Why Can’t You Convince Me?
Modeling Weaknesses in Unpersuasive Arguments

Isaac Persing and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

IJCAI-17
MELBOURNE

Argumentation Mining

- Traditionally concerned with determining the argumentative structure of a text document (i.e., identifying its major claims, claims, and premises and the relationship between them)
- Recently expanded to tasks related to the persuasiveness of arguments
  - **Focus**: how persuasive is your argument?

Example Argument

Motion:
This House would ban teachers from interacting with students via social networking websites.
Assertion:
Acting as a warning signal for children at risk.
Justification:
It is difficult for a child to realize that he is being groomed; they are unlikely to know the risk. After all, a teacher is regarded as a trusted adult. But, if the child is aware that private electronic contact between teachers and students is prohibited by law, the child will immediately know the teacher is doing something he is not supposed to if he initiates private electronic contact. This will therefore act as an effective warning sign to the child and might prompt the child to tell a parent about what is going on.

- Composed of an assertion & a justification in response to a debate motion
  - **Motion**: expresses a stance on the debate’s topic
  - **Assertion**: expresses why author agrees or disagrees with motion
  - **Justification**: explains why author believes her assertion
- Humans can easily determine that this argument is not persuasive
- But it’s equally important to determine why an argument is not persuasive
  - Useful feedback for authors to improve their arguments
  - Policy makers and companies only need to focus on the convincing arguments when understanding why people (dis)like a policy/product

Our Contributions

- Understand why an argument is weak by:
  - defining the errors that negatively impact argument persuasiveness
  - hand-annotate a corpus of arguments with errors and persusasiveness
  - design models for predicting errors and persuasiveness

Errors

- Five errors motivated by theoretical work on argument persuasiveness
  - **Grammar Error** (GE)
    - 1 if argument is hard to understand because of GEs; 0 otherwise
  - **Lack of Objectivity** (LO)
    - 1 if it displays an inappropriate lack of objectivity; 0 otherwise
  - **Inadequate Support** (IS)
    - 0, 1, or 2: whether support is adequate, inadequate or missing
  - **Unclear Assertion** (UA)
    - 2 if assertion is incomprehensible without reading the justification
    - 1 if unclear how assertion is related to the motion w/o justification
    - 0 if assertion is clear
  - **Unclear Justification** (UJ)
    - 2 if justification appears unrelated to assertion
    - 1 if it does not concisely justify the assertion
    - 0 if justification is clear

Corpus and Annotation

- **Corpus:** debates from International Debate Education Association website
  - Debates cover a wide range of topics (politics, economics, science, …)
  - 165 debates and 1208 arguments in total
- **Annotation**: two native English speakers annotated each argument with its persuasiveness score and the five aforementioned errors, if applicable
- **(Simplified) rubric for annotating argument persuasiveness**:
  - 6: a very persuasive argument
  - 5: a persuasive, or only pretty clear argument
  - 4: a decent, or only fairly clear argument
  - 3: a poor, or only most understandable argument
  - 2: a very unpersuasive or very unclear argument
  - 1: unclear what the argument is or author doesn’t make an argument
- **Distribution of errors and argument persuasiveness**

	0	1	2	3	4	5	6
GE	98	02					
LO	76	24					
IS	49	35	16				
UA	57	32	11				
UJ	58	39	03				
AP		03	12	20	21	20	24

Models

- Cast the task of predicting an argument’s error and persuasiveness scores as six independent regression problems
  - Each argument in training set is represented as an instance
    - **Label**: the argument’s gold score for that problem
    - **Learner**: linear support vector regressor (as implemented in SVM-light)
    - **11 features**:
      - # grammar errors per sentence in justification
      - # subjectivity indicators (“morally”, “certain”, “perhaps”) in justification
      - # definite articles in justification: indicators of specificity/subjectivity
      - # 1st person plural pronouns in justification: indicators of subjectivity
      - # citations in justification: is support adequate?
      - # content lemmas only in justification: enough points/support?
      - Assertion length: short assertions could be unclear
      - # content lemmas only in assertion: encodes relevance to justification
      - Justification length: short justifications could be unclear
      - # strong thesis statements in justification: makes justification clearer
      - # subject matches in discourse relation

