A Generic Machine Learning based Approach for Addressee Detection in Multiparty Interaction

Usman Malik
Normandie University, INSA Rouen, LITIS
usman.malik@insa-rouen.fr

Mukesh Barange
Normandie University, INSA Rouen, LITIS
mukesh.barange@insa-rouen.fr

Naser Ghannad
Normandie University, INSA Rouen, LITIS
naser.ghannad@insa-rouen.fr

Julien Saunier
Normandie University, INSA Rouen, LITIS
julien.saunier@insa-rouen.fr

Alexandre Pauchet
Normandie University, INSA Rouen, LITIS
alexandre.pauchet@insa-rouen.fr

ABSTRACT

Addressee detection is one of the most fundamental tasks for seamless dialogue management and turn taking in human-agent interaction. Whereas addressee detection is implicit in dyadic interaction, it becomes a challenging task in multiparty interactions when more than two participants are involved.

Existing research works employ either rule-based or statistical approaches for addressee detection. However, most of these works either have been tested on a single data set or only support a fixed number of participants. In this article, we propose a model based on generic features to predict the addressee in data sets with varying number of participants. The results tested on two different corpora show that the proposed model outperforms existing baselines.

CCS CONCEPTS

• **Human-centered computing** → *Natural language interfaces*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Human-Agent Interaction, Mixed Communities, Multiparty Interaction, Multimodal Interaction, Machine Learning

ACM Reference Format:
Usman Malik, Mukesh Barange, Naser Ghannad, Julien Saunier, and Alexandre Pauchet. 2019. A Generic Machine Learning based Approach for Addressee Detection in Multiparty Interaction. In *IVA 2019*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

A conversational agent has to perform several key tasks during a dialogue, such as speaker identification, intent classification or addressee detection. In dyadic interactions, involving only two participants, addressee detection is straightforward whereas in a multiparty context this detection becomes an issue. In fact, a speaker can address any specific participant, a subset of participants or all the participants. In other words, a conversational agent needs not only to identify if it is being addressed, but also when any other participant is being addressed.

In a typical multiparty interaction, the speaker expresses an intention in the form of a dialogue act (DA). A DA can be defined as the meaning of an utterance at the level of illocutionary force [23]. As per Goffman [6], a DA and the supporting utterance can be addressed to three types of addressees: *over-hearers* who are not concerned by the interaction and whose dialogue states are thus not changed; *participants* whose dialogue states are changed by the speaker utterance but are not directly addressed to, and finally the direct *addressees* of the DA. This article only focuses on the direct addressee detection problem and henceforth the term "*addressee*" further refers to the direct addressee(s) of an utterance. Thus *direct addressees* are defined as "*those ratified participants 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*" [6]. In this way, identifying the current addressee(s) also helps the agent in identifying who should be the next speaker. This is particularly important when the agent is (one of) the addressee(s), to infer when to contribute to the discussion.

To tackle the problem of addressee detection in multiparty interaction, both heuristic and statistical approaches have been developed in the literature. However, most of these works depend on specific settings. Limited amount of training data also makes it difficult to develop a generic addressee detection model. In this article, we hypothesize that a statistical model with generic features could perform well on the addressee detection task in multiple scenarios.

This article is organized into 6 sections. Section 2 reviews some of the existing works on addressee detection. Section 3 presents the model details along with the experimental data sets. Section 4 and 5 describes respectively experiments and implementation. Section 6 concludes the article.

2 LITERATURE REVIEW

This section presents an overview of relevant related works along with approaches and features used for addressee detection.

2.1 Approaches for Addressee Detection

The seminal work by Traum *et al.* proposes a rule based approach exploiting previous utterance, current utterance, immediate previous speaker and current speaker to detect the addressee. The accuracy varies between 65% and 100% on Mission Rehearsal Dataset [26]

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IVA'19, July 02–05, 2019, Paris, France
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Table 1: Existing Addressee Detection Approaches for Multiparty Interaction

