Summarizing Dialogic Arguments from Social Media

Amita Misra, Shereen Oraby, Shubhangi Tandon, Sharath TS,
Pranav Anand and Marilyn Walker
UC Santa Cruz
Natural Language and Dialogue Systems Lab
1156 N. High. SOE-3
Santa Cruz, California, 95064, USA
amisra2|soraby|shtandon|sturuvek|panand|mawalker@ucsc.edu

Abstract

Online argumentative dialog is a rich source of information on popular beliefs and opinions that could be useful to companies as well as governmental or public policy agencies. Compact, easy to read, summaries of these dialogues would thus be highly valuable. A priori, it is not even clear what form such a summary should take. Previous work on summarization has primarily focused on summarizing written texts, where the notion of an abstract of the text is well defined. We collect gold standard training data consisting of five human summaries for each of 161 dialogues on the topics of *Gay Marriage*, *Gun Control* and *Abortion*. We present several different computational models aimed at identifying segments of the dialogues whose content should be used for the summary, using linguistic features and Word2vec features with both SVMs and Bidirectional LSTMs. We show that we can identify the most important arguments by using the dialog context with a best F-measure of 0.74 for gun control, 0.71 for gay marriage, and 0.67 for abortion.

1 Introduction

Online argumentative dialog is a rich source of information on popular beliefs and opinions that could be useful to companies as well as governmental or public policy agencies. Compact, easy to read, summaries of these dialogues would thus be highly valuable. However, previous work on summarization has primarily focused on summarizing written texts, where the notion of an abstract of the text is well defined.

Work on dialog summarization is in its infancy. Early work was domain specific, for example focusing on extracting actions items from meetings (Murray, 2008). Gurevych and Strube (2004) applied semantic similarity to Switchboard dialog, showing improvements over several baseline summarizers. Work on argument summarization has to date focused on monologic data. Ranade et al. (2013) summarize online debates using topic and sentiment rich features, but their unit of summary is a single debate post, rather than an extended conversation. Wang and Ling (2016) generate abstractive one sentence summaries for opinionated arguments from debate websites using an attention-based neural network model, but the inputs are well-structured arguments and a central claim constructed by the editors, rather than user-generated conversations.

PostID	Turn
S1-1:	Gays..you wont let me have everything I want so you must hate me. Spoil child..you wont let me have everything I want so you must hate me.
S2-1:	And who made you master daddy that you think it is your place to grant or disallow anything to your fellow citizens?
S1-2:	Did I say that I was and it is?
S2-2:	You implied it when you compared gays (and their supporters) fighting for rights to spoiled children. For the analogy to work there has to be a parent figure for the gays as well.
S1-3:	The public is the 'parent' figure and the law makers are (or should be) the public's servant .
S2-3:	This then implies that homosexuals are are not part of the public and the law-makers are not their servants as well, and that you do indeed believe it is your right to allow and disallow things to your fellow citizens. That they are lesser group than you. You just proved your hate.
S1-4:	Homosexuals are a deviant minority.

Figure 1: Gay Rights Argument.

To our knowledge there is no prior work on summarizing important arguments from noisy, argumentative, dialogs in online debate such as that in Figure 1. A priori, it is not even clear what form

Summary Contributors	Human Label from Pyramid Annotations	Tier Rank
• S1 says that no one can prove that gun owners are safer than non gun owners. • S1 says no one has been able to prove gun owners are safer than non-gun owners. • S1 points out there is no empirical data suggesting that gun owners are safer than non-gun owners. • S1 states there are no statistics proving owning a gun makes people safer. • S1 believes that there is no proof that gun owners are safer than non-gun owners.	Nobody has been able to prove that gun owners are safer than non-gun owners.	5
• They say that if S2 had a family member die from gun violence it might be more significant to them, • He says if S1 had a personal or family encounter with gun violence, he would feel differently. • that people who have had relatives die from gun violence have a different attitude.	Family encounters with gun violence changes significance.	3
• Pro-gun perspective is: on 9/11, 3000 people died without the ability to defend themselves.	On 9/11, 3000 people died without the ability to defend themselves.	1

Table 1: Example summary contributors, pyramid labels and tier rank in gun control dialogs

such a summary should take. The two conversants in Figure 1 obviously do not agree: should a summary give preference to one person's views? Should a summary be based on decisions about which argument is higher quality, well structured, more logical, or which better follows theories of argumentation?

