Argument Mining: Extracting Arguments from Online Dialogue

Reid Swanson & Brian Ecker & Marilyn Walker
Natural Language and Dialogue Systems
UC Santa Cruz
1156 High St.
Santa Cruz, CA, 95064
rwswanso,becker,mawalker@ucsc.edu

Abstract

Online forums are now one of the primary venues for public dialogue on current social and political issues. The related corpora are often huge, covering any topic imaginable. Our aim is to use these dialogue corpora to automatically discover the semantic aspects of arguments that conversants are making across multiple dialogues on a topic. We frame this goal as consisting of two tasks: argument extraction and argument facet similarity. We focus here on the argument extraction task, and show that we can train regressors to predict the quality of extracted arguments with RRSE values as low as .73 for some topics. A secondary goal is to develop regressors that are topic independent: we report results of cross-domain training and domain-adaptation with RRSE values for several topics as low as .72, when trained on topic independent features.

1 Introduction

Online forums are now one of the primary venues for public dialogue on current social and political issues. The related corpora are often huge, covering any topic imaginable, thus providing novel opportunities to address a number of open questions about the structure of dialogue. Our aim is to use these dialogue corpora to automatically discover the semantic aspects of arguments that conversants are making across multiple dialogues on a topic. We build a new dataset of 109,074 posts on the topics *gay marriage*, *gun control*, *death penalty* and *evolution*. We frame our problem as consisting of two separate tasks:

- **Argument Extraction**: How can we extract argument segments in dialogue that clearly express a particular argument facet?
- **Argument Facet Similarity**: How can we recognize that two argument segments are semantically similar, i.e. about the same facet of the argument?

Parent Post **P**, Response **R**
P1: A person should be executed for kicking a dog? Your neurologically imbalanced attitude is not only worrying, it is psychopathic. How would you prove guilt on somebody who 'kicked a dog'? And, in what way, is kicking a dog so morally abhorrant as to warrant a death sentence for the given act? .... **R1**: Obviously you have issues. Any person who displays such a weakness of character cannot be allowed to contaminate the gene pool any further. Therefore, they must be put down. **If a dog bit a human, they would be put down, so why not do the same to a human?**
P2: So then **you will agree that evolution is useless in getting at possible answers on what really matters, how we got here?** If you concede that then I'm happy to end this discussion. I recall, however, visiting the Smithsonian and seeing a detailed description of how amino acids combined to form the building blocks of life. Evolutionary theory does address origins and its explanations are unsuppported by evidence. **R2**: No, and no. **First, evolution provides the only scientific answers for how humans got here: we evolved from non-human ancestors.** That record is written in both the genes and the fossils. Science might even be able eventually to tell you what the forces of selection were that propelled this evolution.
P3: Do you have any idea how little violent crime involves guns? less than 10%. the US has violance problems, how about trying to controle the violance, not the tools. **R3**: **But most murders are committed with guns.** So if you think it's important to reduce the murder rate, I don't think that guns can be ignored.
P4: Another lie used by people that want to ban guns. Guns as cars were invented to do what the owner uses them for! There is no difference in them. It takes a person to make them dangerous. **R4**: **But guns were made specifically to kill people.** Cars were made to get a person from point A to B. When someone kills a person with a car, it's an accident. When someone kills a person with a gun, it's on purpose.

Figure 1: Sample Argument Segments for Gun Control, Death Penalty and Evolution.

Consider for example the sample posts and responses in Fig. 1. Argument segments that are good targets for argument extraction are indicated, in their dialogic context, in **bold**. Given extracted segments, the argument facet similarity module should recognize that **R3** and **R4** paraphrase the same argument facet, namely that there is a strong relationship between the availability of guns and the murder rate. This paper addresses only the argument extraction task, as an important first step towards producing argument summaries that reflect the range and type of arguments being made,

Proceedings of the SIGDIAL 2015 Conference, pages 217–226,
Prague, Czech Republic, 2-4 September 2015.

on a topic, over time, by citizens in public forums.

Our approach to the argument extraction task is driven by a novel hypothesis, the IMPLICIT MARKUP hypothesis. We posit that the arguments that are good candidates for extraction will be marked by cues (implicit markups) provided by the dialog conversants themselves, i.e. their choices about the surface realization of their arguments. We examine a number of theoretically motivated cues for extraction, that we expect to be domain-independent. We describe how we use these cues to sample from the corpus in a way that lets us test the impact of the hypothesized cues.

Both the argument extraction and facet similarity tasks have strong similarities to other work in natural language processing. Argument extraction resembles the sentence extraction phase of multi-document summarization. Facet similarity resembles semantic textual similarity and paraphrase recognition (Misra et al., 2015; Boltuzic and Šnajder, 2014; Conrad et al., 2012; Han et al., 2013; Agirre et al., 2012). Work on multi-document summarization also uses a similar module to merge redundant content from extracted candidate sentences (Barzilay, 2003; Gurevych and Strube, 2004; Misra et al., 2015).

Sec. 2 describes our corpus of arguments, and describes the hypothesized markers of high-quality argument segments. We sample from the corpus using these markers, and then annotate the extracted argument segments for ARGUMENT QUALITY. Sec. 3.2 describes experiments to test whether: (1) we can predict argument quality; (2) our hypothesized cues are good indicators of argument quality; and (3) an argument quality predictor trained on one topic or a set of topics can be used on unseen topics. The results in Sec. 4 show that we can predict argument quality with RRSE values as low as .73 for some topics. Cross-domain training combined with domain-adaptation yields RRSE values for several topics as low as .72, when trained on topic independent features, however some topics are much more difficult. We provide a comparison of our work to previous research and sum up in Sec. 5.

