And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue

Abstract

We investigate the characteristics of factual and emotional argumentation styles observed in online debates. Using an annotated set of"factual"and"feeling"debate forum posts, we extract patterns that are highly correlated with factual and emotional arguments, and then apply a bootstrapping methodology to find new patterns in a larger pool of unannotated forum posts. This process automatically produces a large set of patterns representing linguistic expressions that are highly correlated with factual and emotional language. Finally, we analyze the most discriminating patterns to better understand the defining characteristics of factual and emotional arguments.

Introduction

Human lives are being lived online in transformative ways: people can now ask questions, solve problems, share opinions, or discuss current events with anyone they want, at any time, in any location, on any topic. The purposes of these exchanges are varied, but a significant fraction of them are argumentative, ranging from hot-button political controversies (e.g., national health care) to religious interpretation (e.g., Biblical exegesis). And while the study of the structure of arguments has a long lineage in psychology BIBREF0 and rhetoric BIBREF1 , large shared corpora of natural informal argumentative dialogues have only recently become available.

Natural informal dialogues exhibit a much broader range of argumentative styles than found in traditional work on argumentation BIBREF2 , BIBREF0 , BIBREF3 , BIBREF4 . Recent work has begun to model different aspects of these natural informal arguments, with tasks including stance classification BIBREF5 , BIBREF6 , argument summarization BIBREF7 , sarcasm detection BIBREF8 , and work on the detailed structure of arguments BIBREF9 , BIBREF10 , BIBREF11 . Successful models of these tasks have many possible applications in sentiment detection, automatic summarization, argumentative agents BIBREF12 , and in systems that support human argumentative behavior BIBREF13 .

Our research examines factual versus feeling argument styles, drawing on annotations provided in the Internet Argument Corpus (IAC) BIBREF6 . This corpus includes quote-response pairs that were manually annotated with respect to whether the response is primarily a factual or feeling based argument, as Section SECREF2 describes in more detail. Figure FIGREF1 provides examples of responses in the IAC (paired with preceding quotes to provide context), along with the response's factual vs. feeling label.

factual responses may try to bolster their argument by providing statistics related to a position, giving historical or scientific background, or presenting specific examples or data. There is clearly a relationship between a proposition being factual versus objective or veridical, although each of these different labelling tasks may elicit differences from annotators BIBREF14 , BIBREF15 , BIBREF16 , BIBREF17 .

The feeling responses may seem to lack argumentative merit, but previous work on argumentation describes situations in which such arguments can be effective, such as the use of emotive arguments to draw attention away from the facts, or to frame a discussion in a particular way BIBREF18 , BIBREF19 . Furthermore, work on persuasion suggest that feeling based arguments can be more persuasive in particular circumstances, such as when the hearer shares a basis for social identity with the source (speaker) BIBREF20 , BIBREF21 , BIBREF22 , BIBREF23 , BIBREF24 . However none of this work has documented the linguistic patterns that characterize the differences in these argument types, which is a necessary first step to their automatic recognition or classification. Thus the goal of this paper is to use computational methods for pattern-learning on conversational arguments to catalog linguistic expressions and stylistic properties that distinguish Factual from Emotional arguments in these on-line debate forums.

Section SECREF2 describes the manual annotations for factual and feeling in the IAC corpus. Section SECREF5 then describes how we generate lexico-syntactic patterns that occur in both types of argument styles. We use a weakly supervised pattern learner in a bootstrapping framework to automatically generate lexico-syntactic patterns from both annotated and unannotated debate posts. Section SECREF3 evaluates the precision and recall of the factual and feeling patterns learned from the annotated texts and after bootstrapping on the unannotated texts. We also present results for a supervised learner with bag-of-word features to assess the difficulty of this task. Finally, Section SECREF4 presents analyses of the linguistic expressions found by the pattern learner and presents several observations about the different types of linguistic structures found in factual and feeling based argument styles. Section SECREF5 discusses related research, and Section SECREF6 sums up and proposes possible avenues for future work.