Fortunately, summarization is something that any native speaker can do without formal training. Thus our gold standard training data consists of 5 human summaries for each dialog from a corpus of dialogs discussing *Gay Marriage*, *Gun Control* and *Abortion*. Arguments that are important to extract to form the basis of summary content are defined to be those that appear in a majority of human summaries, as per the Pyramid model (Nenkova and Passonneau, 2004). We then aim to learn how to automatically extract these important arguments from the original dialogs.

We first define several baselines using off-the-shelf summarizers such as LexRank and SumBasic (Erkan and Radev, 2004a; Nenkova and Vanderwende, 2005). Our experiments explore the effectiveness of combining traditional linguistic features with Word2Vec in both SVMs and Bidirectional LSTMs. We show that applying coreference, and representing the context improves performance. Performance is overall better for the Bidirectional LSTM, but both models perform better when linguistic features and argumentative features are combined with word embeddings. We achieve a best F-measure of 0.74 for gun control, 0.71 for gay marriage, and 0.67 for abortion. We discuss related work in more detail in Section 3 when we can compare it with our approach.

2 Experimental Method

2.1 Data

Our corpus of dialogs and summaries focus on the topics *Gay Marriage*, *Gun Control* and *Abortion* from the the publicly available Internet Argument Corpus (IAC) (Abbott et al., 2016). We used the portion of the IAC containing posts from `http://4forums.com`. We use the debate forum metadata to extract dialog exchanges between pairs of authors with at least 3 turns per author, in order to represent 2 different perspectives on an issue. To get richer and more diverse data per topic containing multiple argumentative claims and propositions, we ensure that the corpus does not contain more than one dialog per topic between any particular pair of authors. The dataset contains 61 gay rights dialogues, 50 gun control dialogues and 50 abortion dialogues.

We adopt a three step process to identify useful sentences for extraction that we briefly summarize here.

- **S1:** Dialogs are read and summarized by 5 pre-qualified workers on Mechanical Turk. Since the dialogs vary in length and content we applied a limit that dialogs with a word count less than 750, must be summarized by the annotators in 125 words and dialogs with

In this task, you will carefully read part of a dialog where two people are discussing the issue of gun control. Several previous workers have each summarized this dialog, and we have related those summaries by grouping together parts of their summaries that roughly describe the same actions in the dialogue. In this task, you will link these action description groups to sentences in the dialogue. Each dialog is automatically divided into sentences. Your job is to provide the best action description group for each sentence.
The action description groups are sets of sentences from several summaries that essentially describe the same action in the dialog in different words. Each group has a unique label and you will select the label that best approximates what is happening in the sentence and select a label using the radio button provided with each sentence.
Please especially note:

- More than one sentence can map to same group. For example, two people may say virtually the same thing multiple times.
- Not all sentences will have a good group, so if you cannot find any similar set for a sentence, then select None of the labels match in the radio button option.
- You are expected to read and comprehend the sentence. Since these come from summaries, the action summaries may use very different words from those used in the dialogs.

Table 2: Directions for Step 3 (**S3** annotation, mapping pyramid labels to sentences.

word count greater than 750 words should be summarized in 175 words.

- **S2:** We train undergraduate linguists to use the Pyramid method (Nenkova and Passonneau, 2004) to identify important arguments in the dialog; they then construct pyramids for each set of five summaries. Repeated elements of the five summaries end up on higher tiers of the pyramid, and indicate the most important content, as shown in Table 1. This results in a ranking of the most important arguments (abstract objects) in a dialog, but the linguistic representation of these arguments is based on the language used in the summaries themselves.

- **S3:** To identify the spans of text in the dialog itself that correspond to the important arguments, we must map the ranked labels from the summaries back onto the dialog text. We recruited 2 graduate students and 2 undergraduates to label each sentence of the dialog with the best set of human labels from the pyramids. Table 2 shows the directions for this task.

We now have one or more labels for each sentence in a dialog, but we are primarily interested in the **tier rank** of the sentences. We group labels by tier and compute the average tier label per sentence. We define any sentence with an average tier score of 3 or higher as **important**. Thus, steps **S1**, **S2** and **S3** above are simply carried out to arrive at a well-motivated and theoretically grounded definition of **important** argument, and the task we address in this paper is binary classification applied to dialogs to select sentences that are important. Table 3 shows the resulting number of important sentences for each topic. The average Cohen's kappa between the annotators is respectable, with a kappa value of 0.68 for gun control, 0.63 for abortion, and 0.62 for gay marriage.