2 Corpus and Method

We created a large corpus consisting of 109,074 posts on the topics *gay marriage* (GM, 22425 posts), *gun control* (GC, 38102 posts), *death penalty* (DP, 5283 posts) and *evolution* (EV, 43624), by combining the Internet Argument Corpus (IAC) (Walker et al., 2012), with dialogues from `http://www.createdebate.com/`.

Our aim is to develop a method that can extract high quality arguments from a large corpus of argumentative dialogues, in a topic and domain-independent way. It is important to note that arbitrarily selected utterances are unlikely to be high quality arguments. Consider for example all the utterances in Fig. 1: many utterances are either not interpretable out of context, or fail to clearly frame an argument facet. Our IMPLICIT MARKUP hypothesis posits that arguments that are good candidates for extraction will be marked by cues from the surface realization of the arguments. We first describe different types of cues that we use to sample from the corpus in a way that lets us test their impact. We then describe the MT HIT, and how we use our initial HIT results to refine our sampling process. Table 2 presents the results of our sampling and annotation processes, which we will now explain in more detail.

2.1 Implicit Markup Hypothesis

The IMPLICIT MARKUP hypothesis is composed of several different sub-hypotheses as to how speakers in dialogue may mark argumentative structure.

The **Discourse Relation** hypothesis suggests that the Arg1 and Arg2 of explicit SPECIFICATION, CONTRAST, CONCESSION and CONTINGENCY markers are more likely to contain good argumentative segments (Prasad et al., 2008). In the case of *explicit* connectives, Arg2 is the argument to which the connective is syntactically bound, and Arg1 is the other argument. For example, a CONTINGENCY relation is frequently marked by the lexical anchor *If*, as in **R1** in Fig. 1. A CONTRAST relation may mark a challenge to an opponent's claim, what Ghosh et al. call *call-out-target* argument pairs (Ghosh et al., 2014b; Maynard, 1985). The CONTRAST relation is frequently marked by *But*, as in **R3** and **R4** in Fig. 1. A SPECIFICATION relation may indicate a focused detailed argument, as marked by *First* in **R2** in Fig. 1 (Li and Nenkova, 2015). We decided to extract only the Arg2, where the discourse argument is syntactically bound to the connective, since Arg1's are more difficult to locate, especially in dialogue. We began by extracting the Arg2's for the connectives most strongly associated with these discourse relations over the whole corpus, and then once we saw what the most frequent connectives were in our corpus, we refined this selection to include only *but*, *if*, *so*, and *first*. We sampled a roughly even distribution of sentences from each category as well as sentences without any discourse connectives, i.e. *None*. See Table. 2.

The **Syntactic Properties** hypothesis posits that syntactic properties of a clause may indicate good argument segments, such as being the main clause (Marcu, 1999), or the sentential complement of mental state or speech-act verbs, e.g. the SBAR

> President Obama had tears in his eyes as he addressed the nation about the horrible tragedy.

This is of no relevance to the discussion.

> President Obama has said before that he supports renewing the assault weapons ban.

Under Connecticut law the riffle that was used in the shooting was a prohibited firearm.

> According to CNN, the killer used an AR-15 which I understand is a version of the M-16 assault riffle used in the military.

That is incorrect. The AR-15 and the M-16 share a similar appearance but they are not the same type of firearm in terms of function.

Table 1: An excerpt of a post that quotes its parent multiple times and the corresponding responses.

in *you agree that SBAR* as in **P2** in Fig. 1. Because these markers are not as frequent in our corpus, we do not test this with sampling: rather we test it as a feature as described in Sec. 3.2.

The **Dialogue Structure** hypothesis suggests that position in the post or the relation to a verbatim quote could influence argument quality, e.g. being turn-initial in a response as exemplified by **P2**, **R3** and **R4** in Fig. 1. We indicate sampling by position in post with **Starts: Yes/No** in Table. 2. Our corpora are drawn from websites that offer a "quoting affordance" in addition to a direct reply. An example of a post from the IAC corpus utilizing this mechanism is shown in Table 1, where the quoted text is highlighted in blue and the response is directly below it.

The **Semantic Density** hypothesis suggests that measures of rich content or SPECIFICITY will indicate good candidates for argument extraction (Louis and Nenkova, 2011). We initially posited that short sentences and sentences without any topic-specific words are less likely to be good. For the topics *gun control* and *gay marriage*, we filtered sentences less than 4 words long, which removed about 8-9% of the sentences. After collecting the argument quality annotations for these two topics and examining the distribution of scores (see Sec. 2.2 below), we developed an additional measure of semantic density that weights words in each candidate by its pointwise mutual information (PMI), and applied it to the *evolution* and *death penalty*. Using the 26 topic annotations in the IAC, we calculate the PMI between every word in the corpus appearing more than 5 times and each topic. We only keep those sentences that have at least one word whose PMI is above our threshold of 0.1. We determined this threshold by examining the values in *gun control* and *gay marriage*, such that at least 2/3 of the filtered sentences were in the bottom third of the argument quality score. The PMI filter eliminates 39% of the sentences from *death penalty* (40% combined with the length filter) and *85%* of the sentences from *evolution* (87% combined with the length filter).

Table 2 summarizes the results of our sampling procedure. Overall our experiments are based on 5,374 sampled sentences, with roughly equal numbers over each topic, and equal numbers representing each of our hypotheses and their interactions.