Pattern Learning for Factual and Emotional Arguments

We first describe the corpus of online debate posts used for our research, and then present a bootstrapping method to identify linguistic expressions associated with factual and feeling arguments.

Data

The IAC corpus is a freely available annotated collection of 109,553 forum posts (11,216 discussion threads). In such forums, conversations are started by posting a topic or a question in a particular category, such as society, politics, or religion BIBREF6 . Forum participants can then post their opinions, choosing whether to respond directly to a previous post or to the top level topic (start a new thread). These discussions are essentially dialogic; however the affordances of the forum such as asynchrony, and the ability to start a new thread rather than continue an existing one, leads to dialogic structures that are different than other multiparty informal conversations BIBREF25 . An additional source of dialogic structure in these discussions, above and beyond the thread structure, is the use of the quote mechanism, which is an interface feature that allows participants to optionally break down a previous post into the components of its argument and respond to each component in turn.

The IAC includes 10,003 Quote-Response (Q-R) pairs with annotations for factual vs. feeling argument style, across a range of topics. Figure FIGREF4 shows the wording of the survey question used to collect the annotations. Fact vs. Feeling was measured as a scalar ranging from -5 to +5, because previous work suggested that taking the means of scalar annotations reduces noise in Mechanical Turk annotations BIBREF26 . Each of the pairs was annotated by 5-7 annotators. For our experiments, we use only the response texts and assign a binary Fact or Feel label to each response: texts with score INLINEFORM0 1 are assigned to the fact class and texts with score INLINEFORM1 -1 are assigned to the feeling class. We did not use the responses with scores between -1 and 1 because they had a very weak Fact/Feeling assessment, which could be attributed to responses either containing aspects of both factual and feeling expression, or neither. The resulting set contains 3,466 fact and 2,382 feeling posts. We randomly partitioned the fact/feel responses into three subsets: a training set with 70% of the data (2,426 fact and 1,667 feeling posts), a development (tuning) set with 20% of the data (693 fact and 476 feeling posts), and a test set with 10% of the data (347 fact and 239 feeling posts). For the bootstrapping method, we also used 11,560 responses from the unannotated data.

Bootstrapped Pattern Learning

The goal of our research is to gain insights into the types of linguistic expressions and properties that are distinctive and common in factual and feeling based argumentation. We also explore whether it is possible to develop a high-precision fact vs. feeling classifier that can be applied to unannotated data to find new linguistic expressions that did not occur in our original labeled corpus.

To accomplish this, we use the AutoSlog-TS system BIBREF27 to extract linguistic expressions from the annotated texts. Since the IAC also contains a large collection of unannotated texts, we then embed AutoSlog-TS in a bootstrapping framework to learn additional linguistic expressions from the unannotated texts. First, we briefly describe the AutoSlog-TS pattern learner and the set of pattern templates that we used. Then, we present the bootstrapping process to learn more Fact/Feeling patterns from unannotated texts.

To learn patterns from texts labeled as fact or feeling arguments, we use the AutoSlog-TS BIBREF27 extraction pattern learner, which is freely available for research. AutoSlog-TS is a weakly supervised pattern learner that requires training data consisting of documents that have been labeled with respect to different categories. For our purposes, we provide AutoSlog-TS with responses that have been labeled as either fact or feeling.

AutoSlog-TS uses a set of syntactic templates to define different types of linguistic expressions. The left-hand side of Figure FIGREF8 shows the set of syntactic templates defined in the AutoSlog-TS software package. PassVP refers to passive voice verb phrases (VPs), ActVP refers to active voice VPs, InfVP refers to infinitive VPs, and AuxVP refers to VPs where the main verb is a form of “to be” or “to have”. Subjects (subj), direct objects (dobj), noun phrases (np), and possessives (genitives) can be extracted by the patterns. AutoSlog-TS applies the Sundance shallow parser BIBREF28 to each sentence and finds every possible match for each pattern template. For each match, the template is instantiated with the corresponding words in the sentence to produce a specific lexico-syntactic expression. The right-hand side of Figure FIGREF8 shows an example of a specific lexico-syntactic pattern that corresponds to each general pattern template.

