Analyzing Natural Language Context in Human-Machine Teaming using Supervised Machine Learning

Bryan A. Barrows[1]
NASA Langley Research Center, Hampton VA, 23681, USA

Lisa R. Le Vie[2]
NASA Langley Research Center, Hampton VA, 23681, USA

James E. Ecker[3]
NASA Langley Research Center, Hampton VA, 23681, USA

B. Danette Allen[4]
NASA Langley Research Center, Hampton VA, 23681, USA

Building a foundation for trustworthiness and trust verification in multi-asset teaming is the research challenge of Autonomy Teaming and TRAjectories for Complex Trusted Operational Reliability (ATTRACTOR). The Design Reference Mission (DRM) for ATTRACTOR is a search and rescue mission objective governed by a multi-member team consisting of human and machine operators. A crucial component to the effort is the communication between humans and autonomous agents throughout both planning and execution stages of the mission. Intuitive communication methods and modalities are posited as critical enablers for certifying trust and trustworthiness. This paper reports on the data collection and analysis conducted in support of the Human Informed Natural-language GANs Evaluation (HINGE) project to attain explainable and trusted communication between human-machine assets. Two identically curated image description datasets were acquired for HINGE, both consisting of two unique input modalities (typed vs. verbal) and retrieved in two distinct contexts (general vs. specific). The gathered datasets were assessed and compared using Parts-of-Speech (POS) features, sentence similarity metrics, and linguistic analysis. Then, the datasets were modeled and tested separately and in combination with one another using machine learning algorithms. The comparison and testing results reveal a superior dataset, by which a preferred context and input is understood, for generating image representations of missing persons using a Generative Adversarial Network (GAN).

I. Nomenclature

AttnGAN	=	Attentional Generative Adversarial Network
ATTRACTOR	=	Autonomy Teaming and TRAjectories for Complex Trusted Operational Reliability
CAS	=	Convergent Aeronautics Solutions

[1] Aerospace Research Engineer, Autonomous Integrated Systems Research Branch, MS 233, AIAA Member
[2] Aerospace Research Engineer, Autonomous Integrated Systems Research Branch, MS 233, AIAA Member
[3] Aerospace Research Engineer, Autonomous Integrated Systems Research Branch, MS 233, Non-Member
[4] NASA Senior Technologist, Autonomous Integrated Systems Research Branch, MS 233, AIAA Associate Fellow

CIDEr	=	Consensus-based Image Description Evaluation
DRM	=	Design Mission Reference
GAN	=	Generative Adversarial Network
GUI	=	Graphical User Interface
HINGE	=	Human Informed Natural-language GANs Evaluation
HMI	=	Human-Machine Interface
NLP	=	Natural Language Processing
NUI	=	Natural User Interface
ML	=	Machine Learning
PASCAL	=	Pattern Analysis, Statistical Modelling and Computational Learning
POS	=	Parts-of-Speech
SAR	=	Search and Rescue
SPICE	=	Semantic Propositional Image Caption Evaluation

II. Introduction

The objective of Autonomy Teaming and TRAjectories for Complex Trusted Operational Reliability (ATTRACTOR), supported by NASA's Convergent Aeronautics Solutions (CAS) project, is to define a baseline concept Design Reference Mission (DRM) for enabling certification of trust and trustworthiness in multi-asset teaming applications. The ATTRACTOR design mission reference is a search and rescue (SAR) task where a person or object is missing in an unmapped outdoor setting. A human-machine team consisting of autonomous agents and a supervising human operator are to search within an area of interest and communicate among team members to achieve improved mission performance. The objective is to strengthen human-agent cohesion through means of reliable context-aware communication that enables greater trust of the system.

In support of ATTRACTOR, the Human Informed Natural-language GANs Evaluation (HINGE) study was conducted to explore human-machine communication for natural and realistic SAR scenarios. [1][2] HINGE is specifically concerned with how human assets communicate descriptive information of images to machines and how that descriptive input is used to create internal representations understood by the autonomous agents. Understanding bi-directional human-machine interaction, utilizing touch, written text, spoken and gestural input may increase the rate of success in human-autonomy teaming. Furthermore, distinguishing between modality and context in human input could offer insight as to how a Natural User Interface (NUI) can be designed to achieve better performance.

