Predicting Group Performance in Task-Based Interaction

Gabriel Murray
University of the Fraser Valley
Abbotsford, Canada
gabriel.murray@ufv.ca

Catharine Oertel
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
catharine.oertel@epfl.ch

ABSTRACT

We address the problem of automatically predicting group performance on a task, using multimodal features derived from the group conversation. These include acoustic features extracted from the speech signal, and linguistic features derived from the conversation transcripts. Because much work on social signal processing has focused on nonverbal features such as voice prosody and gestures, we explicitly investigate whether features of linguistic content are useful for predicting group performance. The conclusion is that the best-performing models utilize both linguistic and acoustic features, and that linguistic features alone can also yield good performance on this task. Because there is a relatively small amount of task data available, we present experimental approaches using domain adaptation and a simple data augmentation method, both of which yield drastic improvements in predictive performance, compared with a target-only model.

KEYWORDS

Group interaction, task performance, multimodal interaction, meetings, social signal processing, data augmentation, domain adaptation, semi-supervised learning

ACM Reference Format:
Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

For most of us, a large part of our everyday life consists of being a member of a group. We are communicating and interacting within the contexts of our families, clubs, and workplace teams. Each of these groups has implicit and explicit goals they would like to achieve, and these goals can vary both in time and regularity. However, they might also vary in terms of whether there is an intrinsic or an extrinsic motivation to meet the goal, and whether all members of the group are aligned in their motivation to meet the goal.

For example, a family might have the goal to leave the house at 7:30 in the morning so that everyone can make it on time to school or to work. This goal is probably a regularly recurring one. It is extrinsically imposed onto the group by constraints imposed by society and the parents are probably more inclined to meet the goal than the children.

For a club, the goal might be to win the next match. Similarly, the goal is regularly recurring. However, the motivation to be in the club and to meet the goal might be more intrinsically motivated and people might be more aligned in meeting this goal, while views on how to reach the goal may vary.

A team at work might have the goal to meet the next deliverable. This goal in a specific form might be a one-time event. It is extrinsically imposed and different members of the team might have contradictory views. However, the interaction between members of a team would probably be more formal and formalized than the one within a family or a club.

All in all, domains and interaction forms vary but successful interaction within groups is an important part of our everyday life, touching very different parts of it. The challenge from a computational point of view is to build models which are not only able to adapt to these different domains and forms of communication, but also account for different personalities of the people involved, and hierarchies which might overlay the group interaction.

Being able to build such models can have many different advantages, especially at the workplace. Here it can provide insights into what communication patterns lead to successful outcomes. Utilizing these models in an online fashion has the advantage of not only making it possible to analyze meetings in hindsight and draw conclusions for future strategies, but to additionally make people aware of disadvantageous patterns during the meeting itself and allow for real-time modification of the group dynamics and strategies. Such models will become even more advantageous and widespread in the future, as robots are becoming an increasing part of our life and can take on the role of optimizing communication [13, 17].

In this work, we build models for predicting how well a group will perform on a well-defined task, where the predictive models use multimodal information in the form of linguistic features and speech features. We show that the best performance is found by combining those multimodal features. We also demonstrate that simple techniques of domain adaptation and data augmentation can dramatically improve performance on this prediction task, and these techniques are needed because of the relatively limited amount of labeled training data. Finally, whereas much work on social signal processing for small groups has focused on nonverbal features, we show that relatively simple verbal features can yield very good prediction results on their own.

The structure of this paper is as follows. Section 2 discusses related work on social signal processing (SSP) and multimodal interaction. Section 3 highlights the key contributions of the current work. Section 4 presents our approach for automatically predicting group performance, including the multimodal features used in these experiments, as well as the domain adaptation and data augmentation methods used. Section 5 describes the experimental setup, including the two corpora used, the machine learning models tested, and our evaluation metrics. Section 6 presents and interprets all of the experimental results, and we conclude in Section 7.

2 RELATED WORK

Social psychological research has been concerned with the understanding of group performance as early as the middle of the last century [15]. A great deal of recent work has also been carried out from a psychological viewpoint [2, 4, 27]. In recent years, there has been a growing interest in the analysis of group interactions from a computational perspective. The main goals of these analyses have been to design computational models that could predict information about the participants as well as the state of the interaction. Examples are the identification of roles within a group [5, 24], the recognition of personality traits [1], and the understanding of first impressions [20].

