What Gets Echoed? Understanding the "Pointers" in Explanations of Persuasive Arguments

David Atkinson and **Kumar Bhargav Srinivasan** and **Chenhao Tan**
Department of Computer Science
University of Colorado Boulder
Boulder, CO
david.i.atkinson, kumar.srinivasan, chenhao.tan@colorado.edu

Abstract

Explanations are central to everyday life, and are a topic of growing interest in the AI community. To investigate the process of providing natural language explanations, we leverage the dynamics of the `/r/ChangeMyView` subreddit to build a dataset with 36K naturally occurring explanations of why an argument is persuasive. We propose a novel word-level prediction task to investigate how explanations selectively reuse, or *echo*, information from what is being explained (henceforth, *explanandum*). We develop features to capture the properties of a word in the explanandum, and show that our proposed features not only have relatively strong predictive power on the echoing of a word in an explanation, but also enhance neural methods of generating explanations. In particular, while the non-contextual properties of a word itself are more valuable for stopwords, the interaction between the constituent parts of an explanandum is crucial in predicting the echoing of content words. We also find intriguing patterns of a word being echoed. For example, although nouns are generally less likely to be echoed, subjects and objects can, depending on their source, be more likely to be echoed in the explanations.

1 Introduction

Explanations are essential for understanding and learning (Keil, 2006). They can take many forms, ranging from everyday explanations for questions such as why one likes Star Wars, to sophisticated formalization in the philosophy of science (Salmon, 2006), to simply highlighting features in recent work on interpretable machine learning (Ribeiro et al., 2016).

Although everyday explanations are mostly encoded in natural language, natural language explanations remain understudied in NLP, partly due to a lack of appropriate datasets and problem formulations. To address these challenges, we leverage `/r/ChangeMyView`, a community dedicated to sharing counterarguments to controversial views on Reddit, to build a sizable dataset of naturally-occurring explanations. Specifically, in `/r/ChangeMyView`, an original poster (OP) first delineates the rationales for a (controversial) opinion (e.g., in Table 1, "most hit music artists today are bad musicians"). Members of `/r/ChangeMyView` are invited to provide counterarguments. If a counterargument changes the OP's view, the OP awards a Δ to indicate the change and is required to *explain why the counterargument is persuasive*. In this work, we refer to what is being explained, including both the original post and the persuasive comment, as the *explanandum*.[1]

An important advantage of explanations in `/r/ChangeMyView` is that the explanandum contains most of the required information to provide its explanation. These explanations often select key counterarguments in the persuasive comment and connect them with the original post. As shown in Table 1, the explanation naturally points to, or *echoes*, part of the explanandum (including both the persuasive comment and the original post) and in this case highlights the argument of "music serving different purposes."

These naturally-occurring explanations thus enable us to computationally investigate the selective nature of explanations: "people rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation" (Miller, 2018). To understand the selective process of providing explanations, we formulate a word-level task to predict whether a word in an

[1]The plural of *explanandum* is explananda.

Original post (OP): CMV: most hit music artists today are bad musicians Now I know, music is art and art has no rules, but this is only so true. Movies are art too but I think most of us can agree the emoji movie was objectively bad. That aside: I really feel like once you remove the persona and performances of the artists from the "top 40" songs and listen to them as just a song, most are objectively bad. They're super repetitive, the lyrics and painfully generic, and there's hardly ever anything new or challenging. And from what I understand most of these artists don't even write their own songs. Of course there are exceptions but I find them to be extremely rare. It seems to me they're only popular become of who they are and how they look/perform. I realize this is probably a very snobbish view which is why I want to be enlightened, so can anyone convince me otherwise? Are they actually good musicians or just good performers? [one more paragraph ...]
Persuasive comment (PC): Music appreciation is a skill, and it's all about pattern recognition. When we're children, we need songs that are really simple, repetitive and with easy to recognize patterns. The younger we are, the simpler the songs. Toddlers like nursery rhymes, lullabies, jingles. Teens like pop music. And teens spend more on music than anyone else. [four more paragraphs ...] Lastly, you have to consider that music can be listened to in different ways and for different purposes. You can listen to it alone on headphones, and think about what it means and how it makes you feel. Or you can dance to it with your friends. Or maybe you need something on in the background during a dinner party, or a house party, or while you study, or are trying to fall asleep, or work out. Pop music is really good in some of these situations, really bad in others. But it serves a definite purpose and isn't bad in any essential way.
Explanation: Δ I guess I never *really looked* at it as *music serving different purposes*. I can *see* how *pop music* fills a certain *purpose*, and I guess the *artist* does *n't* necessarily have to be the *one* to *write* the *song* (although I *appreciate* it when they do).

Table 1: An illustration of the pointers in an example explanation of `/r/ChangeMyView`. We color the words in the explanation based on whether it is used in the original post (e.g., *artist*), in the persuasive comment (e.g., *purpose*), or both (e.g., *music*). We stem all the words before matching and do not color stopwords for readability.

explanandum will be echoed in its explanation.

