Conversational Image Editing: Incremental Intent Identification in a New Dialogue Task

Ramesh Manuvinakurike[1], **Trung Bui**[2], **Walter Chang**[2], **Kallirroi Georgila**[1]
[1]Institute for Creative Technologies, University of Southern California
[2]Adobe Research
[manuvinakurike,kgeorgila]@ict.usc.edu, [bui,wachang]@adobe.com

Abstract

We present "conversational image editing", a novel real-world application domain combining dialogue, visual information, and the use of computer vision. We discuss the importance of dialogue incrementality in this task, and build various models for incremental intent identification based on deep learning and traditional classification algorithms. We show how our model based on convolutional neural networks outperforms models based on random forests, long short term memory networks, and conditional random fields. By training embeddings based on image-related dialogue corpora, we outperform pre-trained out-of-the-box embeddings, for intention identification tasks. Our experiments also provide evidence that incremental intent processing may be more efficient for the user and could save time in accomplishing tasks.

1 Introduction

The development of digital photography has led to the advancement of digital image editing, where professionals as well as hobbyists use software tools such as Adobe Photoshop, Microsoft Photos, and so forth, to change and improve certain characteristics (brightness, contrast, etc.) of an image.

Image editing is a hard task due to a variety of reasons: (1) The task requires a sense of artistic creativity. (2) The task is time consuming, and requires patience and experimenting with various features before settling on the final image edit. (3) Sometimes users know at an abstract level what changes they want but are unaware of the image editing steps and parameters that will result in the desired image. For example, a person's face in a photo may look flushed, but the users may not know that adjusting the saturation and the temperature settings to some specific values will change the photo to match their expectations. (4) Users are not sure what changes to perform on a given image. (5) Users are not fully aware of the features and the functionality that are supported by the given image editing tool.

Users can often benefit from conversing with experts to edit images. This can be seen in action in web services such as the Reddit Photoshop Request forum[1], Zhopped[2], etc. These web services include two types of users: expert editors who know how to edit the photographs, and novice users who post their photographs and request changes to be made. If the editor needs further clarification regarding the requested change, they post their query and wait for a response from the user. The conversational exchanges also happen through edit feedback where the editor interprets the user request and posts the edited photographs. The user can reply with further requests for changes until they are fully satisfied. Due to this message-forum-like setup, users do not have the freedom to request changes in real time (at the same time as the changes are actually being performed), and hence often end up with edited images that do not fully match their requests. Furthermore, the editors are often unable to provide suggestions that could make the photograph fit better the user's narrative for image editing.

In this setup the users can benefit greatly from conversing with an expert image editor in real time who can understand the requests, perform the editing, and provide feedback or suggestions as the editing is being performed. Our ultimate goal is to build a dialogue system with such capabilities.

[1]https://www.reddit.com/r/PhotoshopRequest/
[2]https://zhopped.com

Proceedings of the SIGDIAL 2018 Conference, pages 284–295,
Melbourne, Australia, 12-14 July 2018.

Conversational image editing is a task particularly well suited for incremental dialogue processing. It requires a lot of fine-grained changes (e.g., changing brightness to a specific value), which often cannot be just narrated with a command. In order to perform such fine-grained changes to the user's liking, it is necessary that the editor understands the user utterances incrementally (word-by-word) and in real time, instead of waiting until the user has finished their utterance. For example, if the user wants to increase the brightness, they could utter "more, more, more" until the desired change has been achieved. The changes should occur as soon as the user has uttered "more" and continue happening while the user keeps saying "more, more".

In this paper, our contributions are as follows: (1) We introduce "conversational image editing", a novel dialogue application that combines natural language dialogue with visual information and computer vision. Ultimately a dialogue system that can perform image editing should be able to understand what part of the image the user is referring to, e.g., when the user says "remove the tree". (2) We provide a new annotation scheme for incremental dialogue intentions. (3) We perform intent identification experiments, and show that a convolutional neural network model outperforms other state-of-the-art models based on deep learning and traditional classification algorithms. Furthermore, embeddings trained on image-related corpora lead to better performance than generic out-of-the-box embeddings. (4) We calculate the impact of varying confidence thresholds (above which the classifier's prediction is considered) on classification accuracy and savings in terms of number of words. Our analysis provides evidence that incremental intent processing may be more efficient for the user and save time in accomplishing tasks. To the best of our knowledge this is the first time in the literature that the impact of incremental intent understanding on savings in terms of number of words (or time) is explicitly measured. DeVault et al. (2011) measured the stability of natural language understanding results as a function of time but did not explicitly measure savings in terms of number of words or time.