Topic	Important	Not Important
Gun Control	1010	1041
Gay Marriage	1311	1195
Abortion	849	1203

Table 3: Sentence distribution in each domain.

2.2 Baselines

We use several off-the-shelf extractive summarization engines (frequency, probability distribution and graph based) from the python package sumy [1] to provide a baseline for comparison with our models. To enable direct comparison, we define a sentence as **important** if it appears in the top n sentences in the output of the baseline summarizer, where n is the number of **important** sentences for the dialog as defined by our method.

SumBasic. Nenkova and Vanderwende (2005) show that content units and words that are repeated often are likely be mentioned in a human summary, and that frequency is a powerful predictor of human choices in content selection for summarization. SumBasic uses a greedy search approximation with a frequency-based sentence selection component, and a component to re-weight the word probabilities in order to minimize redundancy.

KL divergence Summary. This approach is based on finding a set of summary sentences

[1] https://pypi.python.org/pypi/sumy

which closely match the document set unigram distribution. It greedily adds a sentence to a summary as long as it decreases the KL Divergence (Haghighi and Vanderwende, 2009).

LexRank. This method is a degree-based method of computing centrality that is used for extractive summarization and has shown to outperform centroid-based methods on DUC evaluation tasks. It computes sentence importance based on eigenvector centrality in a graph where cosine similarity is used for sentence adjacency weights in the graph (Erkan and Radev, 2004a).

Summary Sentences selected by human annotators
Nobody has been able to prove that gun owners are safer than non-gun owners.
You can play around with numbers to make the problem seem insignificant.
I suppose you could also say that only 3,000 people died in 9/11 and use your logic to say that it 's only a small problem.
Perhaps if somebody in your family had died of gun violence you would have a different attitude.
Nobody has been able to prove that non-gun owners are safer than gun owners.
So if you can not prove things one way or the other why try to infringe on my rights?
I did n't say that it ca n't be proven one way or the other.
I just said you ca n't prove that gun owners are safer.
Using illogic , skewed statistics , revisionist history all in an attempt to violate my constitutional rights , that would be you and other gun grabbers who are trying to infringe on law abiding citizens rights.
Show me in the Constitution where it says that making an illogical argument is a violation of somebody 's rights.
You and your ilk are doing everything in your power to implement your " victim disamament " program in " violation " of my civil rights.
No different than " jim crow " laws and other unconstitutional drivel.

Figure 2: Human selected summary sentences for a gun control dialogue.

Figures 2 and 3 show our gold standard summary and the summary sentences selected by LexRank for the same dialog. LexRank identifies many of the important sentences, but it also includes a number of sentences which cannot be used to construct a summary such as *"Wow that is easy"*. The baseline outputs in general suggest that frequency or graph similarity alone leave room for improvement when predicting important sentences in user-generated argumentative dialogue.

Summary sentences selected by LexRank
Show me in the Constitution where it says that making an illogical argument is a violation of somebody 's rights.
Nobody has been able to prove that gun owners are safer than non-gun owners.
I just said you ca n't prove that gun owners are safer.
Wow that is easy.
At least have the courage to say it ... .
Witch hunt.
No different than " jim crow " laws and other unconstitutional drivel.
So if you can not prove things one way or the other why try to infringe on my rights?
Oh, stop your witch hunt.
You can play around with numbers to make the problem seem insignificant.
Using illogic, skewed statistics, revisionist history all in an attempt to violate my constitutional rights, that would be you and other gun grabbers who are trying to infringe on law abiding citizens rights.
I suppose you could also say that only 3,000 people died in 9/11 and use your logic to say that it 's only a small problem.

Figure 3: Lex Rank selected sentences for a gun control dialogue.

2.3 Features

Most formal models of argumentation have focused on carefully crafted debates or face-to-face exchanges. However, as the 'bottom-up' argumentative dialogs in online social networks are far less logical (Gabbriellini and Torroni, 2013; Toni and Torroni, 2012), and the serendipity of the interactions yields less rule-governed conversational turns, ones that violate even the rules of naturalistically grounded argument models (Walton and Krabbe, 1995). This makes it difficult to construct useful theoretically-grounded features. In place of that enterprise, we exploit more conventional summarization, sentiment, word class, and sentence complexity features.

