Incorporating Visual and Textual Cues in Dialogue Generation: An Application to Comic Strips

Aniketh Janardhan Reddy [1] **Ramesh Balaji** [2]

Abstract

Conventional dialogue generation systems mostly leverage textual data and are incapable of detecting and acting on the visual cues while making conversation. Thus, they cannot be used to generate dialogue-oriented compositions such as scripts for television or comic strips which heavily use visual cues. In this work, we try to overcome this obstacle and propose a system which can make use of such cues to generate comic strips. First, we propose a baseline approach based on a conditional variational autoencoder which is only capable of predicting the last speech bubble of a strip. We then model the task as a visual story telling problem and adapt an encoder-decoder style model in order to generate entire comic strips. So as to test this story telling-based approach, we propose new metrics and also perform a qualitative human evaluation on the results. We notice that this model is able to detect the setting of a strip and the characters involved in most cases. It is also able to generate some coherent strips. We believe that the results are promising and that they warrant further research in this area.

1. Introduction

A dialogue is a written or spoken composition in which two or more characters converse. A dialogue can be seen as a complex, dynamic and context-dependant method of creating meaning. Systems which can automatically generate meaningful dialogues could be used for building artificial personal assistants, story generation, etc. Previous research has mostly been focused on dialogue systems which make use of purely textual data. However, many conversations between humans allude to the environment in which the conversation occurs. Similarly, many human compositions such as graphic novels and comic strips use both visual and textual cues to illustrate a conversation. Dialogue generation for use in such media requires systems to detect both textual and visual cues. We carry out our investigations on the medium of comic strips though our ideas could be extended to other similar media. Concretely, our goal was to build a system which can generate sensible and coherent dialogues for an entire comic strip given its visuals. Additionally, the dialogues must progressively build a storyline which fits the image's context.

As a first step towards building such a system, we tried to generate the text in the speech bubble of the last panel of the comic strip, given the transcripts of the dialogues in the first three/four panels using a Conditional Variational Autoencoder (CVAE)-based dialogue generation model (Zhao et al., 2017). We then modeled the task of comic strip generation as a modification of the visual story telling task. This allowed us to effectively use the visual cues present in the comics. It also made it possible to generate entire comic strips using just the images of the panels. The model proposed by Smilevski et al. (Smilevski et al., 2018) was used for this purpose.

Given the nature of the task, it is very hard to assess the performance of a system which aims to complete it. We propose two simple automated metrics based on Latent Dirichlet Allocation (LDA) and character identification capabilities which are indicative of the system's ability to capture visual cues and create relevant content. A human evaluation is also performed in order to ascertain the true quality of the generated strips.

2. Literature Survey

Our problem of predicting the dialogues in a comic strip is related to other problems such as dialogue generation, visual story telling, visual question answering (VQA), and language modeling.

Dialogue generation involves the prediction of a response to a natural language statement. This task is relevant to our work because our goal is to build on vanilla dialogue generation by incorporating visual cues.

[1]AndrewID : ajreddy [2]AndrewID : rbalaji. Correspondence to: Aniketh Janardhan Reddy <ajreddy@cs.cmu.edu>, Ramesh Balaji <rbalaji@cs.cmu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Zhao et al. (Zhao et al., 2017) propose a conditional variational autoencoder (CVAE) based approach for dialogue generation. Their model is capable of generating diverse responses based on both the context in which a dialogue is supposed to be generated and certain meta info about the speaker and the conversation (ex. speaker's attributes, topic of the conversation, dialog acts (Poesio & Traum, 1998), etc.). However, their model can only handle two speaker dialogue generation.

Multi-turn dialogue generation was recently explored by Wu et al. (Wu et al., 2018). In contrast to older approaches which used just the previous sentence (single-turn) to predict the next dialogue, multi-turn dialogue generators incorporate information which was presented in all of the previous sentences while making this prediction. Multi-turn systems are more useful for us because dialogues in comics do not always build on immediately available information but often allude to much older dialogues. Wu et al. proposed an encoder-decoder-based approach which also makes use of attention and a reasoning model. First, each of the previous sentences is passed through a bidirectional recurrent neural network (biRNN)-based encoder which operates at the character level. The hidden states of the biRNN are used to derive various "memories". These memories are then segregated based on the agent who made the statement. The memories generated for each sentence are also passed through another RNN to obtain another set memories which are representative of entire sentences. All of the memories that are generated during the encoding stage are used by a decoder which employs attention and a reasoning model to finally generate the next dialogue. The generated dialogue can then be used when we want to generate the next response.