In addition to the original 17 pattern templates in AutoSlog-TS (shown in Figure FIGREF8), we defined 7 new pattern templates for the following bigrams and trigrams: Adj Noun, Adj Conj Adj, Adv Adv, Adv Adv

Adv, Adj Adj, Adv Adj, Adv Adv Adj. We added these n-gram patterns to provide coverage for adjective and adverb expressions because the original templates were primarily designed to capture noun phrase and verb phrase expressions.

The learning process in AutoSlog-TS has two phases. In the first phase, the pattern templates are applied to the texts exhaustively, so that lexico-syntactic patterns are generated for (literally) every instantiation of the templates that appear in the corpus. In the second phase, AutoSlog-TS uses the labels associated with the texts to compute statistics for how often each pattern occurs in each class of texts. For each pattern INLINEFORM0 , we collect P(factual INLINEFORM1 INLINEFORM2) and P(feeling INLINEFORM3 INLINEFORM4), as well as the pattern's overall frequency in the corpus.

Since the IAC data set contains a large number of unannotated debate forum posts, we embedd AutoSlog-TS in a bootstrapping framework to learn additional patterns. The flow diagram for the bootstrapping system is shown in Figure FIGREF10 .

Initially, we give the labeled training data to AutoSlog-TS, which generates patterns and associated statistics. The next step identifies high-precision patterns that can be used to label some of the unannotated texts as factual or feeling. We define two thresholds: INLINEFORM0 to represent a minimum frequency value, and INLINEFORM1 to represent a minimum probability value. We found that using only a small set of patterns (when INLINEFORM2 is set to a high value) achieves extremely high precision, yet results in a very low recall. Instead, we adopt a strategy of setting a moderate probability threshold to identify reasonably reliable patterns, but labeling a text as factual or feeling only if it contains at least a certain number different patterns for that category, INLINEFORM3 . In order to calibrate the thresholds, we experimented with a range of threshold values on the development (tuning) data and identified INLINEFORM4 =3, INLINEFORM5 =.70, and INLINEFORM6 =3 for the factual class, and INLINEFORM7 =3, INLINEFORM8 =.55, and INLINEFORM9 =3 for the feeling class as having the highest classification precision (with non-trivial recall).

The high-precision patterns are then used in the bootstrapping framework to identify more factual and feeling texts from the 11,561 unannotated posts, also from 4forums.com. For each round of bootstrapping, the current set of factual and feeling patterns are matched against the unannotated texts, and posts that match at least 3 patterns associated with a given class are assigned to that class. As shown in Figure FIGREF10 , the Bootstrapped Data Balancer then randomly selects a balanced subset of the newly classified posts to maintain the same proportion of factual vs. feeling documents throughout the bootstrapping process. These new documents are added to the set of labeled documents, and the bootstrapping process repeats. We use the same threshold values to select new high-precision patterns for all iterations.

Evaluation

We evaluate the effectiveness of the learned patterns by applying them to the test set of 586 posts (347 fact and 239 feeling posts, maintaining the original ratio of fact to feel data in train). We classify each post as factual or feeling using the same procedure as during bootstrapping: a post is labeled as factual or feeling if it matches at least three high-precision patterns for that category. If a document contains three patterns for both categories, then we leave it unlabeled. We ran the bootstrapping algorithm for four iterations.

The upper section of Table TABREF11 shows the Precision and Recall results for the patterns learned during bootstrapping. The Iter 0 row shows the performance of the patterns learned only from the original, annotated training data. The remaining rows show the results for the patterns learned from the unannotated texts during bootstrapping, added cumulatively. We show the results after each iteration of bootstrapping.