2.2 Data Sampling, Annotation and Analysis

Table 8 in the Appendix provides example argument segments resulting from the sampling and annotation process. Sometimes arguments are completely self contained, e.g. **S1** to **S8** in Table 8. In other cases, e.g. **S9** to **S16** we can guess what the argument is based on using world knowledge of the domain, but it is not explicitly stated or requires several steps of inference. For example, we might be able to infer the argument in **S14** in Table 8, and the context in which it arose, even though it is not explicitly stated. Finally, there are cases where the user is not making an argument or the argument cannot be reconstructed without significantly more context, e.g. **S21** in Table 8.

We collect annotations for ARGUMENT QUALITY for all the sentences summarized in Table 2 on Amazon's Mechanical Turk (AMT) platform. Figure 3 in the Appendix illustrates the basic layout of the HIT. Each HIT consisted of 20 sentences on one topic which is indicated on the page. The annotator first checked a box if the sentence expressed an argument, and then rated the argument quality using a continuous slider ranging from hard (0.0) to easy to interpret (1.0).

We collected 7 annotations per sentence. All Turkers were required to pass our qualifier, have a HIT approval rating above 95%, and be located in the United States, Canada, Australia, or Great Britain. The results of the sampling and annotation on the final annotated corpus are in Table 2.

We measured the inter-annotator agreement (IAA) of the binary annotations using Krippendorff's α (Krippendorff, 2013) and the continuous values using the intraclass correlation coefficient (ICC) for each topic. We found that annotators could not distinguish between phrases that *did not express an argument* and *hard* sentences. See examples and definitions in Fig. 3. We therefore mapped unchecked sentences (i.e., non arguments) to zero argument quality. We then calculated the average pairwise ICC value for each rater between all Turkers with overlapping annotations, and removed the judgements of any Turker that did not have a positive ICC value. The ICC for each topic is shown in Table 2. The mean rating across the remaining annotators for each sentence was used as the gold standard for argument quality, with means in the **Argument Quality** (**AQ**) column of Table 2. The effect of the sampling on

argument quality can be seen in Table 2. The differences between *gun control* and *gay marriage*, and the other two topics is due to effective use of the semantic density filter, which shifted the distribution of the annotated data towards higher quality arguments as we intended.

3 Experiments

3.1 Implicit Markup Hypothesis Validation

We can now briefly validate some of the IMPLICIT MARKUP hypothesis using an ANOVA testing the effect of a connective and its position in post on argument quality. Across all sentences in all topics, the presence of a connective is significant ($p = 0.00$). Three connectives, *if*, *but*, and *so*, show significant differences in AQ from no-connective phrases ($p = 0.00,\ 0.02,\ 0.00$, respectively). *First* does not show a significant effect. The mean AQ scores for sentences marked by *if*, *but*, and *so* differ from that of a no-connective sentence by 0.11, 0.04, and 0.04, respectively. These numbers support our hypothesis that there are certain discourse connectives or cue words which can help to signal the existence of arguments, and they seem to suggest that the CONTINGENCY category may be most useful, but more research using more cue words is necessary to validate this suggestion.

In addition to the presence of a connective, the dialogue structural position of being an initial sentence in a response post did not predict argument quality as we expected. Response-initial sentences provide significantly lower quality arguments ($p = 0.00$), with response-initial sentences having an average AQ score 0.03 lower (0.40 vs. 0.43).

3.2 Argument Quality Regression

We use 3 regression algorithms from the Java Statistical Analysis Toolkit[1]: Linear Least Squared Error (LLS), Ordinary Kriging (OK) and Support Vector Machines using a radial basis function kernel (SVM). A random 75% of the sentences of each domain were put into training/development and 25% into the held out test. Training involved a grid search over the hyper-parameters of each model[2] and a subset (2^3-2^9 and the complete set) of the top *N* features whose values correlate best with the argument quality dependent variable (using Pearson's). The combined set of parameters and features that achieved the best mean squared error over a 5-fold cross validation on the training data was used to train the complete model.

We also compare hand-curated feature sets that are motivated by our hypotheses to this simple feature selection method, and the performance of *in-domain*, *cross-domain*, and *domain-adaptation* training using "the frustratingly easy" approach (Daumé III, 2007).

We use our training and development data to develop a set of feature templates. The features are real-valued and normalized between 0 and 1, based on the min and max values in the training data for each domain. If not stated otherwise the presence of a feature was represented by 1.0 and its absence by 0.0. We describe all the hand-curated feature sets below.

Semantic Density Features: *Deictic Pronouns* **(DEI)**: The presence of anaphoric references are likely to inhibit the interpretation of an utterance. These features count the deictic pronouns in the sentence, such as *this*, *that* and *it*.

Sentence Length **(SLEN)**: Short sentences, particularly those under 5 words, are usually hard to interpret without context and complex linguistic processing, such as resolving long distance discourse anaphora. We thus include a single aggregate feature whose value is the number of words.

Word Length **(WLEN)**: Sentences that clearly articulate an argument should generally contain words with a high information content. Several studies show that word length is a surprisingly good indicator that outperforms more complex measures, such as rarity (Piantadosi et al., 2011). Thus we include features based on word length, including the min, max, mean and median. We also create a feature whose value is the count of words of lengths 1 to 20 (or longer).

Speciteller **(SPTL)**: We add a single aggregate feature from the result of Speciteller, a tool that assesses the specificity of a sentence in the range of 0 (least specific) to 1 (most specific) (Li and Nenkova, 2015; Louis and Nenkova, 2011). High specificity should correlate with argument quality.