There is a large body of work which is concerned with the automatic analysis of meetings. Two examples of large, freely available corpora are the AMI (Augmented Multimodal Interaction) corpus [7] and the CHIL (Computers in the Human Interaction Loop) corpus [18]. While the meetings are often structured and have a specific goal, they still remain quite open in terms of conversational dynamics. However, it is not always easy or possible to tell whether these groups have performed well at their given task. For example, the AMI corpus scenario has each group design a remote control, but it is not always clear whether a group did a good or poor job of carrying out that task. As a result, the AMI corpus – despite being a very rich source of meeting data – is not particularly amenable to building models that predict group performance. However, it does feature rich post-meeting questionnaire data, and Lai et al. [14] have used those questionnaires to examine how turn-taking patterns impact group affect.

There is also a smaller body of research which has investigated group dynamics in the context of games [12, 16, 21, 23]. In some cases the performance of the group can be evaluated explicitly, while in other cases implicit measures are taken such as, for example, the analysis of group engagement and individual involvement [21].

Whereas in the above-mentioned domains and datasets group performance can be difficult to measure, there has also been research on group interaction in scenarios where group performance can be very clearly measured. An example of a dataset with clear group performance scores is the ELEA (Emergent Leadership) corpus [25], which uses a survival task where the group has to collectively rank a list of items. This dataset is described in more detail in Section 5. Avci and Aran [3] use this dataset to predict group performance, using nonverbal features. Their work is the most closely related to ours, with key differences being that we use linguistic features in addition to nonverbal features, and we employ two methods for increasing the amount of training data. Later in the paper we explicitly compare our system's performance with theirs. Another example of predicting group performance is by Neubauer et al., who analyze groups tasked with a scenario of disarming a simulated bomb [19].

However, studies concerned with the automatic prediction of group performance are still relatively rare. One reason for this might be that it is still quite costly to record many different groups of people in such a way that their nonverbal cues can be recorded. For predicting the dynamics of a conversation as well as different traits of participants, traditionally eye-gaze (or head-rotation) and prosody [11, 21, 22] have been proven to be useful features. Recording these in a scalable way not only requires access to costly equipment such as eye-trackers, but also involves post-processing that is quite time consuming. Features of linguistic content [10] have not been used as frequently. One of the reasons for this might be that speech recognition has until about five years ago been still relatively poor in terms of word-error rate when applied to group conversational data. Some of the reasons for this are that overlapping speech and diverse accents and dialects make speech processing difficult. Using prosody and other nonverbal features have been proven robust for real-time applications. Using nonverbal features also avoids privacy issues that may arise when analyzing the content of group discussions. Finally, nonverbal features such as gesture, prosody, and gaze can also be unconscious and therefore revealing in terms of individual and group affect.

3 CONTRIBUTIONS

The key focus of our paper is the exploration of the usefulness of verbal/linguistic features for group interaction, which have not previously been explored to a large extent in this domain. We show that we can meet or exceed the performance of a state-of-the-art system on a task that is of interest to the Social Signal Processing community, by using a fairly simple set of verbal features combined with domain adaptation and data augmentation. Below we comment on each of these contributions in more detail.

First, much work on multimodal analysis of group interactions has been concerned with identification of roles, analysis of leadership, and inference of first impressions. There has been less work on the automatic prediction of group performance. We are adding to this small amount of research.

Second, most work on group interactions so far has been focused on speech and nonverbal cues. Linguistic cues have been less explored. We are exploring the usefulness of linguistic content features.

Third, a limiting factor for the analysis of group performance has been the availability of corpora. We are exploring the usefulness of domain adaptation and data augmentation for mitigating the issue of having a small amount of group performance data.

4 MULTIMODAL ANALYSIS OF GROUP PERFORMANCE

In this section, we first describe the multimodal features that were extracted from the group interactions, and subsequently present the domain adaptation and data augmentation methods that were used to increase the amount of training data.