The work presented in this paper is a continuation of prior research within the HINGE study. The primary focus of this work is toward an extended analysis and comparison of two image description datasets intended for informing autonomous systems of missing targets in SAR related missions. Machine Learning (ML) algorithms are trained on the datasets for predicting context via image description quality. The objective of the analysis is to explore and define an approach for providing feedback to users on the quality of their image description data.

III. Background

HINGE hypothesizes that communication between humans and autonomous agents needs to be natural and intuitive in order to facilitate multi-asset teaming. [3] Multimodal interaction is quickly becoming the forefront of human-machine teaming research concepts, making it all the more necessary to explore and compare modality capabilities. [4] Popular modalities for interacting with machine systems include touch, speech, gesture, and gaze. Incorporating some variety of modalities in interface systems may provoke a more natural interaction between humans and autonomous systems.

The awareness of context and input modality provided by the human user is critical for a machine's interpretation of multi-modal or categorical information. A lack of ability to interpret user input and understand user intent presents potential ambiguity to a system's ability to meet performance expectations.

In a traditional SAR mission, the description of the missing person is provided via picture and/or a verbal narrative. People tend to give concise sentences that identify the "who," "what," and "where," while leaving out the information they judge to be less significant. Even so, these descriptions still exhibit consensus with each other. [5] HINGE sought to determine if, and how, modality (spoken vs. written) affects different aspects of the image description given which would inform the interface being designed. Furthermore, the gathered descriptions will be used to build a database to train a Generative Adversarial Network (GAN) whose output is an internal representation

of the missing person being searched for during the mission. [6] A GAN comprises two neural networks that utilize a training set of existing world data to synthesize new data. [7][8] The GAN utilized in this study outputs both an internal latent-space vector representation of the missing person being searched for during the mission and a synthesized image, expanded from the internal representation, for visual explainability of the GAN's understanding of the NLP input. [6] To achieve these outputs, [6] extended the architecture of AttnGAN [7]; a GAN architecture which applies an attention mechanism over a traditional GAN in order to generate images from NLP description input.

IV. Methods

A set of five images were selected from the PASCAL 50S dataset, see Figure 1. [9] This dataset was curated from the former UIUC Pascal Sentence Dataset. [5] Each image contains a single human subject of interest located in a fairly uncluttered outdoor environment. In the first HINGE experiment, conducted within NASA Langley Research Center, 53 participants were asked to provide a total of ten image descriptions with respect to the five chosen PASCAL 50S images. Each participant was randomly chosen to communicate their descriptions as either typed input (into a computer), or verbal input (through a handheld recorder). All descriptions are reviewed and transcribed into text format, if verbally provided. The five images were described twice by each participant. During the first round of descriptions, the participant is asked to use a single sentence to describe the entirety of each image (context 1). In the second round, the participant is told that the individual in each image went missing, and is asked to describe the lost subject in one detailed sentence (context 2) for the purpose of collecting more descriptive text about the human subjects in the images. The data collection took place in a spacious pass-through lobby located outside of the NASA Langley cafeteria and is therefore referred to in this text as the NASA dataset.

An extended HINGE data collection effort was performed later using Amazon's Mechanical Turk (MTurk) on-line crowdsourcing platform. The procedures used in the curation of the NASA dataset were followed for the subsequent MTurk dataset, with the exception that the series of collection processes were performed in an on-line setting. Both the NASA and MTurk datasets are structured identically, such that all five images have the same number of descriptions, each with labels denoting their respective input modality and context.

Several supervised machine learning models were trained and tested using a handful of input variations of the datasets. A set of features consisting of grammatical and structural elements were extracted from the data. Description similarity features were also derived and used from previous studies [2] which offered additional grammatical and semantic context. The set of features remained constant across all learning models and performed tests. The algorithms include K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), and Bootstrap Aggregating Decision Trees (TreeBagger), and were chosen due to their distinguished mathematical approaches for learning generalizations. [11] KNN is an exemplar-based classification algorithm which may learn complex decision boundaries by inferring posterior probabilities from weighted feature distances. The Naïve Bayes method utilizes Bayes' Theorem, however; each feature is conditionally independent, and a final prediction is determined by a maximum a posteriori estimate. Unlike KNN, the Naïve Bayes algorithm is a fast predictor and considered an eager learner, as it may learn a decision function instead of querying exemplars. [12] SVM is a large margin classification algorithm which generates an optimal hyper-plane by finding the maximum error margin between the training data and the decision boundary. [13] The SVM method serves a more complex approach of maximizing class decision boundaries. Finally, the MathWorks' MATLAB® TreeBagger algorithm [11] utilizes an ensemble of decision trees, implemented using random forest, to determine a more frequently obtained class-attribute result. The ensemble voting of TreeBagger is particularly advantageous as it reduces over-fitting on the training data.