Inspired by the observation that words that are likely to be echoed are either frequent or rare, we propose a variety of features to capture how a word is used in the explanandum as well as its non-contextual properties in Section 4. We find that a word's usage in the original post and in the persuasive argument are similarly related to being echoed, except in part-of-speech tags and grammatical relations. For instance, verbs in the original post are less likely to be echoed, while the relationship is reversed in the persuasive argument.

We further demonstrate that these features can significantly outperform a random baseline and even a neural model with significantly more knowledge of a word's context. The difficulty of predicting whether content words (i.e., non-stopwords) are echoed is much greater than that of stopwords,[2] among which adjectives are the most difficult and nouns are relatively the easiest. This observation highlights the important role of nouns in explanations. We also find that the relationship between a word's usage in the original post and in the persuasive comment is crucial for predicting the echoing of content words. Our proposed features can also improve the performance of pointer generator networks with coverage in generating explanations (See et al., 2017).

To summarize, our main contributions are:

- We highlight the importance of computationally characterizing human explanations and formulate a concrete problem of predicting how information is selected from explananda to form explanations, including building a novel dataset of naturally-occurring explanations.
- We provide a computational characterization of natural language explanations and demonstrate the U-shape in which words get echoed.
- We identify interesting patterns in what gets echoed through a novel word-level classification task, including the importance of nouns in shaping explanations and the importance of contextual properties of both the original post and persuasive comment in predicting the echoing of content words.
- We show that vanilla LSTMs fail to learn some of the features we develop and that the proposed features can even improve performance in generating explanations with pointer networks.

Our code and dataset is available at `https://chenhaot.com/papers/explanation-pointers.html`.

2 Related Work

To provide background for our study, we first present a brief overview of explanations for the NLP community, and then discuss the connection of our study with pointer networks, linguistic accommodation, and argumentation mining.

[2] We use the stopword list in NLTK.

The most developed discussion of explanations is in the philosophy of science. Extensive studies aim to develop formal models of explanations (e.g., the deductive-nomological model in Hempel and Oppenheim (1948), see Salmon (2006) and Woodward (2005) for a review). In this view, explanations are like proofs in logic. On the other hand, psychology and cognitive sciences examine "everyday explanations" (Keil, 2006; Lombrozo, 2006). These explanations tend to be selective, are typically encoded in natural language, and shape our understanding and learning in life despite the absence of "axioms." Please refer to Wilson and Keil (1998) for a detailed comparison of these two modes of explanation.

Although explanations have attracted significant interest from the AI community thanks to the growing interest on interpretable machine learning (Doshi-Velez and Kim, 2017; Lipton, 2016; Guidotti et al., 2019), such studies seldom refer to prior work in social sciences (Miller, 2018). Recent studies also show that explanations such as highlighting important features induce limited improvement on human performance in detecting deceptive reviews and media biases (Lai and Tan, 2019; Horne et al., 2019). Therefore, we believe that developing a computational understanding of everyday explanations is crucial for explainable AI. Here we provide a data-driven study of everyday explanations in the context of persuasion.

In particular, we investigate the "pointers" in explanations, inspired by recent work on pointer networks (Vinyals et al., 2015). Copying mechanisms allow a decoder to generate a token by copying from the source, and have been shown to be effective in generation tasks ranging from summarization to program synthesis (See et al., 2017; Ling et al., 2016; Gu et al., 2016). To the best of our knowledge, our work is the first to investigate the phenomenon of pointers in explanations.

Linguistic accommodation and studies on quotations also examine the phenomenon of reusing words (Danescu-Niculescu-Mizil et al., 2011; Giles and Ogay, 2007; Leskovec et al., 2009; Simmons et al., 2011). For instance, Danescu-Niculescu-Mizil et al. show that power differences are reflected in the echoing of function words; Tan et al. (2018) find that news media prefer to quote locally distinct sentences in political debates. In comparison, our word-level formulation presents a fine-grained view of echoing words, and puts a stronger emphasis on content words than work on linguistic accommodation.

Finally, our work is concerned with an especially challenging problem in social interaction: persuasion. A battery of studies have done work to enhance our understanding of persuasive arguments (Wang et al., 2017; Zhang et al., 2016; Habernal and Gurevych, 2016; Lukin et al., 2017; Durmus and Cardie, 2018), and the area of argumentation mining specifically investigates the structure of arguments (Lippi and Torroni, 2016; Walker et al., 2012; Somasundaran and Wiebe, 2009). We build on previous work by Tan et al. (2016) and leverage the dynamics of `/r/ChangeMyView`. Although our findings are certainly related to the persuasion process, we focus on understanding the self-described reasons for persuasion, instead of the structure of arguments or the factors that drive effective persuasion.