We also construct features sensitive to dialogic context. The theoretical literature discusses the ways in which dialogic argumentation shows different speech act uses than in less argumentative genres (Budzynska and Reed, 2011; Jacobs and Jackson, 1992), including the fact that arguments in these conversations are frequently smuggled in via non-assertive speech acts (e.g., hostile questions). Inspired by this, we implement three basic methods for dialogic context: we extract the dialog act tag and some word class class information from the previous sentence; we extract a rough-grained measure of a sentence's position within a turn; and we use coreference chains to resolve

anaphora in a sentence to acquire a (hopefully) more contentful antecedent. Below, we describe these features in more detail.

Google Word2Vec: Word embeddings from word2vec (Mikolov et al., 2013) are popular for expressing semantic relationships between words. Previous work on argument mining has developed methods using word2vec that are effective for argument recognition (Habernal and Gurevych, 2015). We created a 300-dimensional vector by filtering stopwords and punctuation and then averaging the word embeddings from Google's word2vec model for the remaining words.

GloVe Embeddings: GloVe is an unsupervised algorithm for obtaining vector representations for words (Pennington et al., 2014). These pre-trained word embeddings are 100 dimensional vectors and each sentence is represented as a concatenation of word vectors. We use GloVe embeddings to initialize our Long Short-Term Memory (LSTM) models as glove embeddings have been trained on web data, and in some cases work better than Word2Vec (Stojanovski et al., 2016).

Readability Grades: We hypothesized that contentful sentences were more likely to be complex. To measure that, we used readability grades, which calculate a series of linear regression measures based on the number of words, syllables, and sentences. We used 7 readability measures[2] Flesch-Kincaid readability score, Automated Readability Index, Coleman-Liau Index, SMOG Index, Gunning Fog index, Flesch Reading Ease, LIX and RIX.

LIWC: The Linguistics Inquiry Word Count (LIWC) tool has been useful in previous work on stance detenction (Pennebaker et al., 2001; Somasundaran and Wiebe, 2009; Hasan and Ng, 2013), and we suspected it would help to distinguish personal conversation from substantive analysis. It classifies words into different categories based on thought processes, emotional states, intentions, and motivations. For each LIWC category, we computed an aggregate frequency score for a sentence. Using these categories we aim to capture both the style and the content types in the argument. Style words are linked to measures of people's social and psychological worlds while content words are generally nouns, and regular verbs that convey the content of a communication. To capture additional contextual information, we computed the LIWC score of the previous sentence.

Sentiment: Sentiment features have shown to be useful for argumentative claim identification, and here too we suspected that name-calling and the like could be flagged by sentiment features. We used the Stanford sentiment analyzer from (Socher et al., 2013) to compute five sentiment categories (very negative to very positive) per sentence.

Dialog Act of Previous Sentence (DAC): We hypothesized that **important** sentences may be more likely in response to particular dialog acts, like questions, e.g. a question may be followed by an explanation or an answer. To identify if a previous sentence was a question, we combined the tags into two categories indicating whether the previous sentence was a question type or not. We implemented a binary PreviousSentAct feature which used Dialog Act Classification from NLTK (Loper and Bird, 2002).

Sentence position: We divide a turn into thirds and create an integral feature based on which third a sentence is located in the turn.

Coref: In the hope that coreference resolution would help ground utterance semantics, we replaced anaphoric words with their most representative mention obtained using Stanford coreference chain resolution (Manning et al., 2014).

2.4 Machine Learning Models

We reserved 13 random dialogs in each topic for our test set, using the rest as training. Sentences were automatically split. This led to several sentences consisting essentially of punctuation, which were removed (filter for sentences without a verb and at least 3 dictionary words.) For learning, we created a balanced training and test set by randomly selecting an equal number of sentences for each class, giving the following combinations: 1236 train and 462 test sentences for abortion, 1578 training and 534 test for gay marriage and 1352 training and 476 test for gun control. We use two machine learning models.

SVM. We use Support Vector Machines with a linear Kernel from Scikit-learn (Pedregosa et al., 2011) with our theoretically motivated linguistic features and uses cross validation for parameter tuning and the second is a combination Bidirectional LSTM.

CNN + BiLSTM. A combination of Convolutional and Recurrent Neural Networks has been

[2] https://pypi.python.org/pypi/readability

used for sentence representations (Wang et al., 2016) where CNN is able to learn the local features from words or phrases in the text and the RNN learns long-term dependencies. Using this as a motivation, we include a convolutional layer and max pooling layer before the input is fed into an RNN. The model used for binary classification consists of a 1D convolution layer of size 3 and 32 different filters. The convolution layer takes as input the GloVe embeddings. A bidirectional LSTM layer is stacked on the convolutions layer and then concatenated with another layer of bidirectional LSTM: different versions are used with different features and feature combinations as shown in Table 4 and described further below. The outputs of the LSTM are fed through a sigmoid layer for binary classification. LSTM creates a validation set by a 4 to 1 random selection on the training set. Regularization is performed by using a drop-out rate of 0.2 in the drop-out layer. The model is optimized using the Adam (Kingma and Ba, 2014) optimizer. The deep network was implemented using the Keras package (Chollet, 2015).

