Formulating Automated Responses to Cognitive Distortions for CBT Interactions

Ignacio de Toledo Rodriguez, Giancarlo Salton, Robert Ross
School of Computing, Technological University Dublin, Ireland
ignacio.toledo.rodriguez@gmail.com
{giancarlo.salton, robert.ross}@tudublin.ie

Abstract

One of the key ideas of Cognitive Behavioural Therapy (CBT) is the ability to convert negative or distorted thoughts into more realistic alternatives. Although modern machine learning techniques can be successfully applied to a variety of Natural Language Processing tasks, including Cognitive Behavioural Therapy, the lack of a publicly available dataset makes supervised training difficult for tasks such as reforming distorted thoughts. In this research, we constructed a small CBT dataset via crowd-sourcing, and leveraged state of the art pre-trained architectures to transform cognitive distortions, producing text that is relevant and more positive than the original negative thoughts. In particular, the T5 transformer approach to multitask pre-training on a sequence-to-sequence framework, allows for higher flexibility when fine-tuning on the CBT dataset. Human evaluation of the automatically generated responses showcases results that are not far behind from the overall quality of the ground truth scores.

1 Introduction

Recent studies (GDBC, 2018) estimate that approximately 300 million people globally suffer from depression, anxiety and other mental disorders. Cognitive Behavioural Therapy (CBT) is one of the leading practices across the field of psychotherapy (David et al., 2018) and one of the most effective ways of treating mental disorders such as anxiety or depression (Hofmann et al., 2012). CBT focuses on guiding the patients through a series of steps for identifying, analysing and correcting any cognitive distortions that may contribute to their mental health issues.

Traditional in-person CBT techniques applied in counselling sessions can be prohibitive for a large portion of the population due to cost, scarcity of therapists, convenience, stigma or other social considerations. However, in recent years there has been an increase in CBT material delivered online via computers and smartphone applications. In addition, a comprehensive review of these methods shows they can have many of the benefits of face-to-face therapy (Barak et al., 2008; Andersson and Cuijpers, 2009).

Automated agents that can deliver effective treatments represent a clear next step of research for online CBT. However, one of the main challenges here is the lack of publicly available datasets that can be used for training the necessary models. In light of these challenges, this research builds on the idea of a crowd-sourced corpus to generate CBT agent development by focusing on one of the foundational ideas of a CBT exercise, namely, the rewriting of distorted thoughts. Using this dataset, we then develop sequence-to-sequence (seq2seq) models to derive agents that can at least begin to address this central thought-rewriting challenge. While this is only an individual element of a complete CBT agent, it can be seen as a vital step in the study and analysis of the typical properties of CBT. In summary, the main contributions of this study are twofold:

- The creation of a Cognitive Behavioural Therapy dataset [1] that contains key information needed to train automated agents in producing CBT-related content, contributing to the development of Natural Language Processing (NLP) research in this domain.
- The use of modern machine learning techniques that demonstrate the effectiveness of leveraging a small CBT dataset to train a model to transform distorted negative thoughts into more realistic alternatives.

[1] https://github.com/itoledorodriguez/cbt-dataset

Cognitive Distortion	Description
All-or-Nothing Thinking	You see things in black and white categories. If your performance falls short of perfect, you see yourself as a total failure.
Overgeneralization	You see a single negative event as a never-ending pattern of defeat.
Mental Filter	You pick out a single negative detail and dwell on it exclusively so that your vision of reality becomes darkened.
Disqualifying the Positive	You reject positive experiences by insisting "they don't count" for some reason or other.
Jump to Conclusions - Mind Reading	You arbitrarily conclude that someone is reacting negatively to you, and you don't bother to check this out.
Jump to Conclusions - Fortune Teller Error	You anticipate that things will turn out badly, and you feel convinced that your prediction is an already established fact.
Magnification (Catastrophizing) or Minimization	You exaggerate the importance of things (such as your goof-up or someone else's achievements), or you inappropriately shrink things until they appear tiny (your own desirable qualities or the other fellow's imperfection).
Emotional Reasoning	You assume that your negative emotions necessarily reflect the way things really are: "I feel it, therefore it must be true".
Should Statements	You try to motivate yourself with shoulds and shouldnt's, as if you had to be whipped and punished before you could be expected to do anything. The emotional consequence is guilt. When you direct should statements toward others, you feel anger, frustration, and resentment.
Labelling and Mislabelling	This is an extreme form of overgeneralization. Instead of describing your error, you attach a negative label to yourself (eg: "I'm a loser"). When someone else's behaviour rubs you the wrong way, you attach a negative label to him (eg: "He's a goddam louse"). Mislabelling involves describing an event with language that is highly coloured and emotionally loaded.
Personalization	You see yourself as the cause of some negative external event which in fact you were not primarily responsible for.

