Characterizing Online Discussion Using Coarse Discourse Sequences

Amy X. Zhang
MIT CSAIL
Cambridge, MA, USA
axz@mit.edu

Bryan Culbertson
Google
Mountain View, CA, USA
bryan.culbertson@gmail.com

Praveen Paritosh
Google
Mountain View, CA, USA
pkp@google.com

Abstract

In this work, we present a novel method for classifying comments in online discussions into a set of coarse discourse acts towards the goal of better understanding discussions at scale. To facilitate this study, we devise a categorization of coarse discourse acts designed to encompass general online discussion and allow for easy annotation by crowd workers. We collect and release a corpus of over 9,000 threads comprising over 100,000 comments manually annotated via paid crowdsourcing with discourse acts and randomly sampled from the site Reddit. Using our corpus, we demonstrate how the analysis of discourse acts can characterize different types of discussions, including discourse sequences such as Q&A pairs and chains of disagreement, as well as different communities. Finally, we conduct experiments to predict discourse acts using our corpus, finding that structured prediction models such as conditional random fields can achieve an F1 score of 75%. We also demonstrate how the broadening of discourse acts from simply question and answer to a richer set of categories can improve the recall performance of Q&A extraction.

Introduction

As more social interaction takes place online, researchers have become interested in studying the discourse occurring in online social media. From these studies, researchers can examine how people conduct conversations and arguments (Hasan and Ng 2014; Tan et al. 2016) as well as extract information for applications such as search (Cong et al. 2008). While many studies have focused their analyses on metadata surrounding community discussions, other studies have attempted to analyze the textual *content* of discussions. But this can be difficult as language and interactions are complex and variable from discussion to discussion and community to community.

One method for understanding discussion is through analyzing the high level discourse structures inherent within conversations. Much research has demonstrated the power of using *discourse acts*, also known as speech acts, which are categories of utterances that pertain to their role in the discussion (e.g. "question" or "answer"). Researchers have used discourse acts towards applications such as building conversational bots (Allen, Ferguson, and Stent 2001) and summarizing spoken discourse (Murray et al. 2006). However, a great deal of research using discourse acts has focused solely on extracting questions and answers (Hong and Davison 2009) or considered only communities for help or technical support (Kim, Wang, and Baldwin 2010).

In this work, we develop a richer categorization of discourse acts towards characterizing a wide range of discussions from a variety of communities. Our 9 discourse act categories, developed over many iterations with experts and with the crowd, is designed to cover general discourse and be simple enough for crowd annotators to classify. We use as our source of data discussions from the website Reddit, one of the top ten most visited sites in the U.S, according to `Alexa.com`. By sampling over 9,000 discussion threads from the entirety of Reddit, which is comprised of thousands of "subreddits", we can gain information on discourse from many kinds of communities. From these threads, which are made of chains of comments replying to one another, we use a crowd system to annotate each comment within a thread with its discourse act as well as its discourse relation, or the comment to which it is responding. We are releasing this dataset[1], which is to our knowledge the largest manually annotated dataset of discourse acts in online discussions.

From an analysis of the major discourse acts and sequences within our corpus, we uncover patterns of discourse that correspond to well-known blocks of interactions, such as arguments and Q&A. This allows us to identify subreddits that behave much like community Q&A (CQA) sites like Quora, or that are more argumentative in nature.

Through building supervised models for classifying discourse acts, we find that structured prediction models such as conditional random fields (CRFs) achieve the greatest performance, with a 75% average F1 score on our dataset. We also analyze how well our best model classifies question and answer comments compared with models that contain fewer discourse acts. We find that our model with 9 discourse acts has overall better recall and slightly better F1 scores compared with models that only label question and answer comments. This suggests that having an enriched understanding of discourse structure beyond simply questions and answers can improve Q&A extraction.

[1] `https://github.com/dmorr-google/coarse-discourse`

Related Work

Discourse Act Classification Prior work has sought to develop a categorization of discourse acts for the purpose of characterizing discussion. Some early work focused only on conversational speech (Austin 1975; Searle 1969). Since then, researchers have developed standard taxonomies of spoken discourse acts such as DAMSL (Stolcke et al. 2000) and DiAML (Bunt et al. 2010). However, many of these discourse acts for spoken discourse do not translate to online asynchronous mediums. When it comes to online discussion, researchers have developed categories for discussions within e-mail (Cohen, Carvalho, and Mitchell 2004), online classrooms (Feng et al. 2006), newsgroups (Xi, Lind, and Brill 2004), and help forums (Kim, Wang, and Baldwin 2010). Much of this work on taxonomy development informs the final categories that we use. However, we develop a novel categorization that can be applied broadly in unstructured online forums of any topic or function.

Techniques for developing categories in prior work usually involve manual inspection and refinement by a knowledgeable annotator, such as one of the researchers. From there, the annotated dataset is used to build supervised (Kim, Wang, and Baldwin 2010) or semi-supervised (Jeong, Lin, and Lee 2009) models for predicting categories. Some research has attempted to learn categories using an unsupervised approach (Ritter, Cherry, and Dolan 2010). In our work, we chose to use manual annotation of categories from a set of acts refined by the authors. However, since our dataset is an order of magnitude larger than any prior manually annotated dataset, we turn to crowd workers to conduct the annotation. While prior work has had as many as 40 categories, we are limited in the number of categories as well as the level of detail we ask each annotator to provide.

Argumentation and Online Education Another line of work that is relevant to ours is the study of back-and-forth argumentation. Researchers have mined arguments online to learn how people take stances (Hasan and Ng 2014) or have developed systems for structured arguing (Klein 2011). Some of our work has overlap with work identifying argumentation, such as the classification of agreements and disagreements. This allows us to characterize communities by their proportion of and average length of arguments.