Adversarial learning is another common approach to dialogue generation. Li et al. (Li et al., 2017) propose a system which consists of a generator and discriminator which are jointly trained using modified policy gradient training by modeling the problem as a reinforcement learning (RL) task. The generator is used to generate a dialogue and the discriminator outputs a score which measures how indistinguishable the machine-generated dialogue is from the human-generated dialogue. This score serves as a reward to the generator. The generator is based on the Seq2Seq model proposed by Sutskever et al. (Sutskever et al., 2014) and the discriminator is a binary classifier which utilizes a hierarchical encoder to create vector representations of its inputs.

VQA is another task which is relevant to our work. Given an image, a set of natural language questions can be derived from the relationships between the objects in the image. VQA systems generate the answers to such questions. Since such systems need to understand visual cues in order to answer the questions, their design can be incorporated into our framework to better understand the comics. Many VQA systems are also capable of handling textual information. Most recently, Lewis et al. (Lewis & Fan, 2019) proposed such a system which takes a generative approach to question answering. The approach models the joint distribution of questions and answers given the contextual information such as relevant images or text. First, the contextual information is encoded using an RNN-based model if it is textual or is encoded using a CNN-based model if it is visual. Then, a prior over all the possible answers given the context is computed. Using a conditional language model, a distribution over the possible questions given the answer and the context is calculated. Therefore, the joint can now be computed using the inferred distribution over the questions given the answers and the context and the prior distribution over the answers given the context. The model is trained by minimizing the negative log likelihood of the joint distribution of questions and answers given the contextual information. During testing, the answer which has the highest probability given the question and the context is the final output of the model.

Language modeling is another task which we look at as being important in order to accomplish our task. Given the words in a sentence upto timestep i, language models output a probability distribution over the words which occur in the next timestep $i + 1$. Usually, language models encode vector representations for every possible word. These representations or embeddings inform us about the context in which a word occurs and help us in generating a distribution over all possible words given the context. Hence, language models are important tools in sequential text generation. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and ELMo (Embeddings from Language Models) (Peters et al., 2018) are two recent systems which perform remarkably well at modeling language. In contrast to traditional embeddings such as GloVe (Pennington et al., 2014) which generate fixed embeddings for each word, BERT and ELMo generate context dependent embeddings. In other words, they generate different embeddings for the same word based on the sentence in which it occurs. Hence, if a word has more than one meaning, every meaning should ideally be represented by a different context dependent embedding. BERT makes use of several layers of Transformers (Vaswani et al., 2017) to perform this modeling while ELMo makes use of bidirectional Long Short-Term Memories (LSTMs) (Hochreiter & Schmidhuber, 1997). Since these are deep models, every stage of the encoding process generates an embedding for a word. Some of these intermediate embeddings are useful in performing other related tasks such as sentiment analysis. BERT and ELMo can also be fine-tuned so as to perform better on the task at hand.

3. Data collection and Preprocessing

We conduct our experiments using Dilbert strips which are freely available on the web. Drawn by Scott Adams, Dilbert was the first syndicated comic strip to be published for free on the Internet. Around 7000 strips have been published to date, out of which we were able to obtain 5700. All of them contain transcripts that are written by various people and therefore are of variable formats. The strip is drawn in colour and weekday strips have 3 panels each whereas Sunday strips usually have 8 panels each. The comics feature five primary characters and 17 secondary characters.

3.1. Collecting Images and Transcripts

We crawl `dilbert.com` to retrieve the comic strip images by date from 1989 to 2014 by means of a python script. We also retrieve the transcripts corresponding to each day's strip. A typical image and its transcript are given in Figure 1.

Figure 1. Transcript: Dilbert asks Dogbert who is sitting on the bed, "Do you like my new clip-on necktie?" Dogbert replies, "It's very nice. Good colors, nice pattern. Why, with a tie like that, DON'T be surprised if you get an offer to pose for GQ MAGAZINE!" Dilbert says, "I think you crossed that fine line between polite lying and outright sarcasm." Dogbert replies, "The momentum carried me."