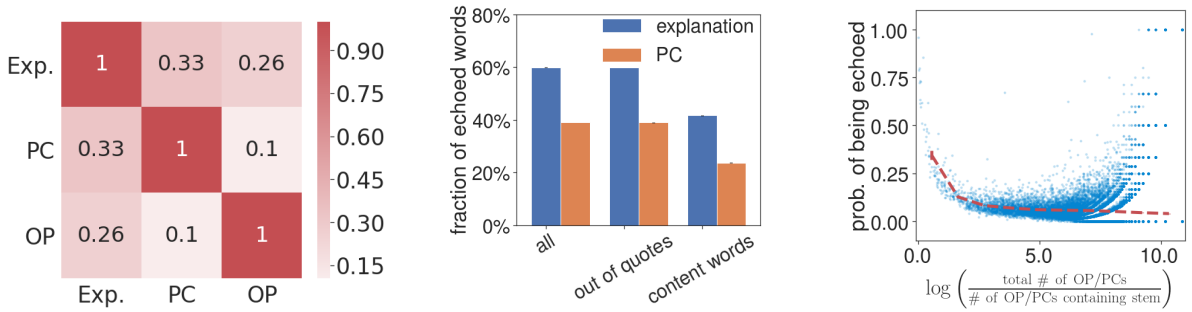
3 Dataset

Our dataset is derived from the `/r/ChangeMyView` subreddit, which has more than 720K subscribers (Tan et al., 2016). `/r/ChangeMyView` hosts conversations where someone expresses a view and others then try to change that person's mind. Despite being fundamentally based on argument, `/r/ChangeMyView` has a reputation for being remarkably civil and productive (CMV moderators, 2019a), e.g., a journalist wrote "In a culture of brittle talking points that we guard with our lives, Change My View is a source of motion and surprise" (Heffernan, 2018).

The delta mechanism in `/r/ChangeMyView` allows members to acknowledge opinion changes and enables us to identify *explanations* for opinion changes (CMV moderators, 2019b). Specifically, it requires "Any user, whether they're the OP or not, should reply to a comment that changed their view with a delta symbol and *an explanation of the change*." As a result, we have access to tens of thousands of naturally-occurring explanations and associated explananda. In this work, we focus on the opinion changes of the original posters.

Throughout this paper, we use the following terminology:

- An **original post (OP)** is an initial post where the original poster justifies his or her opinion. We also use OP to refer to the original poster.
- A **persuasive comment (PC)** is a comment that directly leads to an opinion change on the part



(a) Length correlations.

(b) Fraction of words that are echoed from the explanandum.

(c) Word-level echoing probability vs. document frequency.

Figure 1: Figure 1a shows the pairwise Pearson correlation coefficient between lengths of OP, PC, and explanation (all values are statistically significant with $p < 1e{-}10$). Figure 1b shows the average fraction of words in an explanation that are in its OP or PC, and the fraction of words in a PC that are in its OP. In Figure 1c, the y-axis represents the probability of a word in an OP or PC being echoed in the explanation, while the x-axis shows the inverse document frequency of that word in training data. For each document frequency decile, we calculate the probability of a word in that decile being echoed, and plot those probabilities with the red line. In Figure 1b and Figure 1c, the (small) error bars represent standard errors.

of the OP (i.e., winning a Δ).

- A **top-level comment** is a comment that directly replies to an OP, and `/r/ChangeMyView` requires the top-level comment to "challenge at least one aspect of OPs stated view (however minor), unless they are asking a clarifying question."
- An **explanation** is a comment where an OP acknowledges a change in his or her view and provides an explanation of the change. As shown in Table 1, the explanation not only provides a rationale, it can also include other discourse acts, such as expressing gratitude.

Using `https://pushshift.io`, we collect the posts and comments in `/r/ChangeMyView` from January 17th, 2013 to January 31st, 2019, and extract tuples of (OP, PC, explanation). We use the tuples from the final six months of our dataset as the test set, those from the six months before that as the validation set, and the remaining tuples as the training set. The sets contain 5,270, 5,831, and 26,617 tuples respectively. Note that there is no overlap in time between the three sets and the test set can therefore be used to assess generalization including potential changes in community norms and world events.

Preprocessing. We perform a number of preprocessing steps, such as converting blockquotes in Markdown to quotes, filtering explicit edits made by authors, mapping all URLs to a special `@url@` token, and replacing hyperlinks with the link text. We ignore all triples that contain any deleted comments or posts. We use spaCy for tokenization and tagging (Honnibal and Montani, 2017). We also use the NLTK implementation of the Porter stemming algorithm to store the stemmed version of each word, for later use in our prediction task (Loper and Bird, 2002; Porter, 1980). Refer to the supplementary material for more information on preprocessing.