2 Related Work

Combining computer vision and language is a topic that has recently drawn much attention. Some approaches assume that there are manual annotations available for mapping words or phrases to image regions or features, while other approaches employ computer vision techniques. Research is facilitated by publicly available data sets such as MS COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017). Typically image and language corpora consist of digital photographs paired with crowdsourced captions, and sometimes mappings of words and captions to specific parts of an image.

Yao et al. (2010) is an example of a work relying on manual input. They developed a semi-automatic method for parsing images from the Internet to build visual knowledge representation graphs. On the other hand, the following works did not rely on manual annotations. Feng and Lapata (2013) generated captions from news articles and their corresponding images. Mitchell et al. (2012) and Kulkarni et al. (2013) built systems for understanding and generating image descriptions.

Due to space constraints, below we focus on work that combines computer vision or visual references (enabled through manual annotations) and language in the context of a dialogue task, which is most relevant to our work. Antol et al. (2015) introduced the "visual question answering" task. Here the goal is to provide a natural language answer, given an image and a natural language question about the image. Convolutional neural networks (CNNs) were employed for encoding the images (Krizhevsky et al., 2012). This was later modeled as a dialogue-based question-answering task in Das et al. (2017). These works used images from the MS COCO data set. de Vries et al. (2017) introduced "GuessWhat?!", a two-player game where the goal is to find an unknown object in a rich image scene by asking a series of questions. They used images from MS COCO and CNNs for image recognition.

Paetzel et al. (2015) built an incremental dialogue system called "Eve", which could guess the correct image, out of a set of possible candidates, based on descriptions given by a human. The system was shown to perform nearly as well as humans. Then in the same domain, Manuvinakurike et al. (2017) used reinforcement learning to learn an incremental dialogue policy, which outperformed the high performance baseline policy of Paetzel et al. (2015) in offline simulations based on real user data. Each image was

associated with certain descriptions and the game worked for a specific data set of images without actually using computer vision.

Manuvinakurike et al. (2016a) developed a model for incremental understanding of the described scenes among a set of complex configurations of geometric shapes. Kennington and Schlangen (2015) learned perceptually grounded word meanings for incremental reference resolution in the same domain of geometric shape descriptions, using visual features.

Huang et al. (2016) built a data set of sequential images with corresponding descriptions that could potentially be used for the task of visual storytelling. Mostafazadeh et al. (2016) introduced the task of "visual question generation" where the system generates natural language questions when given an image, and then Mostafazadeh et al. (2017) extended this work to natural language question and response generation in the context of image-grounded conversations.

Some recent work has started investigating the potential of building dialogue systems that can help users efficiently explore data through visualizations (Kumar et al., 2017).

The problem of intent recognition or dialogue act detection has been extensively studied. Below we focus on recent work on dialogue act detection that employs deep learning. People have used recurrent neural networks (RNNs) including long short term memory networks (LSTMs), and CNNs (Kalchbrenner and Blunsom, 2013; Li and Wu, 2016; Khanpour et al., 2016; Shen and Lee, 2016; Ji et al., 2016; Tran et al., 2017). The works that are most similar to ours are by Lee and Dernoncourt (2016) and Ortega and Vu (2017) who compared LSTMs and CNNs on the same data sets. However, neither Lee and Dernoncourt (2016) nor Ortega and Vu (2017) experimented with incremental dialogue act detection as we do.

Regarding incrementality in dialogue, there has been a lot of work on predicting the next user action, generating fast system responses, and turn-taking (Schlangen et al., 2009; Schlangen and Skantze, 2011; Dethlefs et al., 2012; Baumann and Schlangen, 2013; Selfridge et al., 2013; Ghigi et al., 2014; Kim et al., 2014; Khouzaimi et al., 2015). Recently Skantze (2017) presented a general continuous model of turn-taking based on LSTMs. Most related to our work, DeVault et al. (2011) built models for incremental interpretation and prediction of utterance meaning, while Manuvinakurike et al. (2016b) and Petukhova and Bunt (2014) built models for incremental dialogue act recognition.

3 Data

We use a Wizard of Oz setup to collect a dialogue corpus in our image edit domain. The Wizard-user conversational session is set up over Skype and the conversation recorded on the Wizard's system. The screen share feature is enabled on the Wizard's screen so that the user can see in real time the changes requested. There are no time constraints, and the Wizard and the user can talk freely until the user is happy with the changes performed. Users may have varying levels of image editing expertise and knowledge of the image editing tool used during the interaction (Adobe Lightroom).