Table TABREF11 shows that recall increases after each bootstrapping iteration, demonstrating that the patterns learned from the unannotated texts yield substantial gains in coverage over those learned only from the annotated texts. Recall increases from 22.8% to 40.9% for fact, and from 8.0% to 18.8% for feel. The precision for the factual class is reasonably good, but the precision for the feeling class is only moderate. However, although precision typically decreases during boostrapping due to the addition of imperfectly labeled data, the precision drop during bootstrapping is relatively small.

We also evaluated the performance of a Naive Bayes (NB) classifier to assess the difficulty of this task with a traditional supervised learning algorithm. We trained a Naive Bayes classifier with unigram features and binary values on the training data, and identified the best Laplace smoothing parameter using the development data. The bottom row of Table TABREF11 shows the results for the NB classifier on the test data. These results show that the NB classifier yields substantially higher recall for both categories, undoubtedly due to the fact that the classifier uses all unigram information available in the text. Our pattern learner, however, was restricted to learning linguistic expressions in specific syntactic constructions, usually requiring more than one word, because our goal was to study specific expressions associated with factual and feeling argument styles. Table TABREF11 shows that the lexico-syntactic patterns did obtain higher precision than the NB classifier, but with lower recall.

Table TABREF14 shows the number of patterns learned from the annotated data (Iter 0) and the number of new patterns added after each bootstrapping iteration. The first iteration dramatically increases the set of patterns, and more patterns are steadily added throughout the rest of bootstrapping process.

The key take-away from this set of experiments is that distinguishing factual and feeling argumets is clearly a challenging task. There is substantial room for improvement for both precision and recall, and surprisingly, the feeling class seems to be harder to accurately recognize than the factual class. In the next section, we examine the learned patterns and their syntactic forms to better understand the language used in the debate forums.

Analysis

Table TABREF13 provides examples of patterns learned for each class that are characteristic of that class. We observe that patterns associated with factual arguments often include topic-specific terminology, explanatory language, and argument phrases. In contrast, the patterns associated with feeling based arguments are often based on the speaker's own beliefs or claims, perhaps assuming that they themselves are credible BIBREF20 , BIBREF24 , or they involve assessment or evaluations of the arguments of the other speaker BIBREF29 . They are typically also very creative and diverse, which may be why it is hard to get higher accuracies for feeling classification, as shown by Table TABREF11 .

Figure FIGREF15 shows the distribution of syntactic forms (templates) among all of the high-precision patterns identified for each class during bootstrapping. The x-axes show the syntactic templates and the y-axes show the percentage of all patterns that had a specific syntactic form. Figure FIGREF15 counts each lexico-syntactic pattern only once, regardless of how many times it occurred in the data set. Figure FIGREF15 counts the number of instances of each lexico-syntactic pattern. For example, Figure FIGREF15 shows that the Adj Noun syntactic form produced 1,400 different patterns, which comprise 22.6% of the distinct patterns learned. Figure FIGREF15 captures the fact that there are 7,170 instances of the Adj Noun patterns, which comprise 17.8% of all patterns instances in the data set.

For factual arguments, we see that patterns with prepositional phrases (especially NP Prep) and passive voice verb phrases are more common. Instantiations of NP Prep are illustrated by FC1, FC5, FC8, FC10 in Table TABREF13 . Instantiations of PassVP are illustrated by FC2 and FC4 in Table TABREF13 . For feeling arguments, expressions with adjectives and active voice verb phrases are more common. Almost every high probability pattern for feeling includes an adjective, as illustrated by every pattern except FE8 in Table TABREF13 . Figure FIGREF15 shows that three syntactic forms account for a large proportion of the instances of high-precision patterns in the data: Adj Noun, NP Prep, and ActVP.

