Synthesis of Open-ended Survey Responses Using Generative Language Models[1]

Markus Neumann and Burt L. Monroe
The Pennsylvania State University

We describe and evaluate methods for the synthesis of natural language responses to open-ended survey questions. As with numerical and categorical data, objectives might include imputation of missing responses, imputation of responses from counterfactual or small subgroup respondents, imputation of responses for time periods not surveyed, data augmentation, generation of synthetic data-sets that protect respondent privacy while maintaining sufficiently similar statistical and linguistic properties, or inference about the mechanisms of opinion formation and survey response. Our specific target data – our output language sequences – are responses to questions in the Penn State Mood of the Nation Poll. Drawing from existing theories of survey response, we model this process as transduction from an input – representing the question asked, individual attributes, and a media environment to which the respondent pays selective attention – to an output textual response. This makes the task similar, in NLP terms, to problems like machine translation or summarization. We investigate the utility of two specific encoder-decoder transduction models – the first a sequence to sequence model with an attention mechanism and the second a transformer – for this purpose.

1 Introduction

> Get ready, because every Democrat who wants to run for president is about to take that hard turn to appease what is now the radical, extreme, socialist Democratic Party base. Take, for example, possible contender Andrew Cuomo. Now, honest people can have a difference as to when life begins. He just celebrated a bill in New York state making late-term abortion legal. Now, according to, now, New York state, you can get an abortion in the seventh month of pregnancy, the eighth month of pregnancy, and the ninth month of pregnancy. That would be called *infanticide*.
>
> Sean Hannity, *Hannity*, Jan. 28, 2019

> What recently in American politics has made you angry?
>
> Respondent 1: "The coup attempt against President Trump. The democrat lurch toward socialism and *infanticide*."
>
> Respondent 2: "The move by liberals and far left politicians to murder babies thru abortion and *infanticide*."
>
> Respondent 3: "The fact that there are democratic leaders that pass laws that condone killing an infant after it is born, *infanticide*, and then celebrate on TV and act like this is something to celebrate."
>
> Mood of the Nation Poll, Feb. 5-7, 2019

Can modern NLP methods for language generation be used to impute or synthesize artificial natural language responses to open-ended survey questions? As with other types of data, there are many uses to which such a method might be applied. Possible objectives might include imputation of missing responses, imputation of responses from counterfactual or small subgroup respondents, imputation of responses for time periods not surveyed, data augmentation for machine learning, generation of synthetic data-sets that protect respondent privacy while maintaining sufficiently similar statistical and linguistic properties, or inference about the mechanisms of opinion formation and survey response. We presume a plausible method would need to mimic the process by which human respondents come to produce particular free text in response to open-ended questions. Consider the motivating example illustrated by the quotes above.

On January 22, 2019, New York's Reproductive Health Act, decriminalizing abortion after 24 weeks "when the fetus is not viable or a woman's health is at risk," was signed into law. Between then and Donald Trump's State of the Union address on February 5, Fox News devoted more than 6.5 hours to the topic of state abortion laws – 45 times as much as CNN and almost 100 times as much as MSNBC. (Kann and Savillo, 2019) This was frequently framed by the term “infanticide.” (Hagle et al., 2019)

Between February 5 and February 7, the tenth wave of the Mood of the Nation Poll asked Americans to answer open-ended questions about what in politics or recently in the news had made them angry. More than 8% discussed abortion in their answers, a number that had not exceeded 1% in the previous nine waves of the poll dating back to June 2016. Eight respondents, all Republicans, used a term that had never before appeared in the poll: “infanticide.”

This example seems consistent with the literature on opinion formation and survey response described by, e.g, Converse (1964), Zaller (1992), Zaller and Feldman (1992) and Lenz (2009). This literature is unified by the common finding that elites – be it the media or politicians – dominate the formation of attitudes and opinions of citizens. As opposed to the arguably more normatively desirable model, where citizens vote for politicians who support their attitudes and policy preferences, the true causation is reversed: voters adopt these preferences from their favored elites (Lenz, 2009; Broockman and Butler, 2017).

The theory of Zaller (1992) is of particular importance here, as it was developed with the goal of explaining how respondents answer survey questions (Zaller and Feldman, 1992). This “receive-accept-sample” model points towards two important, interacting sources of survey response. The first is elite communication, to which citizens are exposed through the media (broadly construed). The second is comprised of citizen attributes like partisanship and political awareness, which mediate what information is taken in and integrated into the set of considerations from which respondents can sample when forming their “opinion statement” for the survey. Moreover, the process of answering open-ended questions places additional cognitive load on a respondent who must articulate a novel answer in natural language, a process affected by the complexity or ambiguity of the question asked, and respondent attributes like education (Miller and Lambert, 2014).

This theoretical model has parallels in the structures of recently developed encoder-decoder models for natural language “transduction” problems, where the objective is to convert an input, often an input language sequence, into an output language sequence. Examples include machine translation or summarization. In this study, we examine the ability of two of these – the “sequence-to-sequence model with attention” and the “transformer” – to generate high quality synthetic survey responses as a transduction from respondent attributes and the political information environment.

Sequence-to-sequence models use a recurrent neural net architecture to map an input language sequence, in many applications a sentence, into an encoded fixed-length “context vector” or “thought vector” and then decode it into an output. This approach has been very successful in areas such as machine translation (Bahdanau, Cho and Bengio, 2015), chatbots (Vinyals and Le, 2015), style transfer (Zhang et al., 2018) and abstractive summarization (meaning that new text is generated, as opposed to the simple retrieval of tokens from the original corpus in extractive summarization methods such as topic models) (See, Liu and Manning, 2017). For example, in machine translation, the input might be a sentence in English, and the output a French translation. The use of the context vector gives the model flexibility, as the tokens in the input and outputs don't have to be equivalent – i.e. the two sentences can have different length and word order.

Another important ingredient in this model – connecting it more closely to Zaller's theory – is the "attention mechanism." Attention is used to decide which parts of the input are most relevant for the generation of each word in the output. This approach has led to great improvements in performance in traditional natural language processing tasks (Bahdanau, Cho and Bengio, 2015; Luong, Pham and Manning, 2015). An added benefit, which is of particular interest to social scientists, is that models with attention are more interpretable (Ertugrul et al., 2019). In this case, it gives the model the ability to adjudicate the relationship between attributes such as party identification and political awareness, determining which are most pertinent to the output. Our implementation is a modified version of the model presented in See, Liu and Manning (2017), who also give the attention mechanism the ability to copy from the input.

We implement a second model, the transformer (Vaswani et al., 2017), as a comparison. The transformer foregoes the computationally demanding recurrent structure of the sequence-to-sequence approach in favor of a parallelized "self-attention" mechanism, which gives it greater flexibility. This makes it both more computationally efficient to train and more finicky. Despite the lack of explicit sequential architecture, variants have achieved state of the art results on some machine translation tasks.

Our evaluation of the synthesized responses relies on both statistical and human-driven validation. Statistical measures are built based on the concept from the synthetic data literature of general utility – the ability of the artificial data to match the distributional characteristics of the original – and specific utility – the comparability of inferences in downstream tasks (Snoke et al., 2018). We note that Liu et al. (2016) have shown that metrics traditionally used in machine translation, such as ROUGE or BLEU, can be problematic in other applications of sequence-to-sequence models where there is no objectively 'correct' single output. Our human validation is based on the plausible humanness of the response – a Turing test – and the semantic validity of the responses – whether they express a coherent meaning that corresponds to respondent attributes and other inputs similarly to how the original human responses do.

The use of the attention mechanism also provides evidence suggesting a substantive extension to the theory of open-ended survey response. Attributes such as partisanship, education and news interest are generally considered most important in how citizens react to media exposure and respond to surveys. We show that these properties do in fact have the largest impact on the subject they choose to discuss, but finer details of *how* they talk about it, on the other hand, appear to be influenced to a greater degree by age, race, and gender.

2 Data

2.1. The Mood of the Nation Poll

Our data consists of 42,000 open-ended survey questions (with 41,808 text responses) from the first ten waves, June 2016 through February 2019, of the Penn State McCourtney Institute for Democracy Mood of the Nation Poll (hereafter "MOTN"). The poll is fielded by YouGov on its online panel, with sample frame matched and sample subsequently weighted to be representative of the adult US population.[2] Each MOTN wave, 1000 respondents have been given a set of four core prompts about what made them angry, proud, worried, and hopeful, as illustrated by Q1, Q3, Q5, and Q7 in Figure 1.[3] There are two variants of the proud and angry questions, one prompting a response to "American politics today" and the other "recently been in the news" presented randomly to 500 respondents each.

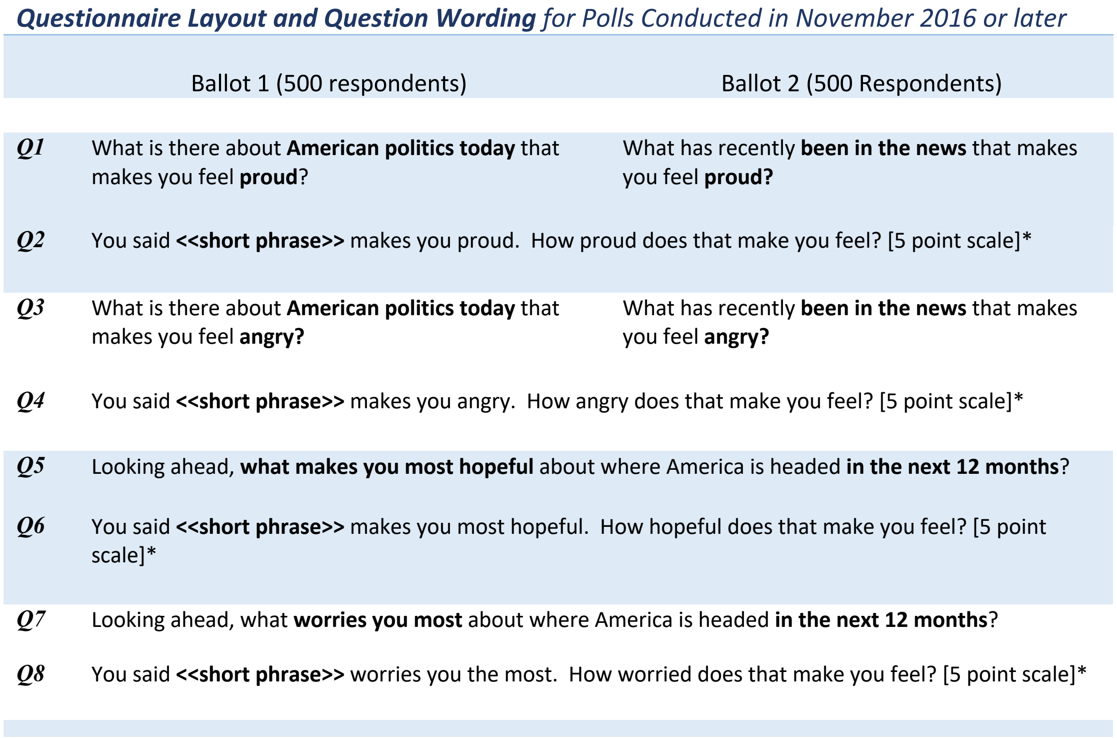
In current MOTN practice, there are three versions of the text produced: raw, pre-processed, and edited. The raw text is as provided from YouGov, edited only to remove technical errors from control characters or encoding. The pre-processed version is case-folded to lower case and punctuation removed, with some minor exceptions, and is otherwise mostly unedited.[4] The edited version applies a bespoke collection of, at this writing, over 5,000 substitution rules to correct spelling (e.g., 71 variants of "president") and grammar where the respondent intent is clear, to identify common multiword entities (e.g., "whitehouse" and "white house" to "white_house"), and to standardize references to common or potentially conflated named entities (e.g., 118 variants of "donald_trump", or "hurricane_harvey" vs. "harvey_weinstein"). This reduces the vocabulary and sparsity and is the version we use here.

We note that these are not all "one sentence" as might typically be the target of a

[2] Further information about the sample and weighting are available in Plutzer (2019).

[3] Waves 1 and 2, in June and September 2016, also contained a prompt about what made the respondent "ashamed," which was subsequently dropped. This accounts for the additional 2000 observations.

[4] Some punctuation is kept (e.g., "!"), recognizable contractions and possessives are collapsed (e.g., "dont"), and a small number of particular pattern substitutions are made (e.g., "U.S." becomes "u_s", some expletives and slurs are obscured).



Questionnaire Layout and Question Wording *for Polls Conducted in November 2016 or later*

	Ballot 1 (500 respondents)	Ballot 2 (500 Respondents)
Q1	What is there about **American politics today** that makes you feel **proud**?	What has recently **been in the news** that makes you feel **proud?**
Q2	You said **<<short phrase>>** makes you proud. How proud does that make you feel? [5 point scale]*	
Q3	What is there about **American politics today** that makes you feel **angry?**	What has recently **been in the news** that makes you feel **angry?**
Q4	You said **<<short phrase>>** makes you angry. How angry does that make you feel? [5 point scale]*	
Q5	Looking ahead, **what makes you most hopeful** about where America is headed **in the next 12 months**?	
Q6	You said **<<short phrase>>** makes you most hopeful. How hopeful does that make you feel? [5 point scale]*	
Q7	Looking ahead, what **worries you most** about where America is headed **in the next 12 months**?	
Q8	You said **<<short phrase>>** worries you the most. How worried does that make you feel? [5 point scale]*	

Figure 1: *Mood of the Nation Poll questionnaire.*

generative language model and do not have a consistent grammar or syntax. Over 11,000 answers consist of a single word or entity, including 3,721 of them the word "nothing" and 1,359 of them "donald_trump." Conversely, some write many sentences; one Wave 11 respondent offered over 650 words for the "angry" question. About a third of the answers are duplicates, and this is of course more likely with shorter answers.