Reference	Approach	Dataset	Salient Features	Accuracy	Accuracy on AMI	Limitations
Traum et al., 2006	Rule Based	Mission Rehearsal Exercise	Current & previous utterance and speaker, current and previous DA	65-100%	36%	Low accuracy, not generic
Akker and Traum, 2009	Rule Based	AMI	Gaze, current & previous speaker, utterance and addressee	65%	65%	Low accuracy
Jovanovic, 2007	Bayesian Network	M4	Current & previous speaker and utterance, topic, gaze, etc.	81%	62%	Works only for 4 participants and fixed positioning
Akker and Akker, 2009	Logistic Model Trees	AMI	Current & previous utterance, current & previous speaker, topic, gaze, etc	92%	92%	Works only for 4 participants, fixed positioning and selected DA
Baba et al, 2011	SVM	Custom using Wizard of Oz	Head Orientation, acoustic features	80.28%	NA	Predicts if an utterance is addressed to agent or not, not Generic
Le et al, 2018	CNN, LSTM	GazeFollow Dataset	Utterance and gaze information	62%	NA	Addressee detection from 3rd party angle

depending upon the DA. The algorithm reports an accuracy of only 36% on the AMI dataset [18]. Akker and Traum [2] then incorporate gaze as one of the foundations of the rules in the algorithm, resulting in an improved accuracy of 65%. The authors have also tested the gaze as the only feature for addressee detection, reporting an accuracy of 57%. In this case, the only rule is that if a speaker looks at a participant for more than 80% of the duration of the utterance, the addressee is that participant, otherwise the utterance is addressed to the whole group.

Concerning statistical approaches, Jovanovic *et al.*'s work [9] uses Baysian Networks to exploit current and previous utterance, speaker, gaze, topic of discussion and other meta features on the M4 multimodal corpus [10] to obtain an accuracy of 81%. This algorithm was also tested by Akker and Traum [2] on the AMI corpus [18], for which they report an accuracy of 62% to illustrate that the algorithm does not generalize well. Akker and Akker [1] propose a statistical model based on logistic regression trees, in order to answer the binary question *are you being addressed*, with a best case accuracy of 92% on AMI [18]. However, the two limits of this work are firstly that it predicts if an utterance is addressed to the agent or not, instead of identifying who is being addressed among all the meeting participants, and secondly that the model depends upon the fixed positioning of the participants and hence cannot be easily generalized to different positioning.

Baba *et al.* [3] exploit human-human-agent triadic conversations through Wizard of OZ experiments to develop a SVM based model that distinguishes whether an utterance is addressed to the human or a robot. They report an accuracy of 80.28%, using text, head orientation and acoustic features to train the model.

Only few works have used deep learning techniques to tackle the addressee detection problem. Le *et al.* [14] propose a convolutional neural network [12] based solution for addressee detection on the *GazeFollow* dataset [21]. One major limitation of this work is that the addressee detection is performed through third party angle, with a final accuracy of 62.5%.

On the addressee detection problem, only few researchers focus on the identification of the best features. To this end, Galley *et al.* [5] hypothesize that the adjacency pair is a good marker for addressee detection. An adjacency pair is a pair of utterances where the second utterance is a response to the first utterance. Also, Vertegal [27] explored the impact of gaze for addressee detection and reports that in 77% of the utterance, the person whom the speaker is looking at is actually the addressee of the utterance.

Discussion

A detailed summary of the existing addressee detection approaches in multiparty interaction, is presented in Table 1. Most existing addressee detection approaches either aim at differentiating between whether an utterance is addressed to an agent or not [3, 14], or depend upon the fixed positioning of the participants [1, 9]. Furthermore most of the systems do not scale well over different numbers of participants owing to non-generic features [1, 9, 26].

In this article, we claim that a model with generic features can be used for addressee detection irrespective of the dataset and the number and positioning of participants. Thus, we consider three model requirements. During training and testing, the model should not depend upon the number of participants in the meeting (r1); the model should not depend upon the participant sitting positions in the room (r2); and the model should predict who is the addressee among all the meeting participants rather than predicting if the utterance is addressed to an agent or not (r3). The rationale behind these three requirements is that the participants who are actually not being addressed should also be aware of who is being addressed, independently of how they are located in the room and how many participants there are. Our work hypothesis is therefore that a model with generic features, and having a number of participants N should achieve at least similar or better classification performance when tested on a dataset with a number of participants equal or less than N.

From existing works, Akker and Traum [2] use multimodal information from the AMI dataset to achieve an accuracy of 65% on the addressee detection detection. This work can be considered as one of the baselines as this is the only work that respects requirements r1, r2 and r3. For a fair comparison between the baseline and the solution proposed in this article, the Akker and Traum's algorithm is also tested on the AMI subset used in this article, as a second baseline. Finally, we have observed that "Group" is the majority class in the AMI dataset. Therefore, a third naive baseline is proposed: all utterances are predicted as "group". Although Akker and Akker [1] report a higher accuracy of 92% on the AMI dataset [18], their

algorithm only supports 4 participants, with fixed sitting positions, hence violating requirements r1 and r2 considered in this article.