Kullback-Leibler Divergence **(KLDiv)**: We expect that sentences on one topic domain will have different content than sentences outside the domain. We built two trigram language models using the Berkeley LM toolkit (Pauls and Klein, 2011). One (*P*) built from all the sentences in the IAC within the domain, excluding all sentences from the annotated dataset, and one (*Q*) built from all sentences in IAC outside the domain. The KL Divergence is then computed using the discrete n-gram probabilities in the sentence from each model as in equation (1).

$$D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \qquad (1)$$

Lexical N-Grams **(LNG)**: N-Grams are a standard feature that are often a difficult baseline to

[1] https://github.com/EdwardRaff/JSAT

[2] We used the default parameters for LLS and OK and only searched hyper-parameters for the SVM model.

Topic	Starts	Total	But	First	If	So	None	ICC	AQ
Gun Control	Yes	826	149	138	144	146	249		0.457
	No	764	149	145	147	149	174		0.500
	Total	1,590	298	283	291	295	423	0.45	0.478
Gay Marriage	Yes	779	137	120	149	148	225		0.472
	No	767	140	130	144	149	204		0.497
	Total	1,546	277	250	293	297	429	0.46	0.484
Death Penalty	Yes	399	60	17	101	100	121		0.643
	No	587	147	20	137	141	142		0.612
	Total	986	207	37	238	241	263	0.40	0.624
Evolution	Yes	609	143	49	147	138	132		0.571
	No	643	142	80	143	138	140		0.592
	Total	1,252	285	129	290	276	272	0.35	0.582

Table 2: Overview of the corpus and Argument Quality (AQ) annotation results.

beat. However they are not domain independent. We created a feature for every unigram and bigram in the sentence. The feature value was the inverse document frequency of that n-gram over all *posts* in the entire combined IAC plus `CreateDebate` corpus. Any n-gram seen less than 5 times was not included. In addition to the *specific lexical* features a set of *aggregate* features were also generated that only considered summary statistics of the lexical feature values, for example the min, max and mean IDF values in the sentence.

Discourse and Dialogue Features: We expect our features related to the discourse and dialogue hypotheses to be domain independent.

Discourse (**DIS**): We developed features based on discourse connectives found in the Penn Discourse Treebank as well as a set of additional connectives in our corpus that are related to dialogic discourse and not represented in the PDTB. We first determine if a discourse connective is present in the sentence. If not, we create a NO CONNECTIVE feature with a value of 1. Otherwise, we identify all connectives that are present. For each of them, we derive a set of *specific lexical* features and a set of generic *aggregate* features.

The *specific* features make use of the lexical (String) and PDTB categories (Category) of the found connectives. We start by identifying the connective and whether it started the sentence or not (Location). We then identify the connective's most likely PDTB category based on the frequencies stated in the PDTB manual and all of its parent categories, for example *but* → CONTRAST → COMPARISON. The *aggregate* features only consider how many discourse connectives and if any of them started the sentence. The templates are:

Specific:{Location}:{String}
Specific:{Location}:{Category}
Aggregate:{Location}:{Count}

For example, the first sentence in Table 8 would generate the following features:

Specific:Starts:but
Specific:Starts:Contrast
Specific:Starts:COMPARISON
Aggregate:Starts:1
Aggregate:Any:1

Because our hypothesis about dialogue structure was disconfirmed by the results described in section 3.1, we did not develop a feature to independently test position in post. Rather the Discourse features only encode whether the discourse cue starts the post or not.

Syntactic Property Features: We also expect syntactic property features to generalize across domains.

Part-Of-Speech N-Grams (**PNG**): Lexical features require large amounts of training data and are likely to be topic-dependent. Part-of-speech tags are less sparse and and less likely to be topic-specific. We created a feature for every unigram, bigram and trigram POS tag sequence in the sentence. Each feature's value was the relative frequency of the n-gram in the sentence.

Syntactic (**SYN**): Certain syntactic structures may be used more frequently for expressing argumentative content, such as complex sentences with verbs that take clausal complements. In `CreateDebate`, we found a number of phrases of the form **`I`** <**`VERB`**> **`that`** <**`X`**>, such as *I agree that*, *you said that*, *except that* and *I disagree because*. Thus we included two types of syntactic features: one for every internal node, excluding POS tags, of the parse tree (NODE) and another for each context free production rule (RULE) in the parse tree. The feature value is the relative frequency of the node or rule within the sentence.

Meta Features: The 3 meta feature sets are: (1) all features except lexical n-grams (**!LNG**); (2) all features that use specific lexical or categorical information (**SPFC**); and (3) aggregate statistics (**AGG**) obtained from our feature extraction process. The **AGG** set included features, such as sentence and word length, and summary statistics about the IDF values of lexical n-grams, but did not actually reference any lexical properties in the

GC	GM	DP	EV
SLEN	SLEN	LNG:penalty	LNG:⟨s⟩,**
NODE:ROOT	NODE:ROOT	LNG:death,penalty	PNG:⟨s⟩,SYM
PNG:NNS	PNG:IN	LNG:death	PNG:⟨s⟩,⟨s⟩,SYM
PNG:NN	Speciteller	LNG:the,death	LNG:**
PNG:IN	PNG:JJ	PNG:NN,NN	PNG:NNS
Speciteller	PNG:NN	NODE:NP	PNG:SYM
PNG:DT	PNG:NNS	PNG:DT,NN,NN	WLEN:Max
LNG:gun	LNG:marriage	KLDiv	WLEN:Mean
KLDiv	WLEN:Max	PNG:NN	NODE:X
PNG:JJ	PNG:DT	WLEN:7:Freq	PNG:IN

Table 3: The ten most correlated features with the quality value for each topic on the training data.

feature name. We expect both **!LNG** and **AGG** to generalize across domains.