2.5 Results

We use standard classification evaluation measures based on Precision/Recall and F measure. Performance evaluation uses weighted average F-score on test set. We first evaluate simple models based on a single feature.

Simple Ablation Models. Table 4, Rows 1A, 1B and 1C show the results for our three baseline systems. The LexRank summarizer performs best across all topics, but overall the results show that summarizers aimed at newswire or monologic data do not work on argumentative dialog.

Row 3 shows that Word2Vec improves over the baseline, but this did not work as well as it did in previous research (Habernal and Gurevych, 2015). One reason could be that averaged Word2Vec embeddings for each word lose too much information in long sentences. Row 2 shows that Dialog Act Classification works better than the random baseline for gun control and gay marriage but not for abortion. Interestingly, Row 6 shows that sentiment by itself beats LexRank across all topics, suggesting a relationship of sentiment to argument that could be further explored.

Each Row has an additional column for each topic indicating what happens when we first run Stanford Coreference to replacing each pronoun with its most representative mention. The results show that coreference improves the F-score for both gun control and abortion.

LIWC categories and Readability perform well across topics.

Feature Combination Models.
We first evaluate SVM with different feature combinations, with details on results in Table 4. For the gun control topic, LIWC categories on the current sentence give an F-score of 0.72. Adding LIWC from the previous sentence improves it to 0.73 (rows 5 and 9, without coref column). In contrast, just doing a coref replacement improves LIWC current sentence score to 0.74 (row 5 for gun control, with and without coref columns). A paired t-test on the result vectors shows that coref replacement provides a statistically significant improvement at ($\mathbf{p < 0.04}$). For the Abortion topic, the overall performance is low as compared to the other two topics suggesting that arguments used for abortion are harder to identify. Both DAC, Word2vec scores are quite low but readability and LIWC do better.
The LSTM models on their own do not perform better than SVM across topics, but adding features to the LSTM models improves them beyond the SVM results. We paired only LSTM (row 8) separately with the best performing model in bold for each topic in Table 4 to evaluate if the combination is significant. Paired t-tests on the result vectors show that the differences in F-score are statistically significant when we compare LSTM to LSTM with features for each topic ($\mathbf{p < 0.01}$) for all topics, indicating that adding contextual features makes a significant improvement. Adding LIWC categories from current and previous utterances to LSTM also improves performance for gun control and abortion. For the gay marriage topic, LSTM combined with LIWC and readability works better than LSTM alone.

2.6 Analysis and Discussion

To qualitatively gain some insight into the limitations of some of the systems, we examined random predictions from different models. One reason that a Graph-based system such as LexRank performs well on DUC might ne that DUC data sets are clustered into related documents by human assessors. To observe the behavior of the method on noisy data, the authors of LexRank added random documents to each cluster to show that LexRank is