Table 1: Definitions of the Cognitive Distortions used in this research. Taken from "Feeling Good: The new Mood Therapy" by Burns, D. 1981

2 Related Work

In the field of task-oriented Dialogue Systems, the technology has vastly improved since the introduction of ELIZA (Weizenbaum, 1966). Modern architectures such as Google Duplex (Leviathan and Matias, 2018) can handle complex goal-oriented conversations without human guidance, and novel approaches to frameworks such as Wizard-of-Oz (Wen et al., 2017) allows for the creation of crowd-sourced human datasets that can be used to train end-to-end agents towards a realistic conversation flow for different scenarios.

In the CBT domain, the highly rated and free of charge Woebot application is helping users around the world to identify and challenge cognitive distortions (Fitzpatrick et al., 2017). It combines template-based rules and modern machine learning techniques to deliver results but it does not, as of the time of writing, fully allow for the flexibility of a natural conversation.

The advancement in the last decade of machine learning, and in particular deep learning techniques for NLP, has made possible the development of automated models that excel at specific language tasks by being trained end-to-end over many iterations of large datasets, without the need for pre-established rules or templates. These techniques build on the seq2seq (Sutskever et al., 2014) and encoder-decoder architectures (Cho et al., 2014) to produce results in tasks such as machine translation, text summarization or sentiment analysis. In particular, the use of attention-based architectures (Bahdanau et al., 2015) that expand on the Transformer model (Vaswani et al., 2017) are widely used in the current state of the art models for NLP tasks.

Transfer Learning, a technique that was originally applied to the fine-tuning of computer vision tasks, has been a recent focus of NLP research, especially since ULMFit (Howard and Ruder, 2018) demonstrated how the weights of a LSTM language model pre-trained on a large dataset could be fine-tuned on a smaller corpus, for both language modelling and additional NLP tasks of the target dataset. Since then, other pre-trained models mostly based on the transformer architecture such as Elmo (Peters et al., 2018), GPT-2 (Radford et al., 2019) or BERT (Devlin et al., 2019), have been producing better results in diverse text generation and classification tasks.

When considering the rewriting of distorted or negative thoughts, this exercise can be compared to a seq2seq style transfer task where the situation or context remains the same, but the negative thoughts passed as inputs to the model are converted into more positive outputs. Shen et al. (2017) successfully demonstrate the effectiveness of style transfer in non-parallel data by mapping the inputs to a style-independent content representation.

3 Key Ideas in Cognitive Behavioural Therapy

A basic CBT interaction outlines a structure where the patients attempts to examine their own thoughts

Situation	Emotions	Negative Thoughts	Cognitive Distortions	Rational Response	Outcome
I had an important meeting that didn't go very well	Anxious 70% Sad 80%	I made a fool out of myself	Labelling Mind-Reading	It's true it wasn't my best meeting, but it's a big leap to label myself a fool just because I had a bad day. Also, you can't know what the rest of the people were thinking. Even if some thought that, they'd probably forget soon enough or do you remember all of the meetings conducted by your colleagues that didn't go that well?	Anxious 30% Sad 40%

Table 2: Daily Record of Dysfunctional Thoughts (Beck, 1979)

in terms of what they perceive to be a negative event, identifying any cognitive distortions and rephrasing them. In that process, the key steps are:

- Recognizing the situation that provoked the patient into experiencing a negative emotion and the intensity of those feelings.
- Writing down the automatic thoughts that accompany such emotions.
- Identifying any negative distortions that may be present in those thoughts (Table 1 shows the list of distortions considered in this study).
- Rewriting each distorted thought, aiming for a more rational or realistic alternative.
- Evaluating the patient feelings after the CBT exercise.

