approach – no note-taking or any number of participants writing simultaneously. This implies that there may be a relationship between the utterance of certain DA types and note-taking thereafter. Thus, it looks encouraging for examining note-taking and summarization further.

The results we gained are important for the development of novel information systems because the insights could be implemented in new tools. Innovative applications could be built that take notes from what was said automatically. This would increase the productivity of the members of staff. If minutes were produced automatically, they would have the most important meeting content readily available.

This would also benefit the organization as a whole, as workers would perform their tasks more efficiently.

7. Conclusions and future work

This paper presented novel studies on the relationship between DAs uttered in the meeting discussion and subsequent meeting participant note-taking of multiple people.

The results of the first experiment revealed that the note-taking can be predicted with 70 per cent accuracy when using the decision tree approach and feature selection as well as a binary distinction of note-taking or no note-taking as class variable. However, in case of a more fine-grained distinction, i.e. 0, 1, 2, 3 or 4 meeting participants writing simultaneously, the best machine learning approach, the Bayesian network, achieved only 30 per cent accuracy. These best models were both significantly better than the random method, which suggests that a certain amount of learning had successfully occurred. The feature selection improved the decision trees and Naïve Bayes only slightly. DAs about *disfluencies* and *assessments* were selected in our approach, but the examination of the suitability of specific DA types as well as note-taking cues as features for automatic meeting summarization requires further research.

From experiment 2, it can also be concluded that there may be a relationship between the six broad DA types and note-taking shortly after their utterance. However, we can predict the note-taking behavior to a moderate extent only, i.e. the accuracy of the best machine learning method was 30 per cent in case of a fine-grained approach for the number of people taking personal notes simultaneously using the Naïve Bayes and feature selection, and about 70 per cent in case of the simple distinction writing or not using decision trees with feature selection. DAs that contain potential facts and salient points are likely to cause writing of multiple participants in the meetings. *Disfluencies* and other DA types that do not contain much meaningful information also appeared to make people write, which may be because of the additional time for note-taking they provide during a meeting discussion. Furthermore, as such disfluencies often occur after decision DAs (Hsueh and Moore, 2007), people are likely to note the decisions at such moments. In information systems that summarize meetings, one could assign weights to the DAs. According to the score, meeting utterances could then be grouped in categories such as "in summary", "not in summary" and "could be in summary" (Bothin and Clough, 2012). Different features, not just DA information, could be used and combined to improve automatic summarization. For example, topic-specific cues or participant role information in the discussion could benefit the automatic creation of meeting summaries.

Overall, our findings provide evidence in support of a link between the speech-act level, i.e. DAs, and people's note-taking activities shortly after these utterances; however, it is necessary to conduct additional work on other meeting topics, e.g. academic meetings, to investigate whether the results generalize well across different domains.

The wider implications of this work are connected to globalization. In business, people often use remote meetings where the participants are anywhere in the world. In this case, it would be very interesting to have novel tools that facilitate the proceedings. Content of international, multi-lingual meetings could be automatically translated. All this is likely to increase the productivity of the participants. Meeting productivity shifts

have been recently examined (Murray, 2015b), but they might still be more interesting for further research in the future.

As mobile communication is important today (Carrascal *et al.*, 2012), one could examine how people remember and retrieve information from informal conversations such as business phone calls or Twitter messages. User-specific applications and summaries may also be created that better fulfill individual information needs.

Future work could also be concerned with generating abstractive summaries (Murray, 2015a), which do not just copy salient information from the document but generate new language. This is what users might prefer (Bothin and Clough, 2014) because of readability issues. The results from this study also help build such type of sophisticated summaries.

A useful application of notes used for displaying important meeting events is to provide shared note-taking areas in a meeting browser environment that display concise keyword summaries, as this would allow insights into other people's understanding of the meeting by being able to see other participants' notes. This would also support computer supported co-operative work. For example, one could build an information system that allows users to edit and label their own notes. These labels would help humans find certain points, but the automatic system would also use them for classification.

Notes and labels can be used as indices into an audio or video recording of a meeting. This makes it easier for people to find things. They can re-listen or re-watch content quickly and accurately, which may prevent businesses from legal difficulties. Highlighted keywords would also help non-native speakers. Clearly-arranged tag-clouds could be produced and presented. A new information system should also provide accurate data storage facilities.

