Argumentation in Dialogues

Yiyi Chen
Computational Argumentation

Demo

- Kialo
  - Death penalty
    https://www.kialo.com/should-the-death-penalty-be-abolished-28302
  - 

- CreateDebate
  - https://www.createdebate.com/
  - 

Tasks

- **Argument Extraction**
  - How can we extract argument segments in dialogues that clearly express a particular argument facet? (such as morality, Second Amendment)

- Argument Facet Similarity
  - How can we recognize that two argument segments are semantically similar, i.e. about the same facet of the argument ?

Data - Internet Argument Corpus (IAC)

- Source : **4forums.com** (a website for political debate and discourse)
- Entire corpus: 390,704 posts in 11,800 discussions (aka threads) by 3,317 authors.
- Import features:
  - CONTEXTUAL AFFORDANCES: A mechanism for quoting another posts (72.3% of all posts contain at least one quote) ->
    - Q-R pair (10,003)
    - P123 (6,797 chains of three posts defined as series P1, P2 and P3 such that P3 is a response to P2 which itself is a response to P1)
  - Additional AFFORDANCES
    - Title, reference URL, breadcrumbs indicating the sub-forum it belongs to
    - Poll information, reply structure, links, formatting, quotes , etc.

(QUOTE: 1,2)

Implicit Markup Hypotheses

- 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.
- **Dialogue Structure**
  - **The position in the post could influence argument quality.**
    - **Indicate sampling by position in post with Starts: Yes/No.**
  - **The quoting affordance of some sites should mirror the target-callout mechanism.**
  - **Example:**
    - **Q**
    -
    - **R**
  -
    - **Q**
    -
    - **R**

> 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.

Implicit Markup Hypotheses

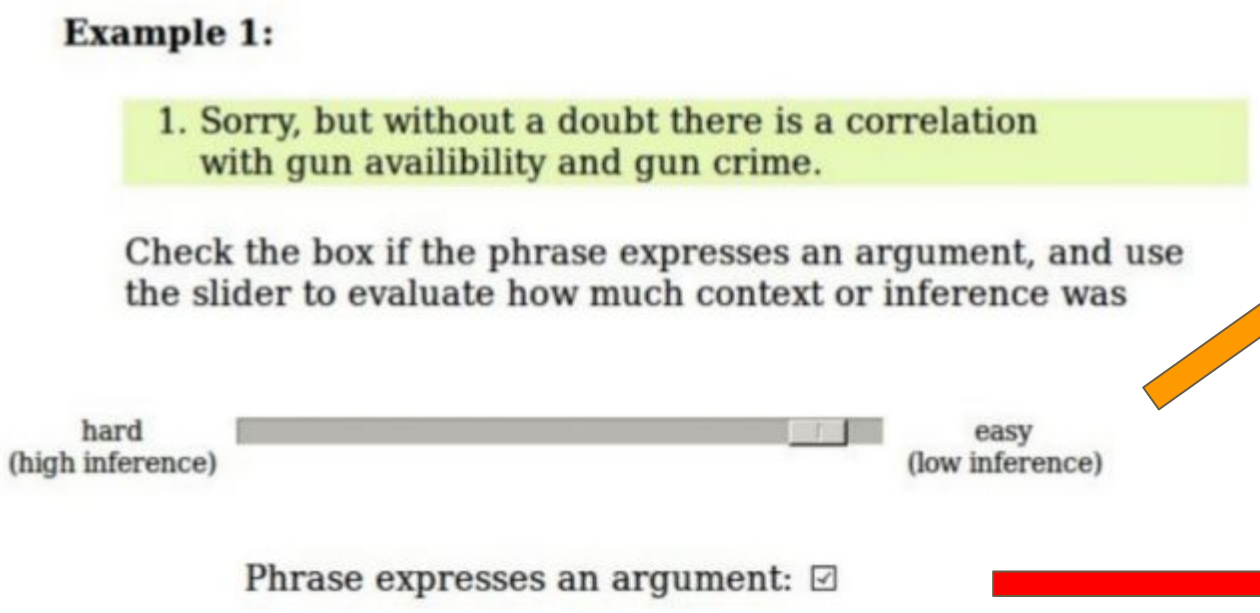
- **Discourse Relation**
  - The Arg1 and Arg2 of explicit SPECIFICATION, CONTRAST, CONCESSION and CONTINGENCY markers are more likely to contain good argumentative segments (Prasad et al.2008)
  - In the case of explicit connectives, Arg2 is the argument to which the connective is syntactically bound, and Arg1 is the other argument.
  - **CONTINGENCY** (***If***)
    - **If** a dog bit a human, they would be put down, so why not do the same to a human? (a response to argue for death penalty)
  - **CONTRAST** (mark a challenge to an opponent's claim, ***But***)
    - **But** most murders are committed with guns. (a response to argue for gun control)
    - **But** guns were made specifically to kill people.
  - **SPECIFICATION** (indicate a focused detailed argument, ***First***)
    - **First**, evolution provides the only scientific answers for how humans got here: we evolved from non-human ancestors. (a response to argue for evolution)