			Gun Control		Gay Marriage		Abortion	
ID	Classifier	**Features**	**F-weight Avg.**	**F-weight Avg. Coref**	**F-weight Avg.**	**F-weight Avg. Coref**	**F-weight Avg.**	**F-weight Avg. Coref**
1A	Baseline	KL-SUM (**KL**)	0.51		0.52		0.47	
1B	Baseline	SumBasic (**SB**)	0.53		0.57		0.49	
1C	Baseline	Lex-Rank (**LR**)	0.58		0.58		0.59	
2	SVM	Dialog Act (**DAC**)	0.61	0.60	0.58	0.58	0.42	0.41
3	SVM	Word2Vec	0.65	0.65	0.63	0.56	0.58	0.58
4	SVM	Readability (**R**)	0.64	0.67	0.68	0.68	0.63	0.64
5	SVM	LIWC current sentence (**LC**)	0.72	**0.74**	**0.69**	0.66	0.64	0.63
6	SVM	Sentiment (**SNT**)	0.66		0.62		0.61	
7	SVM	Sentence Turn (**ST**)	0.61	0.61	0.40	0.40	0.33	0.33
8	Bi LSTM		0.68	0.69	0.63	0.58	0.64	**0.65**
Feature Combinations								
9	SVM	LIWC current + previous (**LCP**)	0.73	0.72	0.66	0.67	0.61	0.61
10	SVM	LCP + R	0.73	0.73	0.70	0.68	0.61	0.60
11	SVM	R+DAC	0.65	0.66	0.68	0.68	0.63	0.63
12	SVM	LCP + DAC + R	0.72	0.73	0.69	0.68	0.61	0.61
13	Bi LSTM	DAC	0.67	0.68	0.69	0.65	0.65	0.66
14	Bi LSTM	ST	0.66	0.66	0.61	0.67	0.64	0.52
15	Bi LSTM	LCP	0.70	0.68	0.52	0.52	0.65	**0.67**
16	Bi LSTM	R	0.70	0.70	0.59	0.63	0.65	0.66
17	Bi-LSTM	LCP+ DAC	0.70	0.71	0.69	0.68	0.61	0.62
18	Bi-LSTM	R+ DAC	0.70	0.68	0.63	0.62	0.60	0.64
19	Bi-LSTM	R+ LCP	0.69	0.68	**0.71**	0.67	0.64	0.66
20	Bi-LSTM	LCP+R +DAC	0.73	**0.74**	0.70	0.69	0.62	0.63

Table 4: Results for classification on test set for each topic. Best performing model in **bold.**

insensitive to some limited noise in the data. However, topic changes are more frequent in dialog and dialogs contain content that is not necessarily related to the argumentative purpose of the dialog.

For example, lexical overlap is important to LexRank, but this resulted in LexRank selecting the two of these sentences *Well it's not going to work.* and *Get to work!*.

One reason that SVM with sentiment features performs well is that positive sentiment predicts the not-important class. It seems that sentiment analyzers classify both phatic communication and sarcastic arguments as positive, both of which can be correctly assigned to the not-important class, as shown by the following examples:

- I 'll be nice ... Out of context sermon.
- You 're a fine one to talk about sliming folks
- Yes it does
- Sounds right to you?

The results show that LIWC performs well and that LIWC used to represent context performs even better. To understand which LIWC features were important, we performed chi-square feature selection over LIWC features on the training set. Content categories were highly ranked across topics, suggesting that the LIWC features are being exploited for a form of within-topic topic detection; this suggests that more general topic modeling could help results.

Table 5 shows the top 5 LIWC categories for each topic based on chi-square based feature selection on the training set for all the three topics. Unsurprisingly, across all topics, the LIWC marker of complexity (Words Per Sentence) appears. In addition, many other topics link commonsense with important facets of these debates – the opposition in abortion between questions of the sanctity of life (biological processes), health of individuals involved. Similarly, with Gay Marriage, we see sides of the debate between personal relationships (family, affiliation) and questions of sexual practice (sexual, drives). The case of Gun Control is somewhat surprising, since one might expect to see LIWC categories relating to life and safety. Instead we see Money category coming from discussions about gun buy back and gun prices. To understand better why coreference resolution was helping, we also examined cases where coreference matters. Coreference resolution can also interact with different features such

as LIWC, i.e. since LIWC calculates a frequency distribution of categories in the text, corefence moves a word from the pronoun to some other category. For example, replacing *it* by *Government* decreases Impersonal Pronouns and Total Pronouns, while increasing Six Letter Words. In several cases these replacements produce correct predictions, e.g. with
Only if it is legal to sell it.

Topic	LIWC Categories
Abortion	*Biological Processes, Health, Second Person, Sexual, Words Per Sentence,*
Gun Control	*First Person Singular, Money, Second Person, Third Person Plural, Words Per Sentence*
Gay Marriage	*Family, Sexual, Words Per Sentence, Affiliation, Drives*

Table 5: Top 5 LIWC categories by chi-square for each topic

3 Related Work

This work builds on multiple strands of research into dialog, summarization and argumentation.

Dialog Summarization. To the best of our knowledge, none of the previous approaches have focused on debate dialog summarization. Prior research on spoken dialog summarization has explored lexical features, and information specific to meetings such as action items, speaker status, and structural discourse features. (Zechner, 2001; Murray et al., 2006; Whittaker et al., 2012; Janin et al., 2004; Carletta, 2007). In contrast to information content, Roman et al. (2006) examine how social phenomena such as politeness level affect summarization. Emotional information has also been observed in summaries of professional chats discussing technology (Zhou and Hovy, 2005). Other approaches use semantic similarity metrics to identify the most central or important utterances of a spoken dialog using Switchboard corpus (Gurevych and Strube, 2004). Dialog structure and prosodic features have been studied for finding patterns of importance and opinion summarization on Switchboard conversations (Wang and Liu, 2011; Ward and Richart-Ruiz, 2013). Additional parallel work is on summarizing email thread conversations using conversational features and dialog acts specific to the email domain (Murray, 2008; Oya and Carenini, 2014).