The work mentioned above is separate from systems that study argumentation within a single piece of text, such as within an opinion article or legal statement. Most of these have annotations at the sentence level. Classifications in this area include Rhetorical Structure Theory (RST) (Mann and Thompson 1988) and the Claim-Premises scheme (Freeman 1991). Because we analyzed text at the comment level as opposed to sentence level, we did not apply these classifications directly, though some categories have overlap.

Discourse acts have also been studied in the context of education (De Wever et al. 2006; Scheuer et al. 2010). Much of this work overlaps with the research on argumentation, as educators seek to understand and identify productive argumentation within the classroom. Researchers have developed novel categorizations to find evidence of critical thinking (Jeong 2003) and have also looked at discourse *sequences*, including adjacency pairs and chains (Lu, Chiu, and Law 2011; Rosé et al. 2008). Like the ARGUNAUT system (McLaren, Scheuer, and Mikšátko 2010), we take a closer look at identifying and understanding "chains of opposition" as well as popular discussion pairs such as "question-answer". However, many of our categorizations are different as we seek to characterize general discussion, while these works focus on classroom discussions and have categories more similar to argumentation systems.

Discourse Acts in Online Discussion In recent years, researchers have become interested in extracting useful information from online discussion. However, many analyses only focus on a particular community (Tan et al. 2016). Additionally, there has been little work analyzing online communities through the lens of high level discourse acts. Research in this area has focused on extraction of Q&A content from online forums (Cong et al. 2008; Hong and Davison 2009) or characterizing the types and quantity of Q&A content on different community platforms (Agichtein et al. 2008; Morris, Teevan, and Panovich 2010). Other research expands beyond Q&A but still focus on areas such as technical help forums (Kim, Wang, and Baldwin 2010). Instead, we characterize a wide range of online communities using a richer classification of discourse acts.

Discourse Act Annotation

Discourse Acts

We developed a set of 9 discourse act categories using a manual iterative process with experts coupled with pilots using the crowd. While there has been prior work on developing discourse acts for online forums, many do not fit our purposes because they are too detailed or too narrow in scope (Kim, Wang, and Baldwin 2010). Also, most have not released their annotated data or details of their coding scheme. In the end, our set of acts most closely resembles efforts such as (Feng et al. 2006), (Fortuna, Rodrigues, and Milic-Frayling 2007), and (Xi, Lind, and Brill 2004).

To build the discourse act categories, the first author randomly sampled threads from Reddit and, using prior work as a guide, classified comments into categories in an iterative process. After achieving a stable set of categories from multiple iterations, the authors then ran three pilots with crowd workers on datasets of 40 threads also randomly sampled, and iterated based on the inter-rater reliability returned. We also solicited qualitative feedback from the crowd workers, who were the same people throughout the annotation process. Some categories were eventually discarded due to too much overlap with other categories (ANECDOTE, FYI), or too low volume (SUMMARY, RESOLUTION).

Discourse Relations

As discourse acts are usually understood in relation to another piece of discourse, we collected both the discourse act of a comment as well as the discourse *relation* of that comment, also known as a link to a prior comment that the comment is responding to, if it exists. For instance, an ANSWER is always related to a prior QUESTION. Some categories may

not always be in relation to another comment, such as a new QUESTION or an ANNOUNCEMENT. In some categorizations of discourse, such as RST (Mann and Thompson 1988), there are *only* discourse relations, and the relations themselves are grouped into categories and named. In our case, we do not name types of discourse relations explicitly, but they are implicitly inferred by the discourse acts they link. For instance, a hypothetical discourse relation "Answers" would always link ANSWER to QUESTION.

Discourse Act Definitions

Detailed information about each discourse act and the relations allowed are given below. For our annotators, we provided a lengthier tutorial and several examples for each act, which we will release with our dataset.

QUESTION: A comment with a question or a request seeking some form of feedback, help, or other kinds of responses. While the comment may contain a question mark, it is not required. For instance, it might be posed in the form of a statement but still soliciting a response. Also, not everything that has a question mark is automatically a QUESTION. For instance, rhetorical questions are not seeking a response.

Relation: This comment might be the first in a thread and have no relation to another comment. Or, it could be a clarifying or follow-up QUESTION linking to any prior comment.

ANSWER: A comment that is responding to a QUESTION by answering the question or fulfilling the request. There can be more than one ANSWER responding to a QUESTION.

Relation: An ANSWER is always linked to a QUESTION.

ANNOUNCEMENT: A comment that is presenting some new information to the community, such as a piece of news, a link to something, a story, an opinion, a review, or insight.

Relation: This comment has no relation to a prior comment and is always the initial post in a thread.

AGREEMENT: A comment that is expressing agreement with some information presented in a prior comment. It can be agreeing with a point made, providing supporting evidence, providing a positive example or experience, or confirming or acknowledging a point made.

Relation: This comment is always linked to a prior comment to which it is agreeing.

APPRECIATION: A comment that is expressing thanks, appreciation, excitement, or praise in response to another comment. In contrast to AGREEMENT, it is not evaluating the merits of the points brought up. Comments of this category are more interpersonal as opposed to informational.

Relation: This comment is always linked to a prior comment for which it is expressing appreciation.

DISAGREEMENT: A comment that is correcting, criticizing, contradicting, or objecting to a point made in a prior comment. It can also be providing evidence to support its disagreement, such as an example or contrary anecdote.

Relation: This comment is always linked to a prior comment to which it is disagreeing.

NEGATIVE REACTION: A comment that is expressing a negative reaction to a previous comment, such as attacking or mocking the commenter, or expressing emotions like disgust, derision, or anger, to the contents of the prior comment. This comment is not discussing the merits of the points made in a prior comment or trying to correct them.

Relation: This comment is always linked to a prior comment to which it is negatively reacting.

ELABORATION: A comment that is adding additional information on to another comment. Oftentimes, one can imagine it simply appended to the end of the comment it elaborates on. One can elaborate on many kinds of comments, for instance, a question-asker elaborating on their question to provide more context, or someone elaborating on an answer to add more information.