Implicit Markup Hypotheses

- **Syntactic Properties**
  - Syntactic properties of a clause may indicate good argument segments, such as being the main clause, or the sentential complement of mental state or speech-act verbs, e.g. the SBAR
  - Example:
    - You will agree **that** evolution is useless in getting at possible answers on what really matters, how we got here? (a parent post for arguing against evolution)
- **Semantic Density**
  - Measures of rich content or SPECIFICITY indicate good candidates for argument extraction.
    - Short sentences and sentences without any topic-specific words are not not likely to be good. => filtered sentences less than 4 words long
    - PMI (pointweise mutual information) - calculate PMI between every word in the corpus appearing more than 5 times and each topic. => only keep those sentences that have at least one word whose PMI is above 0.1

Argument Quality

- Asked 7 annotators to rate how clearly a sentence expresses an argument towards a position



Rate the argument quality using a continuous slider ranging from hard(0.0) to easy to interpret(1.0)

Check if the sentence expressed an argument.
Not checked => AQ=0.0

Argument Clarity Instructions and HIT Layout.

(Quote :3)

Argument Quality - some examples

ID	Topic	AQ	Sentence
S1	GC	0.94	But guns were made specifically to kill people.
S3	GM	0.98	If you travel to a state that does not offer civil unions, then your union is not valid there.
S4	GC	0.57	IF they come from the constitution, they're not natural… It is a statutory right.
S5	EV	0.51	So no, you don't know the first thing about evolution.
S6	GC	0.00	Sorry, but you fail again.

Completely self-contained arguments.

Not explicitly stated
Or requires several steps of inference.

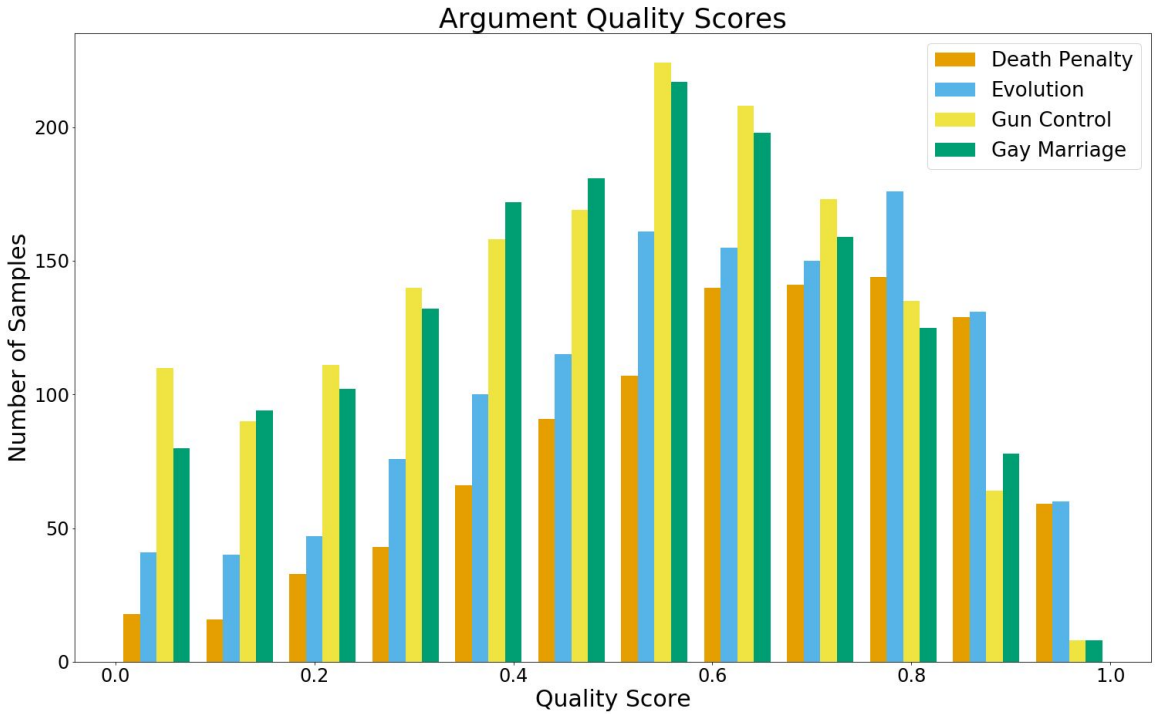
Needs significantly more context

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

(Quote: 3)

Argument Quality Distribution

- The AQ scores are roughly evenly distributed across each topic
- The data is refined accordingly in the more recent extended experiment [4]



Data

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.