It is also encouraging to investigate how potential note-worthy items can be automatically suggested to the meeting participants. The system should learn permanently and adapt to new circumstances. This would decrease the workload and tediousness of detailed note-taking for the individuals, thus, leading to more efficient meeting discussions, which is a step toward facilitating group work. This is of utmost organizational relevance. Excellent employee knowledge increases the business value of companies. Therefore, corporate training, specifically in skills such as computer literacy, note-taking, negotiating, convincing people in meetings, etc., is also important for the long-range success of new information systems. There are also cognitive and social implications for the individuals. People have more time to think and participate in the meeting discussions. This may lead to better motivated employees and subsequently to better performance.

It is also imaginable in the future that applications using virtual characters or robots could be built that interact with people and help them do their work (Yumak and Magnenat-Thalmann, 2016). These tools need to understand the things discussed to assist meeting participants during and after a meeting.

Our findings also have an impact on society. New information systems tend to influence politics and culture. Innovative applications may lead to behavior changes; for example, the advent of sophisticated mobile devices led to more flexibility with audio and video data and digital note-taking using pen-based devices (Lackey *et al.*, 2014a, 2014b). Important meeting content may be displayed on handheld devices and give a concise overview of recent work activities, thus increasing people's efficiency and making the daily job more convenient in general.

References

3M Online Survey (1998), available at: www.3m.com/meetingnetwork/ (accessed 14 July 2007).

Banerjee, S. and Rudnicky, A. (2006), "You are what you say: using meeting participants' speech to detect their roles and expertise", *Proceedings of the Analyzing Conversations in Text and Speech Workshop at HLT-NAACL*.

Banerjee, S. and Rudnicky, A. (2009), "Detecting the noteworthiness of utterances in human meetings", *Proceedings of SIGDIAL 2009, London*, pp. 71-78.

Benson, S. and Standing, C. (2008), *Information Systems: A Business Approach*, John Wiley & Sons, Milton Old.

Bothin, A. and Clough, P. (2012), "Participants' personal note-taking in meetings and its value for automatic meeting summarisation", *Information Technology and Management*, Vol. 13 No. 1, pp. 39-57.

Bothin, A. and Clough, P. (2014), "A user evaluation study: do participants' personal notes help us to summarise meetings?", *Knowledge and Process Management*, Vol. 21 No. 2, pp. 122-133.

Buist, A., Kraaij, W. and Raaijmakers, S. (2004), "Automatic summarization of meeting data: a feasibility study", *Proceedings of the Meeting of Computer Linguistics in the Netherlands 2004, Leiden*.

Carletta, J. (2006), "Announcing the AMI meeting corpus", *The ELRA Newsletter*, Vol. 11 No. 1, pp. 3-5.

Carrascal, J., De Oliveira, R. and Cherubini, M. (2012), "A note paper on note-taking: understanding annotations of mobile phone calls", *Proceedings of the 14th Conference on Human-Computer Interaction with Mobile Devices and Services 2012*.

Castronovo, S., Frey, J. and Poller, P. (2008), "A generic layout-tool for summaries of meetings in a constraint-based approach", *Proceedings of MLMI 2008, Utrecht*, pp. 248-259.

Germesin, S. and Wilson, T. (2009), "Agreement detection in multiparty conversation", *Proceedings of ICSI-MLMI 2009, Cambridge*, pp. 7-13.

Gimenez, G. and Pinel, J. (2013), "A proposed method of group observation and note-taking from a psychoanalytical perspective", *Group Analysis*, Vol. 47, pp. 42-61.

Hillard, D., Ostendorf, M. and Shriberg, E. (2003), "Detection of agreement vs disagreement in meetings: training data with unlabeled data", *Proceedings of NAACL HLT 2003*.

Hsueh, P.-Y. and Moore, J. (2007), "What decisions have you made: automatic decision detection in conversational speech", *Proceedings of NAACL HLT 2007, Rochester*, pp. 25-32.

ICT Results (2010), "The internet of meetings", *Science Daily*, available at: http://sciencedaily.com/releases/2010/04/100423215026.htm (accessed 23 August 2010).