Though several datasets have been used to learn addressee detection models, the AMI dataset [18] is the most relevant since i) it is open source and freely available, ii) it contains multimodal data, iii) it contains annotated data and iv) existing works are evaluated with it. In addition, we also propose to exploit the MULTISIMO dataset [11] to validate our work hypothesis.

3 FEATURE SELECTION AND DATASET ANALYSIS

This section reviews the features selected in the proposed model along with the datasets. Two datasets are exploited: the model is trained on the AMI dataset and the algorithm performance is evaluated on a subset of the AMI dataset [18] and on the MULTISIMO corpus [11]. A detailed statistical analysis of the features is performed on these two datasets in order to evaluate their importance.

3.1 Feature Selection

Previous systems have exploited as features speaker focus, current and previous DA, current and previous speaker and previous addressee as shown in Table 1.

In a previous research work [17], we have proposed an addressee detection model with an additional features called *conjunction*, which checks whether an utterance starts with a conjunction or not. This model achieves an accuracy of 73.44% on AMI. The model proposed in this article once again extends the feature set in [17] by incorporating the focus of all the listeners and the sentence length of the utterance. Furthermore, the addressee detection performance is now tested on two different corpora, and also implemented and distributed as a single component for conversational agents.

Thus, the model is based on the following set of features:

- **Previous and current Speaker:** the participant who uttered the immediate previous utterance and the participant who utters the current one;
- **Previous and current Dialogue Act:** the DA of the previous and current utterances;
- **Previous Addressee:** the previous utterance addressee;
- **You Usage:** whether an utterance contains "You" or not;
- **Conjunction:** whether an utterance starts with a conjunction or not;
- **Speaker Focus:** gaze information of the current speaker;
- **Listener Focus:** gazes of the participants of the meetings;
- **Sentence Length:** word count of the current utterance.

All these features are intended to be generic and thus do not depend upon an underlying scenario: the number of participants and their positioning are not taken into account.

3.2 The AMI Corpus

The AMI corpus [18] is a multimodal interaction corpus consisting of 100 hours of meeting recordings. The corpus includes two types of meetings involving 4 participants: task oriented sessions and open discussions. Task oriented sessions come up with the design of a remote control whereas open discussions have no topic restriction. The participants in the task oriented meetings are PM (Project Manager), UI (User Interface Expert), ID (Industrial Designer) and Marketing Executive (ME) (see Figure 1).

Figure 1: Meeting view from the AMI Corpus (from [4])

Figure 2: Meeting view from the MULTISIMO Corpus ([11])

The corpus contains over 117,000 utterances that have been annotated with DAs, out of which 9,071 utterances have been annotated with speaker focus and 8,874 utterances with addressee information. The number of utterances where the three annotations –speaker focus, addressee, and DA– are available is only 5,628. These utterances belong to the task oriented meetings only. Addressee annotation is not available for open discussion meeting data.

3.3 The MULTISIMO Corpus

The MULTISIMO corpus [11] is a multiparty multimodal corpus where each of the 23 meetings contains 3 participants. The average duration of the meetings is 10mn ($min = 6$, $max = 16$), for a total time of 4 hours. Different meetings in the corpus have been annotated for speech, acoustic, visual, lexical, perceptual and demographic information. However, only two meetings (S02 and S18) in the MULTISIMO corpus are annotated with gaze information.

The corpus is task oriented where one of the 3 participants plays the role of a facilitator while the other 2 participants have the role of players (see Figure 2). The facilitator asks a question for which there are three best answers. The participants have to find the answers and then rank them based on their popularity.

3.4 Data Preprocessing

There are several significant differences between the AMI and MULTISIMO corpora. In the AMI corpus, each meeting has 4 participants:

Table 2: Participants renaming convention

AMI	MULTISIMO	New Name
PM	FC	PM
ME	Left Player	A
ID	Right Player	B
UI	-	C

PM (Project Manager), UI (User Interface Expert), ID (Industrial Designer) and Marketing Executive (ME), while the meetings in the MULTISIMO corpus have 3 participants: Facilitator, Left Player and Right Player.