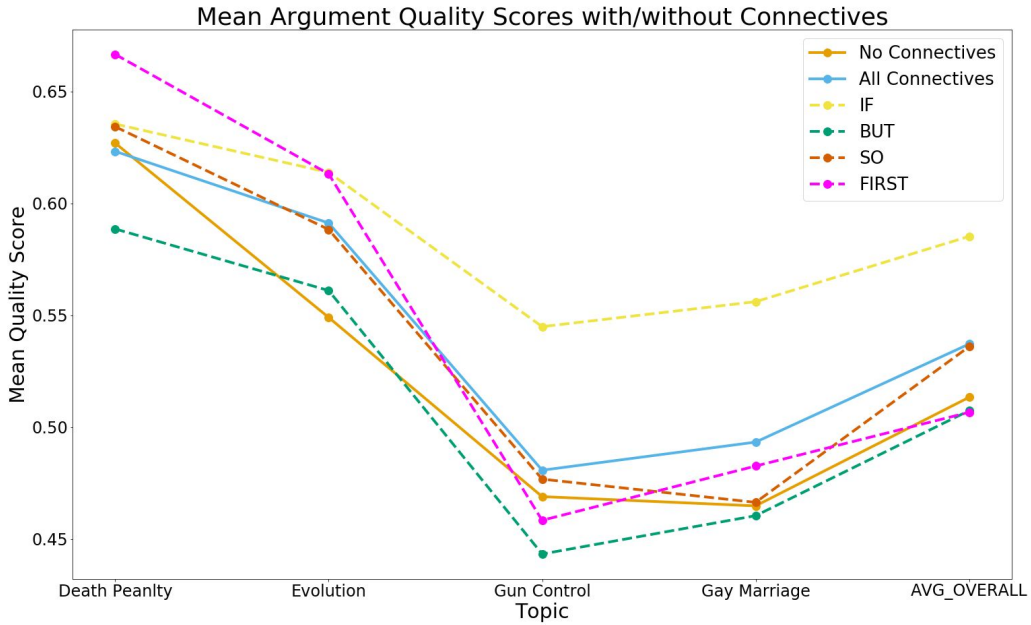
Preliminary Validation of Markup Hypotheses

- ANOVA to test the effect of variables on argument quality
  - Sentences beginning with **if, but, so**, or **First** are significantly better than those with no connective
    -

Connective	p-Value	Difference
if	0.00	+0.11
but	0.01	+0.04
so	0.00	+0.04
first	> 0.05	NS

(Quote: 3)

Implicit Markup Hypothesis Validation



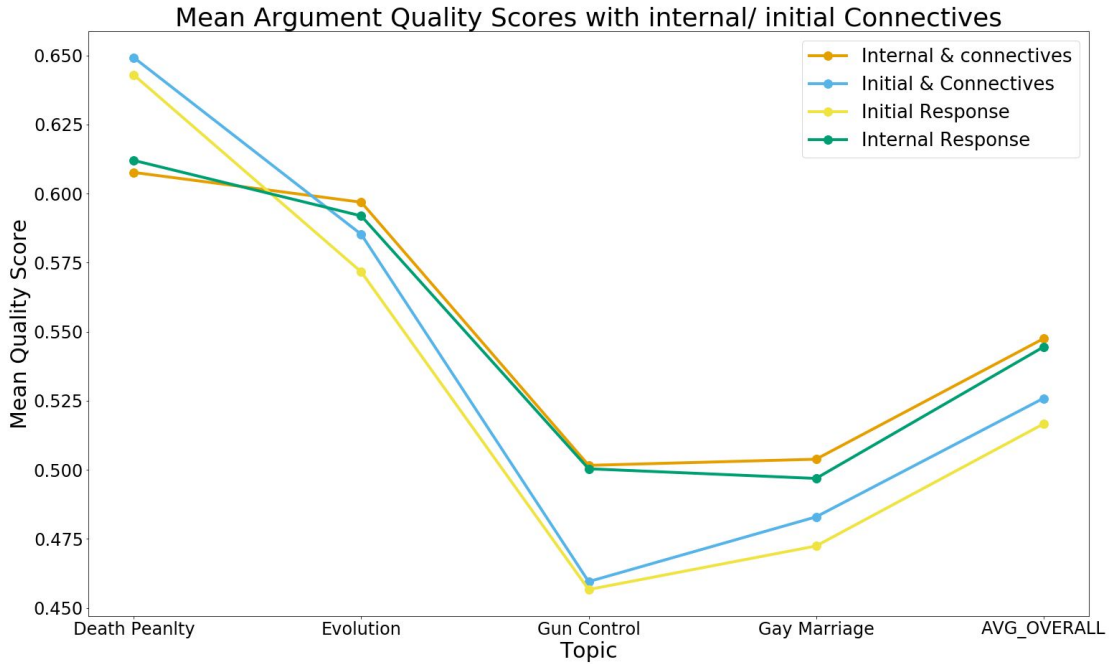
- Overall, the sentences with connectives (if, so, first) shows significant difference in AQ from no-connective.
- In average, all the connectives has 0.02 more in AQ than no connectives.
-
- In average,
  - IF (+ 0.07)
  - SO (+ 0.022)