Relation: This comment is always linked to a prior comment upon which it is elaborating.

HUMOR: This comment is primarily a joke, a piece of sarcasm, or a pun intended to get a laugh or be silly but not trying to add information. If a comment is sarcastic but using sarcasm to make a point or provide feedback, then it may belong in a different category.

Relation: At times, this comment links to another comment but other times it may not be responding to anything.

Data Collection

Sampling Reddit Threads

We randomly sampled from the full Reddit dataset starting from its inception to the end of May 2016, which is made available publicly as a dump on Google BigQuery[2]. We chose to sample from the entire dataset as opposed to a set of subreddits to ensure a wide variety of communities within our dataset. The full dataset of Reddit from this time period contains 238 million threads. However, we performed several filters on the data before sampling as we were interested in collecting substantial back-and-forth discussion.

Minimum Replies: As our goal is to better understand discussion, we chose to only take threads that had at least two reply comments to the initial post so that there was some amount of back-and-forth. Disqualifying these threads decreased the dataset to 87.5 million threads. We took a random sample of 50,000 threads from this dataset, and on this smaller set, we performed the following additional filters.

Deleted Comments: We disqualified any threads that contained a deleted comment or deleted portions of the initial post, as it would be difficult to interpret replies to deleted comments.

Non-English: As our annotators were English-speaking, we ignored any threads coming from subreddits primarily in a different language. We manually went through the most frequent several hundred subreddits in our dataset and added them to a blacklist if their homepage was primarily in a different language. Annotators were also instructed to skip any threads that were in a different language.

NSFW: In order to not subject our annotators to pornography, we additionally blacklisted 693 subreddits labeled Not Safe For Work (NSFW) by a third-party subreddit categorization site[3] that is community-sourced. This does not in-

[2] https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts

[3] http://redditlist.com/nsfw

clude subreddits that *discuss* potentially illegal or explicit content, which are still included in our dataset.

Trading: We also wished to avoid subreddits that were primarily for trading or coordination, mostly in the context of gaming, because these subreddits have little to no actual discussion. Examples include `/r/friendsafari` or `/r/fireteams`. We developed manual rules, such as if the subreddit name ends with the word "swap" or "trade", as well as manually curated a short blacklist.

After conducting filtering, we had 32,728 threads or 65% of our random sample. We chose to sample *link-post* threads, or threads where the body of the post is a link to a picture, video, or webpage, at 10% of our sample, leaving 90% of our sample to be *self-post* threads, or threads where the body of the post is a piece of text written to the community. This was so that we could collect a higher proportion of Q&A-related threads, since this data is particularly valuable in a search and information retrieval context.

In our filtered dataset, we had 10,145 self-post threads (31%) and 22,583 link-post threads (69%). From our filtered dataset, we sampled 9,000 self-posts threads and 1,000 link-post threads. During annotation, some threads were discarded due to bugs that occurred so that in the end, 9,701 threads were fully annotated.

Annotation

Annotation of discourse acts was conducted by crowd workers contracted from a paid crowdsourcing platform. In total, 25 annotators were hired, and they were paid an hourly rate above federal and state minimum wage. They were required to be native English speakers. To divvy up the work, each thread comprised a task, and users were given batches of 40 tasks at a time. For each task, annotators were asked to mark the discourse act of each of the comments within the page as well as the relation of each comment to a prior comment, if it existed. Each comment was annotated by three annotators.

As comments sometimes perform multiple functions, we instructed crowd annotators to consider the content at the comment level as opposed to sentence or paragraph level to make the task simpler. We did allow annotators to add a second discourse category to a comment when it was doing two separate actions in series, such as answering a question and then asking a new question. Secondary categories were annotated infrequently in our dataset (less than 3% of annotations), so we ignore them going forward in our analyses. However, they are available in the released dataset. Comments may also sometimes be responding to multiple other comments, such as thanking multiple ANSWER comments at once. In these cases, we asked annotators to only annotate one relation to the closest comment in terms of thread distance that they were responding to. Finally, we allowed annotators to annotate a comment as OTHER if they could not place it into any of the categories; this was the majority category in 1.8% of comments.

Before workers began annotating, they were presented with an instruction manual that explained each category and showed examples of annotated discussions. They were also given a few warm-up threads to do as practice before beginning annotation. The annotation was done using a Chrome browser extension that allowed annotators to easily click-to-highlight comments, annotate and link comments, and cycle through the different threads in their task.

Category	Krippendorf's Alpha
All Categories	**0.645**
Question	0.823
Answer	0.785
Announcement	0.732
Appreciation	0.611
Agreement	0.426
Elaboration	0.401
Disagreement	0.383
Humor	0.343
Negative Reaction	0.330

Table 1: Inter-rater reliability of the different discourse acts.

Finally, some threads on Reddit have hundreds of comments or more. As this is too much work for an annotator to perform in one sitting, we limited the thread length to 40 replies. This was done using Reddit's default "best" sorting[4] by appending `?limit=40` to every URL in our dataset. These threads represented less than 1% of our dataset.

Annotator Agreement

In Table 1, we present the inter-rater reliability of each discourse act using Krippendorf's Alpha. As can be seen, some acts had more agreement between annotators than others. The least reliable categories were NEGATIVE REACTION and HUMOR. From analyzing comments where annotators disagreed, we noticed examples where, in the case of HUMOR, a comment was being sarcastic or silly but it was not obvious without knowing the context. We also noticed that several categories had some overlap with ELABORATION, such as AGREEMENT and DISAGREEMENT. This was perhaps because annotators did not agree on the degree to which a comment was primarily agreeing or disagreeing with a prior comment versus more neutrally elaborating on a prior comment with more information.