Evaluation

- **Goal**: evaluate our approach to error severity and persuasiveness prediction
- **Three scoring metrics:**
  - **E:** frequency at which a system predicts the wrong score
  - **ME:** mean distance between a system’s prediction and the gold score
  - **PC:** Pearson’s correlation coefficient between predictions and gold scores
- **Six baseline systems:** differ from our approach only w.r.t. the features used
  - **Bag of words (BOW)**
  - **Word n-grams (WNG):** unigrams, bigrams, trigrams
  - **Bag of part-of-speech tags (BOPOS)**
  - **Style:** length, word categories, word complexity, word scores
  - **Duplicated Tan (Tan):** features for predicting success of persuasion
  - **Persing and Ng:** features developed for scoring essay persuasiveness
- **Five-fold cross-validation results:**

	System	GE	LO	IS	UA	UJ	AP
E	WNG	**.022**	**.242**	.650	**.429**	**.426**	.786
	BOW	**.022**	**.242**	.650	**.429**	**.426**	.786
	BOPOS	**.022**	**.242**	.593	**.429**	**.426**	.786
	Style	**.022**	**.242**	.515	**.465**	**.427**	.748
	Tan	**.022**	**.242**	.494	**.456**	**.425**	.744
	P&N	**.022**	**.242**	.531	**.435**	**.441**	.785
	OUR	**.022**	**.242**	**.439**	**.469**	**.431**	**.721**
ME	WNG	.118	.294	.653	**.550**	**.472**	1.218
	BOW	.117	.294	.654	**.551**	**.473**	1.218
	BOPOS	.118	.294	.620	**.551**	**.472**	1.217
	Style	**.106**	**.283**	.547	**.563**	**.476**	1.102
	Tan	**.103**	**.282**	.537	**.517**	**.478**	1.109
	P&N	.115	.293	.607	**.546**	**.476**	1.198
	OUR	.115	.291	**.472**	**.561**	**.474**	**1.036**
PC	WNG	**.006**	.033	.113	.042	.029	.063
	BOW	**-.009**	.082	.124	.060	.036	.073
	BOPOS	**-.070**	.007	.242	.084	.003	.089
	Style	**-.044**	**.221**	.412	**.124**	**.187**	.408
	Tan	**.028**	**.234**	.439	**.169**	**.171**	.398
	P&N	**.034**	.085	.206	.086	**.116**	.252
	OUR	**.004**	**.222**	**.595**	**.241**	**.205**	**.488**

- **Explanatory power of the error classes:**
  - How well can the predicted errors score persuasiveness?
  - Train SVRs with SF (11 features) and EF (predicted error features)

	E		ME		PC	
System	SF	EF	SF	EF	SF	EF
WNG	.786	.786	1.218	1.218	.063	.060
BOW	.786	.786	1.218	1.220	.073	.073
BOPOS	.786	.785	1.217	1.224	.089	.093
Style	.748	.735	1.102	1.094	.408	.426
Tan	.744	.730	1.109	1.106	.398	.410
P&N	.785	.783	1.198	1.195	.252	.259
OUR	**.721**	**.708**	**1.036**	**1.027**	**.488**	**.495**

- **Relative importance of the errors on persuasiveness prediction:**
  - Gold vs. predicted errors as features for training linear SVRs
  - Feature weights when gold errors are used as features for SVR

GE	LO	IS	UA	UJ	Bias
−0.9	−1.0	−0.9	−0.9	−1.0	5.9

  - Feature weights when predicted errors are used as features for SVR

GE	LO	IS	UA	UJ	Bias
−.016	−.154	−1.253	−.822	−1.191	6.009

- **Feature ablation experiments**:
  - In each feature ablation experiment, the SVR was retrained with all but one predicted error feature

Error	E	ME	PC
GE	.708	1.027	.495
LO	.708	1.027	.495
IS	.719	1.063	.448
UA	.729	1.047	.478
UJ	.717	1.028	.494
–	.708	1.027	.495

- **Major sources of error**
  - Predicting UA and UJ errors