Preliminary Validation of Markup Hypotheses

- ANOVA to test the effect of variables on argument quality
  - The position of a sentence predicts the opposite of our expectation (p=0.0)

Position	Avg.Quality
Response Initial	0.40
Response Internal	0.43

Implicit Markup Hypothesis Validation

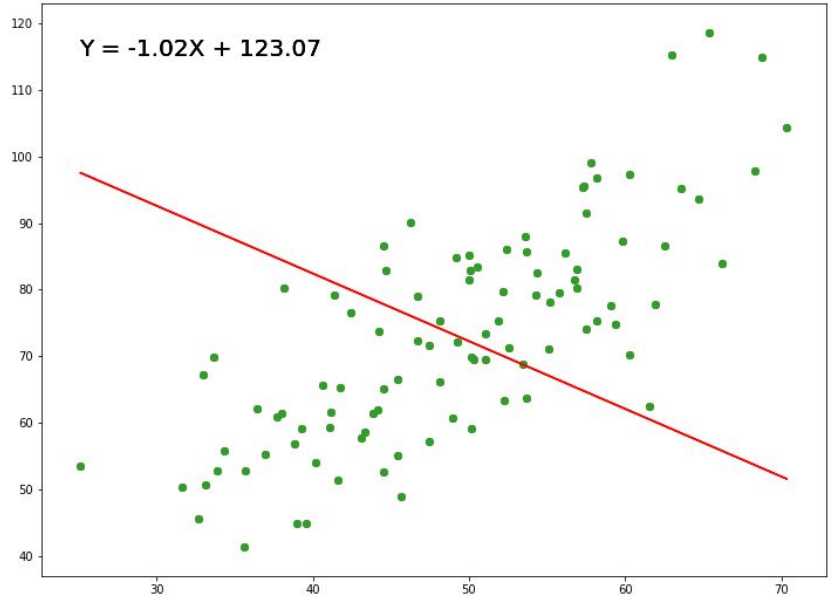


- The dialogue structural position of being an initial sentence in a response did not predict argument quality as expected.
- On the contrary, response-initial sentences provide lower quality arguments
  - Internal & connective (0.547)
  - Internal overall (0.544)
  - Initial & Connective (0.525)
  - Initial overall (0.516)

Experiments- Argument Quality Regression

- Treat the task as a regression problem
  - Linear Least Squared Errors (LLS)
    - A linear approach to modelling the relationship between a dependent variable and one or more independent variables.
  - Ordinary Kriging (OK)
    - A spatial estimation method where the error variance is minimized. [5]
  - Support Vector Machine (SVM)
    - A discriminative classifier formally defined by a separating hyperplane.
    - Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.
- Extend the set of predictive features
- Evaluate the ability of features to generalize across domains

LLS - Linear Regressing using Least Squared Error



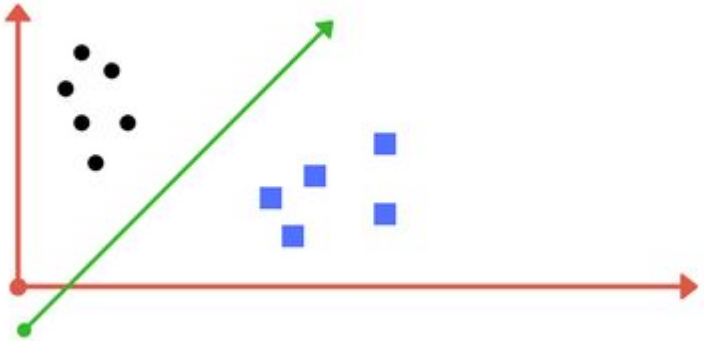
- Y = mX + c
- Use this equation to train the model with a given dataset and predict the value of Y for any given value of X.
- Determine the value of m and c, that gives the minimum error for the given dataset
- Using least squares method