Data statistics. Table 2 provides basic statistics of the training tuples and how they compare to other comments. We highlight the fact that PCs are on average longer than top-level comments, suggesting that PCs contain substantial counterarguments that directly contribute to opinion change. Therefore, we simplify the problem by focusing on the (OP, PC, explanation) tuples and ignore any other exchanges between an OP and a commenter.

Below, we highlight some notable features of explanations as they appear in our dataset.

The length of explanations shows stronger correlation with that of OPs and PCs than between OPs and PCs (Figure 1a). This observation indicates that explanations are somehow better related with OPs and PCs than PCs are with OPs in terms of language use. A possible reason is that the explainer combines their natural tendency towards length with accommodating the PC.

Explanations have a greater fraction of "pointers" than do persuasive comments (Figure 1b). We measure the likelihood of a word in an explanation being copied from either its OP or PC and provide a similar probability for a PC for copying from its OP. As we discussed in Section 1, the words in an explanation are much more likely to come from the existing discussion than

	count	#sentences	#words
Tuples of (OP, PC, Explanations)			
Original Posts	26.3K	16.8	298.8
Persuasive comments	26.3K	12.6	218.3
Explanations	26.3K	5.3	79.8
All of `/r/ChangeMyView` during the training period			
Original posts	93.4k	13.2	172.6
Top-level comments	681.6k	9.1	147.4
All comments	3.6M	6.5	98.9

Table 2: Basic statistics of the training dataset.

are the words in a PC (59.8% vs 39.0%). This phenomenon holds even if we restrict ourselves to considering words outside quotations, which removes the effect of quoting other parts of the discussion, and if we focus only on content words, which removes the effect of "reusing" stopwords.

Relation between a word being echoed and its document frequency (Figure 1c). Finally, as a preview of our main results, the document frequency of a word from the explanandum is related to the probability of being echoed in the explanation. Although the average likelihood declines as the document frequency gets lower, we observe an intriguing U-shape in the scatter plot.[3] In other words, the words that are most likely to be echoed are either unusually frequent or unusually rare, while most words in the middle show a moderate likelihood of being echoed.

4 Understanding the Pointers in Explanations

To further investigate how explanations select words from the explanandum, we formulate a word-level prediction task to predict whether words in an OP or PC are echoed in its explanation. Formally, given a tuple of (OP, PC, explanation), we extract the unique stemmed words as $\mathcal{V}_{\text{OP}}, \mathcal{V}_{\text{PC}}, \mathcal{V}_{\text{EXP}}$. We then define the label for each word in the OP or PC, $w \in \mathcal{V}_{\text{OP}} \cup \mathcal{V}_{\text{PC}}$, based on the explanation as follows:

$$y_w = \begin{cases} 1 & \text{if } w \in \mathcal{V}_{\text{EXP}}, \\ 0 & \text{otherwise.} \end{cases}$$

[3] A similar U-shape exists if we examine the probability of a PC echoing its OP, but does not show up if we compare an OP echoing a different, randomly chosen OP. It is worth noting that PCs can also be viewed as explaining why the OP is problematic. However, constructing a PC involves selecting from a large number of possible counter perspectives (all of which are unobservable). See the supplementary material for a detailed discussion.

Our prediction task is thus a straightforward binary classification task at the word level. We develop the following five groups of features to capture properties of how a word is used in the explanandum (see Table 3 for the full list):

- Non-contextual properties of a word. These features are derived directly from the word and capture the general tendency of a word being echoed in explanations.
- Word usage in an OP or PC (two groups). These features capture *how* a word is used in an OP or PC. As a result, for each feature, we have two values for the OP and PC respectively.
- How a word connects an OP and PC. These features look at the difference between word usage in the OP and PC. We expect this group to be the most important in our task.
- General OP/PC properties. These features capture the general properties of a conversation. They can be used to characterize the background distribution of echoing.

Table 3 further shows the intuition for including each feature, and condensed t-test results after Bonferroni correction. Specifically, we test whether the words that were echoed in explanations have different feature values from those that were not echoed. In addition to considering all words, we also separately consider stopwords and content words in light of Figure 1c. Here, we highlight a few observations:

- Although we expect more complicated words (*#characters*) to be echoed more often, this is not the case on average. We also observe an interesting example of Simpson's paradox in the results for Wordnet depth (Blyth, 1972): shallower words are more likely to be echoed across all words, but deeper words are more likely to be echoed in content words and stopwords.
- OPs and PCs generally exhibit similar behavior for most features, except for part-of-speech and grammatical relation (subject, object, and other.) For instance, verbs in an OP are less likely to be echoed, while verbs in a PC are more likely to be echoed.
- Although nouns from both OPs and PCs are less likely to be echoed, within content words, subjects and objects from an OP are more likely to be echoed. Surprisingly, subjects and objects in a PC are less likely to be echoed, which suggests that the original poster tends to refer back