3.2. Processing Images

One complication is that the transcript is also a part of the image and this is a potential source for data leak, i.e. it is possible that the system learns to the recognize characters in the comic strip image which could potentially lead to better but misleading results. Therefore we obscure the text parts of the image. This is done by detecting the bounding box for text in the image. We use Tesseract (Smith, 2007), an open source OCR library to detect the bounding box for the text regions and then remove the text by superimposing a filled rectangle of similar colour onto the image. The colour of the filled rectangle is chosen from among the nearby pixels to maintain continuity. The image processing is illustrated in Figures 2 and 3.

3.3. Quote and Speaker Extraction

When we examined the 5700 transcripts which we obtained, we noticed that 499 of them were not correctly formatted and many of them did not mention the speakers of various quotes. Hence, these transcripts and their corresponding strips were not used for further analyses. Then, the NeuralCoref [1] coreference resolution system was used to replace the pronouns in the transcripts by the nouns which they referred to. These replacements are not made in quotes so as to retain their structure. StanfordNLP's (Manning et al., 2014) quote extraction system was then used to extract the quotes in the transcripts. It was observed that the speaker of a quote was usually the first noun of the sentence which contained the quote. Hence, every quote was attributed to the first noun in the sentence which contained the quote. Though this leads to some errors, we noticed that this simple method is better than using StanfordNLP's quote attribution feature. We also convert the entire transcript to lower-case and strip it of all punctuation marks.

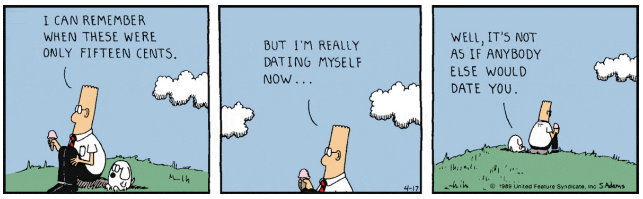

Figure 2. Original image

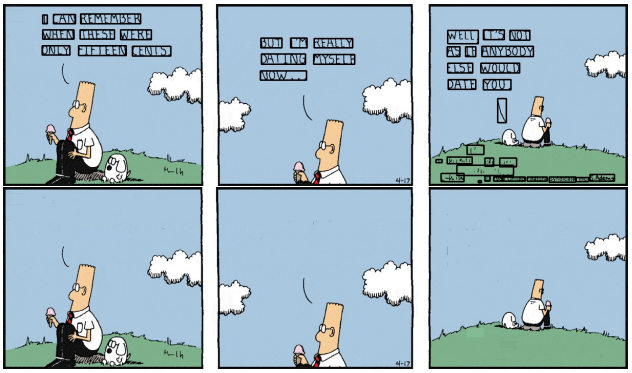

Figure 3. Original image processed to prevent data leak

4. Methods

4.1. CVAE and LDA-based Baseline Approach

In order to ascertain the applicability of current text-based dialogue generation methods to our problem, we developed a baseline system which uses LDA and a CVAE to predict the contents of the last speech bubble of a comic strip given the contents of the previous speech bubbles. This approach consists of two main steps:

[1] https://github.com/huggingface/neuralcoref

1. Determine the topic of a comic strip using LDA.

2. Use a CVAE to predict the last statement of the strip based on this topic, speaker identities and context.

Each of these steps are detailed below:

4.1.1. Document Topic Modelling using Latent Dirichlet Allocation (LDA)

The topic of a given comic strip is not known to us and many dialogue generation models make use of the topic of a conversation to generate relevant dialogues. So, to automatically discern the topic of a comic strip, we use LDA (Blei et al., 2003). More specifically, we assign one of seven possible topics to each strip using the following procedure (we determined that there were 7 topics by looking at the distribution of the number of documents which were assigned a given topic):

1. Each comic strip's transcript is first tokenized.

2. Then, stopwords, words which occur in more than 80% of the strips and words which occur in just one strip are removed from these tokenized transcripts.

3. An LDA model is then trained using all of the transcripts by employing Gensim (Řehůřek & Sojka, 2010).

4. Every transcript's topic distribution is determined using the trained model. The topic with the highest contribution to a strip's transcript is assigned to it.

4.1.2. Dialogue Generation using a CVAE

Our baseline approach uses the model proposed by Zhao et al. (Zhao et al., 2017). The model uses a CVAE to generate diverse responses by capturing the diversity at the level of discourses in the encoder. It does so by learning a distribution over the possible intents of a response using certain latent variables. Then, it uses a greedy decoder to actually generate the response. It is also trained using a novel bag-of-words loss. This model is only capable of generating dialogues which have two speakers. The model is trained using dialogues, speaker meta information and transcript topics. More details about the training and evaluation of this model are in Section 5.1.