A statistical comparison of the two datasets is performed. Both corpora have been processed and the participants have been renamed since (r2) states that the model should not depend upon the sitting positions or roles of the participants.

The Table 2 contains the renaming convention. In the AMI corpus, the PM acts as a moderator similarly to the facilitator in the MULTISIMO corpus. As both of their role is to regulate the meeting, they were given the same label for the sake of uniformity. In MULTISIMO, the two meetings containing gaze information are not annotated with DA and addressee. Therefore, in order to use them for validation of our model, the meetings were manually annotated for addressee and DA by two annotators[1]. Furthermore, in AMI the addressee has not been annotated for 'minor' DA such as stalling, fragment, backchannels and others. Thus, the utterances containing these kinds of DA are removed from MULTISIMO.

Finally, in addition to utterances, the focus of attention is also processed. During the course of a DA, the focus of attention can be any individual, or any object such as laptop, table and slide-screen. The following steps explain the process of focus annotation of each utterance:

(1) if there are more than two participants and/or objects in focus, the focus has been marked as *multiple*;
(2) if the focus is on a participant and an object then
   (a) if the person is first in focus and then the object, the focus is marked as the person name,
   (b) otherwise if the object comes first in focus followed by the person, the focus is marked as multiple.

3.5 Statistical Analysis of MULTISIMO and AMI Corpora

3.5.1 ***Speaker Information.*** The AMI corpus has 4 participant per meeting, while the MULTISIMO corpus has 3. In the AMI corpus, 32.88% of the utterances are spoken by the PM, similar to the 30.87% in the MULTISIMO corpus. A further similarity is that in both corpora, when the current and immediate previous speakers are different, the current addressee is the immediate previous speaker (62 % in the AMI corpus and 63% in the MULTISIMO corpus).

3.5.2 ***Addressee Information.*** Table 3 contains the ratio of speaker vs addressee for the MULTISIMO corpus. The ratio of utterances addressed to individuals is higher than the ratio of utterance addressed to groups. For instance, 58.5% of the utterances of the speaker A are addressed to listener B, while only 3.1% are addressed to the whole group. In the AMI corpus, the ratio of speaker vs addressee is presented in Table 5. In this case, the ratio of utterances addressed to the whole group is higher, compared to the individuals. This difference can be explained by the difference of task in the two corpora. In the AMI corpus, the meetings are more interactive and all the participants have almost an equal role. In the case of MULTISIMO, the PM asks a question and then two participants have a long conversation with each other before they come up with an answer. Hence, more while in MULTISIMO corpus, more utterances are technically dyadic between the participants or with the facilitator.

[1] Available at http://pagesperso.litislab.fr/~jsaunier/consecutive_records11_ms.csv

Table 3: Frequency of Speaker vs Addressee (in percentage) for MULTISIMO Corpus

Speaker/ Addressee	A	B	PM	group
A	0.000	0.585	0.380	0.031
B	0.677	0.000	0.297	0.026
PM	0.144	0.139	0.000	0.716

Table 4: Frequency of Speaker vs Addressee (in percentage) for AMI Corpus

Speaker / Addressee	A	B	C	PM	group
A	0.000	0.098	0.131	0.191	0.579
B	0.106	0.000	0.094	0.197	0.603
C	0.100	0.113	0.000	0.231	0.557
PM	0.126	0.134	0.161	0.000	0.579

Table 5: Frequency of Speaker vs Addressee (in percentage) for AMI Corpus

Speaker / Addressee	A	B	C	PM	group
A	0.000	0.098	0.131	0.191	0.579
B	0.106	0.000	0.094	0.197	0.603
C	0.100	0.113	0.000	0.231	0.557
PM	0.126	0.134	0.161	0.000	0.579

3.5.3 ***Dialogue Act.*** DA is another key feature for a generic model. Statistical analysis of AMI corpus reveals that if the previous speaker is u_1, the previous addressee is u_2, the previous DA is *el.info* or *el.ass* and the current speaker is u_2, then 80% of the utterances are addressed to previous speaker u_1. This percentage rises up to 93% in MULTISIMO. This illustrates the importance of DAs as a feature.

Figure 3 depicts the difference between percentage frequencies of various DAs. The Figure shows that in AMI, *inf* which stands for *inform* has the highest frequency whereas in MULTISIMO, *ass* which stands for *assess* has the highest frequency. This variation is due to the difference of task, and thus interaction, between the two corpora. The detailed information regarding the DAs along with their examples can be found in Table III in Malik et *al* [16].