Each user is given 4–6 images and time to think of ways to edit them to make them look better. The conversation typically begins with the step called *image location*. The user describes the image in a unique manner so that it can be located in the library of photos by the Wizard. If the descriptions are not clear the Wizard can ask clarification questions. Once the image is located, the user conveys to the Wizard the changes they desire. The user and the Wizard have a conversation until the user is happy with the final outcome. In order to capture all the changes that the user wants to achieve in spoken language, the image editing tool is controlled only by the Wizard. Figure 4 in the Appendix shows the Adobe Lightroom interface as seen by the user and the Wizard. Note that users were not explicitly told that they would interact with another human and could not see who they interacted with because the Wizard and the user were in different locations. However, the naturalness of the conversation made it obvious that they were conversing with another human.

The photographs chosen for the study are sampled from the Visual Genome data set (Krishna et al., 2017). For the dialogue to be reflective of a real-world scenario the images sampled should be representative of the images regularly edited by the users. We sampled 200 photoshop requests from the Reddit Photoshop Request forum and Zhopped, and found that the images in those posts fell into eight high-level categories: animals, city scenes, food, nature/landscapes, indoor scenes, people, sports, and vehicles.

Figure 1: Example Wizard-user conversation. The user provides new requests, modifies the requests, provides feedback, and issues a high-level command. The Wizard responds with acknowledgments and provides a clarification. Figure 1a shows the annotation of the dialogue acts for the user utterances.

# users	28
# dialogues	129
# user utterances	8890
# Wizard utterances	4795
# time (raw)	858 min
# user tokens	59653
# user unique tokens	2299
# Wizard tokens	26284
# Wizard unique tokens	1310
# total unique tokens	2650

Table 1: Data statistics.

Figure 1 shows a sample conversation between the user and the Wizard, and Table 1 shows the statistics of the data. Details of the semantics of the conversation are discussed in Section 4. Each dialogue session ranges between 2–30 min (7 min on average). The dialogues were transcribed via crowdsourcing (Amazon Mechanical Turk). We intend to publicly release the data.

4 Dialogue Semantics

The data collected were annotated with dialogue acts. User utterances were segmented at the word level into utterance segments. An utterance is defined as a portion of speech preceded and/or followed by a silence interval greater than 300 msec. Each utterance segment was then assigned a dialogue act. The annotations were performed by two expert annotators. The inter-annotator agreement was measured by having our two annotators annotate the same dialogue session of 20 min, and kappa was found to be 0.81 which indicates high agreement. Below we describe briefly our dialogue act scheme.

Image Edit Requests: The most common dialogue acts used by the user are called "Image Edit Requests (IERs)". These are user requests concerning the changes to be made to the images. IERs are further categorized into 4 groups: IER-New (IER-N), IER-Update (IER-U), IER-Revert (IER-R), and IER-Compare (IER-C). IER-N requests refer to utterances that are concerned with new image edit requests different from the previously requested edits. These requested changes are either abstract ("it's flushed out, can you fix it?") or exact ("change the saturation to 20%"). The Wizard interprets these requests and performs the changes. IER-U labels are used for utterances that request updates to the previously mentioned IER-Ns. These include the addition of more details ("change it to 50%") to the IER-N ("change the saturation"), issuing corrections to the IER ("can you reduce the value again?"), modifiers (more, less), etc. If the users are completely unhappy with the change they can revert the change made (IER-R). The IER-R act is used if the user reverts the complete changes performed, compared to only changing the values. For example, if the user is modifying the saturation of the image and across multiple turns changes the value of saturation from 20% to 30% and back to 20%, the user's action is labeled as IER-U. If the user wants all the saturation changes to be undone, the user's action is labeled as IER-R. Users may also want to compare the changes made across different steps ("can we compare this to the previous update?"), and this action is labeled as IER-C.

Comments: Once the changes are performed the user is typically happy with the change and issues a comment that they like the edit (COM-L), or they are unhappy and issues a comment that they dislike the edit (COM-D). In some cases the users are neutral and neither like nor dislike the edit. Typically such utterances are comments on the images and are labeled as COM-I.

Requests & Responses: The user may ask the Wizard to provide suggestions on the IERs. These are labeled as "Request" acts. "Yes" and "no" responses uttered in response to the Wizard's suggestions are labeled as RS-Y or RS-N.