Feature group	Features and intuitions	Echoed?
Non-contextual properties	Inverse document frequency. As shown in Figure 1c, although document frequency can have non-linear relationships with being copied, the average echoing probability is greater for more common words.	$\downarrow\downarrow\downarrow\downarrow$
	Number of characters. Longer words tend to be more complicated, and may be more likely to be echoed as part of the core argument.	$\downarrow\downarrow\downarrow\downarrow$
	Wordnet depth. Similar to number of characters, the depth in wordnet can indicate the complexity of a word and we expect words with greater depth to be echoed.	$\downarrow\downarrow\downarrow\downarrow^{RC,RS}$
	Echoing likelihood. We also compute the general tendency of a word being echoed in the training data. We expect the feature to be positively correlated with the label.	$\uparrow\uparrow\uparrow\uparrow$
How a word is used in an OP or PC (OP/PC usage)	Part-of-speech (POS) tags. We compute the percentage of times that the surface forms of a stemmed word appear as different part-of-speech tags. We expect nouns and verbs more likely to be echoed. Results: verb in an OP $\downarrow\downarrow\downarrow\downarrow^{RS}$, noun in an OP ($\downarrow\downarrow\downarrow\downarrow$), verb in a PC ($\uparrow\uparrow\uparrow\uparrow$), noun in a PC: $\downarrow\downarrow\downarrow\downarrow^{RC}$.	
	Subjects and objects from dependency labels. We compute the percentage of times that the word appears as subjects, objects, and others. We expect subjects and objects more likely to be echoed. Results: subjects in an OP: $\uparrow\uparrow\uparrow\uparrow$, objects in an OP: $\downarrow\downarrow\downarrow\downarrow^{RC}$, others in an OP: $\uparrow\uparrow\uparrow\uparrow^{RC}$, subjects in a PC: $\downarrow\downarrow\downarrow\downarrow$, objects in a PC: $\downarrow\downarrow\downarrow\downarrow$; others in a PC: $\uparrow\uparrow\uparrow\uparrow$.	
	(Normalized) term frequency. We expect frequent terms to be echoed.	$\uparrow\uparrow\uparrow\uparrow$
	#surface forms. We expect words that have diverse surface forms to be echoed.	$\uparrow\uparrow\uparrow\uparrow$
	Location. For words that never show up in an OP or PC, the default value is 0.5. We expect later words to be echoed. Results: location in an OP: $\uparrow\uparrow\uparrow\uparrow$ (not significant in stopwords); location in a PC: $\uparrow^{RS}$.	
	In quotes. We expect words in quotes to be echoed as they are already emphasized.	$\uparrow\uparrow\uparrow\uparrow$
	Entity. We expect entities to be echoed.	$\uparrow\uparrow\uparrow\uparrow$
How a word connects an OP and PC (OP-PC relation)	Occurs both in an OP and PC.	$\uparrow\uparrow\uparrow\uparrow$
	#Surface forms in an OP but not in the PC.	$\downarrow\downarrow\downarrow\downarrow$
	#Surface forms in a PC but not in the OP.	$\uparrow\uparrow\uparrow\uparrow^{RS}$
	Jensen-Shannon (JS) distance between the OP and PC POS tag distributions of the word.	$\downarrow\downarrow\downarrow\downarrow$
	JS distance between subjects/objects distributions of the word in an OP and PC.	$\downarrow\downarrow\downarrow\downarrow$
General OP/PC properties	OP length.	$\downarrow\downarrow\downarrow\downarrow^{RS}$
	PC length.	$\uparrow\uparrow\uparrow\uparrow$
	Difference in #words.	$\downarrow\downarrow\downarrow\downarrow^{RS}$
	Difference in average #characters in words.	$\downarrow\downarrow\downarrow\downarrow$
	Part-of-speech tags distributional differences between an OP and PC.	$\downarrow\downarrow\downarrow\downarrow$
	Depth of the PC in the thread.	$\uparrow\uparrow\uparrow\uparrow$

Table 3: Features to capture the properties of a word in the context of an explanandum. The last column shows $t-$test results after Bonferroni correction. $\uparrow$ indicates that words that are echoed have a greater value in the feature, while $\downarrow$ indicates the reverse. The number of arrows indicates the level of p-value: $\uparrow\uparrow\uparrow\uparrow$: $p < 0.0001$, $\uparrow\uparrow\uparrow$: $p < 0.001$, $\uparrow\uparrow$: $p < 0.01$, $\uparrow$: $p < 0.05$. ^{RC} and ^{RS} indicate that the direction is flipped in content words and stopwords respectively. We show significance testing results in a condensed format for space reasons. Refer to the supplementary material for the complete testing results.

to their own subjects and objects, or introduce new ones, when providing explanations.