4 Results

Sec. 4.1 presents the results of feature selection, which finds a large number of general features. The results for argument quality prediction are in Secs. 4.2 and 4.3.

4.1 Feature Selection

Our standard training procedure (**SEL**) incorporates all the feature templates described in Sec. 3.2, which generates a total of 23,345 features. It then performs a grid search over the model hyper-parameters and a subset of all the features using the simple feature selection technique described in section 3.2. Table 3 shows the 10 features most correlated with the annotated quality value in the training data for the topics *gun control* and *gay marriage*. A few domain specific lexical items appear, but in general the top features tend to be non-lexical and relatively domain independent, such as part-of-speech tags and sentence specificity, as measured by Speciteller (Li and Nenkova, 2015; Louis and Nenkova, 2011).

Sentence length has the highest correlation with the target value in both topics, as does the node:root feature, inversely correlated with length. Therefore, in order to shift the quality distribution of the sample that we put out on MTurk for the *death penalty* or *evolution* topics, we applied a filter that removed all sentences shorter than 4 words. For these topics, domain specific features such as lexical n-grams are better predictors of argument quality. As discussed above, the PMI filter that was applied only to these two topics during sampling removed some shorter low quality sentences, which probably altered the predictive value of this feature in these domains.

4.2 In-Domain Training

We first tested the performance of 3 regression algorithms using the training and testing data within each topic using 3 standard evaluation measures: R^2, Root Mean Squared Error (RMSE) and Root Relative Squared Error (RRSE). R^2 estimates the amount of variability in the data that is explained by the model. Higher values indicate a better fit to the data. The RMSE measures the average squared difference between predicted values and true values, which penalizes wrong answers more as the difference increases. The RRSE is similar to RMSE, but is normalized by the squared error of a simple predictor that always guesses the mean target value in the test set. Anything below a 1.0 indicates an improvement over the baseline.

Topic	Reg	# Feats	R^2	RMSE	RRSE
GC	LLS	64	0.375	0.181	0.791
GC	OK	ALL	0.452	0.169	0.740
GC	SVM	512	**0.466**	**0.167**	**0.731**
GM	LLS	64	0.401	0.182	0.774
GM	OK	ALL	**0.441**	**0.176**	**0.748**
GM	SVM	256	0.419	0.179	0.762
DP	LLS	16	**0.083**	0.220	0.957
DP	OK	ALL	0.075	**0.221**	0.962
DP	SVM	ALL	0.079	**0.221**	**0.960**
EV	LLS	ALL	0.016	0.236	0.992
EV	OK	ALL	0.114	0.224	0.941
EV	SVM	ALL	**0.127**	**0.223**	**0.935**

Table 4: The performance of in domain training for three regression algorithms.

Table 4 shows that SVMs and OK perform the best, with better than baseline results for all topics. Performance for *gun control* and *gay marriage* are significantly better. See Fig. 2. Since SVM was nearly always the best model, we only report SVM results in what follows.

We also test the impact of our theoretically motivated features and domain specific features. The top half of Table 5 shows the RRSE for each feature set with darker cells indicating better performance. The feature acronyms are described in Sec 3.2. When training and testing on the same domain, using lexical features leads to the best performance for all topics (SEL, LEX, LNG and SPFC). However, we can obtain good performance on all of the topics without using any lexical information at all (!LNG, WLEN, PNG, and AGG), sometimes close to our best results. Despite the high correlation to the target value, sentence specificity as a single feature does not outperform any other feature sets. In general, we do better for *gun control* and *gay marriage* than for *death penalty* and *evolution*. Since the length and domain specific words are important features in the trained models, it seems likely that the filtering process made it harder to learn a good function.

The bottom half of Table 5 shows the results using training data from all other topics, when testing on one topic. The best results for GC are significantly better for several feature sets (**SEL,**

Topic	SEL	LEX	LNG	!LNG	SPTL	SLEN	WLEN	SYN	DIS	PNG	SPFC	AGG
GC	0.73	0.75	0.79	0.79	0.94	0.87	0.93	0.83	0.99	0.80	0.75	0.85
GM	0.76	0.75	0.79	0.81	0.95	0.89	0.91	0.87	0.99	0.83	0.77	0.82
DP	0.96	0.95	0.95	0.99	1.02	1.01	0.98	1.01	1.03	0.98	0.96	0.98
EV	0.94	0.92	0.93	0.96	1.00	0.99	0.99	1.00	1.00	0.96	0.94	0.96
GC^{ALL}	0.74	0.72	0.75	0.81	0.96	0.90	0.94	1.03	0.90	0.82	0.75	0.84
GM^{ALL}	0.72	0.74	0.78	0.79	0.96	0.91	0.92	1.03	0.91	1.02	0.74	0.83
DP^{ALL}	0.97	0.97	1.01	0.98	1.05	1.02	0.98	1.03	1.02	1.03	0.97	0.99
EV^{ALL}	0.93	0.94	0.96	0.97	1.02	1.04	0.98	1.01	1.04	1.01	0.93	0.96

Table 5: The RRSE for in-domain training on each of the feature sets. Darker values denote better scores. **SEL**=Feature Selection, **LEX**=Lexical, **LNG**=Lexical N-Grams, **!LNG**=Everything but LNG, **SPTL**=Speciteller, **SLEN**=Sentence Length, **WLEN**=Word Length, **SYN**=Syntactic, **DIS**=Discourse, **PNG**=Part-Of-Speech N-Grams, **SPFC**=Specific, **AGG**=Aggregate. XX^{ALL} indicates training on data from all topics and testing on the XX topic.