2.2. Television News Corpus

To capture the media content that voters are exposed to, we use transcripts from popular shows on Fox and MSNBC. In an increasingly polarized electorate, media exposure has become equally fractured (Iyengar and Hahn, 2009; Prior, 2013). Conservatives overwhelmingly get their news from Fox, while liberals tend to choose from a wider range of outlets (Mitchell et al., 2014). Consequently, it is important for our corpus to draw from both sides. For conservative media, Fox News is the only viable choice. For liberal media, we selected MSNBC due to its particularly low proportion of conservative viewers (Mitchell et al., 2014). For MSNBC, we scraped transcripts for the following shows from the channel's website (`www.msnbc.com`): Meet the Press Daily, PoliticsNation with Al Sharpton, The 11th Hour with Brian Williams, Hardball with Chris Matthews, All In with Chris Hayes, The Rachel Maddow Show, and For the Record with Greta. Similarly, we scraped transcripts for Hannity, The Ingraham Angle, Tucker Carlson Tonight, Special

Report, and The Five from www.foxnews.com. Figure 2 shows the periods for which transcripts are available for each of these shows, as well as the dates of the Mood of the Nation Poll waves.

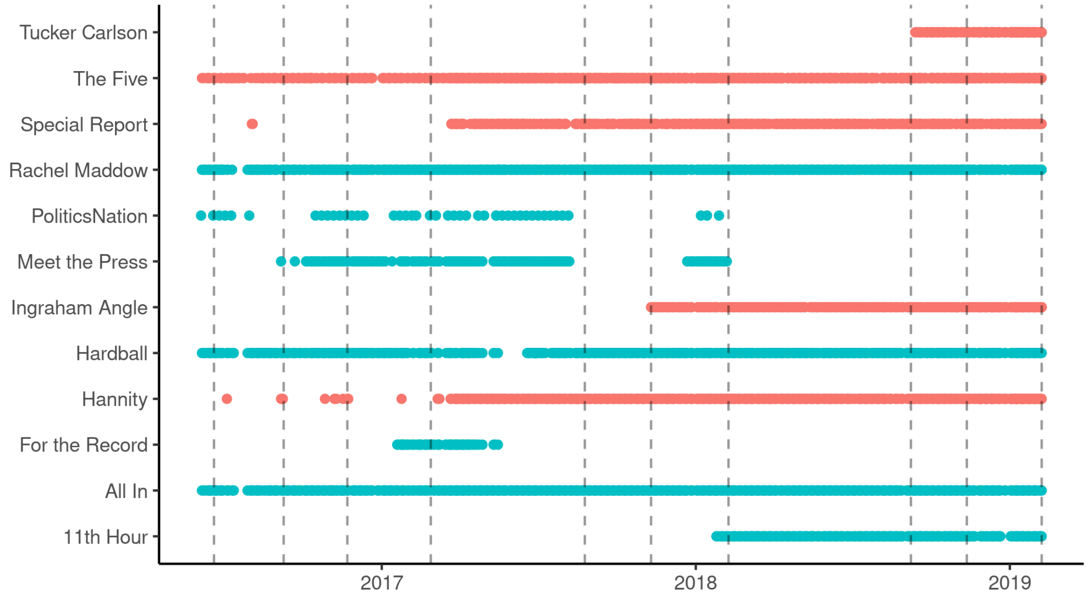


Figure 2: *Availability of TV transcripts by show and date. Red dots are from shows on Fox News, blue dots from MSNBC. The vertical dashed grey lines indicate the dates of the Mood of the Nation Poll waves.*

3 Modeling Attributes, News and Response

3.1. Theoretical Model

We assume, consistent with (Zaller, 1992), that respondents simultaneously draw from both the current media environment as well as their preexisting attitudes in their formulation of a survey answer. We make no assumptions about the exact balance in this relationship: Whereas Zaller emphasizes the importance of elite communication, and attitudes only serve the role of filtering what gets taken in, the later literature on motivated reasoning and selective media exposure (Iyengar and Hahn, 2009; Jerit and Barabas, 2012) places greater weight on the personal characteristics of voters. Rather than taking a side in this debate, we simply let the model decide whether to place greater attention (see section 4.1 for a discussion of the attention mechanism) on the attributes or media environment.

Figure 3 shows a schematic overview of this relationship, along with an example. A white male respondent, age 30-44, with Democratic party identification, high news

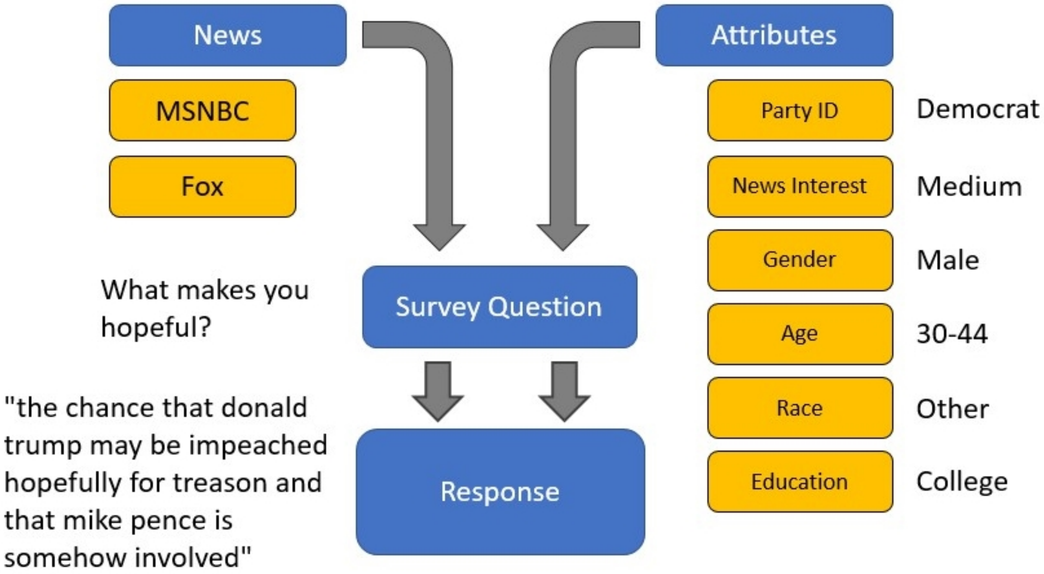


Figure 3: *Theoretical model of survey responses. Respondent attributes filter and interact with news the participant is exposed to, influencing which concerns reside at the top of their head at the time of the survey and how to formulate their answer.*

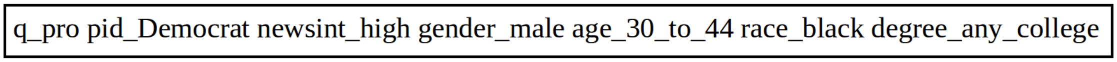
interest and at least some college education is selected for the seventh wave of the Mood Poll, conducted on February 8, 2018. At this point in time, the Russia investigation is in full swing. In the prior months, Michael Flynn and George Papadopoulos had pleaded guilty to lying to federal investigators, and Rick Gates as well as Paul Manafort had turned themselves in to the FBI. Just days before the Mood Poll, the Nunes memo, a previously classified document alleging political motivations behind the start of the Russia investigation, is released to the public. Liberal media outlets cry foul, while Fox News see their suspicions confirmed. In this frenzied and hyper-partisan media environment, the poll respondent is asked what makes him most hopeful about where America is headed in the next 12 months. Due to the media coverage, the Russia investigation is at the top of his head – and he responds in a way consistent with his Democratic party identification and high news interest: "The chance that Donald Trump may be impeached hopefully for treason and that Mike Pence is somehow involved."

Thus, we imagine the poll serves as a kind of transducer. It takes an input, the characteristics of a poll respondent combined with the media environment, and transforms it into an output by asking a question and receiving an answer. This theory then, is analogous to the type of machine learning method we employ – the encoder-decoder model.

3.2. Data Representation

3.2.1 Attributes

In the input vector, the attribute variables are represented as unique tokens (see Figure 4 for an example). For example, a Republican party identification would simply be represented as *pid_Republican* in this string. Consequently, the model considers the variable as just another token type in its vocabulary and it has no intrinsic way of knowing that *pid_Republican* and *pid_Democrat* are two realizations of the same concept. However, since all realizations always appear in the same place in the input string (for example, party identification is always the second token in our input string) and the realizations are mutually exclusive with each other, the model is nevertheless able to learn their significance with respect to each other as well as the output string. See Table 12 in the Appendix for an overview of the six attributes and the distribution of their sub-categories.



q_pro pid_Democrat newsint_high gender_male age_30_to_44 race_black degree_any_college

Figure 4: *Example representation of question and respondent attributes.*

3.2.2 News

The second component of our input string pertains to current events, as reported on by the news media (see Figure 5 for an example). Due to limits in terms of both hardware and model complexity, it is not feasible to dump entire news transcripts into a string and expect the model to understand them. Even if this were possible, it would also be unrealistic – after all, news consumers do not memorize entire broadcasts and recall them perfectly at the time of the survey. Instead, they filter the information stream using their existing political attitudes and keep only what is relevant and consistent with their prior beliefs. Consequently, we represent only a very condensed version of the news environment.[5]

We rely on the "fightin' words" method (Monroe, Colaresi and Quinn, 2008) to filter through the broadcasts. We divide the news leading up to a poll into three 5-day windows (to account for the possibility that some respondents might not have been paying attention directly before the poll, and respond to what they saw 1-2 weeks before instead), and then measure the top 10 fightin words in three ways: (1) those words distinctive for this period

[5] Another upside of this data representation is that even though our news corpus only consists of a number of shows on MSNBC and Fox News, the summarized version doesn't contain as many idiosyncrasies particular to these broadcasters and therefore does not rely on the assumption that survey respondents have watched these particular shows.

versus all others, (2) those words distinctive for Fox News in this period, and (3) those words distinctive for MSNBC in this period. The first measure, comparing this period to all other periods, extracts the information that is most pertinent to the current time. For example, in the five days leading up to the first Mood wave, the Orlando nightclub shooting had just happened. Consequently, the top three words are “orlando”, “ahmed”, and “nightclub” (see Table 13). We then also record the top 10 fightin words in this period on either MSNBC or Fox, compared to both channels during all other points in time. This measure then signals what is particular about this channel at that time. For example, in the same period after the Orlando nightclub shooting, two of the top fightin words on Fox are “islamism” and “infidels”, signifying the channel’s right-wing take on a Muslim mass shooter (see Table 14 for MSNBC and Table 15 for Fox).

By including these three different representations of the news, we hope to capture selective media exposure of our survey respondents (Iyengar and Hahn, 2009; Messing and Westwood, 2012; Bakshy, Messing and Adamic, 2015). The model, through its attention mechanism, can choose which part of the news representation should be paid most attention to in relation to a survey response. We note, however, that this is an extremely difficult learning task, given the Mood Poll’s episodic design. There are only ten observed news environments, each of which is mapped to 4000 responses. As we will see, we find little evidence that this representation offers more for the model to learn from than would a simple “wave_i” labeling scheme for these discrete environments.

woodward kavanaugh op nike anonymous ed brett kaepernick ohr pressley __ patten doe spanberger sam miers omarosa gosling gianno rover abigail __ mccain gillum ohr desantis mcgahn bruce lanny steyer woods gwen <> woodward kavanaugh op brett ed anonymous pressley roe ayanna settled __ patten doe spanberger sam omarosa miers abigail harriet abortion amiri __ mccain gillum mcgahn desantis trial superdelegates mcsally gwen mogilevich ducey <> nike woodward kaepernick ohr anonymous spartacus colin chicago weissmann kavanaugh __ gianno gosling troy rover usama armstrong jan velcro crab roommate __ ohr bruce lanny desantis woods steyer davis tiger gillum antifa

Figure 5: *Example representation of news content. <> is used to separate the different fightin’ words contrasts (time, MSNBC and Fox), __ is used to separate the three five-day windows within these.*

3.3. Tasks

In our first task we envision a setting in which the researcher wishes to impute missing data within an observed wave, generate a synthetic dataset with similar properties, or draw substantive inferences about the data generation process. To imitate the survey response process, the model is given three sets of information in the input: 1) The question the poll respondent was asked, 2) the six respondent attributes (see section 3.2.1) and 3) an abbreviated representation of the news corpus (see section 3.2.2). Consequently, we train on all waves and predict to a held-out test set which also features cases from all waves. Here, our dataset of 41,808 samples is split up as following: 38,000 are randomly selected as training samples, 2,000 as validation samples, and the remaining 1,808 as test samples which we try to predict.