4.1 SPEECH FEATURES

We extracted a large number of acoustic features from the audio recordings in the ELEA and AMI corpora, using the openSMILE software [9]. This is a very large set of features that has previously been used in the Interspeech 2010 Paralinguistic Challenge. In the experiments described herein, we use only a subset of standard deviation features, yielding 76 speech features in total. These include mel-frequency cepstral coefficients (**MFCCs**), associated delta features, jitter, shimmer, PCM loudness, F0 envelope, F0 contour, voicing probability, and log power of Mel-frequency bands.

4.2 LINGUISTIC FEATURES

We extract the following linguistic features from manual transcripts of the meetings in both corpora.

Dependency Parse Features: All sentences are parsed using spaCy's dependency parser[1]. We extract several features, including the branching factor of the root of the dependency tree, the maximum branching factor of any node in the dependency tree, sparse bag-of-relations features, and the type-token ratio for dependency relations.

Part-of-Speech Tags: We use spaCy's part-of-speech tagger, and use a sparse bag-of-tags representation for the most frequent tags, as well as the type-token ratio for tags.

Filled Pauses: We include the number of *filled pauses*, such as 'uh' or 'um.'

Psycholinguistic: We use several psycholinguistic features. All words are scored for their concreteness, imageability, typical age of acquisition, and familiarity[2]. We also derive SUBTL scores for words, which indicate how frequently they are used in everyday life [6].

Sentiment: We use the SO-Cal (Semantic Orientation Calculator) sentiment lexicon [26], which associates positive and negative scores with sentiment-bearing words, indicating how positive or negative their sentiment typically is.

GloVe Word Vectors: Words are represented using GloVe vectors[3], and the vectors are summed over sentences. We then create a document vector that is the average of the sentence vectors. The first five dimensions of the document vectors are used as features.

Lexical Cohesion: We measure cohesion using the average cosine similarity of adjacent sentences in a document, using the GloVe vectors.

Sentence and Document Length: We include the average number of words per sentence, and average number of sentences per meeting.

Bag-of-Words: Finally, we use a bag-of-words representation for the most common 200 non-stopwords in the dataset, and also calculate the type-token ratio for words.

4.3 DATA AUGMENTATION

The goal with data augmentation is to create additional training data by making copies of the original training instances and applying transformations to them. The key is that these transformations should be *label-preserving*. That is, they should augment the original training data but in a way that does not change the outcome or response variable. For example, in the domain of object recognition in image data, the original training images could be copied and augmented through cropping, rotating, and flipping, without fundamentally changing the object in the image [28].

These types of techniques are well-known in the domain of image classification. However, best practices for data augmentation in the domain of natural language processing have not been established. One simple technique is to replace words with their synonyms, also known as thesaurus-based data augmentation [29].

We propose a similarly simple approach that we hypothesize will be effective for conversational data, where the conversations are often loosely structured and may have low information density. We proceed through the conversation transcript, and for each sentence we delete it with some probability p. We do n passes over the entire transcript, creating n augmented copies of the conversation. The augmentations only involve deletion of content, rather than insertion. We have implemented this approach for use with the linguistic features, which are mostly calculated at the level of individual sentences or sentence pairs and then averaged over the meeting. It is hypothesized that this data augmentation method will subtly impact linguistic features such as lexical cohesion, while remaining close enough to the original that prediction of the outcome variable will improve rather than degrade. For these experiments with the data augmentation method, we chose the parameters $p = 0.25$ and $n = 4$; the rationale for these low settings is to avoid having the augmented data differ too drastically from the original data.

4.4 DOMAIN ADAPTATION AND SEMI-SUPERVISED LEARNING

With domain adaptation (also called *transfer learning*), the goal is to leverage data or models from a related source domain in order to improve prediction performance in the target domain. In some cases, the source data may be labeled with the same outcomes of interest as the target domain. In that scenario, there are well-known domain adaptation methods that are simple to implement and can work very well in practice [8]. In other cases, the source data may not be labeled but could still be useful, particularly if it is much larger than the target dataset. In that case, the problem is one of semi-supervised learning.

In our case, our target set of meetings is fairly small and each meeting is associated with a group performance score. We also have access to a much larger meeting set, but the group task is different, and the meetings do not have performance scores. We take a straightforward semi-supervised learning approach, where we train a model on a subset of the target data, use it to make predictions of group performance scores on the source data, then train a new model using the original target data and the automatically labeled source data. This second model can then be used to make predictions on the test set of the target data.