The patients with more experience in CBT techniques will be able to follow these steps by themselves in what is known as a CBT diary, also represented in Table 2. This is an exercise that allows them to immediately and effectively reduce their anxiety levels. However, and especially at the beginning of therapy, it is not always possible for the patients to come up with realistic alternatives that help combat their negative emotions. For that reason, a therapist can assist on guiding the patients through the main steps in the form of a conversation with a clear objective: i.e., reducing their anxiety levels.

When building a CBT dialogue corpus, much of the data needed is publicly available in forums, books or other online content – at least in raw format. It is relatively simple to identify in public forums negative situations where people express both their feelings and the distorted thoughts that accompany them. However, in this online content, there is one piece of information usually missing: the alternative, rational thought that will help patients to combat their negative feelings. This is a key part in CBT exercises, and it is at the same time the more difficult element to source when examining public data.

4 Data Collection

As part of integrating a CBT system within a modern machine learning dialogue framework, the previous section established the key idea of being able to transform irrational or distorted cognitive patterns into more realistic thoughts that are able to alleviate the negative emotions felt by the patients.

Hence, the main focus of our data collection has been the gathering of a series of negative thoughts that are objectively distorted and the use of crowdsourcing resources to obtain realistic counter arguments. More precisely, we build a dataset that contains multiple key value pairs for a single interaction, such as situation, emotions, negative thoughts, and rational response to those thoughts. All this data except the rational response is first prepared and then provided to the users that participate in the study.

As a first step in data preparation, a series of situations, feelings and negative thoughts were collected from a variety of sources such as CBT books, forums and public content aggregators. For those examples where the cognitive distortions contained within the negative thoughts are not mentioned explicitly, those distortions have been annotated manually. Note that the purpose of this research is not the cognitive distortion classification, but rather the rewriting of negative thoughts. The cognitive distortions just provide additional context for the survey users and help them to come up with a realistic counter argument.

The survey respondents were asked to read carefully the instructions and to provide, in their own words, a realistic alternative to the negative thoughts in each of the situations presented. For