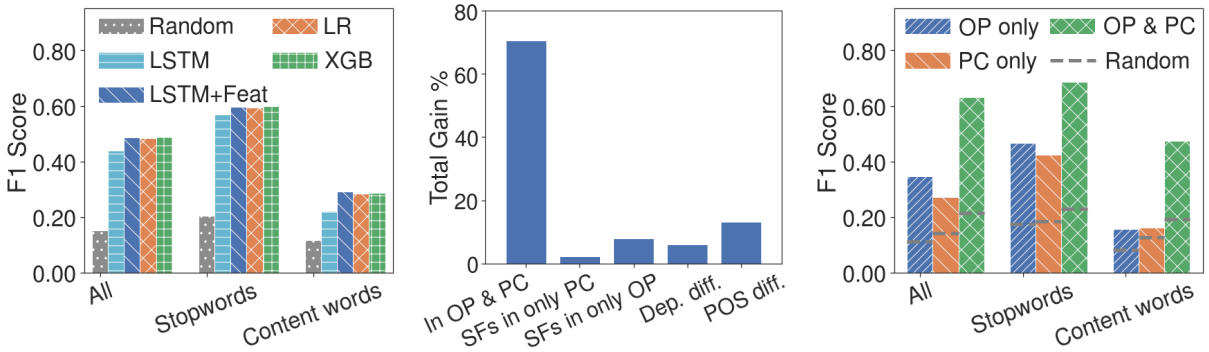
- Later words in OPs and PCs are more likely to be echoed, especially in OPs. This could relate to OPs summarizing their rationales at the end of their post and PCs putting their strongest points last.
- Although the number of surface forms in an OP or PC is positively correlated with being echoed, the differences in surface forms show reverse trends: the more surface forms of a word that show up only in the PC (i.e., not in the OP), the more likely a word is to be echoed. However, the reverse is true for the number of surface forms in only the OP. Such contrast echoes Tan et al. (2016), in which dissimilarity in word usage between the OP and PC was a predictive feature of successful persuasion.

5 Predicting Pointers

We further examine the effectiveness of our proposed features in a predictive setting. These features achieve strong performance in the word-level classification task, and can enhance neural models in both the word-level task and generating explanations. However, the word-level task remains challenging, especially for content words.

5.1 Experiment setup

We consider two classifiers for our word-level classification task: logistic regression and gradient boosting tree (XGBoost) (Chen and Guestrin, 2016). We hypothesized that XGBoost would outperform logistic regression because our problem is non-linear, as shown in Figure 1c.



(a) Overall Performance comparison between models.

(b) Feature importance of an ablated model with OP-PC relation.

(c) Performance vs. word source.

Figure 2: Figure 2a presents the performance of different models. We evaluate the performance of each model on the subset of stopwords and content words. Our features with XGBoost and logistic regression outperform the vanilla LSTM model, and adding our features to the vanilla LSTM model achieves similar performance as XGBoost. Figure 2b shows the normalized total gain of the classifier only based on features in OP-PC relation, while Figure 2c further breaks down the performance based on where the words come from.

To examine the utility of our features in a neural framework, we further adapt our word-level task as a tagging task, and use LSTM as a baseline. Specifically, we concatenate an OP and PC with a special token as the separator so that an LSTM model can potentially distinguish the OP from PC, and then tag each word based on the label of its stemmed version. We use GloVe embeddings to initialize the word embeddings (Pennington et al., 2014). We concatenate our proposed features of the corresponding stemmed word to the word embedding; the resulting difference in performance between a vanilla LSTM demonstrates the utility of our proposed features. We scale all features to $[0, 1]$ before fitting the models. As introduced in Section 3, we split our tuples of (OP, PC, explanation) into training, validation, and test sets, and use the validation set for hyperparameter tuning. Refer to the supplementary material for additional details in the experiment.

Evaluation metric. Since our problem is imbalanced, we use the F1 score as our evaluation metric. For the tagging approach, we average the labels of words with the same stemmed version to obtain a single prediction for the stemmed word. To establish a baseline, we consider a random method that predicts the positive label with 0.15 probability (the base rate of positive instances).

5.2 Prediction Performance

Overall performance (Figure 2a). Although our word-level task is heavily imbalanced, all of our models outperform the random baseline by a wide margin. As expected, content words are much more difficult to predict than stopwords, but the best F1 score in content words more than doubles that of the random baseline (0.286 vs. 0.116). Notably, although we strongly improve on our random baseline, even our best F1 scores are relatively low, and this holds true regardless of the model used. Despite involving more tokens than standard tagging tasks (e.g., Marcus et al. (1994) and Plank et al. (2016)), predicting whether a word is going to be echoed in explanations remains a challenging problem.

Although the vanilla LSTM model incorporates additional knowledge (in the form of word embeddings), the feature-based XGBoost and logistic regression models both outperform the vanilla LSTM model. Concatenating our proposed features with word embeddings leads to improved performance from the LSTM model, which becomes comparable to XGBoost. This suggests that our proposed features can be difficult to learn with an LSTM alone.

Despite the non-linearity observed in Figure 1c, XGBoost only outperforms logistic regression by a small margin. In the rest of this section, we use XGBoost to further examine the effectiveness of different groups of features, and model performance in different conditions.