In our second task, we envision a setting in which the researcher wishes to predict responses for a time when no survey was undertaken, for an information environment in which there are no training examples. Here, we train on the first nine waves, hold out the tenth, and predict to it. The format of input and output are identical to the first task. This is a much harder problem, as the model is required to draw information from the news content in the input that goes beyond identifying the wave. Instead, it actually needs to learn what is pertinent at the time. Returning to our opening example, an important issue in our held-out tenth wave – both in the responses and on Fox News – is late-term abortion, prompted by state legislative actions in New York and Virginia. If the model is capable of learning from our news representation, it should produce a higher proportion of abortion-related synthetic responses and indeed occasionally speak of "infanticide." The dataset consists of 36,000 randomly selected responses for training, 1,827 for validation and the 3,981 responses to wave 10 for testing.

Finally we briefly discuss a third task, using responses to negative questions in order to predict what the same survey participant would have answered to a positive question, and vice versa. This leverages the known common authorship across multiple respondent answers and uses a more recognizable language sequence as its input. This task resembles the style transfer literature in machine learning (see Li et al. (2018); Zhang et al. (2018); Logeswaran, Lee and Bengio (2018); Lample et al. (2019)). This can partly be thought of as exploring the potential for asking half as many questions in a survey, and partly as validation of the sequence-to-sequence model's ability to learn the sequential language structure of Mood survey responses. Negative questions are what makes respondents worried and angry, positive questions are what makes respondents hopeful or proud. Responses to the "ashamed" question in the first two Mood waves are ignored for this purpose. Worried and hopeful, as well as angry and proud, serve as the negative-positive pairs. The dataset consists of 18,000 randomly selected negative-positive pairs for training, 1,000 for validation and 976 for testing.

3.4. Evaluation

In addition to traditional machine metrics used in training the model – see the appendix, Figures 11a-11c, and 12a-12c, for accuracy, cross entropy and perplexity – we construct additional measures and tests to validate the performance of our models for the task objectives.

3.4.1 Novelty

One potential problem in training such a model, especially with relatively little training data, is overfitting. One consequence is the generation of text samples that already exist in the training data. To some extent, this is to be expected naturally. For example, a large portion of responses praise or criticise Donald Trump. There are only so many ways to do that in a handful of words, and many simply type "Donald Trump," "Trump," or "the president." But even so, responses that are effectively "copied" from the training data are less interesting, have potentially less desirable privacy-protecting properties, and definitionally do not "interpolate" answers through any mixing. Consequently we measure the proportion of generated responses that already exist in the original dataset.

Now, one way in which the model could exploit this metric is to generate one original response – and then repeat it across all test samples. To head off this potential problem, we also measure the percentage of duplicates in the outputted samples. Since these two issues – overfitting and underfitting – are negatively correlated and have opposite causes – the former stemming from training too much, the latter from training too little – we also measure the proportion of outputs that are neither in the training data, nor duplicated in the generated outputs.

3.4.2 General utility (comparability of distributions)

XX - Cosine similarity, in the test set between actual and synthesized outputs, of logged (count + 1) of n-grams in partitions of the data on observables. (This is more fine-grained than ROUGE or BLEU in not simply counting the presence or absence of terms, and more appropriate for the context in which there are many correct answers). This partially captures the model's ability to learn bag-of-words and bag-of-n-grams level semantics.

3.4.3 Specific utility (comparability in downstream tasks)

To test how realistically the model outputs emulate the original data semantically, we compare their performance in what might be a potential downstream task. We use the fightin' words method on both the synthesized outputs, as well as the held-out test sets, with respondent partisanship as the contrast. Then, we measure the correlation in z-scores between the words that occur in both. The higher the correlation, the greater the model's ability to imitate the original survey data. (XX - could do on more.)

XX - Our second test measures confidence interval similarity in logistic regressions predicting the use of substantively relevant clusters of terms or entities. Our primary example models the mention of Trump-related words after his election.

3.4.4 Content Validity

XX - Do they make grammatical sense (as much as the originals do)? Under what circumstances don't they?

XX - Do they make sense (as much as the originals do)? Under what circumstances don't they?

XX - Do they make substantive sense given attributes, question, and context? When don't they?

3.4.5 Turing test

XX - MTurk experiment (not yet done). Can people tell which of a set of answers is model-generated? Weak test – Artificial answer will be picked no more often. Strong test – Artificial answer will be picked less often.

XX - Put existing sample test in?

4 Models

The fundamental challenge of text transduction is to transform an input into an output, where there is not necessarily a 1:1 equivalent for each word of the input phrase in the corresponding output phrase. In the canonical case of machine translation, it is generally insufficient to translate each word individually. Grammatical rules dictate that words be in different positions. Furthermore, the number of words required to express a specific concept differs by language. For example, the German translation of the "Association for Subordinate Officials of the Main Maintenance Building of the Danube Steamboat Shipping Electrical Services" is just one word: "Donaudampfschiffahrtselektrizitätenhauptbetriebswerkbauunterbeamtengesellschaft." In summarization, where the express purpose is to create a shorter version of a longer text, the need for a model with inputs and outputs of different length is even more apparent.

4.1. Sequence-to-Sequence Model with Attention

Neural sequence-to-sequence models (which are part of the encoder-decoder family of models) approach this problem by relying on an intermediary, the so-called *context vector* (Sutskever, Vinyals and Le, 2014). The purpose of the context vector is to carry the meaning of the variable-length input, but in a fixed number of dimensions. The encoder (some form of recurrent neural network, in our case, a Long- Short-term Memory, or LSTM, network) is used to encode the input sentence, one token at a time, into the encoder hidden state. The context vector is a weighted sum of the encoder hidden state – the 'gist' of it, so to speak. Then the decoder (another LSTM) makes use of this context vector to generate an output, one token at a time. At each generation step, the decoder only has access to the previous token (or the start token, for the beginning of the phrase), which is used to update its own hidden state. This decoder hidden state, together with the context vector, is used to generate a distribution over the vocabulary, i.e. the words that the model knows. At each step, the word with the highest probability is outputted. When the model generates the end-of-sentence token, it stops.

This basic form of a sequence-to-sequence model works well enough on its own, but a number of improvements have been made since its inception. For one, too much responsibility rests on the shoulders of the fixed-length context vector. Especially for longer inputs, which are common in summarization, this is a problem. Encoding the meaning of an entire news article or scientific paper into a vector of, for example, 500 dimensions, is difficult. To this end, the *attention mechanism* (Bahdanau, Cho and Bengio, 2015) provides some relief. The intuition of this approach is that for expressing the meaning of the input sentence, some words are more important than others, and that this relationship differs for each step of the output generation. For example, in the case of machine translation, the model would be paying greater attention to the direct equivalent of the word it is currently translating.

In addition to making the model better, this technique has another major benefit: it makes neural models more easily interpretable (Ertugrul et al., 2019). In our use case, the attention mechanism points to whichever respondent attribute or piece of news is most relevant to generating an output. For example, the model learns that respondents with low news interest are more likely to respond with something along the lines of "I don't know", or "I don't watch the news". Consequently, when generating this kind of output, the matrix of attention weights shows that the model focuses on the news interest attribute in the input.

Since our approach shares some similarity with summarization, we also use two mechanisms introduced in this field: copy attention and coverage. *Copy attention* (See, Liu and Manning, 2017) is an additional layer which generates a probability of generating directly from the attention distribution rather than the vocabulary. The advantage of this approach is that the model may never have seen specific words – names, for example – in the output and therefore does not know how to deal with them. Consequently they never end up getting outputted. However, if it is aware that a particular word is clearly *relevant* in the current time step, the copy attention mechanism allows it to pull this word straight

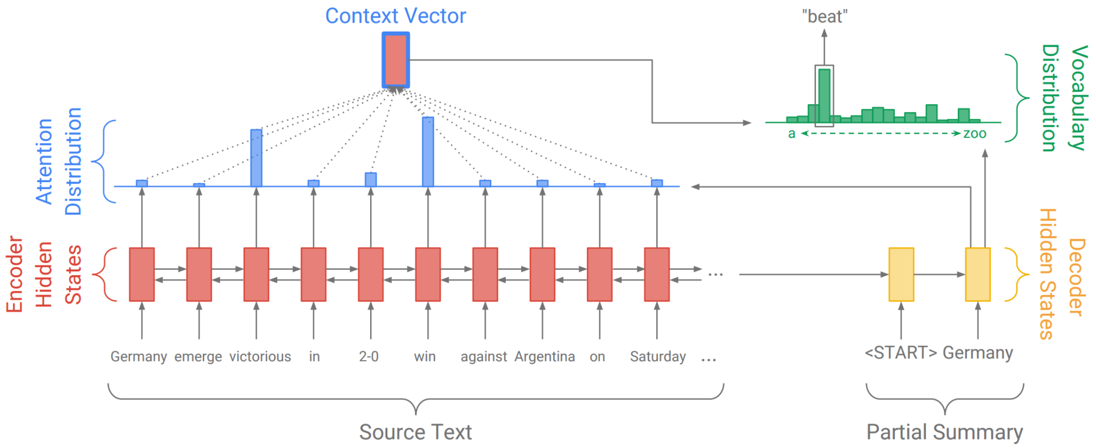


Figure 6: *Illustration of sequence-to-sequence model with attention, as extended with copy attention and coverage for the task of abstractive summarization, by See, Liu and Manning (2017). XX - placeholder - should replace with our own. Source: See, Liu and Manning (2017)*

from the input. In summarization, this allows the model to discuss names, places, etc. it has never seen before. In our case, this would allow copying from the representation of the news corpus. For example it may never have seen a mood respondent discussing the mass shooting in Las Vegas, but it did observe many discussing shootings in Orlando and elsewhere. The idea is that the copy mechanism would then copy “las_vegas” from the input vector.

The *coverage mechanism* (See, Liu and Manning, 2017) prevents a typical failure mode of encoder-decoder models, wherein they get stuck repeating the same word over and over (and over). Coverage downweights words on the basis of previous attention vectors, thus making it less likely for the same word to appear multiple times.

Another important mechanism is *beam search*. This method is not specific to encoder-decoder models, and is common practice in language models more generally. By default, language models operate in a “greedy” manner. They generate one word at a time, and therefore follow a kind of “path dependency”. If the first word generated by the model is “Hillary”, then the next word is very likely to be “Clinton”, and very unlikely to be, for example, “Trump”. Thus, the first word decides the subject of the generated response. However, by the end of the sentence, it may turn out that the full phrase is actually not the optimal fit for the given input. For example, it is possible that the particular combination of attributes in the respondent makes it very likely for them to begin a sentence by talking about Hillary Clinton – because they really hate Hillary Clinton. However, they might also be predisposed to finishing their sentence with “will make America great again.” In the greedy search, the resulting sentence would be “Hillary Clinton will make America great again.”

Beam search allows the model to generate multiple different sentences. At the end, it compares these "hypotheses" and decides on the one that is the best fit. In the above example, with a beam size of two (i.e. two hypotheses) it would compare "Hillary Clinton will make America great again" to "Donald Trump will make America great again" and conclude that the latter is much more plausible. While beam search is frequently used in computer science, it is even more important in the social science use case. From a pure machine learning point of view, it may only matter whether an output sentence makes sense at all. For a political scientist, content is critically important. This is true even more so for the case of generating survey responses – after all, what a respondent will choose to talk about is a crucial component of the model of Zaller (1992).

4.1.1 Model Specification

We use the pytorch-based library `OpenNMT-py` to implement our models. The hyperparameters are selected as follows. The size of the model's hidden state as well as the dimensions of the word embeddings is 500. Both the encoder and the decoder are LSTMs with two layers. This results in a total of 26,826,867 parameters. Gradients are optimized using stochastic gradient descent. We begin with a learning rate of 1.0, then begin to decay it by a factor of 0.5 starting at 50,000 and every 10,000 steps thereafter. To combat overfitting, we use a dropout rate of 0.2. Parameters are initialized from a uniform distribution at 0.1. The model is trained for a total of 60,000 steps. For text generation, we use beam search with a beam size of 10. Output sentences are forced to consist of at least 5 tokens. We use stepwise, length (parameter value is 0.9) and coverage (parameter value is 5) penalties. Input and output vectors are truncated to 170 and 50 tokens, respectively. Vocabulary size is 11,366, which is quite small compared to most machine translation and summarization models.

4.1.2 Semantic validity

First, do the generated results appear on the surface to make sense? To be plausible? Table 1 shows sample outputs of the sequence-to-sequence model, along with the survey prompt and respondent attributes. The first example comes from a black male Democrat in the 30-44 age range, with at least some college education, and high news interest. The real respondent's response to what makes him proud was the "racial injustice discussion". The model generates a response that shares some semantic content (but no individual words) expressing pride about a "black swimmer winning medals in olympics". Evidently, it has learned that this is something a respondent with these characteristics might express in that particular information environment. This prediction expresses the model's best guess over what this respondent might say. The table also shows the model's second and third best guesses – "i feel proud about black panther" and "i feel proud that there is a lot to discuss". The second is similar to the first, again sharing some semantic content with the actual response, while the third appears to be a more vague expression of open-mindedness. Our model, as implemented, currently outputs its top five predictions – but this value can be varied with the beam size in the beam search.