The following section has much more detail on the two corpora, and the leave-one-out evaluation scheme that is used with the domain adaptation approach.

5 EXPERIMENTAL SETUP

In this section we briefly describe our corpora, the machine learning models used, and evaluation methods.

5.1 CORPORA

Our primary corpus of interest is the ELEA corpus [25]. This is a dataset of small group meetings, where each group is engaged in collectively performing a ranking task called the Winter Survival Task[4]. The group members are role-playing that they have crashed-landed in the wilderness in the middle of winter, and cannot expect

[1] https://spacy.io/
[2] http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm
[3] https://nlp.stanford.edu/projects/glove/

[4] There are many variants of this task with the same basic structure but different survival scenarios.

to be rescued right away. The group must decide which items from the airplane wreckage to bring with them into the wilderness. They are given a list of items that were salvaged from the wreckage, and must rank the items as a group, according to how important each item is to their survival. Each group member first ranks the items privately, and the group then ranks the items collectively during a time-constrained meeting.

Part of the appeal of this dataset and the survival scenario more generally is that there is a well-defined task, and the rankings – both the individual rankings and group rankings – can be objectively scored by comparing them with the rankings of survival experts. This gives us a clear measurement of group performance. The scenario also lends itself to analysis of how individual knowledge and performance leads to group knowledge and performance. And because there are no assigned roles, it lends itself to the study of how leaders naturally emerge in groups.

There are 40 meetings in the ELEA corpus, 28 of them in English. Since our predictions are at the level of the group (predicting how well a group will perform on the task), this gives us very few datapoints with which to work. For that reason, we use the AMI meeting corpus [7] as a secondary data source.

Like the ELEA corpus, the AMI corpus involves groups role-playing a fictional scenario. In the case of the AMI corpus, they are members of a company designing and marketing a remote control device. Unlike the ELEA corpus, the members do have defined roles, as project manager, marketing expert, industrial designer, or user interface designer. Also in contrast to the ELEA corpus, there is no clear way to rate group performance in the AMI corpus. The closest we have are post-meeting questionnaires in which the AMI meeting participants rated how well they thought the meeting went, on various criteria. We use the larger AMI corpus as a source domain of meetings which we can exploit to improve predictive models on the ELEA corpus, using the domain adaptation method described above.

5.2 MACHINE LEARNING MODELS

For these experiments, we compare two tree-based methods, Random Forests (RF) and Gradient Boosted Trees (GB), as well as a neural network system. For both tree-based systems, the number of estimators was set at 20. For the neural approach, due to the small amount of training data we employed a single hidden layer of 50 neurons in a fully-connected feed forward network with ReLU activations and the lbfgs solver.

5.3 EVALUATION

For evaluating all systems, we report the mean squared error (MSE) of the group performance predictions. Because of the small amount of target data, we employ a leave-one-out training and testing scheme. For the domain adaptation experiments, this means training on 27 meetings in the training fold, using that model to make predictions on the AMI source data, then training a new model using the 27 meetings in the training fold plus the automatically labeled source data. That second model is then used to make predictions on the held-out meeting. This is repeated for each of the 28 meetings.

6 RESULTS

In the following sections, we report the experimental results, first using only target data, then using domain adaptation, and finally with data augmentation.

6.1 Target Data Only

The first set of results is based on applying several machine learning methods to ELEA performance prediction, using only the ELEA target data. Specifically, we tested neural networks, random forests, and gradient boosted trees on the prediction problem. The resulting MSE scores are shown in Table 1. Unfortunately, all results are substantially worse than the baseline wherein we simply predict the mean value of the outcome variable in the training fold. The tree-based approaches fare significantly better than the neural approach, with the neural approach likely performing poorly simply due to the small amount of available data.

With each machine learning model, the linguistic features are more useful for the predictive task than are the speech features. However, the best score in this set of results is found by using both linguistic and speech features with Gradient Boosted Trees. Overall, Random Forests and Gradient Boosted Trees exhibit similar performance.