LEX, LNG), In general the performance remains similar to the in-domain training, with some minor improvements over the best performing models. These results suggest that having more data outweighs any negative consequences of domain specific properties.

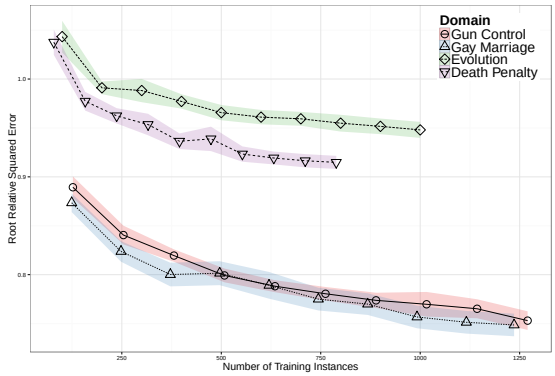

Figure 2: Learning curves for each of the 4 topics with 95% confidence intervals.

We also examine the effect of training set size on performance given the best performing feature sets. See Fig. 2. We randomly divided our entire dataset into an 80/20 training/testing split and trained incrementally larger models from the 80% using the default training procedure, which were then applied to the 20% testing data. The plotted points are the mean value of repeating this process 10 times, with the shaded region showing the 95% confidence interval. Although most gains are achieved within 500-750 training examples, all models are still trending downward, suggesting that more training data would be useful.

Finally, our results are actually even better than they appear. Our primary application requires extracting arguments at the *high* end of the scale (e.g., those above 0.8 or 0.9), but the bulk of our data is closer to the middle of the scale, so our regressors are conservative in assigning high or low values. To demonstrate this point we split the predicted values for each topic into 5 quantiles. The RMSE for each of the quantiles and domains in Table 6 demonstrates that the lowest RMSE is obtained in the top quantile.

%ile	GC	GM	DP	EV
0.2	**0.162**	**0.171**	0.237	**0.205**
0.4	0.184	0.201	0.238	0.242
0.6	0.198	0.181	0.225	0.211
0.8	0.166	0.176	**0.178**	0.208
1.0	**0.111**	**0.146**	**0.202**	**0.189**
ALL	0.167	0.176	0.217	0.220

Table 6: The RMSE for the best performing model in each domain given instances whose predicted quality value is in the given percentile.

4.3 Cross-Domain and Domain Adaptation

To investigate whether learned models generalize across domains we also evaluate the performance of training with data from one domain and testing on another. The columns labeled **CD** in Table 7 summarize these results. Although cross domain training does not perform as well as in-domain training, we are able to achieve much better than baseline results between *gun control* and *gay marriage* for many of the feature sets and some other minor transferability for the other domains. Although lexical features (e.g., lexical n-grams) perform best in-domain, the best performing features across domains are all non-lexical, i.e. **!LNG**, **PNG** and **AGG**.

We then applied Daume's "frustratingly easy domain adaptation" technique (**DA**), by transforming the original features into a new augmented feature space where, each feature, is transformed into a *general* feature and a domain specific feature, *source* or *target*, depending on the input domain (Daumé III, 2007). The training data from both the source and target domains are used to train

		SEL		LNG		!LNG		SPTL		DIS		PNG		AGG	
SRC	**TGT**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**
GC	GM	0.84	0.75	1.00	0.82	0.84	0.94	0.96	0.80	1.01	0.85	0.85	0.76	0.88	0.82
GC	DP	1.13	0.94	1.30	0.97	1.04	1.01	1.13	0.96	1.09	1.02	1.11	0.94	1.08	0.97
GC	EV	1.10	0.92	1.29	0.98	1.05	1.01	1.08	0.97	1.07	0.98	1.09	0.92	1.02	0.96
GM	GC	0.82	0.74	0.96	0.79	0.82	0.94	0.94	0.78	0.99	0.82	0.81	0.74	0.88	0.85
GM	DP	1.13	0.93	1.28	0.97	1.08	1.02	1.11	0.96	1.12	1.01	1.09	0.95	1.07	0.96
GM	EV	1.07	0.93	1.27	0.98	1.03	1.01	1.06	0.96	1.07	0.98	1.02	0.93	1.02	0.96
DP	GC	1.06	0.75	1.01	0.80	1.14	0.96	1.25	0.79	1.28	0.82	1.10	0.74	1.13	0.85
DP	GM	1.04	0.75	1.00	0.83	1.10	0.96	1.23	0.81	1.27	0.87	1.09	0.77	1.10	0.81
DP	EV	0.97	0.91	1.00	0.95	1.00	1.01	1.05	0.95	1.05	1.00	1.00	0.93	0.99	0.96
EV	GC	0.97	0.74	0.97	0.80	1.02	0.95	1.05	0.80	1.13	0.83	1.02	0.74	0.91	0.85
EV	GM	0.96	0.75	0.99	0.82	0.98	0.95	1.04	0.81	1.13	0.87	1.01	0.76	0.91	0.82
EV	DP	1.04	0.95	1.07	0.98	1.01	1.00	1.00	0.98	1.00	1.00	1.00	0.96	1.01	0.98

Table 7: The RRSE for cross-domain training (CD) and with domain adaptation (DA).

the model, unlike the cross-domain experiments where only the source data is used. These results are given in the columns labeled **DA** in Table 7, which are on par with the best in-domain training results, with minor performance degradation on some *gay marriage* and *gun control* pairs, and slight improvements on the difficult *death penalty* and *evolution* topics.