3.5.4 ***Focus of attention.*** The frequency of focus toward individuals is higher in MULTISIMO than in AMI. The ratios of focus

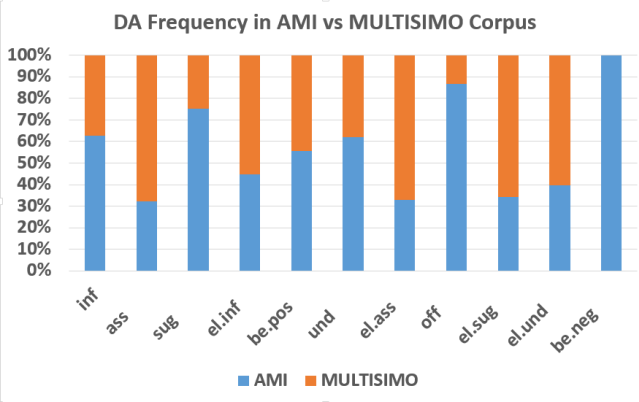

Figure 3: DA frequencies (in percentage) for AMI and MULTISIMO Corpora [11]

Table 6: Frequency of Focus vs Addressee (in percentage) for MULTISIMO Corpus

Focus / Addressee	A	B	PM	group
A	0.585	0.032	0.064	0.319
B	0.000	0.707	0.096	0.192
PM	0.037	0.037	0.916	0.009
multiple	0.121	0.121	0.152	0.606

Table 7: Frequency of Focus vs Addressee (in percentage) for AMI Corpus

Focus / Addressee	A	B	C	PM	group
A	0.525	0.034	0.025	0.031	0.385
B	0.043	0.481	0.027	0.023	0.426
C	0.015	0.036	0.479	0.025	0.446
PM	0.020	0.011	0.034	0.509	0.426
multiple	0.054	0.053	0.076	0.071	0.745
no	0.084	0.101	0.081	0.158	0.576
slide-screen	0.052	0.080	0.065	0.279	0.524
table	0.121	0.074	0.110	0.143	0.552
whiteboard	0.141	0.054	0.079	0.116	0.610

vs addressee for the MULTISIMO and the AMI corpora are given in Tables 6 and 7 respectively. The difference in the ratios of focus can be attributed to the number of participants or objects that a person can look at. In AMI, a person can look at 3 other participants, as well as a slide screen, a table and a whiteboard. In MULTISIMO the participants only have the other 2 participants to look at.

3.5.5 ***You Usage***. Usage of the word "you" in an utterance is a key indicator of the addressee. In MULTISIMO, when an utterance contains the word "you" and the focus is an individual, 70% of the time, that individual is an addressee. In AMI, this number is 42.22%. In cases where "you" is used in the sentence and the focus is multiple people, 85% of the time, the addressee is the group. This percentage is 78% in AMI. The difference may again be attributed to the difference in number of participants and objects that a person can look at in MULTISIMO and the AMI corpora.

3.5.6 ***Conjunction***. In the MULTISIMO corpus, when the previous and current speaker of an utterance are identical and the current utterance starts with a conjunction, the current addressee is the previous addressee 85.91% of the time. This percentage is 90.37% in AMI which depicts the importance of conjunction as a feature.

The next section details our approach regarding the classification of the addressee along with the classifier information and evaluation results.

4 EXPERIMENTS AND RESULTS

As seen in the previous section, the statistical analysis tends to demonstrate that our model features, also based on the literature, are generic and significant for addressee detection in both AMI and MULTISIMO corpora.

4.1 Problem Formalization

Given the set of features mentioned in section 3, the task is to predict, whether the current utterance of the speaker is addressed to listener A, B, C, PM or the whole group (in case of the AMI dataset) or listener A, B, PM or the whole group (in case of the MULTISIMO Dataset).

4.2 Experiments

In order to test our hypothesis that a model with generic features has similar accuracy for addressee detection across different datasets and with varying number of participants, we perform a set of experiments on the two corpora. To respect the requirement r1, the AMI and MULTISIMO datasets are used to train and test the proposed model with varying number of meeting participants. To ensure r2, the roles from the AMI and the participant positions from the MULTISIMO have been mapped to generic participants identifiers PM, A, B and C. Furthermore, to find if the algorithm performance is significantly different on the datasets, p-values are calculated with 0.01 as threshold.