Model	Speech	Ling.	All Feas.
Baseline (Mean Prediction)	· · · **79.3** · · ·		
Neural Network	266.0	142.9	133.3
Random Forests	127.0	90.0	95.9
Gradient Boosted Trees	117.9	91.0	**86.4**

Table 1: MSE: Target Data Only

6.2 Domain Adaptation

We next present results on domain adaptation, using Random Forests models. We found that the domain adaptation results are very sensitive to feature normalization / scaling, and we present results in both normalized and unnormalized conditions. The best overall performance is found by using both linguistic and speech features and normalizing the features (for both the source and target data), yielding an MSE of 64.4, a dramatic improvement compared with the target-only approach reported in the previous section. The full domain adaptation results are shown in Table 2.

Considering each class of features separately, linguistic features are again more useful for this task. Even using domain adaptation, using speech features alone gives performance that is worse than the baseline.

6.3 Data Augmentation

We next present the results using data augmentation, for linguistic features alone and for linguistic + speech features. Because the speech features are conversation-level features (not utterance-level features), we do not present results using data augmentation for speech features alone, as the augmentation method is based on utterance-level deletion.

Model	Speech	Ling.	All Feas.
Baseline (Mean Prediction)	··· 79.3 ···		
Feas. Normalized	114.4	71.9	**64.4**
Feas. Unnormalized	103.2	77.6	99.28

Table 2: MSE: Domain Adaptation (Random Forest)

Model	Ling. Only	All Feas.
Baseline (Mean Prediction)	··· 79.3 ···	
Feas. Normalized	106.1	78.6
Feas. Unnormalized	**69.5**	83.2

Table 3: MSE: Data Augmentation (Random Forest)

Table 3 summarizes the data augmentation results. Using linguistic features alone with data augmentation is best overall, and nearly as good as domain adaptation using all features. This is remarkable, given the relative simplicity of the data augmentation approach.

6.4 Further Discussion of Results

Combining domain adaptation and data augmentation does not yield any further improvement, and in fact is slightly worse than using either approach on its own. We hypothesize that this is because using both approaches simultaneously results in a training set that is too far removed from the original target dataset. We plan to do further experimentation on how to combine domain adaptation and data augmentation without washing out the target data.

Our best reported MSE result of 64.4 is better than the best MSE reported by Avci & Aran on the full set of meetings, which was 71.3. However, their score is based on all 40 ELEA meetings, whereas we apply our model to just the 28 English-language meetings. Avci & Aran also reported results on a smaller subset of 21 meetings that contain video, and show that using video features can result in a much lower MSE on this task. Table 4 summarizes these three sets of results.

Another point of contrast is that the systems of Avci & Aran use the individual task performance scores as input features in their systems when predicting the group task performance scores. None of our systems use the individual performance scores – prediction is entirely based on features of the group interaction. For example, using only linguistic features from the conversation, we can make good predictions of group performance on the task, without any knowledge of how the individuals performed on the task.

System	# Meetings	Fea. Types	Best MSE
Avci & Aran	40	speech, turn-taking	71.3
Avci & Aran	21	as above, plus video	38.0
Our System	28	speech, linguistic	64.4

Table 4: Comparison with Existing Work on ELEA Corpus

Figure 1 shows the predicted scores from our best performing system (domain adaptation using all features), compared with the true task performance scores. One noticeable trend is that the predicted scores are in a much narrower band than the actual distribution of task scores.

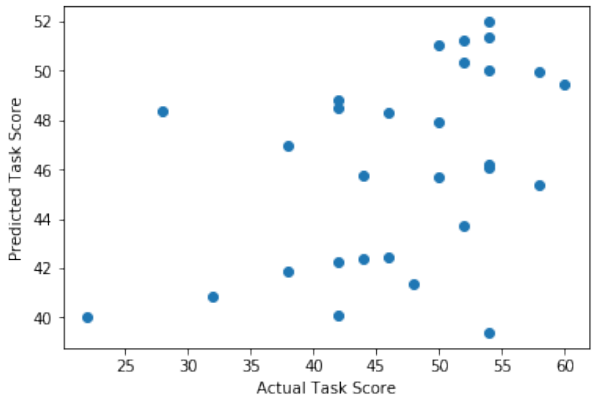

Figure 1: Actual vs. Predicted Task Scores

To explore why the data augmentation approach is improving performance, we calculate feature importance scores for the core linguistic features, where the core linguistic features are just the dense features, with the sparse bag-of features removed. The feature importance score for each feature is based on how much it decreased the MSE, on average, when used as split point in the Random Forest decision trees. Figure 2 shows the top 10 features in terms of importance, when using only the original target data with no augmentation. Figure 3 similarly shows the top 10 features, but this time when using the augmented dataset. Many of the features are same, but with slightly different scores. However, in the second figure we see that cohesion and number of sentences are now amongst the top features, which was not the case originally.