5 Discussion and Conclusions

This paper addresses the **Argument Extraction** task in a framework whose long-term aim is to first extract arguments from online dialogues, and then use them to produce a summary of the different facets of an issue. We have shown that we can find sentences that express clear arguments with RRSE values of .72 for gay marriage and gun control (Table 6) and .93 for death penalty and evolution (Table 8 cross domain with adaptation). These results show that sometimes the best quality predictors can be trained in a domain-independent way.

The two step method that we propose is different than much of the other work on argument mining, either for more formal texts or for social media, primarily because the bulk of previous work takes a supervised approach on a labelled topic-specific dataset (Conrad et al., 2012; Boltuzic and Šnajder, 2014; Ghosh et al., 2014b). Conrad & Wiebe developed a data set for supervised training of an argument mining system on weblogs and news about universal healthcare. They separate the task into two components: one component identifies ARGUING SEGMENTS and the second component labels the segments with the relevant ARGUMENT TAGS. Our argument extraction phase has the same goals as their first component. Boltuzic & Snajder also apply a supervised learning approach, producing arguments labelled with a concept similar to what we call FACETS. However they perform what we call argument extraction by hand, eliminating comments from comment streams that they call "spam" (Boltuzic and Šnajder, 2014). Ghosh et al. also take a supervised approach, developing techniques for argument mining on online forums about technical topics and applying a theory of argument structure that is based on identifying TARGETS and CALLOUTS, where the callout attacks a target proposition in another speaker's utterance (Ghosh et al., 2014b). However, their work does not attempt to discover high quality callouts and targets that can be understood out of context like we do. More recent work also attempts to do some aspects of argument mining in an unsupervised way (Boltuzic and Šnajder, 2015; Sobhani et al., 2015). However (Boltuzic and Šnajder, 2015) focus on the argument facet similarity task, using as input a corpus where the arguments have already been extracted. (Sobhani et al., 2015) present an architecture where arguments are first topic-labelled in a semi-supervised way, and then used for stance classification, however this approach treats the whole comment as the extracted argument, rather than attempting to pull out specific focused argument segments as we do here.

A potential criticism of our approach is that we have no way to measure the recall of our argument extraction system. However we do not think that this is a serious issue. Because we are only interested in determining the similarity between phrases that are high quality arguments and thus potential contributors to summaries of a specific facet for a specific topic, we believe that precision is more important than recall at this point in time. Also, given the redundancy of the arguments presented over thousands of posts on an issue it seems unlikely we would miss an important facet. Finally, a measure of recall applied to the facets of a topic may be irreconcilable with our notion that an argument does not have a limited, enumerable number of facets, and our belief that each facet is subject to judgements of granularity.

6 Appendix

Fig. 3 shows how the Mechanical Turk hit was defined and the examples that were used in the qualification task. Table 8 illustrates the argument quality scale annotations collected from Mechanical Turk.

We invite other researchers to improve upon our results. Our corpus and the relevant annotated data is available at `http://nldslab.soe.ucsc.edu/arg-extraction/sigdial2015/`.

7 Acknowledgements

This research is supported by National Science Foundation Grant CISE-IIS-RI #1302668.

References

E. Agirre, M.Diab, D. Cer, and A. Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proc. of the Sixth Int. Workshop on Semantic Evaluation*, pp. 385–393. ACL.

R. Barzilay. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University.

F. Boltuzic and J. Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proc. of the First Workshop on Argumentation Mining*, pp. 49–58.

F. Boltuzic and J. Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proc. of the Second Workshop on Argumentation Mining*.

A. Conrad, J. Wiebe, and R. Hwa. 2012. Recognizing arguing subjectivity and argument tags. In *Proc. of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pp. 80–88. ACL.

H. Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, June.

D. Ghosh, S. Muresan, N. Wacholder, M. Aakhus, and M. Mitsui. 2014b. Analyzing argumentative discourse units in online interactions. *ACL 2014*, p. 39.

I. Gurevych and M. Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proc. of the 20th Int. conference on Computational Linguistics*, pp. 764–771. ACL.

L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese. 2013. Umbc ebiquity-core: Semantic textual similarity systems. *Atlanta, Georgia, USA*, p. 44.

K. Krippendorff. 2013. *Content analysis: an introduction to its methodology*. Sage, Los Angeles [etc.].

J. J Li and A.Nenkova. 2015. Fast and Accurate Prediction of Sentence Specificity. In *Proc. of the Twenty-Ninth Conf. on Artificial Intelligence (AAAI)*, January.

A. Louis and A. Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proc. of 5th Int. Joint Conf. on Natural Language Processing*, pp. 605–613.

D. Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, pp. 123–136.

D. W. Maynard. 1985. How Children Start Arguments. *Language in Society*, 14(1):1–29, March.

A. Misra, P. Anand, J. E. Fox Tree, and M.A. Walker. 2015. Using summarization to discover argument facets in dialog. In *Proc. of the 2015 Conf. of the North American Chapter of the ACL: Human Language Technologies*.

A. Pauls and D. Klein. 2011. Faster and Smaller N-gram Language Models. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 258–267, Stroudsburg, PA, USA. ACL.