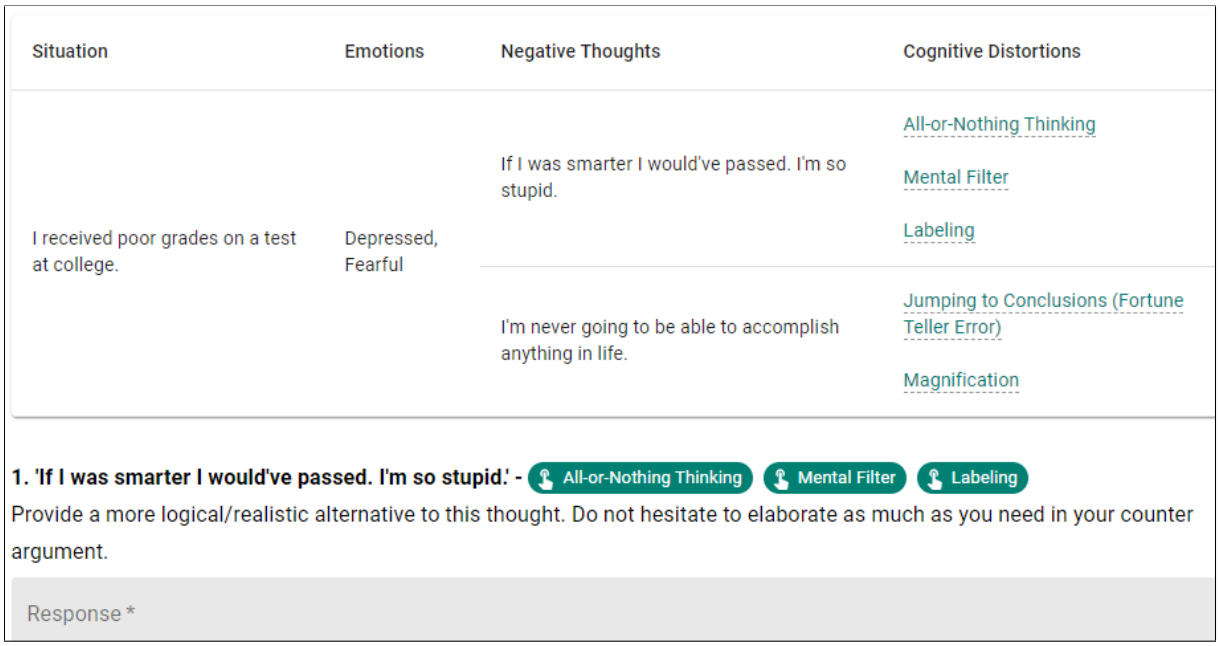

Figure 1: Example situation that contains negative distortions. Participants in the survey will write a more realistic counter-argument to each automatic thought.

	Counts
Situations	108
Type Count (%)	
Work	26.85
Romantic	22.22
Social	12.04
Friends	11.11
Family	10.19
Health	7.41
School and College	5.56
Other	2.78
Bereavement	0.92
Addiction	0.92
Negative Thoughts	200
Participants	114
Responses	442

Table 3: Number of responses gathered during the survey for the situations and negative thoughts that were prepared beforehand. Note that, for some situations, there have been multiple responses collected.

this study, the crowd-sourcing platform of choice has been Prolific [2], linked to a custom website (Fig 1) that loads two random situations for every participant, with an average of two negative thoughts per situation. Table 3 showcases the different situation types and the number of responses collected.

[2]https://www.prolific.co/

5 CBT Response Generation

While creating a new dataset is essential to our goals, the primary objective is to explore the use of modern deep learning architectures to automatically formulate appropriate responses against negative thoughts that can help to counter anxiety and depression. Overall, to do this, a number of seq2seq models that have produced good results in other NLP tasks are examined in this research.

5.1 Modelling Strategies

As a modelling strategy, we concatenate the situation description with each negative thought, forming a single sequence that serves as the input to the models, in a supervised learning approach. The target texts are those responses written by humans as per the crowdsourcing task from last section.

Due to the small dataset collected, and in order to produce significant results when trying to transform distorted thoughts into more realistic alternatives, the use of transfer learning and pre-trained language models is necessary. The responses generated with a model solely trained on the CBT dataset, regardless of the architecture used, do not achieve good results from the point of view of basic literacy or semantic coherence.

However, some of the pre-trained models used during the research, such as a simple transformer architecture, are not nuanced enough to allow for the small CBT dataset to significantly influence the

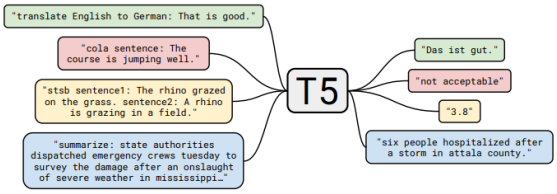

Figure 2: Google T5 Model Overview. The input text contains prefix keywords that allows for parallel training in different downstream tasks.

results produced in a seq2seq cognitive rephrasing task. In order to achieve an effective transfer learning strategy when fine-tuning for the response generation task, this study leverages one of the current state of the art architectures, namely the T5 transformer.

5.2 T5 Transformer

Recently, Google published the T5 text-to-text transformer model (Raffel et al., 2020), trained on a cleaned version of the Common Crawl corpus (C4). The published model checkpoints have been pre-trained in a diverse variety of unsupervised and supervised tasks (language modelling, word embedding, machine translation, text summarization, etc), which allows for the flexibility of fine-tuning smaller datasets in any NLP downstream tasks. During training and evaluation, the model is able to recognize prefix tokens which are added to the input text, in order to distinguish between the different tasks (Figure 2).