With the cohesion feature, there is some intuitive sense to this finding. Cohesion is based on the cosine similarity of adjacent sentence vectors. The augmentation approach involves deleting sentences, and two similar sentences might end up being adjacent to each other when they were not originally. The following toy example illustrates this:

- A. So the remote should be curved.
- C. Um, right.
- B. I like the curved remote idea.

By analyzing adjacent sentences, this toy example would exhibit low lexical cohesion. But if the middle sentence were deleted in the data augmentation approach, cohesion would increase. This intuition also suggests that we can investigate more robust methods of measuring cohesion.

7 CONCLUSION

We have presented experimental results showing that we can predict group performance on a task, using multimodal features of the group interaction. These include a rich set of speech features as well as linguistic content features. Being able to predict task

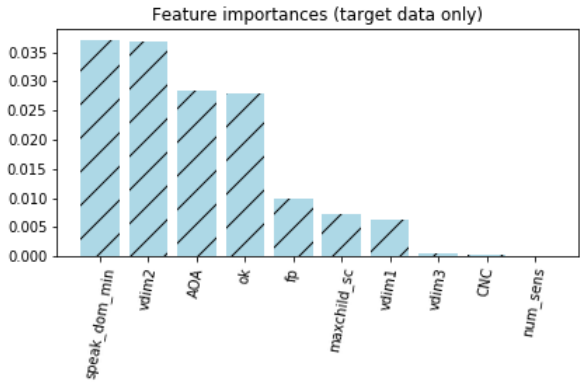

Figure 2: Feature Importance Scores (Target Data Only)

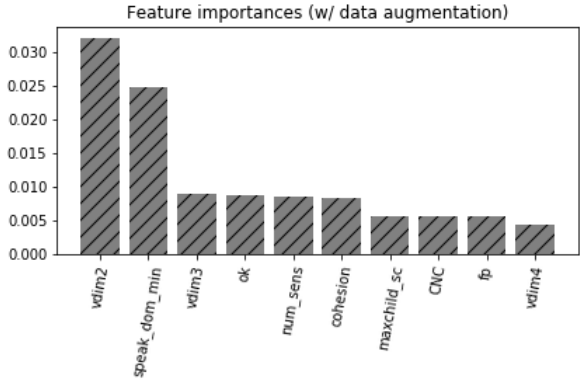

Figure 3: Feature Importance Scores (w/ Data Augmentation)

performance based on the group interaction could be very valuable for providing feedback to collaborative teams in an online fashion. For example, it could allow a group to modify their interactional dynamics during a meeting in order to increase the likelihood of achieving good results on the task.

The target domain – the ELEA corpus and winter survival task – contains a fairly small amount of training data, and building predictive models just on the target data yields poor performance that is worse than the baseline. We therefore explored two approaches for increasing the amount of training data. With domain adaptation and semi-supervised learning, we exploit the AMI meeting corpus as a source domain that can be used to improve performance in the target domain. With data augmentation, we copy and augment meetings from the ELEA corpus using a simple deletion approach, in order to create additional training instances. Both strategies result in dramatic improvement in the MSE scores.

Our system compares very favourably with the system of Avci & Aran in terms of MSE scores, and ours does not require any information about how well the individual participants performed on the task prior to the meeting. However, Avci & Aran showed that prediction performance can be improved with video features, as they demonstrated on a smaller set of meetings. Their approach also has the advantage of being language-independent, whereas our current system has only been applied to the English-language meetings.