Suggestions: This is the most commonly used Wizard dialogue act after "Acknowledgments". When the user does not know what edits to perform, the Wizard issues suggestion utterances with the intention of providing the user with ideas about the changes that could be performed. The Wizard provides new suggestions (S-N), e.g., "do you want to change the sharpness on this image?". The Wizard could also provide update suggestions for the current request under consideration (S-U), e.g., "sharpness of about 50% was better".

Other user actions are labeled as questions about the features supported by the image editing tool, clarifications, greetings, and discourse markers. In total there are 26 dialogue act labels, including the dialogue act "Other (O)" which covers all of the cases that do not belong in the other categories. In this work we are interested in the task of understanding the user utterances only, and in particular, in classifying user utterances into one of 10 labels: IER-N, IER-U, IER-R, IER-C, RS-Y, RS-N, COM-L, COM-D, COM-I, and O.

An agent will eventually be developed to replace the Wizard, which means that the agent will need to interpret the user utterances. The task of understanding the user utterance happens in two phases. In the first step the goal is to identify the dialogue acts. The second step is to understand the user image edit requests IER-N and IER-U at a fine-grained level. For example, when the user says "make the tree brighter to 100", it is important to understand the exact user's intent and to translate this into an action that the image editing tool can perform. For this reason we use action-entities tuples <action, attribute, location/object, value>. The user utterances are mapped to dialogue acts and then to a pre-defined set of image action-entities tuples which are translated into image editing actions. For more information on our annotation framework for mapping IERs to actionable commands see Manuvinakurike et al. (2018). It is beyond the scope of this work to perform the image editing and we intend to pursue this in future work. Table 2 shows an example of the process of understanding the image edit requests.

5 Incrementality

Table 3 shows example utterances for some of the most frequently occurring dialogue acts in the corpus. In these examples it can be seen that, with the exception of 3, all the other dialogue acts can be identified with some degree of certainty without waiting for the user to complete the utterance. Also, Figure 5 in the Appendix shows example IERs. One of the motivations for our work is to identify the right dialogue act at the earliest time.

Utterance	Segments	Dialogue Act	Action	Attribute	Location Object	Value
uh make the tree brighter <sil> to like a 100 <sil> nope too much 50 please	uh	O	-	-	-	-
	make the tree brighter	IER-N	Adjust	brightness	tree	-
	to like a 100	IER-U	Adjust	brightness	tree	100
	nope too much	COM-D	-	-	-	-
	50 please	IER-U	Adjust	brightness	tree	50
perfect <sil> let's work on sharpness	perfect	COM-L	-	-	-	-
	let's work on sharpness	IER-N	Adjust	sharpness	-	-

Table 2: Examples of commonly occurring dialogue acts, actions, and entities.

	Utterance	Tag
1	add a vignette since it's also encircled better	IER-N
2	can we go down to fifteen on that	IER-U
3	go back to .5	IER-U
4	actually let's revert back	IER-R
5	can you compare for me before and after	IER-C
6	I like it leave it there please	COM-L
7	no I don't like this color	COM-D

Table 3: Examples of some of the most commonly occurring dialogue acts in our corpus.

Not only is this more efficient but also more natural. The human Wizard can begin to take action even before the utterance completion, e.g., in utterance 1 the Wizard clicks the "vignette" feature in the tool before the user has finished uttering their request. Another goal is to measure potential savings in time gained through incremental processing, i.e., how much we save in terms of number of words when we identify the dialogue act earlier rather than waiting until the full completion of the utterance, without sacrificing performance.

6 Model Design

For our experiments we use a training set sampled randomly from 90% of the users (116 dialogues for training, 13 dialogues for testing). We use word embedding features whose construction is described in Section 6.1. There are several reasons for using word embeddings as features, e.g., unseen words have a meaningful representation and provide dimensionality reduction.[3]

[3]Figure 6 shows the visual presentation of the utterances embeddings using t-SNE (Maaten and Hinton, 2008).

6.1 Constructing Word Embeddings

We convert the words into vector representations to train our deep learning models (and a variation of the random forests). We use out-of-the-box word vectors available in the form of GloVe embeddings (Pennington et al., 2014) (trained with Wikipedia data), or we employ fastText (Bojanowski et al., 2017) to construct embeddings using the data from the Visual Genome image region description phrases, the dialogue training set collected during this experiment, and other data related to image editing that we have collected (image edit requests out of a dialogue context). From now on these embeddings trained with fastText will be referred to as "trained embeddings".