	Phrase
Attributes	**q_pro pid_Democrat newsint_high gender_male age_30_to_44 race_black degree_any_college**
Human	racial injustice discussion
Prediction 1	black swimmer winning medals in olympics
Prediction 2	i feel proud about black panther
Prediction 3	i feel proud that there is a lot to discuss
Attributes	**q_pro pid_Other newsint_low gender_female age_30_to_44 race_hispanic degree_no_college**
Human	nothing
Prediction 1	nothing about politics makes me proud
Prediction 2	nothing about politics makes me feel proud
Prediction 3	how discrimination can be made
Attributes	**q_ang pid_Other newsint_low gender_female age_45_to_64 race_white degree_no_college**
Human	people hurting and killing each_other in senseless wars
Prediction 1	all the fighting in the world
Prediction 2	the bitterness politicians have toward others and disputes
Prediction 3	all the fighting between politicians it leaves american venerable
Attributes	**q_pro pid_Republican newsint_low gender_female age_18_to_29 race_white degree_any_college**
Human	no
Prediction 1	havent really paid attention lately
Prediction 2	i havent really watched the news recently
Prediction 3	havent really paid attention to the news
Attributes	**q_wor pid_Republican newsint_high gender_male age_65__older race_white degree_no_college**
Human	paid rioters
Prediction 1	the outlaw sanctuary_cities and states
Prediction 2	the demonstrations will become deadly
Prediction 3	that the democrats will keep obstructing and nothing will get done
Attributes	**q_ash pid_Democrat newsint_high gender_female age_65__older race_other degree_any_college**
Human	the stanford rape case the xenophobia of this country and its unwillingness to help [...]
Prediction 1	donald_trump as a serious candidate for president
Prediction 2	donald_trump and his surrogates every single day
Prediction 3	donald_trump and the republican party
Attributes	**q_hop pid_Democrat newsint_high gender_male age_45_to_64 race_hispanic degree_any_college**
Human	america is headed for very dark times if donald_trump isnt reigned in by responsible americans
Prediction 1	the next election in november 2019
Prediction 2	that the president will be impeached
Prediction 3	the next election to be president

Table 1: *Sample outputs from the sequence-to-sequence model. The 'Attributes' line describes the attributes of the respondent for whom responses are synthesized. The 'Human' line shows their actual answer. The 'Prediction 1-3' lines show the top three predictions of the model.*

The other outputs shown in Table 1 show similarly context-appropriate responses. The second and third, both with low news interest and neither Democratic nor Republican party identification, express disillusionment with politics. For the fourth, again a respondent with low news interest, the model directly expresses this attribute: “havent really paid attention lately”. The fifth, a more engaged Republican, complains about sanctuary cities, demonstrations and Democrats – consistent with what might be expected given his attributes, and also similar to the real response: “paid rioters”. Responses six and seven express sentiments that are equally consistent with their Democratic party identification.

This examination of some of the model’s outputs suggests the model answers the questions appropriately, broadly speaking. The lowest bar for the model to clear is the ability to give a response that actually answers each question. Table 1 shows that it is capable in at least these cases of doing so. Not only does it give negative answers to negative questions and positive answers to positive questions, it also distinguishes between them. The predictions for respondent 1 show that when asked what makes them proud, the model begins predictions 2 and 3 with “i feel proud” – consistent with how a human might respond to this prompt. It doesn’t do so when asked what makes the respondent hopeful.

Does the model respond appropriately to changing conditions? Table 2 shows model predictions for a respondent with the same characteristics across all 10 survey waves – a synthetic time series, of a sort. These outputs demonstrate that the model keeps up with the times – in the first wave, the simulated respondent decries Trump’s response to the Orlando shooting, in the third wave, following the presidential election, complains about his victory, and so on. The only outputs which do not seem to fit are the seventh wave – here, the predicted response expresses anger over the fact that “republicans gained seats in the senate”, nine months before the midterms actually happened. The best-ranked predictions also are not always consistent with each other – for example, in the fourth wave, the best prediction is critical of Trump, while the second and third-best answers appear to be defending him.

We can similarly generate synthetic responses for any combination of attributes. Changing one attribute at a time allows us to observe the “marginal” effects on text out. We see, for example, that voters of the top profile who don’t self-identify as either Democrat or Republican nevertheless appear to speak like a partisan, but which type of partisan depends on gender. Conversely, the bottom profile demonstrates a weakness of this approach where there are insufficient data. This raises a caution about bias and error that is likely for synthetic responses based on small populations or subgroups, a caution we return to in our discussion.

Poll Date	Prediction
2016-06-20	donald_trump response to the shooting in orlando
2016-06-20	donald_trump and the tea party
2016-06-20	donald_trump and the republican party
2016-09-09	donald_trump and his racist views
2016-09-09	everything about donald_trump and his supporters
2016-09-09	donald_trump and his racist followers
2016-11-22	the election of donald_trump as president
2016-11-22	the election of donald_trump as our 45th president
2016-11-22	donald_trump being the arrogant prick he is
2017-02-27	donald_trump and his idiot supporters
2017-02-27	the way people are treating our president
2017-02-27	the way people are treating president donald_trump
2017-08-25	the iron rule of the wage form
2017-08-25	the constant revisiting of issues that were dealt with years ago
2017-08-25	immigration policy and its application form
2017-11-10	donald_trump and his idiot supporters
2017-11-10	donald_trump and the republicans who normalize his disgusting behavior
2017-11-10	the divisiveness and obstructionist republicans
2018-02-08	republicans maintain majority in senate
2018-02-08	republicans gained seats in the senate
2018-02-08	donald_trump lying and acting childish
2018-09-08	corporations arent people money isnt speech
2018-09-08	anything and everything in national politics
2018-09-08	donald_trump and his idiot supporters
2018-11-12	donald_trump and the republicans who normalize his disgusting behavior
2018-11-12	the republican party and the moron in the white_house
2018-11-12	the republican party and their worthless president
2019-02-07	donald_trump forced jeff_sessions to resign
2019-02-07	donald_trump and the republicans who normalize his disgusting behavior
2019-02-07	donald_trump and the republicans in general

Table 2: *Predictions for the same person – a white male Democrat, age 30-44, with high news interest and a college education – answering the same question – what makes him angry – across all ten waves of the survey. Top three predictions are shown for each wave.*

	White, 45-64, college, news interest high, 1st prediction	
	Male	**Female**
Dem	(233) the hypocrisy of the republican party	(262) donald_trump and his idiot supporters
Other	(90) the divisiveness and obstructionist democrats	(60) donald_trump and his vile words
Rep	(304) the left and the obstructionist democrats	(168) people who twist the truth

	Black, 18-29, college, news interest medium, 3rd prediction	
	Male	**Female**
Dem	(2) a dishonest guy killed in a 10 year old	(20) my concern is a lot
Other	(0) killing of babies and babies	(1) cant feel angry at all
Rep	(0) the government corruption and over regulation	(4) i cant think of a single thing

Table 3: *Effect of changing assumed party and attributes in two hypothetical respondents answering question about anger in Wave 10 (Feb. 2019). Numbers in parentheses indicate the number of responses observed with those attributes across all waves. In the second case, synthetic answers reflect the model's poor understanding of this region of attribute space.*

4.1.3 Actual and de facto nonresponse

XX - Imputing nonresponses Table 4

qid	Prediction
angry	the government corruption and over regulation
proud	the government shutdown over for now
proud	nfls man of the year chris long
proud	continuing building a mexican border wall
proud	the end of the government shutdown
proud	donald_trump and republicans in congress
proud	the patriotism of our president
proud	not a thing in the world
hopeful	the wall and job growth
hopeful	a crowded democratic primary field
hopeful	that donald_trump will still be president

Table 4: *Imputed answer for those in Wave 10 who did not answer all four mood questions.*

Self-entered open-ended questions, as these are, give the respondent another route to item nonresponse beyond skipping a question. She can simply type the minimum number of characters, or speak a syllable into her phone, and move on having received credit for her answer. Many people literally type "na." XX - Imputing "na" Table 5

Human	Prediction
na	that the democrats control the house of representatives
na	no money and homelessness raising
na	that the election process works as intended
na	nothing i can think of
na	not hopeful about much hope
na	that they are nothing for the people of the country
na	another shutdown another war and the war on drugs
na	continuing divisiveness appointment of conservative judges supreme_court decisions
na	i havent watched the news
na	that the democrats wont work with the republicans
na	the direction the country is headed politically and socially
na	nothing nothing makes me hopeful
na	i really really dont_know so sad
na	nothing that i can see
na	dont care for the seniors of america
na	that they are trying to work for the people

Table 5: *Imputed answer for those in test set who responded with a literal "na".*

4.1.4 Attention and Inference

In addition to the generation of synthetic responses, the model can also be used for inference. To this end, we turn to the attention mechanism. As noted above, for each step in the output sequence, the model relies on attention to determine which parts of the inputs to focus on. Since our input consists of variables indicating respondents' attributes as well as a summarized version of the news corpus, the attention mechanism allows us to draw conclusions about the relative importance of these factors.

Figure 7 illustrates this concept. For each token in the outputted sequence "i dont watch the news", the plot shows the attention weights for the question and respondent attributes (the attention weights for the media content are omitted here). The model begins with the special <start> token (not shown here) and then calculates the probability of the next word. In deciding on "i", it focuses primarily on the question, as signified by the dark red field. This makes sense – all of the sequence's tokens are conditioned on the prior token, so the first token has the most influence over the entire sequence. Consequently it plays a large role in determining its content. For this purpose, the question type – requiring an appropriate response – *should* receive the highest attention weight. For the second token, "dont" the largest weights are on the question and party. Since the question is positive and the respondent's party ID is Democratic, a negative response, such as "dont" is more likely. For the third token, "watch", most of the attention weight is on news interest. Again, this is very intuitive as a respondent's interest in the news directly relates to whether they watch it. For "the", there is heavy emphasis on the respondent's education, as educated respondents are more likely to write longer and more grammatical responses, which correlates with the use of more function words like "the". For the final token, the weights are more spread out, with the question being the most relevant again.

By averaging across the columns of the attention matrices of all outputted responses, we get a view of the relative importance of each attribute in its direct influence on each output word. Figure 8 shows this cumulative attention distribution. For the purpose of simplicity, all attention weights on media content are aggregated.[6]

Somewhat counterintuitively, the variables with the most attention are age, gender and race. This conflicts with the theory, which suggests that party ID and news interest should be more important, as well as logic, which demands that the question be more important. Why does this happen?

To answer this question, Figure 9 shows the same cumulative attention distribution, but only for the *first word* of each output. Here, the weights are exactly as expected. The question is most important, followed by party ID, education and news interest. In light of this evidence, we surmise that party identification and news interest, generally considered to be the most important determinants of survey responses, play a somewhat different role. Due to the path dependency that is inherent to a language modeling approach, the first word is strongly influential for the sentence's topic and the words that follow. Hence,

[6] The attention weights on separator tokens such as <> and __ which we use to delineate different parts of the input, are omitted.

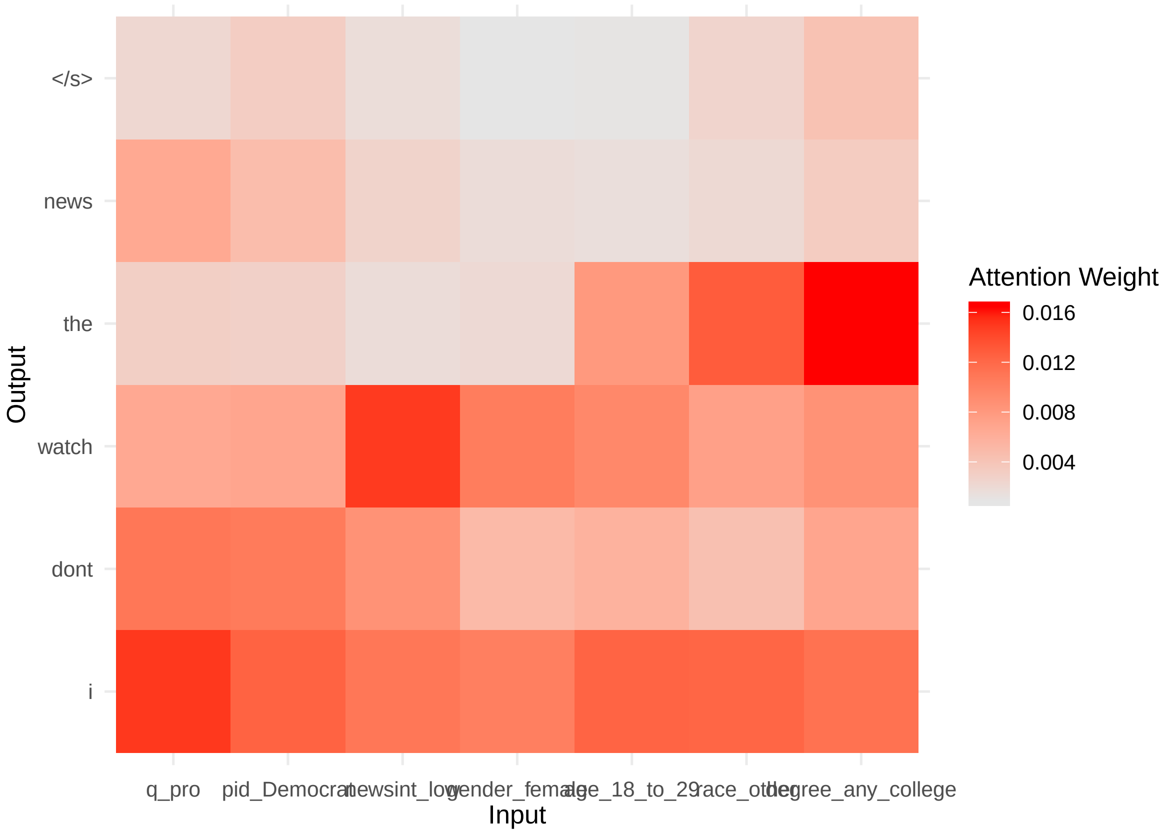


Figure 7: *Illustration of attention weights. When generating the sentence "i dont watch the news" in response to what the respondent is proud of, the model focuses most of its attention on news interest when writing "watch." Note that the attention weights for the news content are omitted here, as the model does not focus on them nearly as much.*

we can conclude that party ID and news interest are very important in determining *what* a respondent will talk about.