As mentioned, we also asked annotators to link each comment to a prior comment to which it was in relation, if such a comment existed. From analyzing the relation annotations, we found that 98.9% of comments had a majority link between the three annotations. Of those comments, the average percent agreement with the majority link was 95.6%. Thus, while some categories had lower agreement, the relation annotations were almost entirely in agreement.

In the end, for the rest of our analyses, we consider the comments that had a majority category that was not OTHER across the three annotators. After removing comments without a majority category, this resulted in 9,131 threads with 101,525 comments (87.5% of the original number of comments), posted from 2,837 communities by 61,174 unique author accounts. While we do not analyze the comments without a majority category, we are releasing individual annotations in our public dataset as potential future work.

[4] `https://redditblog.com/2009/10/15/reddits-new-comment-sorting-system/`

Discourse Act	% Self-Post	% Link-Post	Total %	Total Count
Answer	42.3%	15.8%	41.5%	41658
Elaboration	18.1%	26.1%	18.8%	18927
Question	17.5%	15.2%	17.6%	17681
Appreciation	8.3%	14.1%	8.8%	8807
Agreement	5.0%	5.2%	5.1%	5072
Disagreement	3.3%	4.0%	3.4%	3436
Humor	2.1%	6.6%	2.4%	2409
Announcement	1.6%	8.0%	2.0%	2024
Negative Reaction	1.7%	5.0%	1.9%	1899

Table 2: The percentage and count of comments from each discourse act in our dataset in total, and the percentage broken down by link-post and self-post threads.

Discourse Sequence	% Self-Post	% Link-Post	Total %	Total Count
Ques-Ans	47.4%	18.4%	39.2%	39394
Ans-Elab	6.6%	3.2%	5.5%	5545
Elab-Elab	5.4%	7.3%	4.7%	4749
Ques-Ans-Elab	14.0%	6.5%	13.6%	5271
Ques-Ans-Appr	8.8%	3.6%	8.6%	3322
Ques-Ans-Ques	8.1%	3.2%	7.8%	3036

Table 3: The percentage and count of the three most frequent 2-chain and 3-chain discourse sequences in our dataset in total, and percentage broken down by self-post and link-post threads.

Data Analysis

We now present analyses of the discourse acts and sequences in our annotated dataset that contained a majority annotated discourse act. In Table 2, we present the proportion of each discourse act in our dataset, as well as broken down by threads started with link-posts and self-posts, and in Table 3, we present the most frequent 2-chain and 3-chain discourse sequences in our dataset, where a chain constitutes a series of replies. As can be seen, QUESTIONs and ANSWERs make up a large portion of the dataset, partially due to sampling more heavily from self-posts. However, even the least frequent discourse act, NEGATIVE REACTION, has nearly 2,000 comments, which is on its own larger than many entire datasets (Kim, Wang, and Baldwin 2010). We also have more AGREEMENTs than DISAGREEMENTs, echoing prior work on blogs (Gilbert, Bergstrom, and Karahalios 2009).

Next we consider the discourse relations that were annotated. Of the annotations that had a majority relation and were not the first comment in the thread, 98.3% of these relations were to the direct parent of the comment, as designated by Reddit's threaded structure. Thus, on Reddit, the reply relation available via the site is already a close approximation of the proper discourse relation.

Discourse Sequences

We analyze prevalent discourse sequences to better understand the major types of discussion in our dataset. For instance, the first comment in a thread can signal what happens in the rest of the thread. In our dataset, 78% of threads started as a QUESTION, while 22% of threads started as an ANNOUNCEMENT. However, threads starting out as QUESTIONs are concentrated among the self-post threads, with 82% of self-post threads starting out as a QUESTION, while only 17% of link-post threads start out as a QUESTION.

Most Questions	Total %		Most Answers	Total %
iama	44%		askwomen	69%
casualiama	44%		weddingplanning	65%
fakeid	34%		shittyadvice	64%
jailbreak	34%		askreddit	64%
techsupport	31%		explainlikeimfive	63%
buildapcforme	30%		manga	63%
feedthebeast	29%		music	62%
tipofmytongue	28%		anime	62%

Table 4: Subreddits with highest proportion of QUESTION comments and ANSWER comments.

Q&A Q&A pairs are well-studied discourse sequences in research (Cong et al. 2008; Morris, Teevan, and Panovich 2010) because of their applications to information retrieval and relation to CQA sites. Using our dataset, we can look at discourse sequences that go beyond Q&A pairs to provide richer information about discussions that begin with a question. Focusing on the 7,150 threads in our dataset that start as questions, 88% of immediate replies to the first comment are ANSWERs and 6% are follow-up QUESTIONs. The QUESTIONs that are in response to a QUESTION may be of interest as "clarifying" questions for overly broad requests.

As seen in Table 3, Q&A pairs are followed primarily by ELABORATION (33%), APPRECIATION (21%), and QUESTION (18%). ELABORATIONs could be seen as extensions or augmentations of the ANSWER comment, which could be useful for informational retrieval applications. APPRECIATIONs could be seen as an additional signal of quality, on top of signals such as "accepted" answers in some CQA sites such as Yahoo! Answers or community upvotes. In our dataset, 73% of APPRECIATION comments in response to ANSWERs were by the question-asker.

While other works have estimated the number of questions and answers in other social platforms (Morris, Teevan, and Panovich 2010; Paul, Hong, and Chi 2011), we can provide the first estimate towards the Reddit corpus. Because of our filters, we cannot provide an estimate of the number of total QUESTIONs, including unanswered ones. Instead we can make an estimate of around 29.4 million ($\pm$0.3 million) self-post threads and 8.8 million ($\pm$1.5 million) link-post threads that start with a QUESTION and have at least two replies, using a 95% confidence interval. Given our other filters, this estimate is a lower bound on the entirety of Reddit.