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

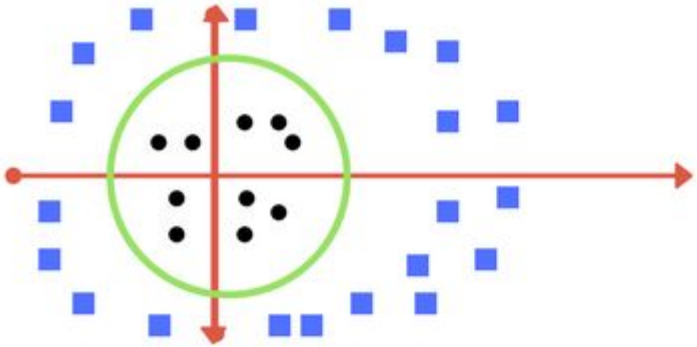
$$c = \bar{y} - m\bar{x}$$

(Quote : 6)

SVM (support vector machine)



Sample cut to divide into two classes.

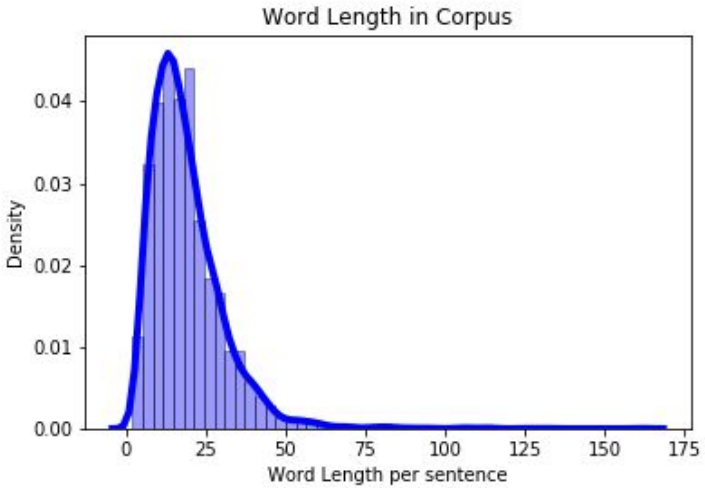


Transforming back to x-y plane, a line transforms to circle.

(Quote: 7)

Features

- **Semantic Density Features**
  - *Deictic Pronouns* (reference must be fixed through the context of the utterance)
    - But the way i look at it is that whatever one person does to another, especially if it causes death or extreme hospital bills, they deserve to have it done to them as well.
  - *Sentence Length* (short sentences are usually hard to interpret without context and complex linguistic processing)
    - “So what about murderers?” (death penalty)
    - “But the materialist (!)” (Evolution)



Features

- **Semantic Density Features**
  - Lexical N-grams (for every unigram and bigram)
    - Exclude any n-gram seen less than 5 times
  - Speciteller
    - Result from Speciteller
    - Assess the specificity of a sentence
    - 0 (least specific) to 1 (most specific)

(Quote : 8)

Newly labeled general sentences	Newly labeled specific sentences
1. Edberg was troubled by inconsistent serves. 2. Demands for Moeller's freedom have been a feature of leftist demonstrations for years. 3. But in a bizarre bit of social engineering, U.S. occupation forces instructed Japanese filmmakers to begin showing on-screen kisses. 4. Although many of the world's top track and field stars are Americans, the sport has suffered from a lack of exposure and popularity in the United States.	1. Shipments fell 0.7 percent in September. 2. Indian skipper Mohammed Azharuddin won the toss and decided to bat first on a slow wicket. 3. He started this week as the second-leading rusher in the AFC with 1,096 yards, just 5 yards behind San Diego's Natrone Means. 4. The other two, Lt. Gen. Cedras and Brig. Gen. Philippe Biamby, resigned and fled into self-imposed exile in Panama two days before Aristide's U.S.-backed homecoming on Oct. 15.

Table 2: Examples of general and specific sentences newly labeled during the co-training procedure.

Features

- *Kullback-Leibler Divergence:*
  - It is a measure of dissimilarity or distance between distributions
  - Sentences on one topic domain will have different content than sentences outside the domain.
- **Discourse and Dialogue Features (domain-independent)**
  - **Discourse:**
    - Connectives
    - Whether it starts the sentence or not

Features

- **Syntactic Property Feature (across domain)**
  - **Part-of-Speech N-Grams (PNG)**
  - **Syntactic**
    - **I <verb> that <X>**
      - **I agree that,**
      - **You said that**
      - **I disagree because**
  - **Meta Features**
    - All non lexical features
    - All features that use directly use lexical tokens
    - Aggregate statistics (sentence length, word length, summary statistics)

Results - Feature Selection

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.