Figure 1: HINGE Images 1-5, in order from left to right [9]

V. Results

In the first phase of the HINGE experiment, concerning only the NASA dataset, the quality of the image descriptions and their similarity to one another are deeply evaluated. [1] Parts-of-Speech (POS) features and similarity metrics are relied on to better understand the sentence qualities within the dataset. The Penn Treebank [14] POS tagset is used throughout the work discussed in this paper. The second curated dataset collected through the MTurk platform correlates strongly with the NASA dataset after comparing corpus linguistic analysis findings and sentence similarities. Both datasets are cleaned of irrelevant description outliers, as well as incomplete and unidentifiable words. Commingling the MTurk dataset with the NASA dataset triples the amount of data from 525 descriptions to 1,605 descriptions in total. This data is used to train and test prediction models for distinguishing qualitative descriptions typically found in context 2 from input text that lacked descriptive information of the primary human subject typically found in context 1.

The success of the ML models are determined by the accuracy of the models in predicting the preferred context 2 type. Likewise, informative methods for providing user feedback on system performance are determined by the simplification of feedback features in a NUI. These indicator features are used for estimating input description quality and disclosing system interpretability to the user. The features derived for training the ML models are metrics of success in themselves, as these features are independently utilized in determining the user feedback component.

A. Corpus Linguistic Analysis

The second HINGE data collection is conducted using Amazon's Mechanical Turk (MTurk) crowdsource platform. Similar to the first HINGE data collection, participants of the MTurk data collection are asked to either type or record and then transcribe their own responses using Amazon's online MTurk user interface. The effort produced a set of 1080 total descriptions across both modalities, contexts, and all five images combined, as compared to the original NASA dataset of 525 total descriptions. Overall, the Amazon MTurk platform collected twice as much data in a fraction of the time of the original HINGE effort, and was able to reach users from a wider range of geographical settings.

The total corpus of the MTurk data consists of 16,260 words, which is roughly 1.3 times larger than the original NASA dataset (12,673 words), and the unique word count of 1,459 for the MTurk dataset is 1.43 times larger than the original NASA dataset unique word count (1,021 unique words). The MTurk and NASA dataset average description lengths are 15.1 and 24.2 words per sentence, respectively. Despite this difference, the datasets share a strong positive correlation with frequency of POS occurrences per description. For each of the five images, the image data subsets between the datasets have a POS frequency correlation of 0.94 and higher. Figure 2 illustrates the average frequency of occurrence for all five images of the eight most common POS tags in the two datasets. Both datasets share similar frequency trend behaviors across the POS elements, with the difference in magnitude being a direct result of their respective vocabulary sizes. In summary, the MTurk description data presents higher variability in word choice and vocabulary. However, the nearly identical POS occurrence suggests that the data may be examined together to further refine an understanding of language use and communication styles used in SAR missions. Moreover, the MTurk collection effort offers a scalable and efficient solution for further collection of human description responses for use in SAR-related NLP studies.

Three influential variables are primarily responsible for the composition of both datasets. Modality type, context type, and the image set each contribute significantly to the elicitation of the data that was collected in this study. The highest description categorizing tier, which happens to be the least impactful to the data in this study, is the chosen image set. As in the comparison between both full datasets, the trends of POS occurrences across description sets from each of the five chosen images still have strong positive correlations. The most significant differences found when comparing descriptions between images are the word lengths and POS frequency of occurrence. Image 4 has at least 16% more words on average per description in the NASA dataset than the other images. Similarly in the MTurk dataset, Image 5 has 12% more words on average than the other four images. The combined datasets reveal no outlying image with significantly greater average description lengths. These findings have little to no impact on the overall quality of the datasets. However, it is worth noting that Image 4 and Image 5 have more background artifacts and explainable scenery than the other three images in the chosen set.