As we can see in Table 4, for models E (LSTMs) and I (CNNs) we use word embeddings trained with fastText on the aforementioned data sets. The Vanilla LSTM (model D) does not use GloVe or trained embeddings, i.e., there is no dimensionality reduction. Model H (CNN) uses GloVe embeddings. The vectors used in this work (both GloVe and trained embeddings) have a dimension of 50. For trained embeddings, the vectors were constructed using skipgrams over 50 epochs with a learning rate of 0.5.

Recent advancements in creating a vector representation for a sentence were also evaluated. We used the Sent2Vec (Pagliardini et al., 2018) toolkit to get a vector representation of the sentence and then used these vectors as features for models G and J. Note that LSTMs are sequential models where every word needs a vector representation and thus we could not use Sent2Vec.

6.2 Model Construction

We use WEKA (Hall et al., 2009) for the Naive Bayes and Random Forest models, MALLET

	Model	Accur
A	Baseline (Majority) *	0.32
B	Naive Bayes *	0.41
C	Conditional Random Field *	0.51
D	LSTM (Vanilla) *	0.53
E	LSTM (trained word embeddings) *	0.55
F	Random Forest *	0.72
G	Random Forest (with Sent2Vec)	0.73
H	CNN (GloVe embeddings)	0.73
I	CNN (trained word embeddings)	0.74
J	CNN (Sent2Vec)	0.74

Table 4: Dialogue act classification results for perfect segmentation. * indicates significant difference ($p < 0.05$) between the best performing models (I and J) and the other models.

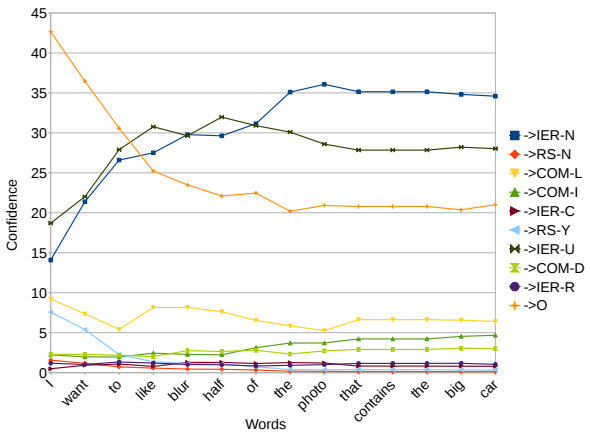

Figure 2: Confidence contours based on every word. The correct tag is IER-N. The confidence contours at the word level take time to stabilize.

(McCallum, 2002) for the CRF model (linear chain), and TensorFlow (Abadi et al., 2016) for the LSTM and CNN models. The models B, C, D, and F in Table 4 use bag-of-words features. The CNN has 2 layers, with the first layer containing 512 filters and the second layer 256 filters. Both layers have a kernel size of 10 and use ReLU activation. The layers are separated by a max pooling layer with a pool size of 10. The dense softmax is the final layer. We use the Adam optimizer with the categorical cross entropy loss function. The LSTM cell is made up of 2 hidden layers. We use a dropout with keep_prob = 0.1. We put the logits from the last time steps through the softmax to get the prediction. We use the same optimizer and loss function as for the CNN since they were found to be the best performing.

Table 4 shows the dialogue act classification accuracy for all models on our test set. Here we assume that we have the correct utterance segmentation for both the training and the test data. Note that because of the "Other" dialogue act all words in a sentence will belong to a segment and a dialogue act category. We hypothesize that the poor performance of the sequential models (CRF and LSTM) is due to the lack of adequate training data to capture large context dependencies.

6.3 Incrementality

Table 5 shows the savings in terms of overall number of words and average number of words saved per sentence, for each dialogue act in the corpus.

Figure 2 shows the confidence curves for predicting the dialogue act with the progression of every word. From this figure it is clear that after listening to the word "photo" the classifier is confident enough that the user is issuing the IER-N command. Here the notion of incrementality is to predict the right dialogue act as early as possible and evaluate the savings in terms of the number of words. While from this example it is clear that the correct dialogue act can be identified before the user completes the utterance, it is not clear when to commit to a dialogue act. The trade-off involved in committing early is often not clear. Table 5 shows the maximum savings that can be achieved in an ideal scenario where an oracle (an entity informing if the prediction is correct or wrong as soon as the prediction is made) identifies the earliest point of predicting the correct dialogue act.