Ablation performance (Table 4). First, if we only consider a single group of features, as we hypothesized, the relation between OP and PC is crucial and leads to almost as strong performance in content words as using all features. To further

	content	stop
all features	0.286	0.600
random	0.116	0.205

	forward		backward	
	content	stop	content	stop
Non-contextual prop.	0.177	**0.582**	0.285	**0.561**
OP usage	0.191	0.527	0.281	0.599
PC usage	0.233	0.520	**0.275**	0.598
OP-PC relation	**0.280**	0.542	0.289	0.600
General OP/PC prop.	0.153	0.266	0.285	0.598

Table 4: Ablation performance with XGBoost on content words and stopwords (each ablated model is tuned based on performance on all words). "forward" refers to only using a group of features, while "backward" refers to only removing a group of features.

understand the strong performance of OP-PC relation, Figure 2b shows the feature importance in the ablated model, measured by the normalized total gain (see the supplementary material for feature importance in the full model). A word's occurrence in both the OP and PC is clearly the most important feature, with distance between its POS tag distributions as the second most important. Recall that in Table 3 we show that words that have similar POS behavior between the OP and PC are more likely to be echoed in the explanation.

Overall, it seems that word-level properties contribute the most valuable signals for predicting stopwords. If we restrict ourselves to only information in either an OP or PC, how a word is used in a PC is much more predictive of content word echoing (0.233 vs 0.191). This observation suggests that, for content words, the PC captures more valuable information than the OP. This finding is somewhat surprising given that the OP sets the topic of discussion and writes the explanation.

As for the effects of removing a group of features, we can see that there is little change in the performance on content words. This can be explained by the strong performance of the OP-PC relation on its own, and the possibility of the OP-PC relation being approximated by OP and PC usage. Again, word-level properties are valuable for strong performance in stopwords.

Performance vs. word source (Figure 2c). We further break down the performance by where a word is from. We can group a word based on whether it shows up only in an OP, a PC, or both OP and PC, as shown in Table 1. There is a striking difference between the performance in the three categories (e.g., for all words, 0.63 in OP & PC vs. 0.271 in PC only). The strong performance on words in both the OP and PC applies to stopwords and content words, even accounting for the shift in the random baseline, and recalls the importance of occurring both in OP and PC as a feature.

	content	all	random
noun	0.354	0.361	0.130
adverb	0.342	0.411	0.127
verb	0.306	0.466	0.122
proper noun	0.280	0.336	0.109
adjective	0.237	0.289	0.111

Table 5: Performance on five non-function part-of-speech tags (sorted by performance within content words). As a comparison, we also show the performance of the random baseline on content words, which is relatively stable across part-of-speech tags.

Furthermore, the echoing of words from the PC is harder to predict (0.271) than from the OP (0.347) despite the fact that words only in PCs are more likely to be echoed than words only in OPs (13.5% vs. 8.6%). The performance difference is driven by stopwords, suggesting that our overall model is better at capturing signals for stopwords used in OPs. This might relate to the fact that the OP and the explanation are written by the same author; prior studies have demonstrated the important role of stopwords for authorship attribution (Raghavan et al., 2010).

Nouns are the most reliably predicted part-of-speech tag within content words (Table 5). Next, we break down the performance by part-of-speech tags. We focus on the part-of-speech tags that are semantically important, namely, nouns, proper nouns, verbs, adverbs, and adjectives.

Prediction performance can be seen as a proxy for how reliably a part-of-speech tag is reused when providing explanations. Consistent with our expectations for the importance of nouns and verbs, our models achieve the best performance on nouns within content words. Verbs are more challenging, but become the least difficult tag to predict when we consider all words, likely due to stopwords such as "have." Adjectives turn out to be the most challenging category, suggesting that adjectival choice is perhaps more arbitrary than other parts of speech, and therefore less central to the process of constructing an explanation. The important role of nouns in shaping explanations resonates with the high recall rate of nouns in memory tasks (Reynolds and Flagg, 1976).

	ROUGE-1	ROUGE-2	ROUGE-L
w/o features	18.91	4.12	17.05
with features	22.01	3.93	19.02

Table 6: ROUGE scores (F1) on the test dataset (Lin, 2004). The differences in ROUGE-1 and ROUGE-L are statistically significant with $p < 1e{-}10$.

5.3 The Effect on Generating Explanations

One way to measure the ultimate success of understanding pointers in explanations is to be able to generate explanations. We use the pointer generator network with coverage as our starting point (See et al., 2017; Klein et al., 2017) (see the supplementary material for details). We investigate whether concatenating our proposed features with word embeddings can improve generation performance, as measured by ROUGE scores.