- For GC and GM, sentence length has the highest correlation with the target value.
- =>
  - Remove all sentences shorter than 4 in DP and EVO.
  - PMI also only applied to DP and EVO.

In-Domain Training

Measures:

- Support vector machines
- Best features using simple features selection techniques
- Grid Search for best parameters on training set
- Performance on test set

Topics	# Features	R^2	RMSE	RRSE
Gun Control	512	0.466	0.167	0.731
Gay Marriage	256	0.419	0.179	0.762
Death Penalty	ALL	0.079	0.221	0.960
Evolution	ALL	0.127	0.223	0.935

Comparison of Features for In-Domain Training

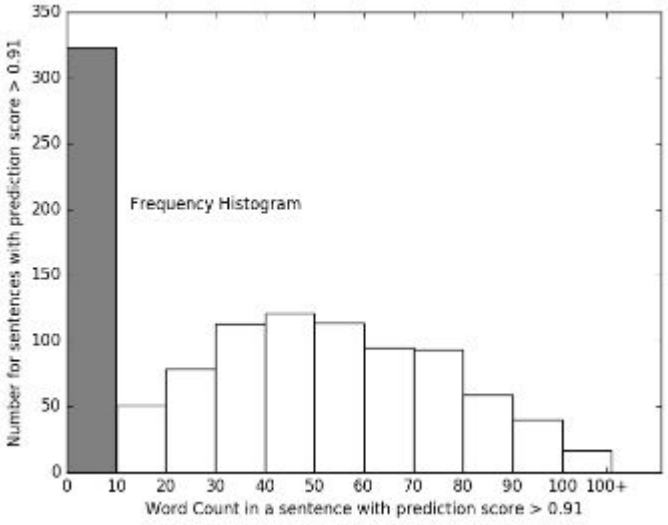
- Lexical features (SEL, LEX, LNG, SPEC) work the best
  - 0.75 RRSE for gun control and gay marriage
- Still do very well without lexical n-grams
  - 0.79 for gun control and 0.81for gay marriage
- Speciteller is highly correlated but not a good predictor alone
  - 0.75 for gun control and 0.77 for gay marriage
- Simple features like sentence length do surprisingly well
  - 0.87 for gun control and 0.89 for gay marriage (
- Since the length and domain specific words are important features in the trained models, the filtering process for DP and EV might make it harder to learn a good function.

Conclusion

- **Argument Extraction**
  - **Features significantly outperform the baseline**
  - **The predicted high quality sentences look good qualitatively**
- **Discourse connectives and sentence position were not as strong a predictor as expected**
  - **Possible explanations**
    - **Anaphora make some sentences hard to interpret in isolation**
    - **Connectives may signal an argument is near, but not necessarily in the same sentence**
    - **Identifying both arguments to a connective is difficult and is not well studied in dialogic text (in this work it only extracted Arg2)**

Applying the regressor - refine Argument Quality[4]

- New scoring criteria
- Filtering with high PMI n-grams might result in diversity issues (that many sentences given high AQ scores were very similar)