The architecture of the T5 transformer is very similar to that of the original transformer, keeping a stack of encoder and decoder blocks, each composed of self-attention layers and feed-forward networks. There are some modifications around layer normalization and position embedding, but what makes the pre-trained models excel at transfer learning is the multi-task approach to pre-training when applied to their large C4 dataset, and scaling up the number of parameters of the model.

While the largest model checkpoint - T5-11B, with 11 billion parameters, is able to exceed previous benchmarks in tasks such as GLUE or SQAD, for the purposes of this research and, due to processing constraints, the T5-Large model with 770 million parameters has been leveraged when fine-tuning the CBT task.

5.3 Baseline

We also make use of a non-trivial baseline model consisting of a seq2seq framework where both the encoder and decoder are composed of a BERT base architecture (Devlin et al., 2019). As one of the first transformer models fully pre-trained on a plain text corpus, mostly on the English Wikipedia and the BooksCorpus (Zhu et al., 2015), we choose BERT to contrast and showcase the advancement, in a relatively short span of time, of these type of pre-trained architectures when fine-tuning on a smaller dataset.

6 Experiments

To evaluate the quality of our automatically constructed responses we make use of both quantitative metrics and survey-based human evaluation, comparing aspects such as fluency, positive sentiment and overall quality of the text produced.

6.1 Metrics

As discussed in previous sections, rewriting negative thoughts into a more positive or realistic version can be considered a style transfer task. With this in mind, we have considered metrics that have been commonly applied to the style transfer objective. (Yang et al., 2018).

Here we have specifically made use of **Perplexity**, **BLEU** (Papineni et al., 2002) and **METEOR** (Banerjee and Lavie, 2005). These metrics are also commonly applied across a range of other NLP tasks such as machine translation and image captioning. Each of them can be thought of as providing an assessment of how much our predicted text seems to match that of the original labels.

Sentiment Analysis. We also make use of a pre-trained sentiment classifier model which we fine-tune for the CBT dataset. The classifier determines whether the alternative responses produced by the text-to-text transformers are considered positive or negative, obtaining an average accuracy for each of the models. The model we chose for this sentiment analysis has been RoBERTa (Reimers and Gurevych, 2019), pre-trained on the Yelp dataset reviews and fine-tuned on the CBT dataset, where all of the inputs or negative thoughts are considered as negative and all of the targets as positive.

The evaluation set has been used to compute the automatic metrics score. Except for perplexity, the experiments generate thirty different responses for each of the inputs in the target, averaging the

	Automatic Metrics				Human Evaluation		
	Perplexity	BLEU	METEOR	Sentiment	Rel.	Sentiment	Quality
Google T5	**18.21**	**0.016**	**0.094**	70.29%	**3.74**	**3.77**	**3.60**
BERT Seq2Seq	71.09	0.012	0.077	**90.72%**	2.74	3.30	2.56
Human	-	-	-	-	4.01	4.18	4.05

Table 4: Evaluation results for the CBT dataset, for both automatic metrics and human assessment. For reference, the table also includes the human evaluation of the dataset labels.

results to obtain the BLEU, METEOR and sentiment scores. The model also uses nucleus sampling (Holtzman et al., 2019) with a top-p value of 0.95, to allow for diversity in the responses.

6.2 Human Evaluation

In order to subjectively evaluate the responses generated by the models under study, and to contrast them against the original human labels, a number of surveys have been sent to users of the Prolific crowd-sourcing platform. The participants are restricted to those that have English as their first language.

The surveys were divided in three groups of 20 participants each. One group evaluated the original human targets from the dataset, while the other two groups received the responses from the T5 and BERT models. The survey in each group contains the same five different situations picked from the evaluation set, along with their initial negative thoughts, and the only difference is the generated responses.