Consistent with results in sequence tagging for word-level echoing prediction, our proposed features can enhance a neural model with copying mechanisms (see Table 6). Specifically, their use leads to statistically significant improvement in ROUGE-1 and ROUGE-L, while slightly hurting the performance in ROUGE-2 (the difference is not statistically significant). We also find that our features can increase the likelihood of copying: an average of 17.59 unique words get copied to the generated explanation with our features, compared to 14.17 unique words without our features. For comparison, target explanations have an average of 34.81 unique words. We emphasize that generating explanations is a very challenging task (evidenced by the low ROUGE scores and examples in the supplementary material), and that fully solving the generation task requires more work.

6 Concluding Discussions

In this work, we conduct the first large-scale empirical study of everyday explanations in the context of persuasion. We assemble a novel dataset and formulate a word-level prediction task to understand the selective nature of explanations. Our results suggest that the relation between an OP and PC plays an important role in predicting the echoing of content words, while a word's non-contextual properties matter for stopwords. We show that vanilla LSTMs fail to learn some of the features we develop and that our proposed features can improve the performance in generating explanations using pointer networks. We also demonstrate the important role of nouns in shaping explanations.

Although our approach strongly outperforms random baselines, the relatively low F1 scores indicate that predicting which word is echoed in explanations is a very challenging task. It follows that we are only able to derive a limited understanding of how people choose to echo words in explanations. The extent to which explanation construction is fundamentally random (Nisbett and Wilson, 1977), or whether there exist other unidentified patterns, is of course an open question. We hope that our study and the resources that we release encourage further work in understanding the pragmatics of explanations.

There are many promising research directions for future work in advancing the computational understanding of explanations. First, although `/r/ChangeMyView` has the useful property that its explanations are closely connected to its explananda, it is important to further investigate the extent to which our findings generalize beyond `/r/ChangeMyView` and Reddit and establish universal properties of explanations. Second, it is important to connect the words in explanations that we investigate here to the structure of explanations in pyschology (Lombrozo, 2006). Third, in addition to understanding what goes into an explanation, we need to understand what makes an explanation effective. A better understanding of explanations not only helps develop explainable AI, but also informs the process of collecting explanations that machine learning systems learn from (Hancock et al., 2018; Rajani et al., 2019; Camburu et al., 2018).

Acknowledgments

We thank Kimberley Buchan, anonymous reviewers, and members of the NLP+CSS research group at CU Boulder for their insightful comments and discussions; Jason Baumgartner for sharing the dataset that enabled this research.

References

Colin R Blyth. 1972. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Proceedings of NeurIPS*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of KDD*.

CMV moderators. 2019a. CMV media coverage. https://changemyview.net/subreddit/#media-coverage. [Online; accessed 27-Apr-2019].

CMV moderators. 2019b. The Delta System. https://www.reddit.com/r/changemyview/wiki/deltasystem. [Online; accessed 27-Apr-2019].

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: Linguistic style accommodation in social media. In *Proceedings of WWW*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of NAACL*.

Howard Giles and Tania Ogay. 2007. Communication accommodation theory. *Explaining communication: Contemporary theories and exemplars*, pages 293–310.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Proceedings of ACL*.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of EMNLP*.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher R. 2018. Training Classifiers with Natural Language Explanations. In *Proceedings of ACL*.

Virgina Heffernan. 2018. Our best hope for civil discourse online is on ... Reddit. *Wired*.

Carl G Hempel and Paul Oppenheim. 1948. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Benjamin D Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adali. 2019. Rating reliability and bias in news articles: Does ai assistance help everyone? In *Proceedings of ICWSM*.

Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL*.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of FAT**.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of KDD*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wang Ling, Phil Blunsom, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, Fumin Wang, and Andrew Senior. 2016. Latent predictor networks for code generation. In *Proceedings of ACL*, pages 599–609, Berlin, Germany. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of EACL*.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313–330.

Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Richard E Nisbett and Timothy D Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Goldberg Yoav. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of ACL (short papers)*.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(2):130–137.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of ACL (short papers)*, pages 38–42.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of ACL*.

Allan G Reynolds and Paul W Flagg. 1976. Recognition memory for elements of sentences. *Memory & Cognition*, 4(4):422–432.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of KDD*.

Wesley C Salmon. 2006. *Four decades of scientific explanation*. University of Pittsburgh press.

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

Matthew P Simmons, Lada A Adamic, and Eytan Adar. 2011. Memes online: Extracted, subtracted, injected, and recollected. In *Proceedings of ICWSM*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*, WWW '16, pages 613–624, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. Event-place: Montral, Qubec, Canada.

Chenhao Tan, Hao Peng, and Noah A. Smith. 2018. You are no Jack Kennedy: On media selection of highlights from presidential debates. In *Proceedings of WWW*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proceedings of NeurIPS*, pages 2692–2700.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of NAACL*.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*.

Robert A. Wilson and Frank Keil. 1998. The Shadows and Shallows of Explanation. *Minds and Machines*, 8(1):137–159.

James Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of NAACL*.