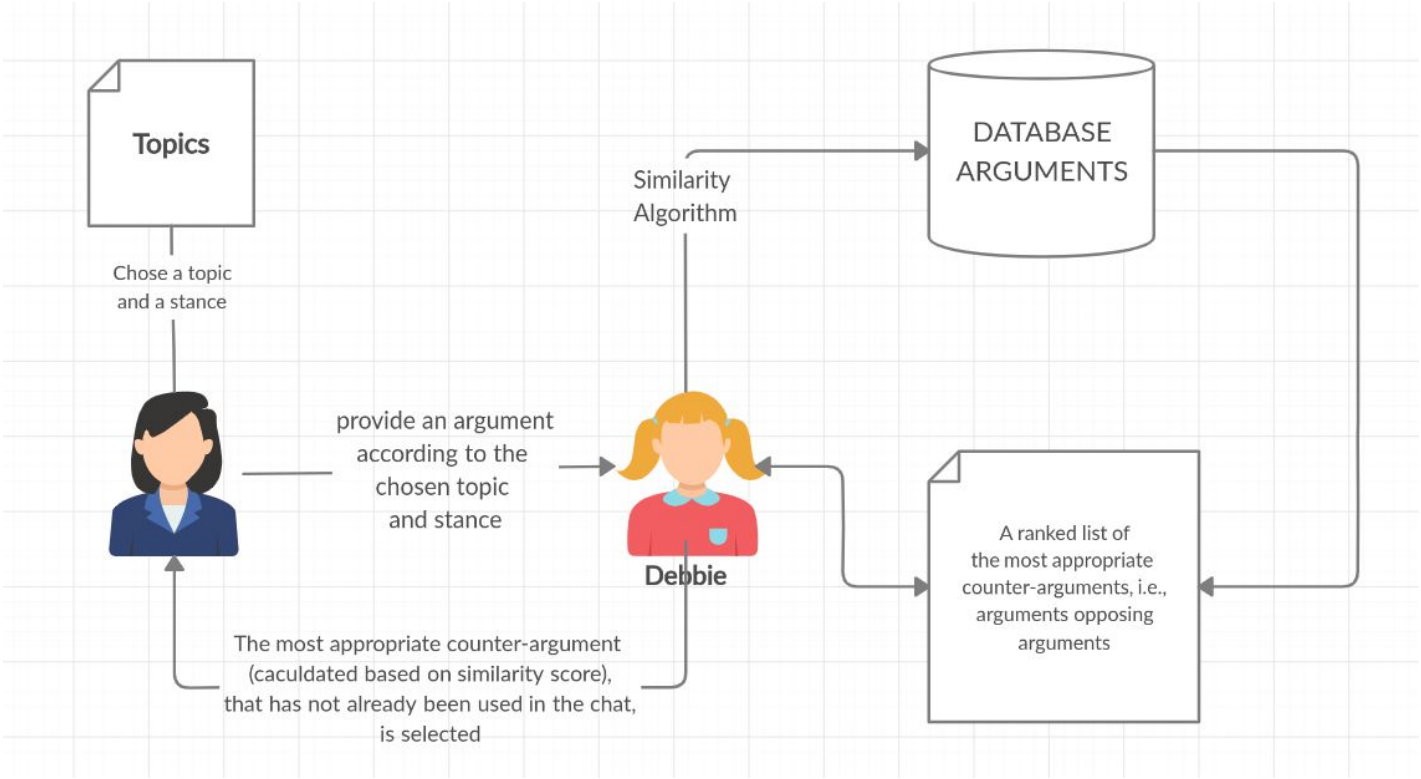
- Word count distribution for argument quality prediction scores >0.91
- The first bin shows that many sentences with less than 10 words are predicted to be high quality, but many of these sentences consisted of only few elongated words (e.g. HAHAHA..)
- The upper part of the distribution with more than 70 words => multiple sentences without punctuation
- Refine: deduplicate sentences, rescoring sentences without a verb and with less than 4 words to AQ=0, restrict samples to between 10 and 40 tokens

Debbie Debate Bot - Data

- Start from AQ regressor from [3], which predicts a quality score for each sentence, AQ>0.55, diversity and quality in the argument
- Keep only stance bearing statements from the IAC [1,2]
- Evaluate the prototype with hand labeled 2000 argument quality sentence pairs for the topic of DP from [4]
- Test the model for both appropriateness of responses and response times.

(Quote: 9)

Debbie Debate Bot - Workflow



Debbie Debate Bot - speedup

- Generating Clusters
  - Generate a distance matrix of similarity scores (between 0 and 1) for each topic and stance.
  - Using agglomerative clustering from scikit-learn => 15 clusters
  - Identify the head of a cluster - the argument within each cluster, that best represents all of the statements within the cluster
    - Finding the average distance of each statement in the cluster to all the statements in the cluster and chose the one with the minimum average as the head.

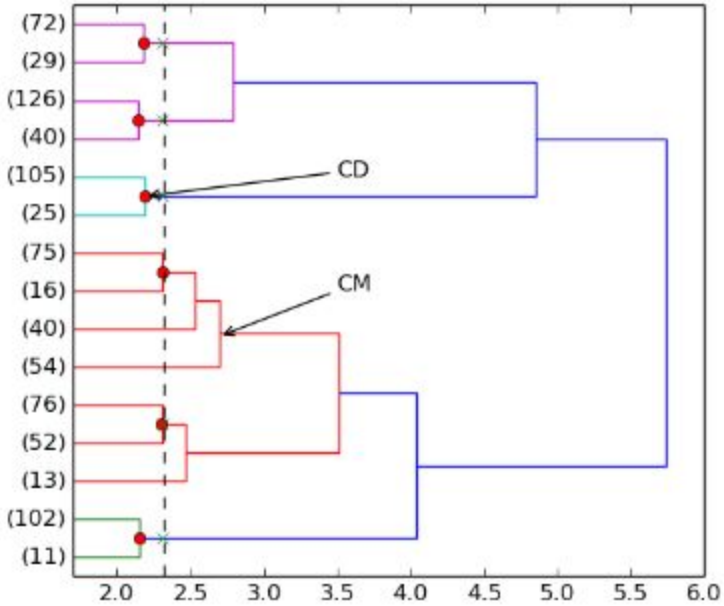


Figure 1: Dendrogram for the "Marijuana" topic (the dashed line shows the 10-clusters cut)