Conditional on that, however, *how* they talk about it is, in a sense, influenced more strongly by age, gender and race. This suggests a theoretical expansion from the roles these variables traditionally are considered to play. The theory doesn't make any distinctions between topic and style, and considers party ID and news interest to be uniformly most important across the entire sentence. Of course, the process of averaging across the attention matrices can also be further stratified. For example, it would be feasible to analyze how Democratic and Republican respondents compare in terms of attention, with the same being true for any of the other attributes as well.

4.1.5 Novelty

We have argued that the most important indicator for the quality of the model is its ability to generate realistic and grammatically coherent outputs, tailored to the respondents'

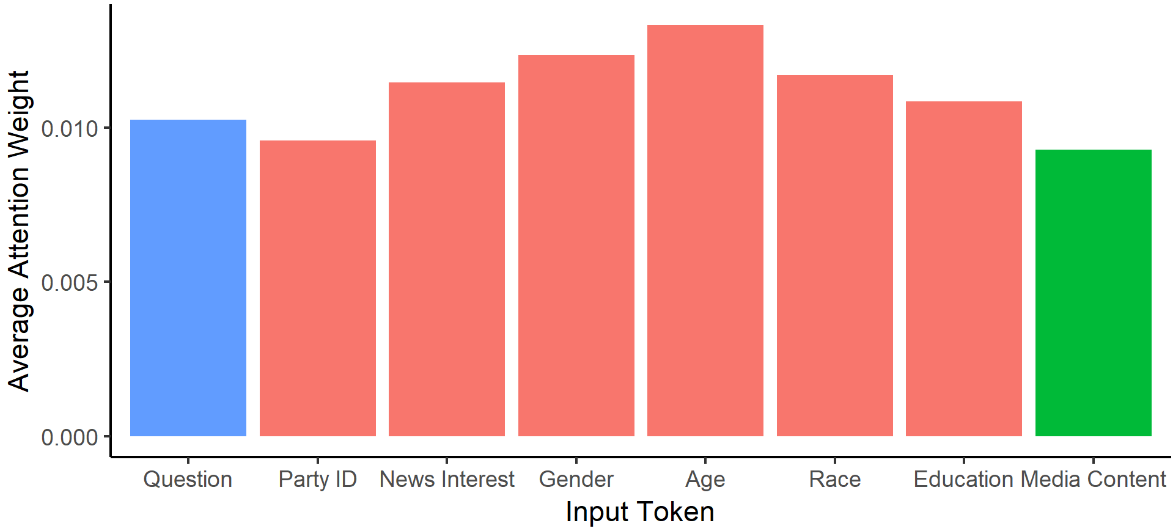


Figure 8: *Average attention per input token for the entire generated sequence, with all news tokens aggregated. The figure suggests that the model pays the most attention to age, gender, and race, but this is in all cases conditional on prior words.*

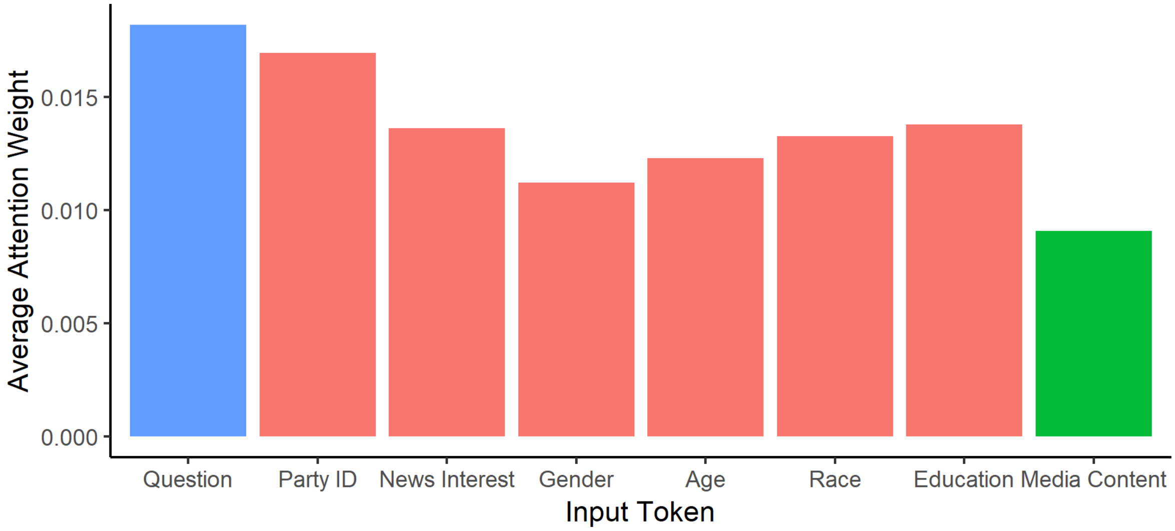


Figure 9: *Average attention for the first generated word. The figure shows that for deciding the word that has the strongest influence on the topic of the output, the model pays the most attention to the question, party identification, education and news interest.*

attributes and answering the question asked. However, as described in section **??**, one arguably undesirable way for the model to accomplish this would be to simply copy its answers from the training data. In other use cases of encoder-decoder models, such as machine translation, this might be intended behavior. For synthesizing survey responses however, this is less useful.

To test whether the model engages in such overfitting, we measure the proportion of predictions that already occur in the training data, the proportion of predictions that are themselves duplicates of other predictions, and the proportion that are neither. All of these indicators are correlated with the number of training iterations. Since the first measure gets slightly worse over the course of training (see Figure 13a in the Appendix), while the second improves considerably (see Figure 13b in the Appendix), we focus on the third, which optimizes them jointly.

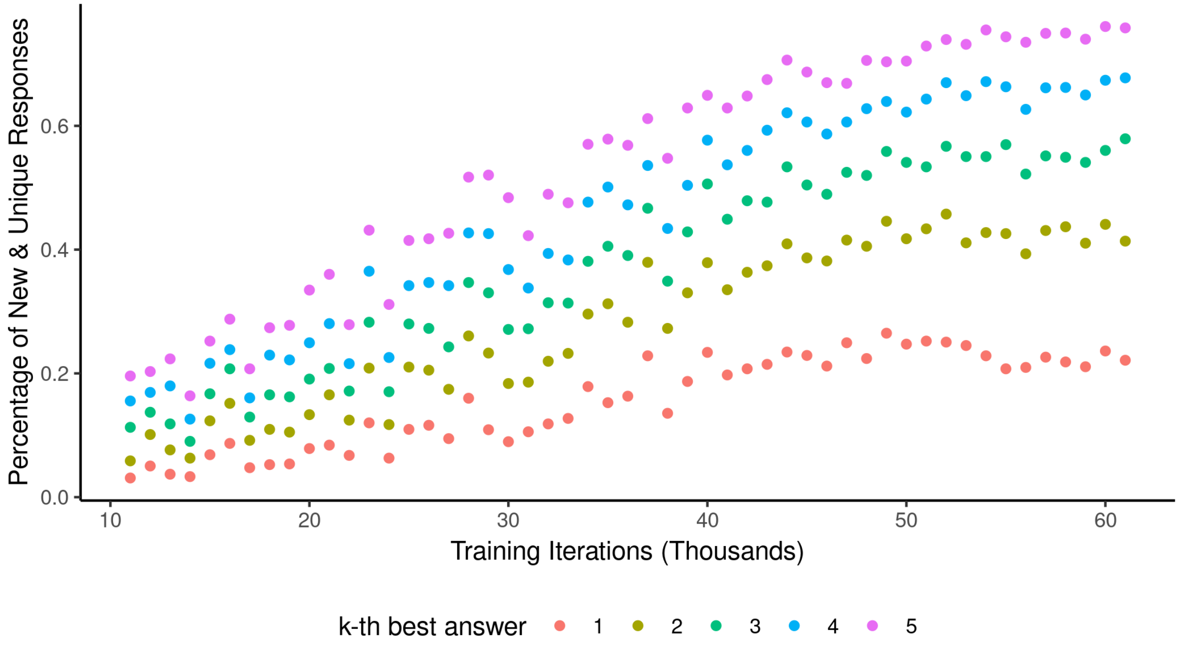


Figure 10: *Proportion of generated responses that are both different from the training data and unique among each other, by number of training samples and k-th best answer.*

Figure 10 shows that the proportion of generated responses that are both new and unique increases with the number of training iterations, but levels off at around 50,000. Much more influential than the length of training however, appears to be whether the model's best guess, or any of its further predictions are used. As explained above, beam search is used to output not just the model's best output, but also further predictions. It appears that the model is much more likely to be overfitting to full training data answers with the highest-ranked answer. Even at 60,000 training iterations, only about 20% of first-choice answers are both new and unique. By contrast, this proportion is about 80% for the model's fifth-ranked prediction.

4.1.6 General utility - distributional similarity

We receive some insight into this phenomenon when we turn to our measure of general utility; cosine similarity of the logged (and smoothed) counts of unigrams and bigrams in partitions and subgroups of the data, between our synthetic responses and the actual test set. Results are shown in Table 6. We see, for example, the (k+1)th best answer tends to perform slightly better than the kth on this measure, as we saw with novelty.

		Sequence-to-Sequence							**Transformer**						
		(unigrams)					(bigrams)		(unigrams)					(bigrams)	
Grouping	**N**	**p1**	**p2**	**p3**	**p4**	**p5**	**p1**	**p5**	**p1**	**p2**	**p3**	**p4**	**p5**	**p1**	**p5**
Aggregate	1	0.66	0.66	0.67	0.68	0.68	0.31	0.34	0.70	0.71	0.71	0.72	0.72	0.31	0.34
Question	5	0.56	0.57	0.58	0.58	0.59	0.21	0.24	0.57	0.58	0.59	0.60	0.61	0.20	0.22
Hopeful		0.63	0.62	0.63	0.63	0.65									
Worried		0.59	0.61	0.60	0.60	0.61									
Proud		0.55	0.58	0.59	0.58	0.58									
Angry		0.57	0.58	0.58	0.60	0.59									
Ashamed		0.44	0.44	0.48	0.47	0.46									
Party ID	3	0.60	0.60	0.50	0.61	0.62	0.24	0.26	0.62	0.63	0.64	0.65	0.65	0.23	0.26
Race	4	0.54	0.55	0.55	0.57	0.56	0.17	0.20	0.56	0.57	0.58	0.58	0.58	0.18	0.19
White		0.64	0.64	0.65	0.65	0.65									
Black		0.53	0.55	0.53	0.54	0.53									
Hispanic		0.54	0.55	0.54	0.55	0.55									
Other		0.46	0.47	0.48	0.52	0.48									
Gender	2	0.62	0.62	0.63	0.64	0.64	0.26	0.29	0.65	0.66	0.67	0.68	0.67	0.26	0.29
Age	4	0.57	0.59	0.60	0.60	0.60	0.22	0.24	0.60	0.61	0.62	0.62	0.63	0.21	0.24
Degree	2	0.63	0.63	0.64	0.65	0.64	0.26	0.29	0.66	0.67	0.67	0.68	0.68	0.27	0.29
News Interest	3	0.61	0.61	0.61	0.62	0.61	0.24	0.26	0.62	0.64	0.64	0.65	0.65	0.23	0.26

Table 6: *Cosine similarities with human answers (of log(count+1) of unigrams and bigrams) across groupings of responses for sequence-to-sequence and transformer model. p1-p5 indicate the 1st through 5th prediction of the models. Where sequence matters less – unigrams, lower probability predictions, or in the transformer – these distributions match the originals more closely. For the most part, groupings with fewer observations are harder to match.*

These are related, in that lower-ranked answers are lower-ranked because their sequence is less likely. By accepting a less likely sequence of words, the model is freer to build that sequence out of novel groupings of individual pieces that match the overall distribution of those pieces. So, as it builds less likely sequences, the model experiments with deleting words, repeating words, substituting words found in similar contexts, and recombining them.

Some of these will be minor variations on other answers. Some will be novel variations that constitute plausible interesting appropriate answers. Some will be nonsense. The likelihood of nonsense goes up when seeking novel (low probability) answers in a sparse region of the input space (rarer attributes and combinations of attributes). We include some examples in Appendix XX.