In order to overcome the shortcomings of this model - the fact that it cannot handle more than two speakers and its inability to use visual information in order to generate entire comic strips from scratch, we now propose another visual story telling based model.

4.2. Incorporating visual information in the dialogue generation process

Our second approach is to treat the dialogues for a particular image as the caption for that particular scene and treat the problem as a Visual Storytelling problem. Visual story telling is an area of active research that seeks to make sense of visual input to tie disparate moments together as they give rise to a cohesive narrative of events through time. The overall structure of the model architecture is that of an encoder-decoder module as given in Figure 4. It has the following components-

Encoder: This is composed of two individual encoders - one, for the input image sequence and another, for the previous predicted sentence. The Image encoder consists of a Convolutional Neural Network (CNN) which outputs image embeddings corresponding to the series of images. These embeddings are then fed at each timestep into a visual-encoder Recurrent Neural Network (RNN). In this work we use a GRU (Gated Recurrent Unit) Network which is a type of RNN that can capture long term dependencies in sequences and at the same time uses lesser number of parameters as compared to an LSTM (Long short term Memory) Network. The sentence-encoder RNN takes as input the previous generated sentence and generates word level embeddings by passing it through an embedding layer. These word-level embeddings serve as input to the encoder which is again a GRU Network. Both the visual-encoder and sentence-encoder produce a fixed length vector as output.

Decoder: The decoder is a language model that acts on a series of word level embeddings to produce sentences. It is again a GRU network that uses the concatenation of the outputs of the two encoders as its initial hidden state.

Figure 4 shows the complete architecture of the encoder-decoder network.

5. Experimental Setup

5.1. Training and Evaluating the CVAE-based Baseline

Out of the 5201 strips which we retained, we further exclude 2117 strips because they contain more than two speakers (as mentioned before, this model can only handle two speaker conversations). Of the 3084 strips which remain, we use 2158 strips (∼70%) for training the model, 309 strips (∼10%) for validation and 617 (∼20%) strips for testing the model. For all of these strips, we additionally replace the original speaker of a quote by a "generic speaker" tag if the speaker of the quote does not appear in 10 or more strips. This is because we want to avoid overfitting the model to characters whose talking styles are not fully defined. This also helps us in overcoming some of the parsing errors which might have occurred in the speaker attribution