Summarization. Document summarization is a mature area of NLP, and hence spans a vast range of approaches. The graph and clustering based systems compute sentence importance based on inter and intra-document sentence similarities (Mihalcea and Tarau, 2004; Erkan and Radev, 2004a; Ganesan et al., 2010). (Carbonell and Goldstein, 1998) use a greedy approach based on Maximal Marginal Relevance. (McDonald, 2007) reformulated this as a dynamic programming problem providing a knapsack based solution. The submodular approach by (Lin and Bilmes, 2011) produces a summary by maximizing an objective function that includes coverage and diversity.

Recently there has been a surge in data-driven approaches to summarization based on neural networks and continuous sentence features. An encoder decoder architecture is the main framework used in these types of models. However, one major bottleneck to applying neural network models to extractive summarization is that the generation systems need a huge amount of training data i.e., documents with sentences labeled as summary-worthy. (Nallapati et al., 2016; Rush et al., 2015; See et al., 2017) used models trained on the annotated version of the Gigaword corpus and paired the first sentence of each article with its headline to form sentence-summary pairs. Such newswire models did not work well here; the neural summarization model from OpenNMT framework (Klein et al., 2017) very often generated <UNK >tokens for our data. (Iyer et al., 2016) train an end to end neural attention model using LSTMs to summarize source code from online programming websites. Pairing the post title with the source code snippet from accepted answers gives a large amount of training data that can be used to generate summaries.

Our approach is similar in spirit to (Li et al., 2016). In this work, RST elementary discourse units (EDU's) are used as SCU's for extractive summarization of news articles. However, we observed in debate dialogs, that the same argumentative text can be used by interlocutors on opposite sides of an issue, and hence could not be considered in isolation as a summary unit. Barker et al. (2016) describe a corpus of original Guardian articles along with associated content (comments, groups, summaries and backlinks). However, the comment data is different from conversational dialogic debates (it is less strongly threaded, less directly dialogic, and less argumentative) and they

do not present a computational model for argument summary generation. Misra et al. (2015) use pyramid annotation of dialog summaries on online debates to derive SCUs and labels, but they go on to work with the **human-generated labels** of the pyramid annotation. Our task, using raw sentences from social media dialogs, is appreciably harder.

Argumentation. Argumentative dialog is a highly challenging task with creative, analytical and practical abilities needed to persuade or convince another person, but what constitutes a "good argument" is still an open ended question (Jackson and Jacobs, 1980; Toulmin, 1958; Sternberg, 2008; Walton et al., 2008). The real world arguments found in social media dialog are informal, unstructured and so the well established argument theories may not be a good predictor of people's choice of arguments (Habernal et al., 2014; Rosenfeld and Kraus, 2016). In this work, we propose pyramid based summarization to rank and select arguments in social media dialog, which to the best of our knowledge is a novel method for ranking arguments in conversational data.

4 Conclusion and Future Work

We presented a novel method for argument summarization of dialog exchanges from social media debates with our results significantly beating the traditional summarization baselines. We show that adding context based features improves argument summarization. Since we could find both topic specific and topic independent features, we plan to explore unsupervised topic modeling that could be used to create a larger and more diverse dataset and build sequential models that could generalize well across a vast range of topics.

Acknowledgments

This work was supported by NSF CISE RI 1302668. Thanks to the three anonymous reviewers for helpful comments.

References

Robert Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proc. of the LREC2016*.

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtic, Mark Hepple, and Robert J. Gaizauskas. 2016. The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proc. of the SIGDIAL 2016* .

Katarzyna Budzynska and Chris Reed. 2011. Speech acts of argumentation: Inference anchors and peripheral cues in dialogue. In *in Proc. of the 10th AAAI Conference on Computational Models of Natural Argument*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In *Proc. of the LREC 2007*.

François Chollet. 2015. Keras.

Günes Erkan and Dragomir R Radev. 2004a. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

S Gabbriellini and P Torroni. 2013. Ms dialogues: Persuading and getting persuaded. a model of social network debates that reconciles arguments and trust. In *Proc. of the 10th ArgMAS 2013*.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proc. of the 23rd COLING 2010*.