Next, we further examine the NP Prep patterns since they are so prevalent. Figure FIGREF19 shows the percentages of the most frequently occurring prepositions found in the NP Prep patterns learned for each class. Patterns containing the preposition "of" make up the vast majority of prepositional phrases for both the fact and feel classes, but is more common in the fact class. In contrast, we observe that patterns with the preposition “for” are substantially more common in the feel class than the fact class.

Table TABREF20 shows examples of learned NP Prep patterns with the preposition "of" in the fact class and "for" in the feel class. The "of" preposition in the factual arguments often attaches to objective terminology. The "for" preposition in the feeling-based arguments is commonly used to express advocacy (e.g., demand for) or refer to affected population groups (e.g., treatment for). Interestingly, these phrases are subtle indicators of feeling-based arguments rather than explicit expressions of emotion or sentiment.

Related Work

Related research on argumentation has primarily worked with different genres of argument than found in IAC, such as news articles, weblogs, legal briefs, supreme court summaries, and congressional debates BIBREF2 , BIBREF30 , BIBREF31 , BIBREF0 , BIBREF3 , BIBREF4 . The examples from IAC in Figure FIGREF1 illustrate that natural informal dialogues such as those found in online forums exhibit a much broader range of argumentative styles. Other work has on models of natural informal arguments have focused on stance classification BIBREF32 , BIBREF5 , BIBREF6 , argument summarization BIBREF7 , sarcasm detection BIBREF8 , and identifying the structure of arguments such as main claims and their justifications BIBREF9 , BIBREF10 , BIBREF11 .

Other types of language data also typically contains a mixture of subjective and objective sentences, e.g. Wiebe et al. wiebeetal2001a,wiebeetalcl04 found that 44% of sentences in a news corpus were subjective. Our work is also related to research on distinguishing subjective and objective text BIBREF33 , BIBREF34 , BIBREF14 , including bootstrapped pattern learning for subjective/objective sentence classification BIBREF15 . However, prior work has primarily focused on news texts, not argumentation, and the notion of objective language is not exactly the same as factual. Our work also aims to recognize emotional language specifically, rather than all forms of subjective language. There has been substantial work on sentiment and opinion analysis (e.g., BIBREF35 , BIBREF36 , BIBREF37 , BIBREF38 , BIBREF39 , BIBREF40) and recognition of specific emotions in text BIBREF41 , BIBREF42 , BIBREF43 , BIBREF44 , which could be incorporated in future extensions of our work. We also hope to examine more closely the relationship of this work to previous work aimed at the identification of nasty vs. nice arguments in the IAC BIBREF45 , BIBREF8 .

Conclusion

In this paper, we use observed differences in argumentation styles in online debate forums to extract patterns that are highly correlated with factual and emotional argumentation. From an annotated set of forum post responses, we are able extract high-precision patterns that are associated with the argumentation style classes, and we are then able to use these patterns to get a larger set of indicative patterns using a bootstrapping methodology on a set of unannotated posts.

From the learned patterns, we derive some characteristic syntactic forms associated with the fact and feel that we use to discriminate between the classes. We observe distinctions between the way that different arguments are expressed, with respect to the technical and more opinionated terminologies used, which we analyze on the basis of grammatical forms and more direct syntactic patterns, such as the use of different prepositional phrases. Overall, we demonstrate how the learned patterns can be used to more precisely gather similarly-styled argument responses from a pool of unannotated responses, carrying the characteristics of factual and emotional argumentation style.

In future work we aim to use these insights about argument structure to produce higher performing classifiers for identifying factual vs. feeling argument styles. We also hope to understand in more detail the relationship between these argument styles and the heurstic routes to persuasion and associated strategies that have been identified in previous work on argumentation and persuasion BIBREF2 , BIBREF0 , BIBREF4 .

Acknowledgments

This work was funded by NSF Grant IIS-1302668-002 under the Robust Intelligence Program. The collection and annotation of the IAC corpus was supported by an award from NPS-BAA-03 to UCSC and an IARPA Grant under the Social Constructs in Language Program to UCSC by subcontract from the University of Maryland.