We can also examine Q&A at the community level to find sites that behave much like CQA sites like StackOverflow or Quora. We focus our analysis on the 186 communities that have 100 or more comments in our dataset. Looking at Figure 4 we show the subreddits with the highest proportion of QUESTIONs and ANSWERs. While some subreddits are clearly dedicated to Q&A, such as `/r/askwomen` or `/r/explainlikeimfive`, other subreddits such as

Community	Total %	Avg Chain Len
canada	21.7%	1.4
changemyview	20.0%	1.5
politicaldiscussion	17.1%	1.8
smite	17.1%	1.3
dndnext	12.3%	1.6
reddevils	10.9%	1.1
politics	10.5%	1.1
atheism	10.1%	1.7

Table 5: Subreddits with highest proportion of DISAGREEMENT comments, shown with their average length of chains of DISAGREEMENT.

`/r/weddingplanning` or `/r/manga` are not obviously about Q&A from their name but may operate like a CQA site for a specific domain. We also found that the top subreddits for ANSWER are different than the top subreddits for QUESTION. This suggests that some subreddits may have more ANSWERs per QUESTION on average than others. In our dataset, QUESTIONs that appear as the initial post received on average 3.99 ANSWERs (SD=3.57). This signal could be useful towards the task of predicting whether a particular question or a community overall is informational or conversational (Harper, Moy, and Konstan 2009).

Arguments Another sequence of interest is the "chain of disagreement" or sequence of DISAGREEMENT comments replying to each other, which signify an argument occurring (McLaren, Scheuer, and Mikšátko 2010). Overall, we had 2,712 chains of DISAGREEMENT of size 1 to 7 comments, with 17% of chains longer than 1 comment, and an average chain length of 1.23 comments. From this data, we can analyze what concludes arguments if anything. Focusing only on the pages where there were 40 comments or fewer, so no comments were excluded from annotation, we found that 61% of DISAGREEMENTs were followed by nothing. Arguments followed with an ELABORATION 18% of the time, which can be interpreted as a comment elaborating on the arguments of a prior comment, or continuing the argument. DISAGREEMENT chains ended with AGREEMENT only 7% of the time, which may characterize a concession in the disagreement or a resolution. In Table 5, we show the subreddits with the highest proportion of DISAGREEMENT comments out of the communities with over 100 comments in our dataset. We also calculate average chain length, finding that some subreddits that are more dedicated to debate, such as `/r/changemyview` or `/r/politicaldiscussion` have longer arguments than other subreddits such as `/r/politics`.

Announcements There is less research into the kinds of discussions that start out as an ANNOUNCEMENT. However, these threads do constitute a large portion of Reddit, given that 59% of the Reddit corpus with 2 replies is link-post threads, and over 80% of link-post threads in our dataset begin with ANNOUNCEMENT. To understand the major types of discussion in reply to an ANNOUNCEMENT, we cluster the 2,024 threads by their proportion of discourse acts in the replies, using k-means with 4 clusters. The best silhouette score (Rousseeuw 1987) determined the cluster number.

- **Appreciation (18%)**: One cluster has a high average proportion of APPRECIATION comments at 62%. Threads in this cluster come primarily from subreddits such as `/r/keto` (related to the ketogenic diet) and `/r/stopdrinking` (about abstaining from alcohol), where people post updates on their personal goals and receive encouragement.
- **Arguments (21%)**: Another cluster has a higher proportion of HUMOR (37%), DISAGREEMENT (35%), and AGREEMENT (34%) comments. Threads in this cluster come from subreddits like `/r/politics` and `/r/atheism`, where most announcements are news articles, and arguments and jokes occur in the replies.
- **Q&A (29%)**: A third cluster is predominantly threads with Q&A, at 35% of discourse pairs. Some notable subreddits represented include `/r/pcmasterrace` (related to PC gaming) and `/r/ultrahardcore` (related to a mode in the Minecraft game), where announcements more readily lead to requests for more information.
- **Elaboration (32%)**: The final cluster is primarily ELABORATION comments, at 85%. The predominant subreddit in this cluster is a gaming community, `/r/leagueoflegends`. In this cluster, users might pass around stories, tips, or opinions building on each other regarding a particular topic.

Predicting Discourse Acts

We investigate how well supervised models for extracting discourse acts perform, experimenting with both structured and unstructured models. Because our annotated dataset has shown that discourse relations map well to the existing Reddit reply structure, we focus only on the discourse act multiclass classification task.

Features

Content + Punctuation: We collect unigrams, bigrams, and trigrams from the text of the comment. If the comment has a title, in the case of the initial post, then the n-grams of the title are counted separately from the n-grams of the body. We use a word tokenizer that tokenizes punctuation instead of stripping it so that we count potentially important punctuation like question marks or exclamation points. We use TF-IDF weighting and set a minimum document frequency of 50 comments.

Structure: We calculate several features related to the structure of the comment and its position. One feature is the depth of the comment according to Reddit's threaded structure, which we collect as both a raw count and normalized by the number of comments in the discussion. We also calculate number of sentences, number of words, and number of characters of both the body and the title of the comment. We computed these values for both the current comment and the parent comment.

Author: We collect features about the author of the comment, including a binary feature for whether the current

Model	Accuracy	Precision	Recall	F1
All Answers	0.43	0.16	0.41	0.23
Q-Mark & Answers	0.55	0.30	0.52	0.38
LogReg	0.672	0.657	0.672	0.648
SVM-HMM	0.708	0.673	0.708	0.680
CRF	**0.763**	**0.747**	**0.763**	**0.747**

Table 6: Results of the models to predict discourse acts.

commenter is also the commenter of the initial post and a binary feature for whether the current commenter is the same as the parent commenter.

Thread: We calculate features that are the same across all comments in the thread. One feature is the total number of comments in the discussion. Another is the number of unique branches in the discussion tree. We also record whether the discussion originated as a self-post or a link-post. Finally, we collect the average length of all the branches or threads of discussion in the discussion tree.

Community: We have a feature naming the subreddit that the thread came from, as some subreddits have a greater portion of some types of discourse and not others.