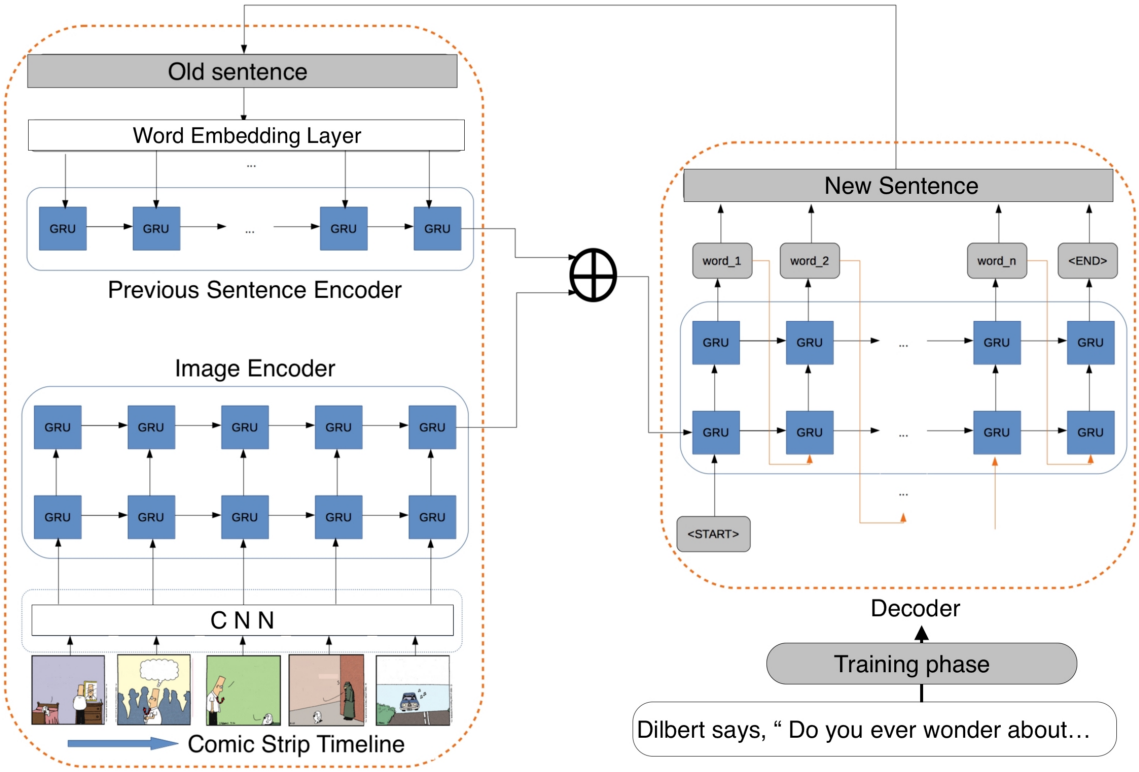

Figure 4. Architecture of the proposed model adapted and modified from (Smilevski et al., 2018)

process.

Then, the topics of these 3084 strips are determined using LDA. The quotes, the speakers and the topics of the strips in the training set are supplied as training data to the model. We then evaluate our model using the test set.

5.2. Training and Evaluating VIST (Visual Storytelling) style captioning

In the data preprocessing step we divide the quotes present in the transcripts into approximately three equal parts so that we have a 1 to 1 correspondence between every panel and caption. We construct the model as described in section 4.2. The encoder creates a hidden state representation of the images which is used by the decoder to generate the dialogue. We employ the word-level Cross-Entropy loss to train the network, which is computed using the output of the decoder and the original transcripts . We use 4395 comic strips for training and 444 strips for testing. This is after excluding the custom layout Sunday strips. We use AlexNet (Krizhevsky et al., 2012) as our CNN model. We use `Glove` (Pennington et al., 2014) for getting the word embeddings. The parameters used for training the network are given in table 1. In the decoder component, during the training phase the input at each time step is a word from the transcript, whereas during evaluation phase the input is the output of the previous time step.

5.3. Evaluation Metrics

5.3.1. Automated Evaluation Metrics

Since we try to predict the last dialogue of a comic strip using our baseline approach, it makes sense to use the BLEU score (Papineni et al., 2002) in order judge the performance of the baseline. Furthermore, 10 possible responses which are supposed to be equivalent to a given target response are generated. We report two different types of BLEU-3 scores based on these responses - average recall BLEU-3 score and average precision BLEU-3 score. The former is computed as the average of the modes of the BLEU-3 scores of the 10 responses across the test set. The latter metric is computed as the average of the means of the BLEU-3 scores of the 10 responses across the test set. These measures were described by Zhao et al (Zhao et al., 2017).

The evaluation of the outputs of the story telling-based model is much harder. We propose an LDA-based metric called $AgOLDA$ to automatically assess its performance. It is computed as follows:

1. The transcripts of the comic strips are tokenized. Stop-words, frequent words (occur in more than 80% of the strips) and words which occur in just one strip are

Parameter	Value	Description
batch_size	13	Batch size
epochs	50	Number of epochs to train for.
image_encoder_latent_dim	1024	Latent dimensionality of the encoding space.
sentence_encoder_latent_dim	1024	Latent dimensionality of the encoding space.
word_embedding_size	300	Size of the word embedding space.
num_of_stacked_rnn	2	Number of Stacked RNN layers
cell_type	GRU	RNN unit type
optimizer	Adam	Optimizer type
learning_rate	0.0001	Learning rate for Adam
gradient_clip_value	5	Ceiling for gradient Update

Table 1. Parameters of the visual story telling network

Score	Description
0	Unintelligible- no structure to sentences
1	1-2 dialogues are done well
2	All dialogues are done well
3	Dialogues are coherent with the picture.
4	Dialogues form a coherent narrative.
5	Perfect, a human could have written it

Table 2. Rubric used for giving the generated output a score between 0-5

removed from these tokenized transcripts.

2. Using Gensim (Řehůřek & Sojka, 2010), an LDA model is then trained using the original transcripts of comics in the training set. The number of topics is set to four.

3. The topic distribution for every original transcript in the testing set is determined using the trained model. The same is done for every generated transcript.