S.T. Piantadosi, H. Tily, and E. Gibson. 2011. Word lengths are optimized for efficient communication. *Proc. of the National Academy of Sciences*, 108(9):3526–3529, March.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proc. of the 6th Int. Conf. on Language Resources and Evaluation (LREC 2008)*, pp. 2961–2968.

P. Sobhani, D. Inkpen, and S. Matwin. 2015. From argumentation mining to stance classification. In *Proc. of the Second Workshop on Argumentation Mining*.

M.A. Walker, P. Anand, R. Abbott, and J. E. Fox Tree. 2012. A corpus for research on deliberation and debate. In *Language Resources and Evaluation Conf. , LREC2012*.

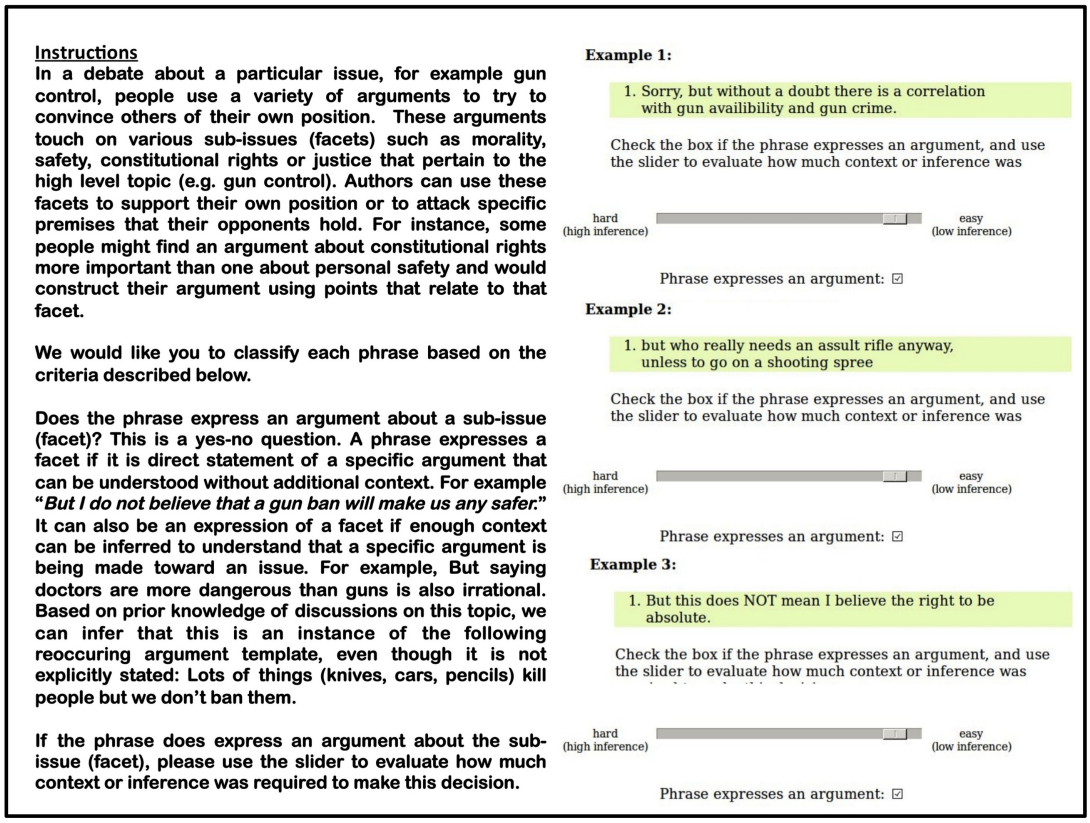

Figure 3: Argument Clarity Instructions and HIT Layout.

ID	Topic	Argument Quality	Sentence
S1	GC	0.94	But guns were made specifically to kill people.
S2	GC	0.93	If you ban guns crime rates will not decrease.
S3	GM	0.98	If you travel to a state that does not offer civil unions, then your union is not valid there.
S4	GM	0.92	Any one who has voted yes to place these amendments into state constitutions because they have a religious belief that excludes gay people from marriage has also imposed those religious beliefs upon gay people.
S5	DP	0.98	The main reasons I oppose the death penalty are: #1) It is permanent.
S6	DP	0.97	If a dog bit a human, they would be put down, so why no do the same to a human?
S7	EV	0.97	We didn't evolve from apes.
S8	EV	0.95	Creationists have to pretty much reject most of science.
S9	GC	0.57	IF they come from the Constitution, they're not natural... it is a statutory right.
S10	GC	0.52	This fear is doing more harm to the gun movement than anything else.
S11	GM	0.51	If it seems that bad to you, you are more than welcome to leave the institution alone.
S12	GM	0.50	Nobody is trying to not allow you to be you.
S13	DP	0.52	Why isn't the death penalty constructive?
S14	DP	0.50	But lets say the offender decides to poke out both eyes?
S15	EV	0.51	so no, you don't know the first thing about evolution.
S16	EV	0.50	But was the ark big enough to hold the number of animals required?
S17	GC	0.00	Sorry but you fail again.
S18	GC	0.00	Great job straight out of the leftard playbook.
S19	GM	0.00	First, I AIN'T your honey.
S20	GM	0.00	There's a huge difference.
S21	DP	0.03	But as that's not likely to occur, we fix what we can.
S22	DP	0.01	But you knew that, and you also know it was just your try to add more heat than light to the debate.
S23	EV	0.03	marc now resorts to insinuating either that I'm lying or can't back up my claims.
S24	EV	0.00	** That works for me.

Table 8: Example sentences in each topic domain from different sections of the quality distribution.