The methodology followed for choosing the generated responses included in the survey was to produce three different responses for each of the situations, and pick the subjective best one from those. Appendix A includes all three responses generated for each situation by the T5 model.

The questions asked in the survey attempt to evaluate the generated responses in terms of relevance, positive sentiment and, finally, semantic quality and coherence of the text. The participants are asked to rank each of these metrics from 1 to 5, lowest to highest score.

6.3 Configurations

All experiments have been run using the Simpletransformers library [3], which leverages the more popular HuggingFace's Transformers repository [4] allowing for a fast setup and a fine-tuning of many pre-trained transformer architectures.

This research uses the T5 large model for the tests, comprising a total of 770 million parameters and 24 layers for both the encoder and decoder, along with a 16-head attention mechanism. During fine-tuning and evaluation, the max sequence length has been restricted to 64 tokens.

The baseline seq2seq model uses a BERT uncased pre-trained model with 110 million parameters, 12 layers and a 12-headed attention, for both the encoder and the decoder. The quantitative and qualitative results are better than those produced by the larger BERT model with 336 million parameters; we believe that this is likely due to the small size of the CBT dataset.

6.4 Results

Table 4 summarizes the results obtained for both the automatic and human evaluations, which also includes the results of the original dataset targets, for contrast.

The BLEU and METEOR scores are very low due to the large probability space when generating responses and the use of p-sampling to obtain more diverse and fluent results. This, coupled with the small size of the dataset in comparison with the pre-training corpus, affects the score by producing text which diverges substantially from the original labels.

The automated sentiment analysis by the fine-tuned RoBERTa model shows a higher positive sentiment for the baseline BERT model, but this doesn't reflect the subjective quality of the text produced by both models which is subjectively much better for the T5 architecture, as seen in the human evaluation results.

When the automatically generated responses are judged by participants in the survey, the BERT model falls behind significantly, specially in terms of relationship to the situation and overall quality. In fact, the scoring of the T5 in these two metrics is much closer to the original human written responses, showcasing BERT's inability to directly address the situation, often producing text with low

[3] https://github.com/ThilinaRajapakse/simpletransformers
[4] https://github.com/huggingface/transformers

semantic coherence.

7 Next Steps

One of the limitations in this research is the small size of the dataset, with just about a hundred different situations, so the obvious course of action would be to continue expanding on them by gathering new situations via crowd-sourcing. With a bigger corpus of data, along with other architectural improvements and transfer learning mechanisms, the results obtained in this study can be improved significantly.

Ultimately though, the aim of the research is to incorporate all of the key value pairs of the dataset - such as situations, emotions, negative thoughts, cognitive distortions and alternative responses - into a full fledged dialogue framework with the clear task of guiding patients through all of the steps of a CBT interaction.

8 Conclusion

This research focuses on the key CBT idea of transforming negative thoughts that contain cognitive distortions into more realistic alternatives, in order to provide automatic and therapeutic assistance to patients experiencing anxiety and depression.

Although there have been previous research within the NLP and CBT domains, especially in distortion and emotion classification (Rojas-Barahona et al., 2018), this study appears to be the first that manufactures a full CBT dataset, and attempts to apply modern machine learning architectures to automatically convert an initial negative thought into a more positive or realistic alternative.

Existing crowd-sourcing platforms represent a practical way for collecting human responses against distorted or negative thoughts and, in the future, they may also prove to be an effective source for gathering new situations.

The results obtained show how effective transfer learning can be when using state of the art transformers architectures to fine-tune a small CBT dataset. In particular, and specially when considering human evaluation, the Google T5 transformer model produces quality responses that are more realistic, while still being relevant to the situations and thoughts causing anxiety.

Future work will focus on expanding the existing CBT dataset, while trying to incorporate it into a more complete dialogue system framework.

References

Gerhard Andersson and Pim Cuijpers. 2009. Internet-based and other computerized psychological treatments for adult depression: A meta-analysis. *Cognitive behaviour therapy*, 38:196–205.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Azy Barak, Liat Hen, Meyran Boniel-Nissim, and Na'ama Shapira. 2008. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. *Journal of Technology in Human Services*, 26:109–160.