Other experiments we conducted were with features such as word overlap between parent and current comment, discourse act priors for each author across the training set, number of replies to the current comment, and sentiment analysis. These are omitted here for space reasons and because they did not lead to improvements in performance.

Data and Models

For comparison with our models, we designed two baselines, one where all questions are labeled as ANSWER (All Answers) and a slightly more sophisticated one where all initial posts and also comments containing a question mark in the text are labeled as QUESTION, while all other comments are labeled as ANSWER (Q-Mark & Answers).

Our first model is a standard logistic regression model using L2 regularization and the LibLinear optimization, implemented in scikit-learn[5]. Our next two models are structured prediction models that take into account the sequence of comments. The first is a hidden Markov model with second-order transition dependencies and no emission dependencies, using the SVM^{hmm} library (Joachims 2008). Finally, we build a conditional random field, using CRFSuite (Okazaki 2007), with the Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) training algorithm and L1 regularization. These models were chosen because prior work has suggested that models such as these that capture structural dependencies within a sequence of labels provide important information for identifying discourse acts (Ding et al. 2008; Kim, Wang, and Baldwin 2010).

To split our training and testing sets, we conducted stratified 10-fold cross validation, splitting our data at the thread level. This is so all comments from a single discussion are in the same training or testing group. Additionally, the structured prediction models require items provided in a sequence. However, discussions on Reddit branch outwards like a tree instead of being append-only. Thus, for those models, we constructed for each discussion tree all possible branches as individual sequences. If a comment has multiple replies, each of those would be part of a separate sequence of comments. When it came to evaluating the results of a structured prediction model, as a comment may be represented in multiple sequences and thus tagged multiple times, we collect all the predicted tags for a given comment and assign the most common tag to that comment. Thus, our evaluation metrics only count each comment in the test dataset once.

[5] http://scikit-learn.org

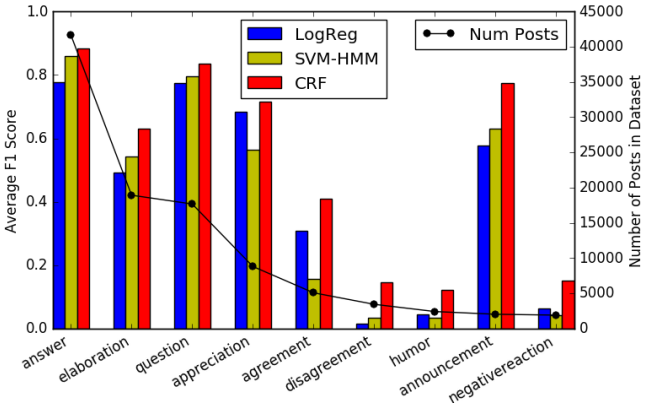

Figure 1: Average F1 scores of each model broken down by each discourse act along with their prevalence in the dataset.

Results

We report results of our experiments in Table 6. The metrics shown are all for prediction at the comment level, as opposed to at the thread level. As can be seen, the CRF model performs the best overall, achieving an average F1 score of 75%. Both structured prediction models perform better than the logistic regression model, demonstrating that the context, encapsulated by the preceding comment classifications, is important towards determining the discourse act of the current comment. Finally, all models perform better than the two baselines given.

In Figure 1, we break down the average F1 scores by discourse act. The CRF model performs best across all the categories, though some categories such as DISAGREEMENT, HUMOR, and NEGATIVE REACTION have relatively low F1 scores. These categories may be more difficult to distinguish for humans as well, as these had the lowest inter-rater reliability in Table 1. A test for correlation between inter-rater reliability and CRF F1 score yields a strong positive correlation (Spearman's rank, ρ=0.917, p<0.001).

Given the differences in performance across the discourse acts and also different frequencies of discourse acts in different communities, we would expect that some subreddits would have higher performance overall. Considering only the subreddits with over 100 predictions made, we saw 4 subreddits have an average F1 score above 90% across the 10 folds, with the highest being /r/boardgames with

Features	Accuracy	Precision	Recall	F1
All	**0.763**	**0.747**	**0.763**	**0.747**
All - Author	0.762	0.746	0.763	0.744
All - Thread	0.761	0.747	0.760	0.746
All - Community	0.743	0.738	0.742	0.739
All - Content	0.587	0.538	0.588	0.550
All - Structure	0.540	0.564	0.539	0.507

Table 7: Results of feature ablation experiments, removing one feature group at a time.

94% average F1. There was also 38 subreddits with an average F1 between 80% and 90%. On the other hand, 16 subreddits had an average F1 below 70%, with the lowest being `/r/funny` at 45% average F1.

Feature Ablation We examine the importance of the different feature groups we used by performing a set of feature ablation experiments. Removing one of the feature groups but retaining the rest for each of the feature groups, we calculate evaluation metrics using the best-performing CRF model. As can be seen in Table 7, the most important feature groups included the structural and content features. In contrast, the thread and author feature groups had the least impact on the classification accuracy.

Comparison to Q&A-Only Models Finally, we focus on our best model's performance on predicting the categories of QUESTION and ANSWER due to their outsize importance in information retrieval research. In many works, these are the only discourse acts considered when performing Q&A prediction. However, the introduction of additional categories could worsen performance on Q&A prediction by introducing more confusion between Q&A and the other labels. We consider how our models would perform if they only had to predict QUESTION or ANSWER comments, or both, with an OTHER category signifying the rest.

In Table 8, we can see how well our CRF model performs with regards to classifying QUESTIONs, when we vary the number of discourse act labels. The best precision is achieved when the model is only a binary classifier between QUESTION and OTHER. However, the best recall and F1 on QUESTION prediction is achieved when all 9 of our discourse acts are used as labels. As the difference in F1 is small between the Q+A+Other model and the 9 categories model, we conduct an 1-way ANOVA test using the 10 cross validation folds from evaluations with each model. From this test, we find there is a statistically significant difference between the F1 scores of the two models ($F=10.97$, $p<0.005$).