A conventional machine learning pipeline is followed for experimentation. After feature selections, the categorical features are converted into one hot encoded vectors. Table 8 depicts one hot encoded vector for focus of Person A in AMI and MULTISIMO corpora respectively. For example in the first row of the left Table (AMI), the person A is looking at listener B, therefore there is 1 in the column for B and 0 for the rest of the columns. Since, MULTISIMO corpus does not have participant C, the focus for participant C will always be a vector of zeros in case of MULTISIMO data points.

Then, we used six of the most commonly used machine learning classifiers: Multilayer Perceptron (MLP) [13], Random Forest (RF) [15], Logistic Regression (LR) [8], Support Vector Machines (SVM) [7], Naive Bayes (NB) [22] and K Nearest Neighbours (KNN) [28]. The details of the classifiers used along with the parameters are shown in Table 9.

Parameter selection was performed through grid search [25]. For the hyper-parameters that are not mentioned, default values are used as specified in Python's Sklearn library. Five folds cross validation is performed to ensure that the models do not overfit and

Table 8: Examples of one hot encoded vector for focus of A in AMI corpus (left) and MULTISIMO (right)

B	C	PM	Multiple
1	0	0	0
0	0	1	0
0	1	0	0
0	0	0	1

B	C	PM	Multiple
1	0	0	0
0	0	1	0
0	0	0	1

Table 9: Classifiers along with parameter values

Classifier	Parameters
MLP [13]	activation = 'tanh', max_iter=100, hidden_layer_sizes = (100), alpha = 0.05, learning_rate = 'constant', solver ='adam'
NB [22]	No hyper parameters
SVM [7]	kernel='rbf', C=100, gamma = 0.01
LR [8]	penalty='l2', C =100
RF [15]	bootstrap = True, criterion = 'gini', max_features = 'sqrt', n_estimators = 500
KNN [28]	n_neighbors=12

the obtained results are stable. Finally accuracy and F1 measure have been used to evaluate performance of the algorithms. The F1 measure is preferred since the class distribution is irregular in both datasets [19].

4.3 Results

The table 10 shows the result for the algorithm trained and tested on the AMI corpus. To conduct tests on the MULTISIMO corpus, the algorithms are trained on the whole set of AMI. The table 10 contains the average results of the 5-folds cross-validation along with standard deviation. It shows that for AMI, Multilayer Perceptron (MLP) achieves the highest accuracy of 74.26% and F1 score of 0.74. The table also shows that the algorithms trained on AMI show similar results on MULTISIMO. Highest accuracy of 77.19% and F1 score of 0.77 is achieved for MULTISIMO using the Naive Bayes model (NB) with standard deviation of 0.004 for 5 runs. The slight increase in accuracy, compared to other algorithms, can be attributed to the fact that most of the interactions in MULTISIMO are dyadic with a focus mostly on participants, as there are no objects in the experimental setting.

Similarly, the results for 5 folds cross validation for the algorithms trained and tested on the MULTISIMO corpus are shown in Table 11. The results show that the highest accuracy of 84.80% is achieved with logistic regression on the MULTISIMO corpus. The model trained on the complete MULTISIMO corpus has been further tested on a subset of the AMI corpus. The highest average accuracy of 56.27% is achieved using the Multilayer Perceptron. The reason for the low accuracy can be attributed to the fact that MULTISIMO corpus contains three participants and hence the trained model has no information about the fourth participant as it is the case with AMI. Furthermore, the MULTISIMO corpus contains no information about objects in focus such as table, slide screen, whiteboard. Finally, the MULTISIMO corpus -with all required information- only contains 657 utterances, therefore, there is not enough information in the data to be transferred to a totally new dataset.