A.T. Beck. 1979. *Cognitive Therapy of Depression*. Guilford clinical psychology and psychotherapy series. Guilford Publications.

David D. Burns. 1981. *Feeling Good: The New Mood Therapy*. Penguin Books.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Daniel David, Ioana Cristea, and Stefan Hofmann. 2018. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in Psychiatry*, 9:4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health*, 4:e19.

GDBC. 2018. Global burden of disease study 2017.

Stefan Hofmann, Anu Asnaani, Imke Vonk, Alice Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36:427–440.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Yaniv Leviathan and Yossi Matias. 2018. Google duplex: An ai system for accomplishing real-world tasks over the phone.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gašić. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 44–54, Brussels, Belgium. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

J. Weizenbaum. 1966. Eliza - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9:36–45.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Y. Zhu, Ryan Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Appendix: Generated Responses by T5 during Evaluation

The following tables showcase the different situations presented in the human evaluation survey. For each situation, three alternative responses are generated automatically by the T5 model. The responses in bold were hand-picked to be included in the survey.

Situation	You are a psychiatrist at the University of Pennsylvania. You are attempting to revise your manuscript on depression after meeting with your editor in New York. Although your editor seemed extremely enthusiastic, you notice you are feeling nervous and inadequate due to your thoughts.
Negative Thought	"They made a terrible mistake when they chose my book! I won't be able to do a good job."
T5 Alternative Responses	**1. "I am trying to revise my book to make it better. You have to find the time to relax and not feel nervous. The better the rest of the week the better the results will be."** 2. "You have made a horrible mistake and you must find another way of expressing your opinion. You can not make all mistakes at the same time and you must be willing to improve your writing." 3. "No matter how much you understand about your client's personality, it is important that you get the answers to the questions before your editor."

Situation	You are about to give a lecture and you notice that your heart is pounding.
Negative Thought	"My mind will blank out. I'll make a fool of myself."
T5 Alternative Responses	**1. "You will do your best to not make a fool of yourself. I will have the confidence in myself to do what I need to."** 2. "No one wants to make a fool of themselves, just because the thought in their head is negative – they won't realize how many people out there really believe in them." 3. "You are not able to think through that and will let you make mistakes. So I will not be so focused to speak, if I recite the lesson I can forget."

Situation	Starting to have a panic attack on the subway.
Negative Thought	"What if I pass out on the subway?"
T5 Alternative Responses	**1. "You will not feel like you will pass out on the subway, and it is not likely to occur to you. You will never know what happens."** 2. "You can't run out of oxygen in the subway." 3. "There are tons of things that can go wrong when you're panicking on the subway, and you don't have to worry about the fact that you are going to pass out in public."

Situation	Friend walks past you in the street, without acknowledging you.
Negative Thought	"I'm a terrible person that isn't funny nor charismatic. I don't have any real friends."
T5 Alternative Responses	**1. "People should be interested in you and you aren't looking for friendship. Friendship is about being yourself e.g. making new friends is important to gaining new relationships and it is also important that friendship develops over time so you continue to grow, develop, and improve."** 2. "I should just tell someone to make friends for themselves." 3. "This is not my personal issue I have no friends and that does not change the fact of it. It is not a private matter of me deciding to make friends, or to be friends only, but to be friends with one another to share your life experiences and thoughts if they choose to."

Situation	My friend has left several messages for me about going to lunch.
Negative Thought	"I am so lazy and weak that I can't do a simple thing like calling my friend."
T5 Alternative Responses	**1. "I should call back again, it's because I am so lazy and weak that I can't do simple things. I will try something but I will make it easy for my friend to call and see if I can go."** 2. "I can contact him even when the friend can't reach me." 3. "It is not easy and I won't know if this will happen to her. Maybe the way we communicate, when it happened with a friend, isn't enough to fix the problem."