In Table 9, we show results for classifying ANSWER comments. Precision is relatively low when only ANSWER and OTHER categories are used and is best when classifying QUESTION, ANSWER, and OTHER. This may be because ANSWER comments are dependent on having a preceding QUESTION, and QUESTION comments may be easier to identify. The model with all 9 discourse acts has the best recall and shares the best F1 score with the Q&A model.

Altogether, for predicting QUESTION and ANSWER, the CRF model containing all 9 discourse acts performs better or on par with a model predicting only QUESTION or ANSWER, or both, due mainly to improvements in recall. The improvements may be because having a richer discourse act categorization would allow for more fine-grained transition probabilities. On the other hand, overall precision decreases slightly, due to the greater number of classifications for which a comment could be mistakenly classified.

Categories	Precision	Recall	F1
Question + Other	**0.877**	0.791	0.832
Q + A + Other	0.875	0.784	0.827
All 9 categories	0.854	**0.823**	**0.837**

Table 8: Results for predicting QUESTION using the CRF model and varying the number of discourse acts represented.

Categories	Precision	Recall	F1
Answer + Other	0.793	0.837	0.815
Q + A + Other	**0.87**	0.898	**0.885**
All 9 categories	0.855	**0.917**	**0.885**

Table 9: Results for predicting ANSWER using the CRF model and varying the number of discourse acts represented.

Discussion

In this work, we present a new coarse discourse act categorization for online discussion as well as a new dataset of discussions labeled with discourse acts and relations from a diversity of communities. We demonstrate how discourse acts can tell us more about common sequences of discourse and isolate CQA-like communities. We show that using structured models such as CRF, we can build classifiers to predict discourse acts at a 75% F1 score. Our model with 9 categories also improves in recall over models with only Q&A labels for the tasks of QUESTION and ANSWER prediction.

New applications become possible with the ability to tag comments with discourse acts. For instance, labeled discourse acts could help moderators know whether existing questions have been answered (Kim, Li, and Kim 2010) or step in to resolve lingering disputes. Users with questions could be routed towards more CQA-like communities when there may be several subreddits dedicated to the same topic, such as `/r/askscience` versus `/r/science`.

Another area that could use discourse acts is discussion summarization (Murray et al. 2006; Rambow et al. 2004). Most automatic summarization techniques are built for long individual documents as opposed to a sequence of discourse acts. It is also unclear what an ideal summary for a discussion would look like. One consideration is that different types of discussions could warrant different types of summaries. For instance, an argument might be summarized by summarizing the arguments on one side followed by the arguments on the other side. Quantities might also be useful for the summary of an argument, for instance how many comments had one stance versus another. Our discourse acts for AGREEMENT and DISAGREEMENT may be useful for stance classification (Rosenthal and McKeown 2015). Alternatively, the summary for a QUESTION followed by a

series of ANSWERs might instead be a short sentence extracted from the QUESTION comment and the highest voted answer, the most frequent answer, or a series of common answers, depending on the nature of the question. Knowing the discourse structures may help determine what kind of summary is needed and from which comments to pull sentences, if the summarization strategy is extractive. This information could also help support existing systems for manual discussion summarization (Zhang, Verou, and Karger 2017).

Finally, this dataset can be useful to improve search engines and natural dialogue systems such as chat bots and virtual assistants. Search engines and virtual assistants that gather answers to queries from documents on the web can use discourse acts to better characterize community search results. For instance, snippets could be taken from ANSWER comments as opposed to other portions of the thread. Queries that return more conversational Q&A threads with many answers to a question could trigger a different interface or interaction than more informational queries, such as clusters of answers grouped by stance, sentiment, or topic. Answers that are controversial, meaning they are followed by an argument, could be marked as such.

Future Work

We conducted our analysis using the site Reddit, which has some particular characteristics that may not transfer to communities on other sites. For instance, Reddit is a threaded discussion forum while many forums are append-only. In the future, we plan to analyze a non-threaded forum like TripAdvisor. Expanding also allows us to look beyond the overall Reddit community, which has biases compared to the average internet user.

This dataset and analysis was based on a discourse act classification that we developed, which may not be suitable for particular tasks. For instance, some researchers may desire a more fine-grained categorization for a particular discourse act. Future work could expand our classification to create a taxonomy and augment our dataset with more detailed annotations, or use other datasets (Sameki, Barua, and Paritosh 2016) in concert with ours. For instance, our set of QUESTION comments could be further labeled into informational and conversational questions (Harper, Moy, and Konstan 2009). Future work could also build on our dataset by collecting annotations at the sub-comment level or collecting additional tags or relations per comment.

We imagine empirical analyses of online discussions could be furthered using this dataset. Prior studies on question-answering (Harper et al. 2008), argumentation (Tan et al. 2016), echo chambers (Gilbert, Bergstrom, and Karahalios 2009), and gratitude (Spiro, Matias, and Monroy-Hernández 2016) have used datasets significantly smaller than our dataset or focused on only one or a few communities. Other work includes observing how characteristics of communities and authors relate to discourse structures, such as the role of social and administrator moderation in shaping discourse or how structural properties such as community size, diversity of users, and age can cause discourse to vary.

Finally, our dataset suggests further work in question-answering. For instance, much research looks at Q&A at the start of a thread. However, as shown in our dataset, many Q&A pairs exist deeper in discussion threads. Future work could work on determining which Q&A pairs can be understood on their own, and finding ways to resolve ambiguity and bring in context for Q&A pairs that require context.

Conclusion

Using a novel discourse act categorization, we present one of the largest manually annotated datasets of threads of discussion sampled from thousands of communities on Reddit, with each comment in each thread annotated with its discourse act and relation. From our dataset, we observe common patterns of discourse sequences, including Q&A and arguments, and use these signals to characterize communities. Finally, we conduct experiments on classification of discourse acts, with a structured CRF model achieving a 75% F1 score. We additionally demonstrate how our use of 9 discourse acts overall improves recall of Q&A detection over a model that only labels questions and answers.