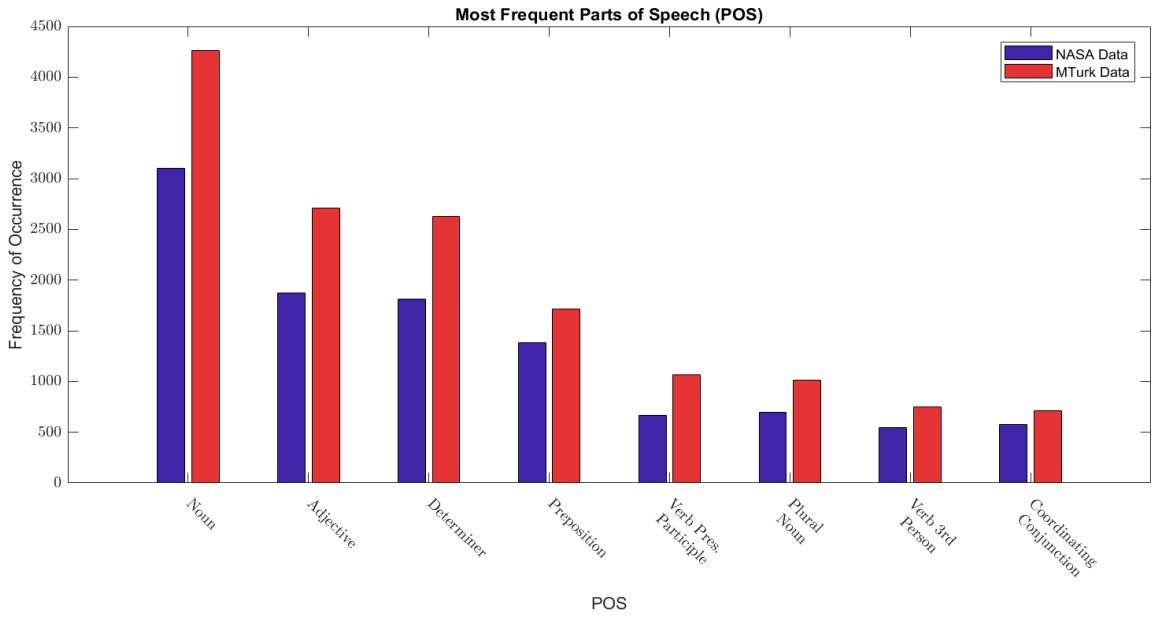

Figure 2: Parts of Speech (POS) Frequency of Occurrence similarities found between the MTurk and NASA dataset compositions.

Modality is the second most influential description category. Independent of their context, descriptions given by verbal input, on average, had higher word counts, greater usage of verb POS, addressed more descriptively what the subjects' were doing within the image, and were more informative of the subjects' clothing. [2] In addition to higher description length, verbal descriptions combined in both datasets had roughly six percent more unique words than the typed descriptions. Typed descriptions had 25 percent more adjectives per sentence, with no other significant difference in POS, aside from the previously mentioned verb usage. In the interest of SAR, these results make the typed modality the more desirable choice for supplying informative descriptions to the GAN. However, they are not distinguishable enough from verbal descriptions to generalize.

Lastly, the context category is by far the most influential category to the information quality of descriptions collected for this research. To better understand the impact of context on description quality, statistical metrics for description composition, word choice, and POS usage were calculated. A strong detectable difference in descriptions with informative words and word phrases, denoted by POS tags, is seen in Table 1. As with the modality and individual image analyses, the frequency of POS occurrence between both datasets correlates positively when the descriptions are categorized according to context type. What stands out almost immediately in these findings is the significantly increased use of descriptive adjectives and coordinating conjunctions for context 2 descriptions. This suggests that characterizing sentences for each image by POS occurrence and frequency can determine description context more noticeably than determining modality type. In the interest of sentence classification with respect to modality and context, the number of word classification types is greatly reduced from over 1000 unique words in each corpus to the Penn Treebank POS tagset size of 36 unique types. Furthermore, the MTurk dataset has a higher average occurrence of these POS tags in its context 2 descriptions than that of the NASA dataset. As expected, context 2 for both datasets combined has much higher average description lengths, by nearly double, than in context 1. Detailed descriptions of subjects, including adjectives and attributes of objects, is desirable for higher AttnGAN performance. [7] However, a potential drawback of the context 2 type is its large average description length. This causes high output variance from multiple semantic differences between words, and over-enlarges the attention vector, especially given a small dataset.