4.1.7 Specific utility in downstream tasks

XX - Confidence interval overlap in logits

XX - FW comparison

D (Human)	Score	D (Prediction)	Score	R (Human)	Score	R (Prediction)	Score
donald_trump	6.58	donald_trump	6.31	they	4.40	hillary_clinton	4.69
nothing	3.04	his	4.51	hillary_clinton	3.47	great	3.93
republican	2.71	as	3.14	country	3.22	democrats	3.66
one	2.47	republican	2.70	our	2.99	media	3.49
war	2.45	be	2.63	economy	2.90	economy	3.39
in	2.35	republicans	2.50	we	2.63	country	3.10
his	2.21	won	2.44	what	2.59	way	3.07
hopeful	2.20	impeached	2.42	jobs	2.58	their	3.05
house	2.20	really	2.35	media	2.58	between	2.92
congress	2.17	office	2.25	do	2.40	division	2.79

Table 7: *Top fightin' words in a Democratic-Republican contrast, applied to both the human responses to the test sample, as well as the machine-generated outputs. The correlation of scores between real outputs and their generated equivalents is 0.68 for the top 30 words, and 0.34 for all words that occur in both. Ergo, the two are fairly similar, indicating that the model is imitating the real data well.*

4.1.8 Out of sample wave task

Training on the first nine waves and predicting to the 10th, as outlined in section **??**, is a more difficult task. Table 8 shows results on some of the held-out test samples from the final wave of the survey. While the model continues to do well in predicting content that had already featured in the first nine waves, it fails to absorb events that are particular to the tenth. As an example for the former, the second response shows the respondent's reaction to the results of the midterm elections, which happened prior to the ninth wave. Hence, this comment continues to be valid in the tenth.

It does worse in other regards however. For example, the fourth respondent lamented the Democrats' purported agenda of killing infants (the sixth response is similar) – the hot-button issue with conservatives at the time of the tenth wave we have discussed. The model does not manage to imitate this real answer, and complains about Democrats in general. The first prediction of the seventh answer – "the appeal of roe_v_wade abortion decision" however indicates that the model may have understood that abortion was

	Phrase
Attributes	**q_ang pid_Democrat newsint_high gender_female age_65__older race_white degree_any_college**
Human	everything that donald_trump lies about
Prediction 1	donald_trump and his arrogant attitude
Prediction 2	the president of the united_states
Prediction 3	donald_trump and his constant harangue of the media
Attributes	**q_hop pid_Democrat newsint_high gender_female age_45_to_64 race_other degree_any_college**
Human	that mr donald_trump will be impeached the congress can get safer gun_control/laws passed [...]
Prediction 1	the only thing that makes me feel hopeful is that the democrats have the house right now
Prediction 2	the only thing that makes me feel hopeful is that the democrats have the house right now [...]
Prediction 3	the only thing that makes me feel hopeful about the next 12_months
Attributes	**q_wor pid_Democrat newsint_high gender_male age_65__older race_white degree_any_college**
Human	donald_trump starting world_war_iii to avoid being impeached or sent to trial
Prediction 1	donald_trump will do something very stupid and dangerous
Prediction 2	the president will do something very stupid and dangerous
Prediction 3	donald_trump will destroy the fabric of our constitution
Attributes	**q_wor pid_Republican newsint_high gender_female age_45_to_64 race_white degree_any_college**
Human	the democrats agenda raise taxes kill infants
Prediction 1	the democrats refusing to do anything
Prediction 2	that the democrats will continue to work against the president agenda
Prediction 3	the violence in our country
Attributes	**q_ang pid_Democrat newsint_high gender_male age_65__older race_white degree_any_college**
Human	republican legislatures of those states planning to cancel the will of the people
Prediction 1	the lies and misinformation that is spread as gospel
Prediction 2	donald_trump and the republicans who normalize his disgusting behavior
Prediction 3	donald_trump and the republican party
Attributes	**q_ang pid_Republican newsint_high gender_female age_45_to_64 race_other degree_any_college**
Human	the move by liberals and far left politicians to murder babies through abortion and infanticide
Prediction 1	liberals and democrats try to destroy
Prediction 2	there is so much corruption and deceit
Prediction 3	liberals and the democrats for trying to obstruct the constitution
Attributes	**q_hop pid_Republican newsint_high gender_male age_65__older race_white degree_no_college**
Human	president donald_trump building a wall making peace with north_korea [...]
Prediction 1	the appeal of roe_v_wade abortion decision
Prediction 2	a vast red wave in november
Prediction 3	the president and his views for the country

Table 8: *Sample outputs from the sequence-to-sequence model, trained on the first nine waves, predicting to the tenth (see section* **??***). While the model continues to do well in predicting content that had already featured in previous waves – see the first three respondents – it fails to forecast events that are particular to the tenth wave – abortion and the aftermath to the gubernatorial election in Wisconsin.*

relevant at the time, even if it didn't manage to talk about it in terms of "infanticide", or "murdering babies."

The human answer of the fifth response presents an even harder case, as it pertains to the gubernatorial election in Wisconsin, when Republican legislators limited the office's powers after losing it to their Democratic opponents. It is unsurprising that the model does not make this prediction – the issue wasn't sufficiently important to appear in our representation of the news and thus there was no way for it to learn about this. However, it is doubtful whether it could have done so even if it had featured in our news representation – this issue is simply far too complex and too unique to be distinguished within the ten distinct information environments.

Table 9 repeats the FW test applied earlier and, as expected, shows less correspondence to the original data.

D (Human)	Score	D (Prediction)	Score	R (Human)	Score	R (Prediction)	Score
in	4.82	donald_trump	9.14	democrats	7.97	democrats	10.41
donald_trump	4.64	be	7.76	economy	5.61	what	7.84
nothing	4.09	his	6.41	wall	5.00	doing	7.33
house	3.76	united_states	5.63	abortion	4.26	he	7.32
women	3.67	up	4.80	dont	4.24	great	5.77
of	3.63	right	4.71	immigration	3.87	who	5.32
election	3.61	will	4.52	american	3.80	respect	5.32
his	3.53	still	4.43	country	3.76	obamacare	4.87
office	3.43	that	4.40	media	3.73	lack	4.86
administration	3.27	about	4.14	border	3.69	media	4.67

Table 9: *Top fightin' words in a Democratic-Republican contrast, applied to both the human responses to the 10th survey wave, as well as the machine-generated outputs. The correlation of scores between real outputs and their generated equivalents is 0.52 for the top 30 words, and 0.3 for all words that occur in both. The two are not as similar as in Table 7, indicating that the model not as good at predicting to an out-of-sample survey wave.*

4.2. Transformer

In terms of computation, all recurrent neural networks suffer from being inherently sequential – the hidden state of any one timestep is based on the hidden state of the previous timestep, and can thus only be computed in sequence. This limits the potential for parallelization, imposing computational limitations on model complexity. Thus, researchers are forced to use models with only a few layers – in many cases just one (two in our case). Vaswani et al. (2017) address this problem by introducing the transformer, which does away with recurrence in favor of self-attention. Their encoder and decoder are neural networks with six layers (we use a smaller version with two layers) each, as well as multiple forms of attention. In addition to the global attention also used by sequence-to-sequence models, the self-attention mechanism gives both the encoder as well as the decoder the ability to focus on different positions in their respective previous layers. This leads to both improved performance as well as better parallelizability.

4.2.1 Model Specification

The model parameters for our transformer model are chosen as follows. The size of the model's hidden state as well as the dimensions of the word embeddings is 256. Both the encoder and the decoder use two layers. The number of model parameters is 11,619,687. For optimization, we use Adam. We begin with a learning rate of 0.00001, then increase it during the warm-up phase (8000 steps) and decay it after this point. Similar to our sequence-to-sequence architecture, this model also uses copy attention. Similar to the sequence-to-sequence model, a dropout rate of 0.2 is used. Parameters are initialized with values from a uniform distribution as described in Glorot and Bengio (2010). The model is trained for a total of 60,000 steps. Parameters for generating text are identical to the sequence-to-sequence model.

4.2.2 Results

Example model outputs for the same respondents as in Table 1 can be found in Table 10. The predictions of the transformer appear to be of similar quality as those of the sequence-to-sequence model, often revolve around a comparable topic, and in some cases, are even identical. For example, the number 1 prediction of the sequence-to-sequence model for the first respondent, "black swimmer winning medals in olympics" appears as it its second-best prediction, preceded by "feel proud to see another day".

Judging by training-related metrics such as accuracy, cross entropy and perplexity – see Figures 11a to 11c for the sequence-to-sequence model, and Figures 12a to 12c for the transformer – the latter appears to be slightly better. This is further reinforced by the cosine similarity results, shown in Table 6. The transformer's unigram distributions are slightly, but consistently, more similar to originals than are those of the sequence-to-sequence model. This is, however, at least partly due to word order mattering less, as we've seen with lower probability predictions. This is borne out by the very slightly worse performance in matching bigram distributions. We also see marginally lower quality semantics under the transformer (XX)m with examples in Table 10 like "nothing for the death of my children"/"nothing for the death ! !" being illustrative.

	Phrase
Attributes	**q_pro pid_Democrat newsint_high gender_male age_30_to_44 race_black degree_any_college**
Human	racial injustice discussion
Prediction 1	feel proud to see another day
Prediction 2	black swimmer winning medals in olympics
Prediction 3	feel proud about another day
Attributes	**q_pro pid_Other newsint_low gender_female age_30_to_44 race_hispanic degree_no_college**
Human	nothing
Prediction 1	nothing for the death !
Prediction 2	nothing for the death of my children
Prediction 3	nothing for the death ! !
Attributes	**q_ang pid_Other newsint_low gender_female age_45_to_64 race_white degree_no_college**
Human	people hurting and killing each_other in senseless wars
Prediction 1	too much talk not much action
Prediction 2	all the fighting in the world
Prediction 3	all the shootings people dying
Attributes	**q_pro pid_Republican newsint_low gender_female age_18_to_29 race_white degree_any_college**
Human	no
Prediction 1	havent really paid attention lately
Prediction 2	havent really paid attention lately in the news
Prediction 3	havent really paid the news recently
Attributes	**q_wor pid_Republican newsint_high gender_male age_65__older race_white degree_no_college**
Human	paid rioters
Prediction 1	that globalist will screw it up
Prediction 2	how divided our country is
Prediction 3	the outlaw sanctuary_cities and states
Attributes	**q_ash pid_Democrat newsint_high gender_female age_65__older race_other degree_any_college**
Human	the stanford rape case the xenophobia of this country and its unwillingness to help [...]
Prediction 1	constant bickering and inability for candidates to get along
Prediction 2	the way donald_trump disrespects people how can someone like this be a president candidate
Prediction 3	the way donald_trump disrespects people how can someone like this be president candidate
Attributes	**q_hop pid_Democrat newsint_high gender_male age_45_to_64 race_hispanic degree_any_college**
Human	america is headed for very dark times if donald_trump isnt reigned in by responsible americans
Prediction 1	the next election in november 2019
Prediction 2	that there will be just a check within the legislative
Prediction 3	mueller completing his report and hopefully donald_trump impeachment

Table 10: *Sample outputs from the transformer model. The 'Attributes' line describes the attributes of the respondent for whom responses are synthesized. The 'Human' line shows their actual answer. The 'Prediction 1-3' lines show the top three predictions of the model.*

5 Discussion

The models examined here do appear to be capable of learning how to generate artificial survey responses that meet some reasonable standards, with the sequence-to-sequence model providing arguably higher quality outputs. The model's highest ranked outputs are for the most part grammatically and semantically coherent, respond appropriately to the survey questions and correlate with the attributes of survey participants. Furthermore, through the use of beam search and the right number of training iterations, the models can be calibrated to reach a higher level of originality in the generated responses, as well as alignment with the term distributions of the human responses, albeit with some tradeoff of semantic quality and human-ness. The attention mechanism connects our approach to the relevant theory and allows the model to be used for the purpose of inference. When predicting within the time frames in which the surveys were conducted, the model also responds appropriately to current issues and understands, for example, that during the time of the first two waves, Donald Trump is running for president, whereas later on, he is the president.

The model's main weakness lies in predicting to windows in time during which no survey was conducted. Predicting from the first nine waves to the tenth, the model fails to generate the larger number of observed responses on abortion that were observed in the held-out tenth wave. In particular, it doesn't capture the stark language used by primarily conservative respondents, who talked about abortion in terms of "infanticide" and "killing babies". This is in spite of the fact that the term "infanticide" does appear in our representation of the news (specifically Fox News) at the time (see Table 15). The problem here is that the model never had a chance to learn what the term means, as it has never observed it in conjunction with anything else before. Hence, the word embedding of the word isn't located in the 'abortion space' it has learned from reading prior mood answers on the topic. It is clear that our sample and representation of the news generally isn't sufficiently informative for the model. This can, for example, be gleaned from the fact that the attention mechanism rarely focuses on the media, and that the words it does focus on (see Table 16) differ little from the ordering assigned to them by the fightin' words method.

In comparison to the canonical machine translation and summarization use cases, the size and structure of our dataset poses additional challenges. Machine translation training data often consists of hundreds of thousands, if not millions, parallel texts. By contrast, we only have 41,808 open-ended survey responses available to us. Deep neural networks thrive on large amounts of data, and 41,808 samples is definitely at the low end of what is necessary to build a functioning model. As a consequence, our model cannot be trained for as long without resorting to overfitting. However, in reality, the model has even less information available to it than the 41,808 samples suggest. Machine translation generally relies on parallel corpora, where each input sample has one corresponding output sample.

In our case, there is much less variance in the input sample. There are 516 different combinations of the six respondent attributes in the data (of the 576 that are possible), and only 10 different news environments. With the addition of the question type, this means there are 12,219 unique inputs mapped to 41,808 outputs. This means that each of these inputs can lead to an average of 3.42 different outputs – and the model does not have any information on why they can be 'translated' in these different ways. One possible theoretical solution is to dramatically increase the temporal granularity of the Mood Poll itself by, for example, surveying 100 people every day, which is unlikely to be a practical solution to any practical problem.

We can, however, demonstrate that the Mood responses on their own are sufficient as both input and output sequences for the sequence-to-sequence model to learn from some of them how to generate others of them. XX - Neg to pos task.