Acknowledgements

We would like to thank Ka Wong, Akihiro Matsukawa, Olivia Rhinehart, and Nancy Chang for their input and assistance, as well as our annotators.

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *WSDM '08*, 183–194. ACM.

Allen, J.; Ferguson, G.; and Stent, A. 2001. An architecture for more realistic conversational systems. In *IUI '01*, 1–8. ACM.

Austin, J. L. 1975. *How to do things with words*. Oxford University Press.

Bunt, H.; Alexandersson, J.; Carletta, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Lee, K.; Petukhova, V.; Popescu-Belis, A.; Romary, L.; et al. 2010. Towards an iso standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Cohen, W. W.; Carvalho, V. R.; and Mitchell, T. M. 2004. Learning to classify email into"speech acts". In *EMNLP '04*, 309–316.

Cong, G.; Wang, L.; Lin, C.-Y.; Song, Y.-I.; and Sun, Y. 2008. Finding question-answer pairs from online forums. In *SIGIR '08*, 467–474. ACM.

De Wever, B.; Schellens, T.; Valcke, M.; and Van Keer, H. 2006. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & education* 46(1):6–28.

Ding, S.; Cong, G.; Lin, C.-Y.; and Zhu, X. 2008. Using conditional random fields to extract contexts and answers of questions from online forums. In *ACL '08*, volume 8, 710–718. Citeseer.

Feng, D.; Shaw, E.; Kim, J.; and Hovy, E. 2006. Learning to detect conversation focus of threaded discussions. In *NAACL HLT '06*, 208–215. ACL.

Fortuna, B.; Rodrigues, E. M.; and Milic-Frayling, N. 2007. Improving the classification of newsgroup messages through social network analysis. In *CIKM '07*, 877–880. ACM.

Freeman, J. B. 1991. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter.

Gilbert, E.; Bergstrom, T.; and Karahalios, K. 2009. Blogs are echo chambers: Blogs are echo chambers. In *HICSS '09*, 1–10. IEEE.

Harper, F. M.; Raban, D.; Rafaeli, S.; and Konstan, J. A. 2008. Predictors of answer quality in online q&a sites. In *CHI '08*, 865–874. ACM.

Harper, F. M.; Moy, D.; and Konstan, J. A. 2009. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *CHI '09*, 759–768. ACM.

Hasan, K. S., and Ng, V. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP '14*, 751–762.

Hong, L., and Davison, B. D. 2009. A classification-based approach to question answering in discussion boards. In *SIGIR '09*, 171–178. ACM.

Jeong, M.; Lin, C.-Y.; and Lee, G. G. 2009. Semi-supervised speech act recognition in emails and forums. In *EMNLP '09*, 1250–1259. ACL.

Jeong, A. C. 2003. The sequential analysis of group interaction and critical thinking in online. *The American Journal of Distance Education* 17(1):25–43.

Joachims, T. 2008. Svm-hmm: Sequence tagging with structural support vector machines.

Kim, J.; Li, J.; and Kim, T. 2010. Towards identifying unresolved discussions in student online forums. In *NAACL HLT '10*, 84–91. ACL.

Kim, S. N.; Wang, L.; and Baldwin, T. 2010. Tagging and linking web forum posts. In *CoNLL '10*, 192–202. ACL.

Klein, M. 2011. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*.

Lu, J.; Chiu, M. M.; and Law, N. W. 2011. Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. *Computers in Human Behavior* 27(2):946–955.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3):243–281.

McLaren, B. M.; Scheuer, O.; and Mikšátko, J. 2010. Supporting collaborative learning and e-discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence in Education* 20(1):1–46.

Morris, M. R.; Teevan, J.; and Panovich, K. 2010. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *CHI '10*, 1739–1748. ACM.

Murray, G.; Renals, S.; Carletta, J.; and Moore, J. 2006. Incorporating speaker and discourse features into speech summarization. In *NAACL HLT '06*, 367–374. ACL.

Okazaki, N. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Paul, S. A.; Hong, L.; and Chi, E. H. 2011. Is twitter a good place for asking questions? a characterization study. In *ICWSM '11*. AAAI.

Rambow, O.; Shrestha, L.; Chen, J.; and Lauridsen, C. 2004. Summarizing email threads. In *NAACL HLT '04*, 105–108. ACL.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In *NAACL HLT '10*, 172–180. ACL.

Rosé, C.; Wang, Y.-C.; Cui, Y.; Arguello, J.; Stegmann, K.; Weinberger, A.; and Fischer, F. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning* 3(3):237–271.

Rosenthal, S., and McKeown, K. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 168.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20:53–65.

Sameki, M.; Barua, A.; and Paritosh, P. 2016. Rigorously collecting commonsense judgments for complex question-answer content. In *HCOMP '16*. AAAI.

Scheuer, O.; Loll, F.; Pinkwart, N.; and McLaren, B. M. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning* 5(1):43–102.

Searle, J. R. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press.

Spiro, E. S.; Matias, J. N.; and Monroy-Hernández, A. 2016. Networks of gratitude: Structures of thanks and user expectations in workplace appreciation systems. In *ICWSM '16*. AAAI.

Stolcke, A.; Coccaro, N.; Bates, R.; Taylor, P.; Van Ess-Dykema, C.; Ries, K.; Shriberg, E.; Jurafsky, D.; Martin, R.; and Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26(3):339–373.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *WWW '16*, 613–624.

Xi, W.; Lind, J.; and Brill, E. 2004. Learning effective ranking functions for newsgroup search. In *SIGIR '04*, 394–401. ACM.

Zhang, A. X.; Verou, L.; and Karger, D. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. In *CSCW 17*. ACM.