I. Gurevych and M. Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proc. of the 20th ACL 2004*.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proc. of the 2015 EMNLP*.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Proc. of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proc. of HLT:NAACL 2009*.

Kazi Saidul Hasan and Vincent Ng. 2013. Frame semantics for stance classification. In *Proc. of the CoNLL* 2013.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2016. Summarizing source code using a neural attention model. In *Proc. of the 54th Annual Meeting of the ACL 2016*.

Sally Jackson and Scott Jacobs. 1980. Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*.

Scott Jacobs and Sally Jackson. 1992. Relevance and digressions in argumentative discussion: A pragmatic approach. *Argumentation*.

A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, et al. 2004. The icsi meeting project: Resources and research. In *Proc. of the 2004 ICASSP NIST Meeting Recognition Workshop*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proc. of the 3rd ICLR 2014*.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proc. of the SIGDIAL 2016*.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proc. of the 49th Annual Meeting of the ACL:HLT 2011*.

E. Loper and S. Bird. 2002. NLTK: The natural language toolkit. In *Proc. of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of 52nd ACL: System Demonstrations 2014*.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proc. of the 29th European Conference on IR Research*, ECIR'07. Springer-Verlag.

R. Mihalcea and P. Tarau. 2004. TextRank: Bringing order into texts. In *Proc. of 2004 Conference on EMNLP*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 2013*.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in dialog. In *Proc. of the 2015 NAACL:HLT*.

G. Murray, S. Renals, J. Carletta, and J. Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proc. of the main conference on HLT of the NAACL*.

Gabriel Murray. 2008. Summarizing spoken and written conversations. In *in Proc. of the EMNLP 2008*.

Ramesh Nallapati, Bowen Zhou, and Bowen Zhou. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proc. of the CoNLL 2016*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proc. of the Joint Annual Meeting of HLT/NAACL*.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*.

Tatsuro Oya and Giuseppe Carenini. 2014. Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In *Proc. of the 15th Annual Meeting of SIGDIAL*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

James W. Pennebaker, L. E. Francis, and R. J. Booth, 2001. *LIWC: Linguistic Inquiry and Word Count*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. of the 2014 Conference on EMNLP*.

Sarvesh Ranade, Jayant Gupta, Vasudeva Varma, and Radhika Mamidi. 2013. Online debate summarization using topic directed sentiment analysis. In *Proc. of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, ACM.

N. Roman, P. Piwek, P. Carvalho, and M. B. R. Ariadne. 2006. Politeness and bias in dialogue summarization: two exploratory studies. In J. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing attitude and affect in text: theory and applications*, volume 20 of *The Information Retrieval Series*. Springer.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Trans. Interact. Intell. Syst.*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proc. of the EMNLP 2015*.

Abigail See, Christopher Manning, and Peter Liu. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of the ACL 2017*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of the 2013 Conference on EMNLP*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proc. of the 47th Annual Meeting of the ACL*.

R. Sternberg. 2008. *Cognitive Psychology*. Cengage Learning.

Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski. 2016. Finki at semeval-2016 task 4: Deep learning architecture for twitter sentiment analysis. In *SemEval@ NAACL-HLT*.

Francesca Toni and Paolo Torroni. 2012. Bottom-up argumentation. In *Proc. of the First International Conference on Theory and Applications of Formal Argumentation*. Springer-Verlag.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

D. Walton and E. Krabbe. 1995. *Commitment in Dialogue: Basic concept of interpersonal reasoning*. State University of New York Press.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proc. of the HLT-NAACL*.

Dong Wang and Yang Liu. 2011. A pilot study of opinion summarization in conversations. In *Proc. of the 49th Annual Meeting of the ACL: HLT 2011-Volume 1*

Xingyou Wang, Weijie Jiang, and Zhiyong Luo. 2016. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proc. of the COLING 2016*.

Nigel G Ward and Karen A Richart-Ruiz. 2013. Patterns of importance variation in spoken dialog.

S. Whittaker, V. Kalnikaité, and P. Ehlen. 2012. Markup as you talk: establishing effective memory cues while still contributing to a meeting. In *Proc. of the ACM 2012 conference on Computer Supported Cooperative Work*.

Klause Zechner. 2001. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*.

L. Zhou and E. Hovy. 2005. Digesting virtual geek culture: The summarization of technical internet relay chats. In *Proc. of the 43rd Annual Meeting on ACL*.