Negative (Human)		Positive (Human)	Positive (Prediction)
elections	–>	donald_trump is still running [...]	ha ! hopeful ? whats that
democrats with power	–>	not now that democrats have house	the economy is doing better
everything	–>	nothing	no hope at this time
donald_trump d	–>	mrs hillary_clinton	hillary_clinton surge in the polls
the charlottesville tragedy	–>	the manchester charity concert	the march in boston counter protesting racism
people are passing school buses [...]	–>	our daughter is engaged	that president donald_trump is doing a great job !
donald_trump and the gop policies [...]	–>	mueller and his investigations	mueller investigation and democratic house
support for hillary_clinton	–>	the problems with immigrants	donald_trump to stick to his words
x	–>	x	dont follow the news x

Table 11: *Sample output predicting what a respondent would say to a positive question (hopeful, proud), given their answer to a negative one (worried, angry).*

The sequence-to-sequence model and the transformer perform similarly, as both manage to produce good texts that adhere to grammatical rules and correlate appropriately with their input strings. Consequently it is worth asking which one should be used preferentially. On the one hand, the advantage of the sequence-to-sequence approach is that it is simpler, more software implementations are available and that its training is less sensitive to the correct parameters. When (multiple) GPUs are available to the researcher, the strengths of the transformer become clearer, as the ability to build more powerful models becomes apparent.

Furthermore, our particular research design does not fully leverage the strengths of the sequence-to-sequence approach. Our input vector does not carry the complex sequential relationships of natural language and is always of the same length (although depending on how the news is represented, it could be varied). We still have some sequential information, as the model does not have any intrinsic understanding of what varibles are, and has to learn that, for example, *pid_Democrat* and *pid_Republican* are two realizations of the same concepts from the fact that they always appear in the same place. There is also sequential information to be gained from the separators we use to keep the attributes apart from the media (and since the model uses its attention mechanism to focus on these separators, we can infer that it uses this information), and the three different forms of media representation different five-day windows apart from each other. Nevertheless, despite the fact that this includes some sequential information, it is not nearly as complex as the relational properties of natural language.

Consequently, the computationally expensive sequential properties of recurrent neural networks are not fully leveraged in our use of the decoder. This is not a problem for the transformer, which doesn't use any recurrence to begin with. This may not matter as much for the purpose of synthesizing survey responses, but if the goal is inference through the attention mechanism, it could be an advantage. Here, the parallelizability of the transformer might also allow for the bootstrapping of confidence intervals, even though this would still be very expensive.

In addition to the sequence-to-sequence and transformer models, we also experimented with the OpenAI GPT-2 model, a very large transformer, pre-trained on a massive dataset scraped from outgoing links on Reddit (Radford et al., 2018). This model has been shown to generate human-like natural language sufficient to write not just a few sentences, but entire articles. At its core, OpenAI GPT-2 is a language model and therefore still needs to be fine-tuned to fit any individual task, such as the response to an input sentence of respondent attributes. We found that even though the model is capable of generating outputs of similar quality to those in our other models, it is considerably harder to control. Furthermore, since its authors were concerned about its potential use for malicious purposes such as the generation of fake news, only smaller versions of the pre-trained model and no source code are available. Consequently it is somewhat of a black box. Since the model does not appear to perform any better than our model (in spite of the fact that, unlike our models, it should have seen training data from time frames outside of the Mood polls), and has a number of downsides, we elected not to explore it further here. However, pre-trained language models, in conjunction with transfer learning, would likely lead to improvements in this case.

While the primary purpose of our research is the synthesis of survey responses, our model can also be used for inference by interpreting the attention weights. In this sense, it has some similarity to topic models[7], which are also generative models of text. The advantage here is that it is not a bag-of-words model, and can therefore take sequence into

[7] Especially the structural topic model (Roberts et al., 2014), but also LDA (Blei, Ng and Jordan, 2003), which can be used to analyze covariates by stratifying the document-topic matrix.

account. This advantage is demonstrated in our example, as we show that respondent attributes matter in a different way for the first token, compared to the sentence as a whole. The downside, compared to structural topic models, is that the attention weights don't have any polarity – they only indicate whether a variable is relevant, not whether it has a positive or negative impact. At this point it should also be noted that if inference is the goal, overfitting, as discussed above, is less of a concern.

XX - Caution about counterfactual and low probability synthetic data. Also note that this problem interacts with the privacy problem.

In future work, we hope to improve on the current approach and fix its main shortcoming – its inadequate ability to predict to time frames it has not observed before. We believe that, beyond changes in the implementation of the Poll itself, the biggest room for improvement lies in the representation of media content. The fightin' words method (alternatively, tf-idf produces a similar representation) performs adequately in identifying the words most pertinent in describing a specific political context. For any politically engaged human, the words "woodward kavanaugh op nike" are enough to evoke a point in time – when the Kavanaugh confirmation hearings were about to happen, Bob Woodward had just released his book on the Trump administration, and Nike had run a controversial ad with Colin Kaepernick. For a human, these words are sufficient to place them in a certain context, because we were exposed to the news at the time and already had an understanding of the political context necessary to make sense of it. A model which learns language from scratch however, lacks this context. One possible alternative might be representing news content in a more informative manner by relying on methods developed in the machine-coding event literature. In addition to improving the representation of the media data, we also believe that relying on transfer learning could provide our model with better context. At present, we do not use pretrained word embeddings, as this is not common practice in sequence-to-sequence learning, where models are generally able to learn these embeddings by themselves. However, our case might be an exception due to the small size of the survey training data, and the even smaller variation in the news data. Be it through pretrained word embeddings, or a pretrained language model such as OpenAI GPT-2 (Radford et al., 2018), we suspect that transfer learning is the key to solving the model's architectural shortcomings.

References

Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*. pp. 1–15.

Bakshy, Eytan, Solomon Messing and Lada A Adamic. 2015. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348(6239):1130–1132.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Broockman, David E. and Daniel M. Butler. 2017. "The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication." *American Journal of Political Science* 61(1):208–221.

Converse, Philip E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. David Apter. University of Michigan.

Ertugrul, Ali Mert, Yu Ru Lin, Wen Ting Chung, Muheng Yan and Ang Li. 2019. "Activism via attention: interpretable spatiotemporal learning to forecast protest activities." *EPJ Data Science* 8(1).

Glorot, Xavier and Yoshua Bengio. 2010. "Understanding the difficulty of training deep feedforward neural networks Xavier." *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* pp. 249–256.

Hagle, Courtney, Grace Bennett, Nick Fernandez and Julie Tulbert. 2019. "Right-wing media are flat-out lying about later abortions being "infanticide".".
URL: *https://mediamatters.org/sean-hannity/right-wing-media-are-flat-out-lying-about-later-abortions-being-infanticide*

Iyengar, Shanto and Kyu S. Hahn. 2009. "Red media, blue media: Evidence of ideological selectivity in media use." *Journal of Communication* 59(1):19–39.

Jerit, Jennifer and Jason Barabas. 2012. "Partisan Perceptual Bias and the Information Environment." *The Journal of Politics* 74(3):672–684.

Kann, Sharon and Rob Savillo. 2019. "Fox News almost single-handedly manufactured anti-abortion outrage before Trump's State of the Union.".
URL: *https://mediamatters.org/donald-trump/fox-news-almost-single-handedly-manufactured-anti-abortion-outrage-trumps-state-union*

Lample, Guillaume, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc ′ Aurelio Ranzato and Y-Lan Boureau. 2019. "Multiple-Attribute Text Rewriting." *ICLR* (i):1–20.

Lenz, Gabriel S. 2009. "Learning and opinion change, not priming: Reconsidering the priming hypothesis." *American Journal of Political Science* 53(4):821–837.

Li, Juncen, Robin Jia, He He and Percy Liang. 2018. "Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer.".
URL: *http://arxiv.org/abs/1804.06437*

Liu, Chia-Wei, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin and Joelle Pineau. 2016. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation.".

Logeswaran, Lajanugen, Honglak Lee and Samy Bengio. 2018. "Content preserving text generation with attribute controls." (Nips).
URL: *http://arxiv.org/abs/1811.01135*

Luong, Minh-Thang, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP*.

Messing, S. and S. J. Westwood. 2012. "Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online." *Communication Research* 41(8):1042–1063.
URL: *http://crx.sagepub.com.ejournals.um.edu.mt/content/41/8/1042*

Miller, Angie L and Amber D Lambert. 2014. "Open-ended survey questions: Item nonresponse nightmare or qualitative data dream." *Survey Practice* 7(5):1–11.

Mitchell, Amy, Jeffrey Gottfried, Jocelyn Kiley and Katerina Eva Matsa. 2014. "Political Polarization & Media Habits.".
URL: *http://www.journalism.org/interactives/media-polarization/*

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4 SPEC. ISS.):372–403.

Plutzer, Eric. 2019. "The McCourtney Institute Mood of the Nation Poll: How MOTN polls are conducted.".
URL: *https://democracys.psu.edu/research/mood-of-the-nation-poll-1/about-the-poll*

Prior, Markus. 2013. "Media and Political Polarization." *Annual Review of Political Science* 16:101–27.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. 2018. "Language Models are Unsupervised Multitask Learners.".

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4):1064–1082.

See, Abigail, Peter J. Liu and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*.

Snoke, Joshua, Gillian M. Raab, Beata Nowak and Chris Dibben. 2018. "General and specific utility measures for synthetic data." *Journal of the Royal Statistical Society, Series A* 181(3):663–688.

Sutskever, Ilya, Oriol Vinyals and Quoc V Le. 2014. "Sequence to sequence learning with neural networks." *Advances in Neural Information Processing Systems (NIPS)* pp. 3104–3112.
URL: *http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. "Attention Is All You Need." (Nips).
URL: *http://arxiv.org/abs/1706.03762*

Vinyals, Oriol and Quoc Le. 2015. "A Neural Conversational Model.".

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Zaller, John and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36(3):579–616.

Zhang, Zhirui, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou and Enhong Chen. 2018. "Style Transfer as Unsupervised Machine Translation.".

6 Appendix

	Category 1	Category 2	Category 3	Category 4
Party ID	Democrat (4665)	Republican (3353)	Other (1970)	
News Interest	High (5053)	Med (2654)	Low (2281)	
Race	White (7138)	Black (1104)	Hispanic (1048)	Other (698)
Gender	Female (5364)	Male (4624)		
Age	18 to 29 (1697)	30 to 44 (2825)	45 to 64 (3600)	65 Older (1866)
Education	Any College (6193)	No College (3795)		

Table 12: *Summary statistics for the politically-relevant attributes used in our representation of survey respondents.*

Poll Date	Fightin Words
2016-06-20	orlando ahmed nightclub ireland gay lewandowski mateen alton corey convention
2016-09-09	forum intrepid commander birtherism iraq birther certificate baghdadi libya kingston
2016-11-22	mccrory albert alt registry remnick negroes helms thurmond hail hamilton
2017-02-27	cpac perez gaggle oscar ix laverne moonlight issa remnick fsb
2017-08-25	afghanistan arpaio phoenix pakistan corpus taliban christi icahn afghan pardon
2017-11-10	gillespie moore brazile donna northam virginia roy schiller alabama delegates
2018-02-08	porter memo parade steele fisa dossier shutdown clearance raj eagles
2018-09-08	woodward kavanaugh op nike anonymous ed brett kaepernick ohr pressley
2018-11-12	broward whitaker recount snipes ballot county brenda sinema supervisor nelson
2019-02-07	northam blackface fairfax bezos ami herring yearbook ralph tyson queso

Table 13: *Top 10 Fightin Words of the 5 days up to the poll, compared to all other days. Used together with two additional sets of words for the preceding 10 days in the news data representation.*

Poll Date	Fightin Words
2016-06-20	orlando ahmed nightclub ireland gay lewandowski mateen alton corey convention
2016-09-09	forum intrepid birtherism commander birther certificate baghdadi kingston bondi iraq
2016-11-22	mccrory albert registry remnick negroes helms hail thurmond hamilton strom
2017-02-27	cpac perez gaggle ix laverne oscar fsb remnick novaya gazeta
2017-08-25	afghanistan phoenix arpaio icahn pakistan pardon afghan taliban linton charlottesville
2017-11-10	gillespie moore northam virginia schiller delegates roy icahn alabama ralph
2018-02-08	porter memo parade clearance shutdown booster raj rob market kelly
2018-09-08	woodward kavanaugh op brett ed anonymous pressley roe ayanna settled
2018-11-12	whitaker recount broward sinema matt whittaker ballot ftc nelson county
2019-02-07	bezos ami northam enquirer inaugural blackface fairfax herring whitaker ralph

Table 14: *Top 10 Fightin Words of the 5 days up to the poll on MSNBC, compared to all other days on both MSNBC and FOX. Used together with two additional sets of words for the preceding 10 days in the news data representation.*

Poll Date	Fightin Words
2016-06-20	hurtle islamism infidels jaws malignancy accusatory math infidel bisexual enlightenment
2016-09-09	bleachbit header austan coughing allah libya concussion eboni jihadists server
2016-11-22	alt mukasey whitelash colton rosie huckster mistruths grenell ric mercedes
2017-02-27	beatty correspondents la wicker predominately encryption audiobook moonlight selfies prom
2017-08-25	corpus christi gust harrigan espn storm arpaio pakistan taliban phoenix
2017-11-10	brazile donna moore gillespie tezlyn lanny rig negligence louis sportsmanship
2018-02-08	fisa steele dossier memo unverified application shearer sidney bulk emojis
2018-09-08	nike woodward kaepernick ohr anonymous spartacus colin chicago weissmann kavanaugh
2018-11-12	broward snipes recount brenda county ballot supervisor crenshaw palm florida
2019-02-07	blackface fairfax northam tyson queso infanticide herring yearbook ralph socialism