In contrast with much work on social signal processing, which often focuses on nonverbal features, we have demonstrated that features of linguistic content can be extremely useful for this task. The best results are from a system that uses the full set of multimodal features. And linguistic features on their own, when derived using the data augmentation method, yield very good performance. We believe that this provides solid motivation for more SSP work that incorporates lexical, syntactic, and psycholinguistic features.

Future work will look at how to best combine the domain adaptation and data augmentation methods to yield further improvement. We are also collecting a new corpus of meetings using the winter survival task, to supplement the ELEA corpus meetings.

ACKNOWLEDGMENTS

Catharine Oertel is supported by the leading house DUAL-T research project funded by the Swiss State Secretariat for Education, Research and Innovation (SERI). Gabriel Murray is supported by an NSERC Discovery Grant.

REFERENCES

[1] Oya Aran and Daniel Gatica-Perez. 2013. One of a kind: Inferring personality impressions in meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 11–18.

[2] John R Austin. 2003. Transactive memory in organizational groups: the effects of content, consensus, specialization, and accuracy on group performance. *Journal of applied psychology* 88, 5 (2003), 866.

[3] Umut Avci and Oya Aran. 2016. Predicting the performance in decision-making tasks: From individual cues to group interaction. *IEEE Transactions on Multimedia* 18, 4 (2016), 643–658.

[4] Bernard M Bass, Bruce J Avolio, Dong I Jung, and Yair Berson. 2003. Predicting unit performance by assessing transformational and transactional leadership. *Journal of applied psychology* 88, 2 (2003), 207.

[5] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. 2016. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 317–324.

[6] Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods* 41, 4 (2009), 977–990.

[7] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 28–39.

[8] Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. *ACL 2007* (2007), 256.

[9] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[10] Heather Friedberg, Diane Litman, and Susannah BF Paletz. 2012. Lexical entrainment and success in student engineering groups. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 404–409.

[11] Yuyun Huang, Emer Gilmartin, Benjamin R Cowan, and Nick Campbell. 2016. A Preliminary Exploration of Group Social Engagement Level Recognition in Multiparty Casual Conversation. In *International Conference on Speech and Computer*. Springer, 75–83.

[12] Hayley Hung and Gokul Chittaranjan. 2010. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 879–882.

[13] Martin Johansson, Gabriel Skantze, and Joakim Gustafson. 2013. Head pose patterns in multiparty human-robot team-building interactions. In *International conference on social robotics*. Springer, 351–360.

[14] Catherine Lai, Jean Carletta, and Steve Renals. [n. d.]. Modelling participant affect in meetings with turn-taking features.

[15] Harold J Leavitt. 1951. Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology* 46, 1 (1951), 38.

[16] Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1421–1431.

[17] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. 2015. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language* 33, 1 (2015), 1–24.

[18] Djamel Mostefa, Nicolas Moreau, Khalid Choukri, Gerasimos Potamianos, Stephen M Chu, Ambrish Tyagi, Josep R Casas, Jordi Turmo, Luca Cristoforetti, Francesco Tobia, et al. 2007. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language resources and evaluation* 41, 3-4 (2007), 389–407.

[19] Catherine Neubauer, Joshua Woolley, Peter Khooshabeh, and Stefan Scherer. 2016. Getting to know you: a multimodal investigation of team behavior and resilience to stress. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 193–200.

[20] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. 2014. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE transactions on multimedia* 16, 4 (2014), 1018–1031.

[21] Catharine Oertel and Giampiero Salvi. 2013. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 99–106.

[22] Catharine Oertel, Stefan Scherer, and Nick Campbell. 2011. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Twelfth Annual Conference of the International Speech Communication Association*.

[23] Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gatica-Perez. 2011. An audio visual corpus for emergent leader analysis. *ICMI-MLMI), Multimodal Corpora for Machine Learning, Nov* (2011), 14–18.

[24] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces* 7, 1-2 (2013), 39–53.

[25] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.

[26] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.

[27] Daan Van Knippenberg, Carsten KW De Dreu, and Astrid C Homan. 2004. Work group diversity and group performance: an integrative model and research agenda. *Journal of applied psychology* 89, 6 (2004), 1008.

[28] Jason Wang and Luis Perez. 2017. *The effectiveness of data augmentation in image classification using deep learning*. Technical Report.

[29] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*. 649–657.