The method used for calculating the savings is shown in Table 6. In this example for the utterance "I think that's good enough", we feed the classifier the utterances one word at a time and get the classifier confidence. The class label with the highest score is obtained. Here the oracle tells us that we could predict the correct class COM-L as soon as "I think that's good" was uttered and thus the word savings would be 1 word.

However, in real-world scenarios the oracle is not present. We use several confidence thresholds and measure the accuracy and the savings achieved in predicting the dialogue act without the oracle. For the predictions in the test set we get the accuracy for each of the thresholds. Then if the

Tag	% Overall Word Savings	Average Word Savings per Utterance
IER-N	37	3.96
IER-U	39	2.72
IER-R	41	1.63
IER-C	40	1.69
COM-L	36	1.13
COM-D	41	1.38
COM-I	37	2.56
RS-Y	28	0.34
RS-N	37	0.69
O	47	3.95

Table 5: Percentage of overall word savings and average number of words saved per utterance, for each dialogue act.

Utterance	Max conf	Class
I	0.2	O
I think	0.3	O
I think that's	0.3	O
I think that's good	**0.5**	**COM-L**
I think that's good enough	**0.5**	**COM-L**

Table 6: Example incremental prediction. The correct label is COM-L. Columns 2 and 3 show the maximum confidence level and model prediction after each word is uttered.

predictions are correct, we calculate the savings. Thus Figure 3 shows the word savings for each confidence threshold when the predictions are correct for that threshold.

So in the example of Table 6, for a confidence threshold value of 0.4, we extract the class label assigned for the utterance once the max confidence score exceeds 0.4. In this case once the word "good" was uttered by the user the confidence score assigned (0.5) was higher than the threshold value of 0.4 and we take the predicted class as COM-L. The word savings in this case is 1 word and our prediction is correct. But for a confidence threshold value of 0.2, our prediction would be the tag O which would be wrong and there would be no time savings. Figure 3 shows that as the confidence threshold values increase the accuracy of the predictions rises but the savings decrease.

Researchers have used simulations (Paetzel et al., 2015) or a reinforcement learning policy (Manuvinakurike et al., 2017) to learn the right

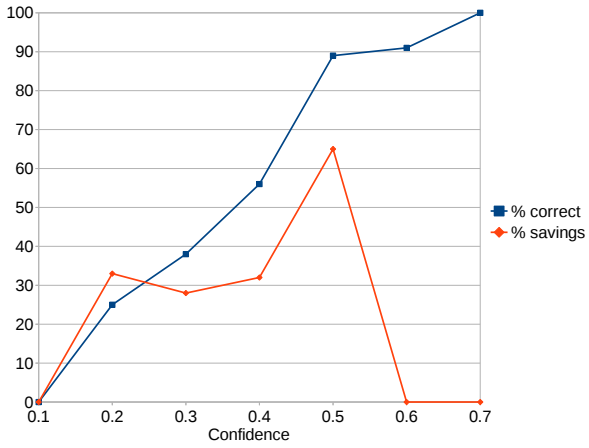

Figure 3: % savings (for correct predictions) and accuracy (% correct) of incremental predictions of dialogue acts as a function of confidence level.

points of interrupting the user which are dependent on the language understanding confidence scores. Here we do not focus on learning such policies. Instead, our work is a precursor to learning an incremental system dialogue policy.

7 Conclusion

We presented "conversational image editing", a novel real-world application domain, which combines dialogue, visual information, and the use of computer vision. We discussed why this is a domain particularly well suited for incremental dialogue processing. We built models for incremental intent identification based on deep learning and traditional classification algorithms. We calculated the impact of varying confidence thresholds (above which the classifier's prediction is considered) on classification accuracy and savings in terms of number of words. Our experiments provided evidence that incremental intent processing could be more efficient for the user and save time in accomplishing tasks.

Acknowledgments

This work was supported by a generous gift of Adobe Systems Incorporated to USC/ICT, and the first author's internship at Adobe Research. The first and last authors were also funded by the U.S. Army Research Laboratory. Statements and opinions expressed do not necessarily reflect the position or policy of the U.S. Government, and no official endorsement should be inferred.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Wuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, Georgia, USA.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of ICCV*, Santiago, Chile.

Timo Baumann and David Schlangen. 2013. Open-ended, extensible system utterances are preferred, even if they require filled pauses. In *Proceedings of SIGDIAL*, Metz, France.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of CVPR*, Honolulu, Hawaii, USA.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012. Optimising incremental dialogue decisions using information density for interactive systems. In *Proceedings of EMNLP–CoNLL*, Jeju Island, Korea.