Table 15: *Top 10 Fightin Words of the 5 days up to the poll on FOX, compared to all other days on both MSNBC and FOX. Used together with two additional sets of words for the preceding 10 days in the news data representation.*

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8
Wave 1	orlando	ahmed	nightclub	ireland	lgbt	mateen	gay	gay
Wave 2	forum	intrepid	commander	birther	birtherism	iraq	certificate	mexico
Wave 3	mccrory	ifill	transition	bannon	kobach	albert	bolton	christie
Wave 4	fsb	remnick	issa	moonlight	hall	semitism	panda	laverne
Wave 5	supremacist	nazis	neo	monument	barcelona	spain	pedestrian	nazi
Wave 6	brazile	clovis	bergdahl	guantanamo	papadopoulos	gillespie	gitmo	donna
Wave 7	dossier	fisa	shutdown	steele	clearance	parade	raj	memo
Wave 8	doe	spanberger	gianno	patten	gosling	sam	omarosa	miers
Wave 9	snipes	ballot	county	broward	recount	brenda	whitaker	sinema
Wave 10	ralph	tyson	queso	bezos	yearbook	ami	herring	fairfax

Table 16: *Top attention words, by survey wave.*

Violence

a mother murdered her two children running for president
parents killing their kids and animals
the unnecessary killing of some of our leaders
a cop murdered nearly every celebrity instead of a real politician
im worried about the homosexual movement going to war with north_korea and china

Sports

ted cruz and gregg abbott won in the superbowl
the eagles winning the superbowl carson wentz and bradley cooper got engaged
skateboarding in the executive branch
the olympics and the way they get away with it

Democracy

i dont see the government
what worries me the most is if the republicans and democrats get together and help the american people
the democracy is in miami !
that we are on the road to electing a border wall

Trump

that the russia investigation will find enough evidence to oust donald_trump at the border
we have a president that is trying to do what he promised and puts america first instead of letting america do it
democrats that refuse to work with him to make america a socialist country

The left

the fact that we have liberals and pokemon
at least we have free stuff we still have
the socialism between the parties and the flag

Protest

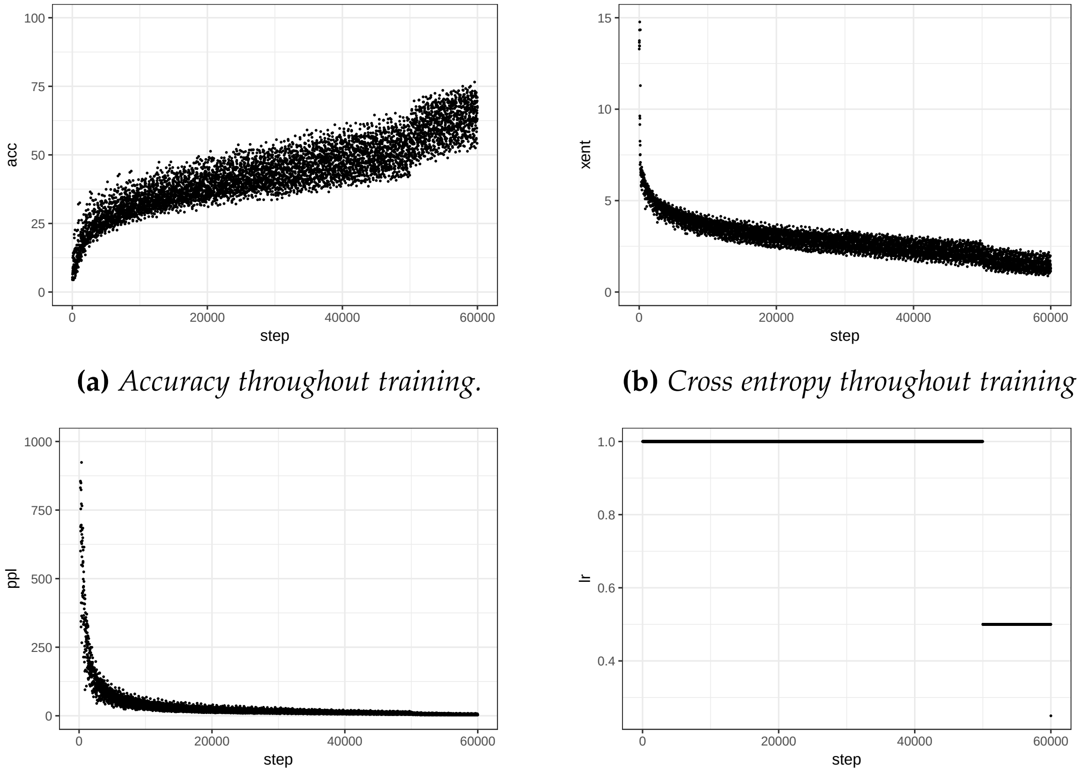
i dont fight for what is right
the number of people who speak out about their asses

Cognitive Dissonance

i dont feel proud of our politics but im proud of our politics
democrats winning the house against democrats
im so happy and upset
anything and everything in particular
nancy_pelosi offering other options as far as nancy_pelosi
the vocal minority will realize that the majority of americans are vocal

Table 17: *Selected synthetic responses, sequence-to-sequence model. Lower ranked predictions, especially in low probability areas of the attribute space, can demonstrate novelty, but can also reveal a mad-libsian understanding of concepts and how they compose into coherent thoughts.*

Figure 11: *Training the sequence-to-sequence model*



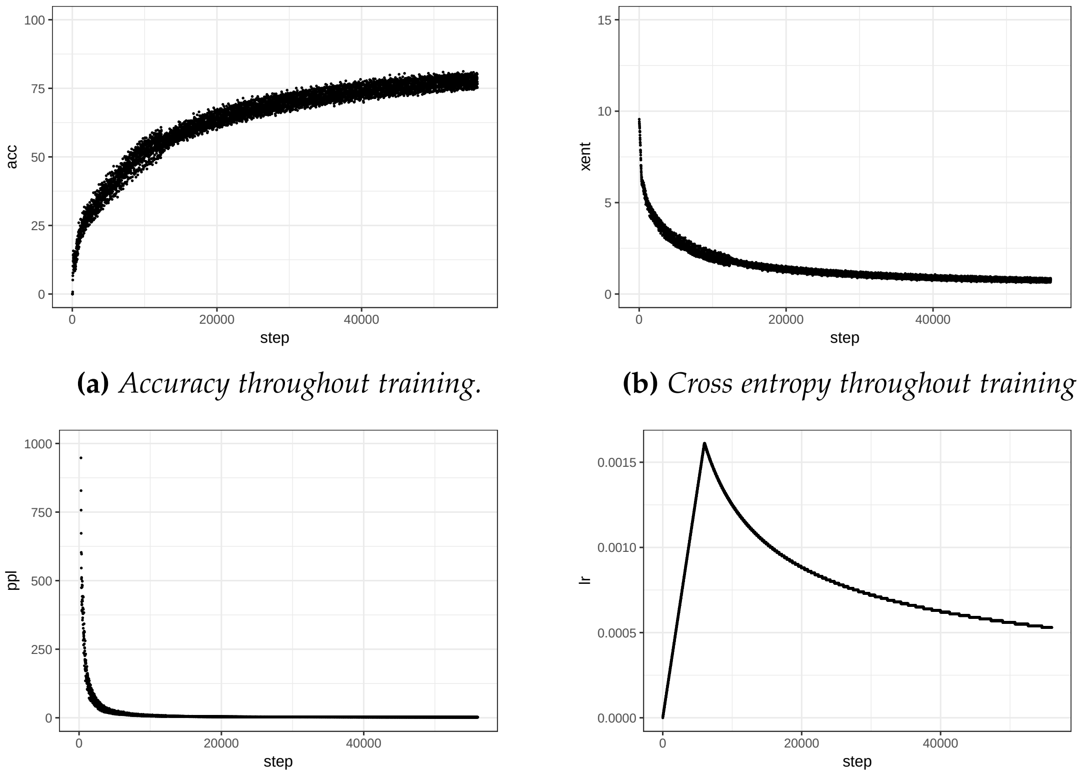
(a) *Accuracy throughout training.*

(b) *Cross entropy throughout training.*

(c) *Perplexity throughout training.*

(d) *Learning rate throughout training.*

Figure 12: *Training the transformer*



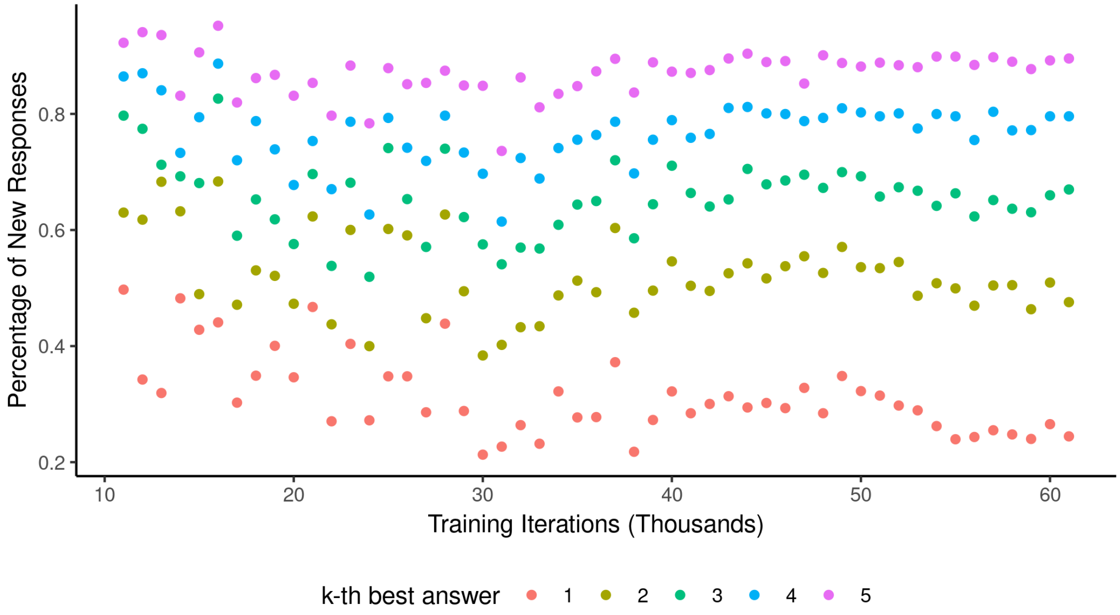
(a) *Accuracy throughout training.*

(b) *Cross entropy throughout training.*

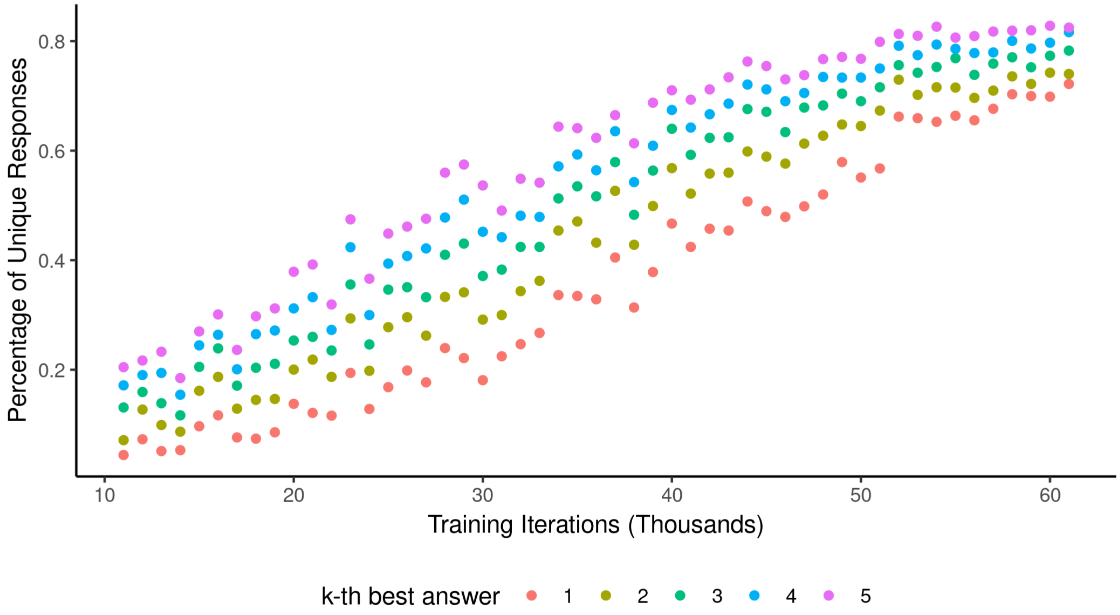
(c) *Perplexity throughout training.*

(d) *Learning rate throughout training.*

Figure 13: *Novelty in generated data*



(a) *Proportion of generated responses that are different from the training data, by number of training samples and k-th best answer.*



(b) *Proportion of generated responses that unique among each other, by number of training samples and k-th best answer.*