Table 10: Percentage Accuracy for Algorithms Trained on AMI and Tested on AMI and MULTISIMO Corpora (F1 in Brackets)

Classifier	AMI Results	MULTISIMO Results
MLP	74.26 ± 0.02 (0.74)	75.61 ± 0.01 (0.76)
RF	72.57 ± 0.02 (0.72)	73.35 ± 0.02 (0.73)
LR	74.09 ± 0.03 (0.73)	76.66 ± 0.01 (0.77)
SVM	73.09 ± 0.02 (0.73)	70.91 ± 0.00 (0.70)
NB	64.12 ± 0.03 (0.63)	77.19 ± 0.00 (0.77)
KNN	68.65 ± 0.02 (0.67)	70.91 ± 0.00 (0.69)
Always group	54%	–
Baseline [2]	65%	–
[2]: same AMI subset	60%	–

Table 11: Percentage Accuracy for Algorithms Trained on MULTISIMO and Tested on AMI and MULTISIMO Corpora (F1 in Brackets)

Classifier	MULTISIMO Results	AMI Results
MLP	83.82 ± 0.02 (0.85)	56.27 ± 0.003 (0.56)
RF	82.5 ± 0.02 (0.83)	47.02 ± 0.02 (0.47)
LR	84.80 ± 0.03 (0.85)	54.64 ± 0.00(0.55)
SVM	84.34 ± 0.04 (0.83)	46.66 ± 0.00 (0.44)
NB	75.89 ± 0.03 (0.75)	25.0 ± 0.00 (0.29)
KNN	80.34 ± 0.03 (0.80)	43.12 ± 0.00 (0.41)

4.4 Discussion

The results obtained from the model trained on the AMI corpus outperforms the three baselines mentioned in Table 10 and is also generic enough to produce at least comparable results on a totally unknown (*i.e.* not trained with) dataset with less number of participants as in the training dataset, such as the MULTISIMO corpus.

The performance of the algorithms trained on AMI and tested on AMI and MULTISIMO is similar between both corpora. This validates our hypothesis which states that a model with generic features,and having a number of participants N should achieve at least similar or better classification performance when tested on a dataset with a number of participants equal or less than N.

Similarly, the model trained and tested on a lower number of participants has higher accuracy than the model trained and tested on higher number of participants. For instance, the performance of models trained and tested on MULTISIMO (3 participants) is higher (84.80%) than the models trained and tested on the AMI corpus (74.26% with four participants). The reason between the performance difference can be attributed to the fact that (i) the default probability of addressee detection in meetings with three participants is 0.25 for each of the three participants and the group class, compared to 0.20 in case of 4 participants and the group class, and (ii) in the MULTISIMO corpus, the ratio of dyadic utterance is higher with focus on participants only.

Figure 4: System Implementation

The results from the best performing algorithm (logistic regression) are interpreted with the help of logistic regression coefficients [20]. Absolute Mean value 0.28 is obtained for the coefficients of all the features in the data set. The results show that the features *previous speaker*, *previous addressee*, *current speaker* and *current and previous focus* and *current DA* have coefficient values greater than the mean coefficient values and hence can be regarded as the top contributors to the performance of the algorithm.

The difference in the performance of two data sets can be partially attributed to the contrasting nature of the features. For instance, the experiments show that focus is a critical feature. In MULTISIMO, all the utterances are annotated with focus, furthermore meeting participants can be the focus of attention. However, in AMI some of the utterances are not annotated with speaker focus. Furthermore, the focus of attention can be furniture such as table, side-screen and whiteboard, resulting in a decreased addressee detection rate.

5 IMPLEMENTATION

The proposed model is implemented as a component in the AgentSlang platform [24], in order to be exploited by virtual agents and robots during human-agent interactions. The AgentSlang platform consists in a collection of components integrating several existing and original algorithms to provide a development environment for interactive systems. The platform is based on a data and component oriented design, that integrates into a unified system the concepts of Feedback Management, Dialogue Management *etc.* An example of supported system is, for instance, a virtual environment with humans and agents interacting together. The AgentSlang platform is freely distributed.

A view of the implemented system is presented in Figure 4. It uses a pipeline of three components: DA Annotator, Focus Detector and Addressee detection. The DA Annotator module predicts the dialogue acts of the current and previous utterances. The Focus detector detects the gaze information of the speaker and listeners with the help of individual cameras. The addressee detection module contains the trained algorithm to predict the addressee of the current utterance. This prediction is added to the outputs from the DA annotator and Focus Detector along with other input features such as current and previous speaker, sentence length *etc.* to predict the addressee of the current utterance. This information is then used to update the beliefs of the agent, and determine whether it is the next speaker.

6 CONCLUSION AND PERSPECTIVES

In this article, a generic addressee detection model for multiparty interaction has been proposed. The proposed model outperforms an existing baseline. The results show that the proposed model is capable of addressee detection (requirement r3) on multiple datasets, with a varying number of participants (requirement r1), and without depending upon the sitting position or roles of the participants (requirement r2). Also, our model returns best case accuracy of 74.26% which is better than 65% obtained by the best baseline model [2].