4. The topic with the highest contribution to a strip's transcript is assigned to it. Then, the $AgOLDA$ metric is computed as the ratio between the number of comics in the testing set for which the generated and original transcripts are assigned the same topic and the total number of comics in the testing set.

This metric is indicative of the ability of the model to capture visual cues since the only information shared between the original and generated transcripts is in the form of the images of strip. Assuming that these images are indicative of the topic of the conversation (which is usually the case), this metric allows us to gauge how well the model is able to detect and act on visual cues as it tells us if the model is broadly able to preserve context.

Another metric which is indicative of this capability is based on the correct identification of the characters in the scene. This metric which we call the $CharLap$ score is measured as the average of the ratio between the number of characters which have dialogues in both the generated and original transcripts for a given strip and the total number of characters in the generated transcript which have a dialogue across the testing set.

5.3.2. Human evaluation

We randomly select 50 samples from the generated output (for the story-telling model only) and score them individually by hand. The rubric that we use for scoring is given in table 2.

6. Results and Discussion

6.1. Qualitative

Figures 5 and 6 give two examples of generated samples. Qualitatively we can say a few things about the generated output.

1. **Identifying characters correctly:** Since the output is in the same format of the input we can verify if the system is able to recognize the characters in the comic strip correctly. In the first example it recognizes Dilbert and Dogbert and in the second example it recognizes Dilbert and terms the unknown character 'man'. It is to be noted that this is a non-recurring character and the system still manages to give him an appropriate label.

2. **Identifying mood/setting correctly:** In the first example the setting is a park in which two friends sit talking about something. The colouring and the image suggests that this is a contemplative scene where the characters are likely to talk about something philosophical. The generated output conversation matches this mood correctly in that it focuses on abstract concepts. The second example in contrast to the first is an office setting and here the conversation is focused on office work and uses vocabulary including 'new org meeting', 'product change', etc. Thus, we can say that in these two examples the system seems to have learned how to adapt the conversation to the mood conveyed by the images.

3. **Humour:** It is difficult to consider any of the two examples to be humorous as such and this can be an interesting direction for future work. Our attempt at using textual-style transfer (Shen et al., 2017) to try and capture humour from the training transcripts was not successful.

6.2. Quantitative

The quantitative results are summarized in Table 3.

Model	Metric	Value
LDA + CVAE	Avg. Recall BLEU-3	0.372140
	Avg. Precision BLEU-3	0.313544
Story Telling	$AgOLDA$	61.04%
	$CharLap$	59.05%
	Human Eval Avg. Score	1.02