For overall quality of data, it was found that the MTurk dataset was preferred over the NASA dataset for several reasons. The dataset collected from MTurk consisted of twice as many descriptions, of which had broader national demographic coverage, as opposed to the data collected at NASA LaRC, in Virginia. From an analytical standpoint, the MTurk descriptions generally contain more descriptive and distinguishable POS tags between the two contexts, as shown in Table 1. Of the six most frequently occurring POS tags, adjectives and coordinating conjunctions were more abundant in context 2 for MTurk. Upon further analysis of the NASA dataset, it is found the NASA

descriptions are sufficiently descriptive, but contain significant wordiness. [1][2] These observation, in addition to the shorter average description lengths of the MTurk descriptions, make the MTurk dataset more desirable for classifier training and validation.

Average Context 2 / Context 1 Part of Speech Usage of All 5 Images

Dataset	Adjectives	Coordinating Conjunctions	Determiners	Nouns	Prepositions	Verbs
NASA	3.83	3.19	0.94	1.49	0.91	0.87
MTurk	6.02	7.99	0.94	1.52	0.89	0.92

Table 1: Parts of Speech (POS) Frequency of Occurrence similarities found between the MTurk and NASA dataset compositions.

B. Machine Learning

As mentioned in the previous section, the word choice variance creates substantial complexity for the classification of description context type. An alternative approach is a POS tagset and derived statistical metrics to help quantize the number of data labels and reduce feature set dimensionality. Feature selection plays a crucial role in the accuracy of the supervised learning prediction models. Several informative POS types, such as Nouns, Verbs, and Prepositions have the same frequency of occurrence across both context and modality categories. These important speech components need to be distinguishable across the context type in order to achieve higher prediction accuracy. To address this, POS bigrams and trigrams are computed for obtaining desired POS tags which neighbored other significant POS tags more commonly found in context 2. As in Figure 3, the "Adjective-Noun" is one bigram found to have a large ratio of context 2 to context 1 occurrence, as well as a high rate of frequency throughout both data sets. "Preposition-Determiner," "Noun-Verb," and "Determiner-Adjective-Noun" tag combinations are shown to also be more specific to context 2. Other useful tag combinations not shown, due to lower frequencies of occurrence, are "Adjective-Adjective" and "Noun-Coordinating Conjunction." Since adjectives, among other tags and tag phrases, relate well to context 2 by their descriptive notion, a feature for quantity of adjectives per description was utilized. In both datasets combined, 51.1% and 96.3% of context 1 and context 2 sentences contain at least one adjective, respectively. Additionally, context 1 and context 2 sentences on average contain 0.94 and 4.78 adjectives per description, respectively. Similarly, POS ratios were found to be useful features, as context 1 has a ratio of 0.23 adjectives-to-nouns and context 2 has a ratio of 0.93 adjectives-to-nouns, per description. Other features considered include CIDEr and SPICE description similarity metrics. [2][15][16] Both CIDEr and SPICE algorithms are relative feature metrics, in that they measure grammatical and semantic properties of a text sample relative to other text samples. Details of their application to the NASA dataset is discussed in [2]. Of these two data-dependent features, only CIDEr was chosen and used in the ML model. SPICE leverages semantic contexts of a description, in addition to that of grammatical properties, as performed in CIDEr, and does not score well across the full dataset consisting of descriptions for all five image.