David DeVault, Kenji Sagae, and David Traum. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue and Discourse*, 2(1):143–170.

Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

Fabrizio Ghigi, Maxine Eskenazi, M. Ines Torres, and Sungjin Lee. 2014. Incremental dialog processing in a task-oriented dialog. In *Proceedings of INTERSPEECH*, Singapore.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of NAACL–HLT*, San Diego, California, USA.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of NAACL–HLT*, San Diego, California, USA.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the ACL Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of ACL*, Beijing, China.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING*, Osaka, Japan.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. 2015. Optimising turn-taking strategies with reinforcement learning. In *Proceedings of SIGDIAL*, Prague, Czech Republic.

Dongho Kim, Catherine Breslin, Pirros Tsiakoulis, Milica Gašić, Matthew Henderson, and Steve Young. 2014. Inverse reinforcement learning for micro-turn management. In *Proceedings of INTERSPEECH*, Singapore.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, Lake Tahoe, Nevada, USA.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Abhinav Kumar, Barbara Di Eugenio, Jillian Aurisano, Andrew Johnson, Abeer Alsaiari, Nigel Flowers, Alberto Gonzalez, and Jason Leigh. 2017. Towards multimodal coreference resolution for exploratory data visualization dialogue: Context-based annotation and gesture identification. In *Proceedings of SemDial*, Saarbrücken, Germany.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of NAACL–HLT*, San Diego, California, USA.

Wei Li and Yunfang Wu. 2016. Multi-level gated recurrent neural network for dialog act classification. In *Proceedings of COLING*, Osaka, Japan.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Ramesh Manuvinakurike, Jacqueline Brixey, Trung Bui, Walter Chang, Doo Soon Kim, Ron Artstein, and Kallirroi Georgila. 2018. Edit me: A corpus and a framework for understanding natural language image editing. In *Proceedings of LREC*, Miyazaki, Japan.

Ramesh Manuvinakurike, David DeVault, and Kallirroi Georgila. 2017. Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of SIGDIAL*, Saarbrücken, Germany.

Ramesh Manuvinakurike, Casey Kennington, David DeVault, and David Schlangen. 2016a. Real-time understanding of complex discriminative scene descriptions. In *Proceedings of SIGDIAL*, Los Angeles, California, USA.

Ramesh Manuvinakurike, Maike Paetzel, Cheng Qu, David Schlangen, and David DeVault. 2016b. Toward incremental dialogue act segmentation in fast-paced interactive dialogue systems. In *Proceedings of SIGDIAL*, Los Angeles, California, USA.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of EACL*, Avignon, France.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of IJCNLP*, Taipei, Taiwan.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of ACL*, Berlin, Germany.

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. In *Proceedings of SIGDIAL*, Saarbrücken, Germany.

Maike Paetzel, Ramesh Manuvinakurike, and David DeVault. 2015. "So, which one is it?" The effect of alternative incremental architectures in a high-performance game-playing agent. In *Proceedings of SIGDIAL*, Prague, Czech Republic.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of NAACL–HLT*, New Orleans, Louisiana, USA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, Doha, Qatar.

Volha Petukhova and Harry Bunt. 2014. Incremental recognition and prediction of dialogue acts. In *Computing Meaning*, pages 235–256. Springer.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proceedings of SIGDIAL*, London, UK.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue and Discourse*, 2(1):83–111.

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. 2013. Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of SIGDIAL*, Metz, France.

Sheng-syun Shen and Hung-yi Lee. 2016. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. In *arXiv preprint arXiv:1604.00077*.

Gabriel Skantze. 2017. Towards a general continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of SIGDIAL*, Saarbrücken, Germany.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A generative attentional neural network model for dialogue act classification. In *Proceedings of ACL – Short Papers*, Vancouver, Canada.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! visual object discovery through multi-modal dialogue. In *Proceedings of CVPR*, Honolulu, Hawaii, USA.

Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. 2010. I2T: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508.

Figure 4: The interface as seen by the user and the Wizard. We use Adobe Lightroom as the image editing program.

Tag	User Edit Requests
IER-N	I want to um add more focus on the boat
IER-N	can you make the water uh nicer color
IER-N	uh can we crop out uh little bit off the bottom
IER-N	is there a way to add more clarity
IER-N	can we adjust the shadows
IER-U	more [saturation]
IER-U	can we get rid of the hints of green in it
IER-U	bluer
IER-U	little bit more from the left [crop]
IER-R	can you unfocus it
IER-C	can you show me before and after

Figure 5: Example user edit requests. Only two bounding boxes are labeled in the image for better reading. The actual images have more extensive object labels.