(Quote: 9, 10)

Debbie Debate Bot - speedup

- Using the cluster
  - Compare the user's input to the head of each cluster -> get the head with highest similarity score and calculate the similarity score of each response within that cluster to return the most similar response.
  - Optimize by a graph-based comparison.

topic	stance	baseline	cluster	graph
DP	for	60.6	3.9	8.1
DP	against	55.5	7.9	5.0
GC	for	70.7	25.2	24.3
GC	against	73.5	22.9	15.3
GM	for	62.8	10.0	9.2
GM	against	62.8	3.2	2.9

Table 1 Average response times in seconds

(Quote: 10)

Debbie Debate Bot - Example & Future Work

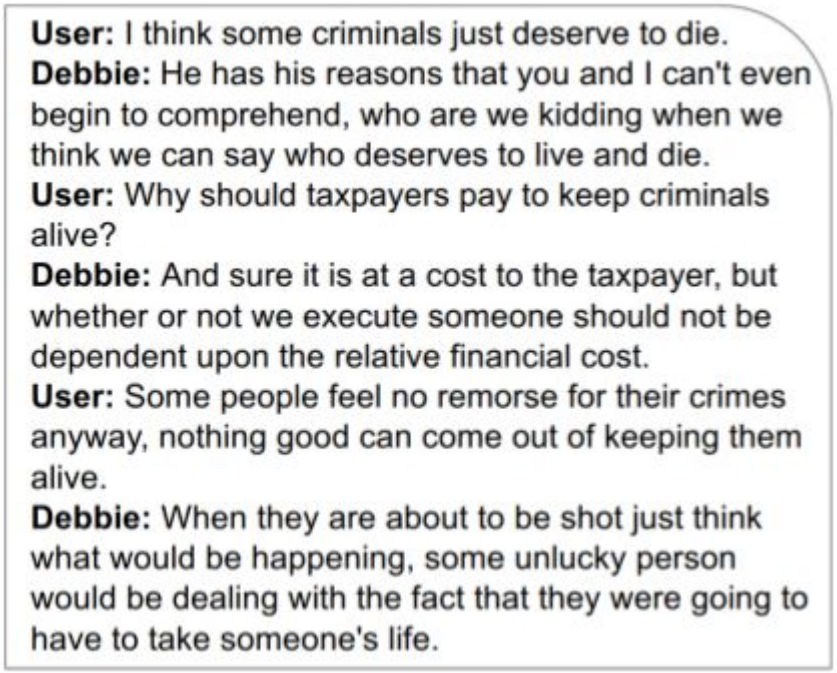


Fig. 1 Chat where Debbie is against the death penalty

- Detect the arguments of low argumentative quality from user
  - You're just wrong
  - I don't think so.
- Enhancing user's experience by entraining Debbie's personality with respect to the user's personality
- Improve retrieval time
  - Recursively employ the graph method within the clusters
  - INDEXING
  - Other algorithms to calculate similarity scores

(Quote: 9)

References

1. Marilyn A. Walker, Pranav Anand, Jean E. Fox Tree, Rob Abbott, Joseph King. "A Corpus for Research on Deliberation and Debate." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012.*
2. Rob Abbott, Brian Ecker, Pranav Anand, Marilyn A. Walker. "Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it." In *Language Resources and Evaluation Conference (LREC), Portorož, Slovenia, 2016.*
3. Argument Mining: Extracting Arguments from Online Dialogue. Reid Swanson, Brian Ecker, Marilyn Walker. *In the 15th Annual SIGdial Meeting on Discourse and Dialogue, Prague, CZ*
4. Amita Misra, Brian Ecker, and Marilyn A. Walker, “Measuring the Similarity of Sentential Arguments in Dialog” , *Proceedings of the SIGDIAL 2016 Conference*
5. https://webcam.srs.fs.fed.us/impacts/ozone/spatial/kriging.shtml

Reference

6. https://towardsdatascience.com/linear-regression-using-least-squares-a4c3456e8570

7. https://towardsdatascience.com/linear-regression-using-least-squares-a4c3456e8570

8. Junyi Jessy Li & Ani Nenkova 2015, Fast and Accurate Prediction of Sentence Specificity, Twenty-Ninth Conference on Artificial Intelligence (AAAI) https://github.com/jjessyli/speciteller

9. Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, Marilyn Walker, 2017, Debbie, the Debate Bot of the Future

10. Filip Boltuzic and Jan Snajder, Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity

Thanks for your attention!

Merry Christmas and Happy New Year~