Though the proposed model outperforms the existing baseline and is generic enough to show results with multiple datasets, the model has still been only tested with 3 and 4 participants. A perspective is to test these features with a higher number of participants. Another limitation is the small size of the dataset, which is a major difficulty in the use of more advanced deep learning techniques. Thus, the next step would be to perform an experiment to collect a larger dataset.

7 ACKNOWLEDGEMENTS

This work was supported by the DAISI project, cofunded by the European Union with the European Regional Development Fund (ERDF), by the French Agence Nationale de la Recherche and by the Regional Council of Normandie.

REFERENCES

[1] Harm Akker and Rieks Akker. 2009. Are You Being Addressed?-real-time addressee detection to support remote participants in hybrid meetings. In *Proceedings of the SIGDIAL 2009 Conference*. 21–28.
[2] Rieks op den Akker and David Traum. 2009. A comparison of addressee detection methods for multiparty conversations. In *Workshop on the Semantics and Pragmatics of Dialogue*. 99–106.
[3] Naoya Baba, Hung-Hsuan Huang, and Yukiko I Nakano. 2011. Identifying Utterances Addressed to an Agent in Multiparty Human–Agent Conversations. In *International Workshop on Intelligent Virtual Agents*. 255–261.
[4] Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation* 41, 2 (2007), 181–190.
[5] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of ACL'04*. 669.
[6] Erving Goffman. 1981. Forms of talk. University of Pennsylvania publications in conduct and communication.
[7] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *Intelligent Systems and their applications* 13, 4 (1998), 18–28.
[8] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398.
[9] Natasa Jovanovic. 2007. To Whom It May Concern-Addressee Identification in Face-to-Face Meetings. (2007).
[10] Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006. A corpus for studying addressing behaviour in multi-party dialogues. *LREC'06* 40, 1 (2006), 5–23.
[11] Maria Koutsombogera and Carl Vogel. 2018. Modeling collaborative multimodal behavior in group dialogues: the MULTISIMO Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
[13] Rudolf Kruse, Christian Borgelt, Frank Klawonn, Christian Moewes, Matthias Steinbrecher, and Pascal Held. 2013. Multi-layer perceptrons. In *Computational*

Intelligence. 47–81.

[14] Thao Minh Le, Nobuyuki Shimizu, Takashi Miyazaki, and Koichi Shinoda. 2018. Deep Learning Based Multi-modal Addressee Recognition in Visual Scenes with Utterances. *arXiv preprint arXiv:1809.04288* (2018).

[15] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[16] Usman Malik, Mukesh Barange, Julien Saunier, and Alexandre Pauchet. 2018. Performance Comparison of Machine Learning Models Trained on Manual vs ASR Transcriptions for Dialogue Act Annotation. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 1013–1017.

[17] Usman Malik., Mukesh Barange., Julien Saunier., and Alexandre Pauchet. 2019. Using Multimodal Information to Enhance Addressee Detection in Multiparty Interaction. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,*. INSTICC, SciTePress, 267–274. https://doi.org/10.5220/0007574602670274

[18] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. 2005. The AMI meeting corpus. In *Proc. of the 5th International Conference on Methods and Techniques in Behavioral Research*, Vol. 88. 100.

[19] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[20] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *The journal of educational research* 96, 1 (2002), 3–14.

[21] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Adv. in Neural Information Processing Systems*. 199–207.

[22] Irina Rish et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. IBM New York, 41–46.

[23] John Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*.

[24] Ovidiu Şerban and Alexandre Pauchet. 2013. Agentslang: A fast and reliable platform for distributed interactive systems. In *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 35–42.

[25] Selmar K Smit and Agoston E Eiben. 2009. Comparing parameter tuning methods for evolutionary algorithms. In *Proc of CEC'09*. 399–406.

[26] David R Traum, Susan Robinson, and Jens Stephan. 2006. *Evaluation of Multi-Party Reality Dialogue Interaction*. Technical Report. University of Southern California Marina Del Rey CA Inst For Creative Technologies.

[27] Roel Vertegaal. 1998. Look Who's Talking to Whom. *Mediating Joint Attention in multiparty* (1998).

[28] Min-Ling Zhang and Zhi-Hua Zhou. 2005. A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, Vol. 2. IEEE, 718–721.