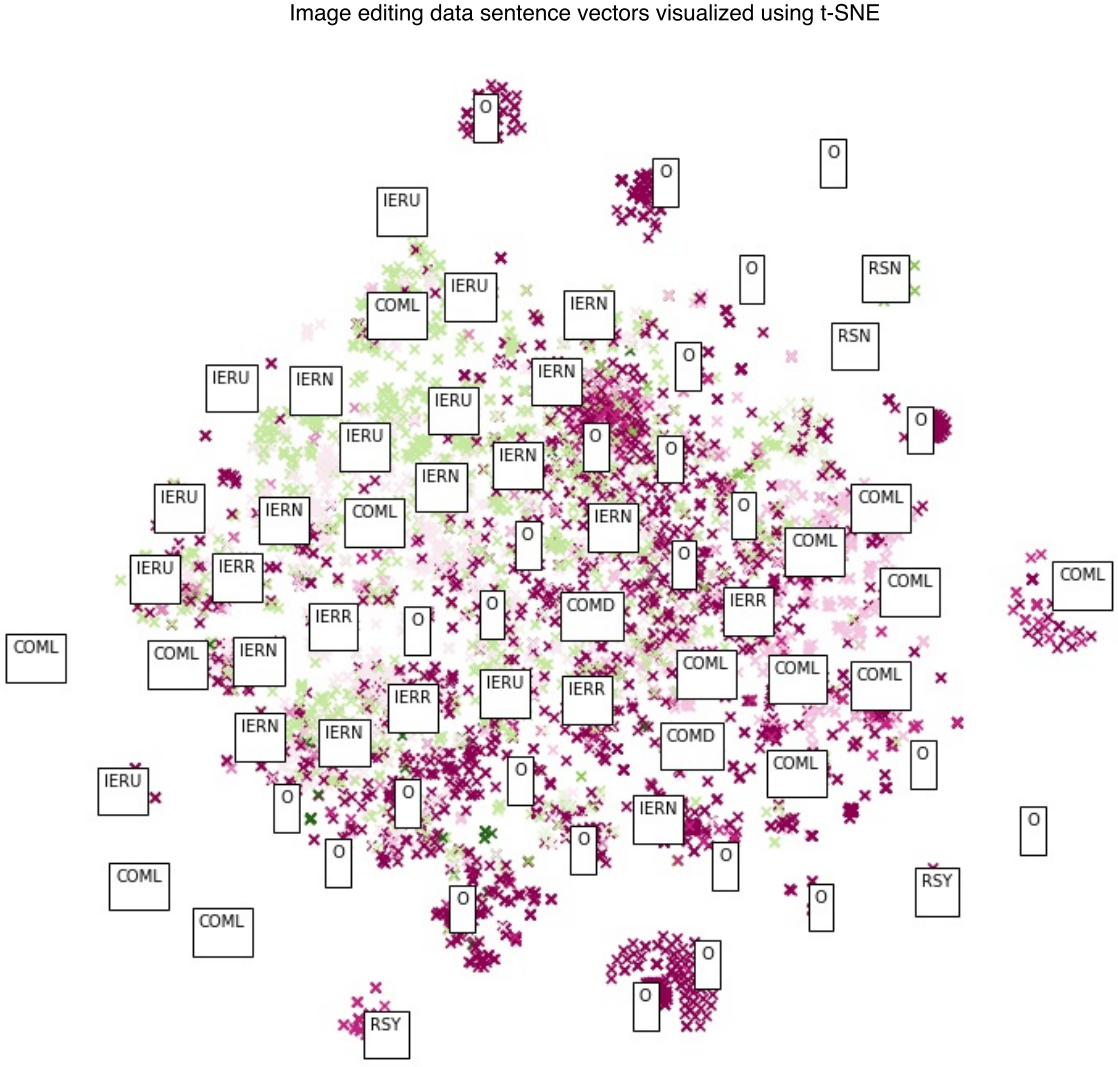

Figure 6: Visualization of the sentence embeddings of the user utterances used for training. The t-SNE visualizations after half-way through the utterances are shown. The utterances that have the same dialogue acts can be seen grouping together. This shows that the complete utterance is not always needed to identify the correct dialogue act.