Janin, A., Baron, D., Edwards, D., Ellis, D., Gelbart, N., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. and Wooters, C. (2003), "The ICSI meeting corpus", *Proceedings of ICASSP 2003, Hong Kong*.

Kalnikaite, V. and Whittaker, S. (2008), "Social summarisation: does social feedback improve access to speech data?", *Proceedings of CSCW 2008, San Diego, CA*.

Khan, F. (1992), "A survey of note-taking practices", Technical Report, Hewlett Packard Laboratories.

Lackey, A., Moshiri, M., Pandey, T., Lall, C., Lalwani, N. and Bhargava, P. (2014a), "Productivity, part 1: getting things done, using e-mail, scanners, reference managers, note-taking applications, and text expanders", *Journal of the American College of Radiology*, Vol. 11 No. 5, pp. 481-489.

Lackey, A., Pandey, T., Moshiri, M., Lalwani, N., Lall, C. and Bhargava, P. (2014b), "Productivity, part 2: cloud storage, remote meeting tools, screencasting, speech recognition software, password managers, and online data backup", *Journal of the American College of Radiology*, Vol. 11 No. 6, pp. 580-588.

Mitchell, T. (1997), *Machine Learning*, The McGraw-Hill Companies, New York, NY.

Murray, G. (2015a), "Abstractive meeting summarization as a Markov decision process", *Lecture Notes in Computer Science, Advances in Artificial Intelligence*, Springer, pp. 212-219.

Murray, G. (2015b), "Analyzing productivity shifts in meetings", *Lecture Notes in Computer Science, Advances in Artificial Intelligence*, Springer, pp. 141-154.

Murray, G. and Renals, S. (2007), "Term-weighting for summarisation of multi-party spoken dialogues", *Proceedings of MLMI 2007, Brno*.

Niekrasz, J. and Purver, M. (2005), "A multimodal ontology for meeting understanding", *Proceedings of MLMI 2005, Edinburgh*.

Piolat, A., Olive, T. and Kellogg, R. (2005), "Cognitive effort during note-taking", *Applied Cognitive Psychology*, Vol. 19, pp. 291-312.

Quinlan, R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers, San Mateo, CA.

Renals, S. (2010), "Recognition and understanding of meetings", *Proceedings of HLT 2010, Los Angeles, CA*, pp. 1-9.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Survey*, Vol. 34 No. 1, pp. 1-47.

Wang, L. and Cardie, C. (2013), "Domain-independent abstract generation for focussed meeting summarization", *Proceedings of ACL*.

Weigand, E. (2016), "The dialogic principle revisited: speech acts and mental states", *Interdisciplinary Studies in Pragmatics, Culture and Society*, Vol. 4, pp. 209-232.

Wellner, P., Flynn, M. and Guillemot, M. (2004), "Browsing recorded meetings with ferret", *Proceedings of MLMI, Martigny*, pp. 12-21.

Whittaker, S., Laban, R. and Tucker, S. (2005), "Analysing meeting records: an ethnographic study and technological implications", *Proceedings of MLMI 2005, Edinburgh*.

Whittaker, S., Tucker, S., Swampillai, K. and Laban, R. (2008), "Design and evaluation of systems to support interaction capture and retrieval", *Personal and Ubiquitous Computing*, Vol. 12, pp. 197-221.

Witten, I. and Frank, E. (2005), *Data Mining – Practical Machine Learning Tools and Techniques*, Morgan Kaufman Publishers, San Francisco, CA.

Wrede, B. and Shriberg, E. (2003), "The relationship between dialogue acts and hot spots in meetings", *Workshop on Automatic Speech Recognition and Understanding*, pp. 180-185.

Yu, Z. and Nakamura, Y. (2010), "Smart meeting systems: a survey of state-of-the-art and open issues", *ACM Computing Surveys*, Vol. 42 No. 2.

Yumak, Z. and Magnenat-Thalmann, N. (2016), "Multimodal and multi-party social interactions", *Context Aware Human-Robot and Human-Agent Interaction*, pp. 275-298.

Corresponding author
Antje Bothin can be contacted at: a.bothin@sheffield.ac.uk