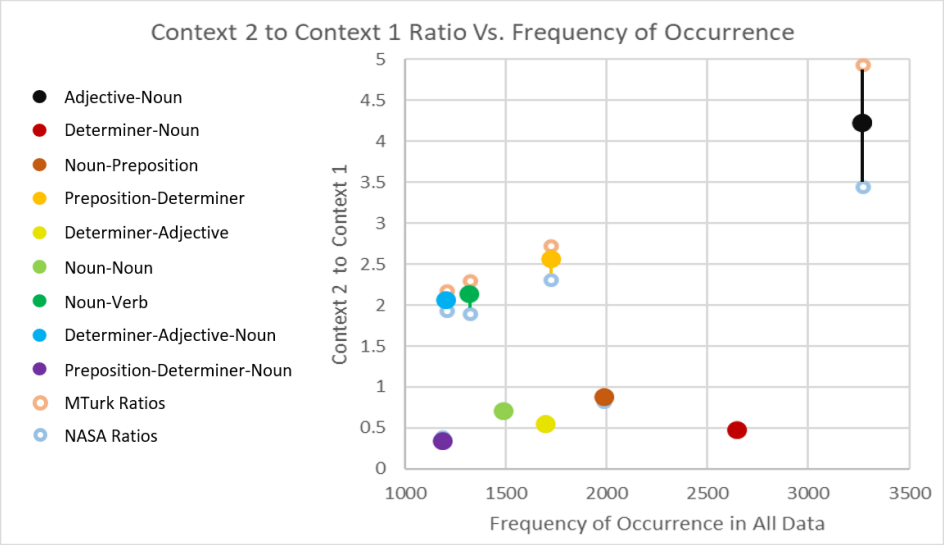

Figure 3: Frequency of occurrence differences across description context categories in the combined dataset

Specifically, three machine learning approaches are explored to better understand the similarity and predictability of the context 2 descriptions. The approaches are defined by the training and testing of both datasets in three different configurations. These ML configurations include: training the models on the NASA dataset and testing with the MTurk dataset (approach 1), training the models on the MTurk dataset and testing with the NASA dataset (approach 2), and training and testing the models with an 80/20 validation set split of the combined MTurk and NASA datasets (approach 3). Four classical machine learning models (KNN, Naïve Bayes, TreeBagger, and SVM) are selected for each configuration and stay consistent across all three approaches. Of the three configurations, all four machine learning models perform consistently better when the models are trained on a single commingled dataset, as in approach 3. Mainly, this is due to the increased number of training samples after combining and splitting apart the validation data subset. On average, approach 1, of training the prediction model on the NASA dataset and testing using the MTurk dataset outperforms the opposite approach of training on the MTurk dataset and testing using NASA dataset (approach 2). The validation results of approach 3 outperforms both of the other approaches, by within three percent for context 1 and by greater than nine percent for the desired context 2 prediction, as seen in Table 2. Of the four ML algorithms tested, the SVM is consistently found to have the greatest context 2 prediction accuracy, by over 2 percent on average. The SVM algorithm is desirable for the classification problem in this work because of its ability to create non-linear decision boundaries to predict a ranging variety of context 2 descriptions.

Algorithm	Context 1 *Precision*	Context 1 *Recall*	Context 2 *Precision*	Context 2 *Recall*
KNN	89.6	86.0	86.6	90.1
Naïve Bayes	90.9	85.6	86.0	91.1
TreeBagger	91.5	85.2	85.5	91.6
SVM	90.2	88.1	88.8	90.9

Table 2: Comparison of context 1 and context 2 predictions for 4 algorithms: K-Nearest Neighbors (KNN), Naïve Bayes, Bootstrap Aggregating Decision Trees (TreeBagger), Support Vector Machine (SVM).

Predictions from the ML models are intended for relaying context 2 descriptions to the GAN system as well as feedback to the user. Some datasets consisting of images with human subjects and a sufficient number of descriptions per image could not be compiled for training an AttnGAN model for this research. Alternatively, an AttnGAN model was trained as in [7] using the CUB dataset, which consists of 200 bird species, to explore the impact of description context differences on generated image representations. In order to validate ML features, HINGE results are compared with findings from the published CUB dataset. The average description length across the full CUB dataset is 15.4 words. The average relative number of significant POS tags per description are 3.51 (nouns), 3.93 (adjectives), 2.39 (determiners), 0.16 (prepositions), 1.01 (verbs), and 1.46 (coordinating conjunctions), which has a strong positive correlation to that of context 2 description types in both the MTurk and NASA datasets. Most significantly, in testing, the varied use of adjectives and nouns generally always affected the quality of generated bird image representations. This is supportive to HINGE, as the most statistically significant discriminators for description context are adjective and noun occurrences. The black-box behavior of the AttnGAN and the limited data for testing made it difficult to accurately quantify the direct impacts that each POS has on the quality of the generated images.