Table 3. Quantitative results

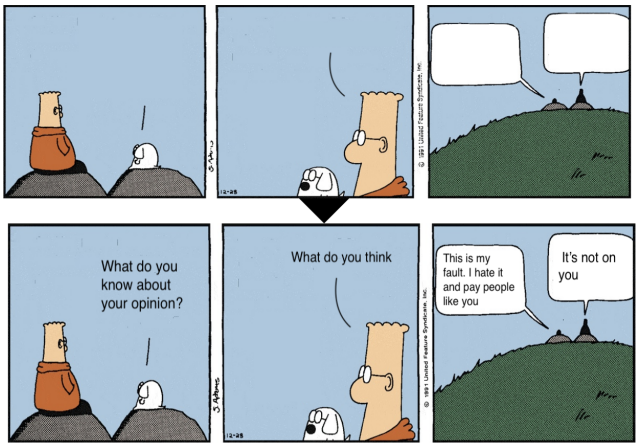

Figure 5. Result Example 1

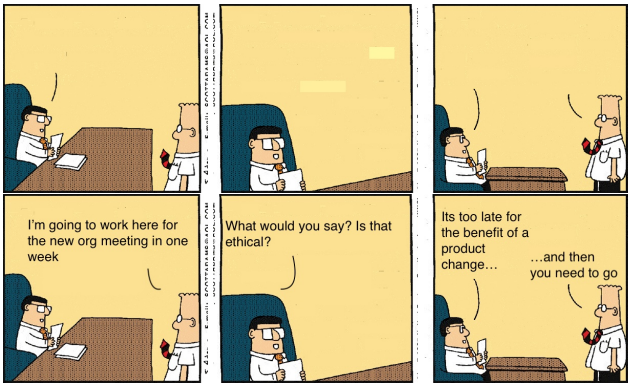

Figure 6. Result Example 2

6.2.1. Results obtained using the baseline approach

We evaluated the baseline approach described in 4.1.2 using the metrics described in Section 5.3.1. The CVAE-based model achieves an average recall BLEU-3 score of 0.372140 on the test set and it achieves an average precision BLEU-3 score of 0.313544 on the test set. Given that the authors of the original paper reported an average precision BLEU-3 score of 0.310 and an average recall BLEU-3 score of 0.318 using the Switchboard (SW) 1 Release 2 Corpus [2] which is bigger and much cleaner than our own dataset, these results are very encouraging although the authors considered the harder task of generating every response instead of just the last response. However, when we manually analyzed the generated responses, we observed that responses were not very sensible and coherent. This can be attributed to the model's inability to incorporate visual cues. Lack of a large

[2] https://catalog.ldc.upenn.edu/LDC97S62

amount of clean training data could be another issue.

6.2.2. Results obtained using the story telling approach

When we used the $AgOLDA$ score to evaluate our story telling-based model, we saw that 61.04% of the generated transcripts and their corresponding original transcripts were assigned the same topic by LDA. This shows that the model is indeed capable of capturing some visual cues.

Our model achieved a $CharLap$ score of 59.05%, indicating that it is quite good at identifying the characters in the scenes. We even noticed that some rare characters like Ratbert were also correctly identified by our model.

The results obtained using the human evaluation scheme described in section 5.3.2 can be seen in the figure 7. We find that the majority of the generated strips achieve a score of 1 i.e. 1-2 dialogues generated well. Also it is worth noting that none of the strips achieve a score of 4 and higher, which illustrates the difficulty of the problem.

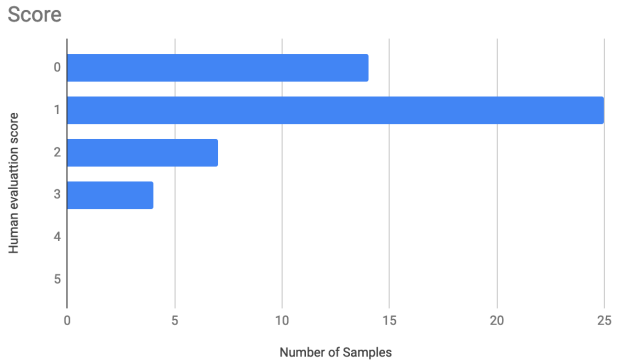

Figure 7. Human evaluation results

These metrics hint at the fact that visual storytelling models can be quite good at performing comic strip completion/generation. We believe that the main bottleneck is the amount of clean training data. We also believe that choosing comics which have more action in the panels could lead to better results.

7. Conclusion

In this work we have introduced the new and exciting task of comic strip completion. We have cast this problem in the mould of Visual Storytelling and used an encoder-decoder approach to solve it. The noteworthy aspects of this system are that it is able to recognize the characters in the comic strips and ascribe dialogues to them correctly 59% of the time. It is also able to adapt the conversation to the mood conveyed by the visuals. We have shown this qualitatively using two special examples and using LDA which gives a topic match of 61%. As far as we know, automated comic strip generation has never been attempted before. We believe that this work has the potential to be developed further in various directions. One such direction involves capturing humour in the generated dialogues. The other is to incorporate the style of a particular character explicitly to generate dialogues using the character's favoured vocabulary and manner.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lewis, M. and Fan, A. Generative question answering: Learning to answer the whole question. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkx0RjA9tX.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2157–2169. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1230. URL http://aclweb.org/anthology/D17-1230.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association*

for computational linguistics, pp. 311–318. Association for Computational Linguistics, 2002.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

Poesio, M. and Traum, D. Towards an axiomatization of dialogue acts. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues (13th Twente Workshop on Language Technology*. Citeseer, 1998.

Řehůřek, R. and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pp. 6830–6841, 2017.

Smilevski, M., Lalkovski, I., and Madjarov, G. Stories for images-in-sequence by using visual and narrative components. *CoRR*, abs/1805.05622, 2018. URL http://arxiv.org/abs/1805.05622.

Smith, R. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pp. 629–633. IEEE, 2007.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL http://arxiv.org/abs/1409.3215.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Wu, X., Martinez, A., and Klyen, M. Dialog generation using multi-turn reasoning neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2049–2059. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1186. URL http://aclweb.org/anthology/N18-1186.

Zhao, T., Zhao, R., and Eskenazi, M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 654–664, 2017.