VI. Conclusion

The first analysis of HINGE image description data revealed that the methods of interaction between humans and machine assets changed for each context and modality. It was determined then that the context 2 and typed modality groups of image descriptions provided description information that was more strongly fitted for SAR target description. For SAR, collecting relevant data efficiently and unambiguously is critical for achieving operational success.

This work contributes to the verification of a desired context and input modality for language information sharing in human-machine teaming. The analyses performed also demonstrate how freeform description input can be deciphered by context, and modeled using classical machine learning techniques for its use in SAR target description. Future work will adopt the findings in this study for use with a GAN feedback system embedded in a natural human-machine user interface. Additional work will implement the findings discussed in this paper into a prototype GAN feedback system embedded in a natural human-machine user interface for use within the ATTRACTOR project.

Acknowledgments

The authors would like to thank the Autonomous Integrated Systems Research Branch, Benjamin Kelley, Erica Meszaros, Miranda Smith, the ATTRACTOR team and the experiment participants for their support.

References

[1] Le Vie, L. R., Last, M. C., Barrows, B. A., and Allen, B. D., “Towards Informing an Intuitive Mission Planning Interface for Autonomous Multi-Asset Teams via Image Descriptions,” *AIAA Aviation Forum*, Atlanta, GA. 2018.

[2] Meszaros, E. L., Le Vie, L. R., Barrows, B. A., Last, M. C., Smith, M., and Allen, B. D., “Evaluating Communication Modality for Improved Human/Autonomous System Teaming,” *AIAA SciTech*, San Diego, CA. 2019.

[3] Meszaros, E. L., Le Vie, L. R., and Allen, B. D., “Trusted Communication: Utilizing Speech Communication to Enhance Human-Machine Teaming Success,” *2018 Aviation Technology, Integration, and Operations Conference*. P. 4014. 2018.

[4] Ohneiser, O., Jauer, M.-L., Gürlük, H., and Uebbing-Rumke, M., “TriControl—A Multimodal Air Traffic Controller Working Position,” *In Proceedings of the Sixth SESAR Innovation Days*, Delft, The Netherlands, 8–10 November 2016.

[5] Farhadi, A., Hejrati, M., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D., “Every Picture Tells a Story: Generating Sentences for Images”. *European conference on computer vision*, pp. 15-29. Springer, Berlin, Heidelberg, 2010.

[6] Ecker, J., and Allen, B. D., “Goal Detection via Mental Representation,” In *2018 Aviation Technology, Integration, and Operations Conference*, p. 4012. 2018.

[7] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." arXiv preprint arXiv:1711.10485. 2017

[8] Wang, J. K., and Robert Gaizauskas. "Cross-Validating Image Description Datasets and Evaluation Metrics." *Proceedings of the 10th Language Resources and Evaluation Conference*. European Language Resources Association, 2016.

[9] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., and Zisserman, A., "The PASCAL Visual Object Classes Challenge: A Retrospective." International journal of computer vision, 111(1), 2015, pp.98-136.

[10] Houston, V. E., Manuel, W. J., Gizzi, E., and Barrows, B. A., "Advancing Aircraft Operations in a Net-Centric Environment with the Incorporation of Increasingly Autonomous Systems and Human Teaming," June 2019.

[11] MathWorks TreeBagger Class https://www.mathworks.com/help/stats/treebagger-class.html

[12] McCallum, A., and Nigam, K., "A comparison of event models for naive bayes text classification," *IN AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[13] Cortes, C., and Vapnik, V. N. (1995). "Support-vector networks," *Machine Learning. 20* (3): 273–297.

[14] Marcus, M., Santorini, B., and Marcinkiewicz, M.A., "Building a large annotated corpus of English: the Penn Treebank." *Computational Linguistics*, Vol 19, 1993

[15] Vedantam, R., Lawrence Zitnick, C., and Parikh, D., "Cider: Consensus-based image description evaluation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566-4575.

[16] Anderson, P., Fernando, B., Johnson, M., and Gould, S., "Spice: Semantic propositional image caption evaluation," in European Conference on Computer Vision. Springer, 2016, pp. 382-398