EasyChair Preprint

№ 3850

Automated Claim Detection in Argumentative Essays and their Relationship with Writing Quality

Qian Wan, Scott Crossley, Laura Allen and Danielle McNamara

July 13, 2020

Automated Claim Detection in Argumentative Essays and Relationship with Writing Quality

Qian Wan[1], Scott Crossley[1], Laura Allen[2], Danielle McNamara[3]

[1]Department of Applied Linguistics and ESL, Georgia State University

[2]Department of Psychology, University of New Hampshire

[3]Department of Psychology, Arizona State University

Author Note

The authors declare that there no conflicts of interest with respect to this preprint.

This research was supported in part by the Institute for Education Sciences (IES R305A180261, R305A180144), the Office of Naval Research (N00014-17-1-2300), the Bill & Melinda Gates Foundation, the Chan Zuckerberg Initiative, and Schmidt Futures. Ideas expressed in this material are those of the authors and do not necessarily reflect the views of our funders.

Correspondence should be addressed to Qian Wan, Georgia State University, Department of Applied Linguistics and ESL; 15th floor, 25 Park Place, Atlanta, GA 30303, United States. Email: qwan1@gsu.edu

Abstract

This study extracted content and structural features to predict human annotations for claims and non-claims in argumentative essays. The evaluation of classification models indicated the Random Forest classifier yielded the most balanced identifications of claims and non-claims. We used the model to make predictions in a validation corpus that included human ratings of writing quality. The number of claims and the average position of non-claims in essay were significant indicators of essay quality.

Keywords: argument mining, claim detection, essay quality

Automated Claim Detection in Argumentative Essays and Relationship with Writing Quality

Argumentative essays are an important element of writing assessment. Various techniques have been applied to automatically identify argumentative structures including claims (Burstein et al., 1998, 2001, 2003; Stab & Gurevych, 2014, 2017; Nguyen & Litman, 2015, 2016; Persing & Ng, 2015). In this study, we developed and tested algorithms to automatically identify claims and non-claims in argumentative essays based on n-grams, part of speech n-grams, and positional data. We then applied the best performing algorithm to an independent corpus to assess how well the incidence of claims could predict essay quality. These research questions guide this study:

1. To what extent do (1) the frequency of n-grams (bigrams and trigrams), (2) the frequency of part-of-speech (POS) n-grams (bigrams and trigrams), and (3) positional (structural) information of sentences predict whether or not the sentence is a claim?

2. What are the relations between the number, percentage, and positionality of predicted claims/non-claims in an essay and the quality of the essay?

Method

Data

Human annotated argumentative essays developed by Stab and Gurevych (2014, 2017) were split into a training (N = 329) and testing (N = 90) sets, which were used to train and test the claim detection algorithm. The training and testing set were annotated by human annotators for argument components (major claim, claim, and premises). The overall inter-rater agreement among the annotators was .72 and .68, respectively. Our validation corpus consisted of student essays (N = 2269) with human ratings of essay quality based on SAT rubric (1-6 scale. Interrater

reliability for the human ratings of essay quality was greater than Cohen's Kappa = .60 and r = .70.

Data pre-processing

Corpora standardization. We merged the tags of "major claim" and "claim" in the training and testing corpora and treated them as claims. Any sentences that did not fall into the category of claim was treated as non-claim. We then unified the formats of the two corpora and added structural tags for each sentence.

N-gram tokenization. We tokenized the sentences within each corpus into bigrams and trigrams. All the punctuations within the sentences were removed before tokenization. We used the NLTK package for n-gram and part-of-speech tokenization.

Feature Development

Frequency features. We calculated raw frequency and normalized frequency for each n-gram and POS (part of speech) n-gram in the training corpus. Keyness values, which provided evidence of whether a n-gram was more common in claims or non-claims, were also calculated. For any n-gram that appeared in both claims and non-claims, if the keyness value was greater than 3.84 (equivalent to $p < .05$), we considered it more likely to occur in claims over non-claims (or vice-versa).

Positional features. We also calculated the raw and normalized position of a sentence in the whole essay, the raw and normalized paragraph position of the sentence, and the raw and normalized position of the sentence in the paragraph where the sentence occurred.

Analyses

We developed a number of different machine learning algorithms to predict claims and non-claims in the training set and tested them on the test set. We then used the best performing algorithm to predict the discourse type for each sentence in the validation corpus. The predicted number and percentage of claims and non-claims and positional features for the predicted claims and non-claims were then used as independent variables in a regression analysis to predict the human scores of the essay. Prior to analyses, all variables were checked for normality and multicollinearity.

Results

In Table 1, we list the top 10 n-grams or POS n-grams with highest keyness values found in claims and non-claims, respectively. By analyzing the significant n-grams and POS n-grams, we found that the significant n-grams in claims were generally more abstract and less prompt-specific, while the significant n-grams in non-claims were more concrete. Specifically, abstract verbs (e.g., *believe, think*), modal words (*MD*), to infinity (*TO*), adjectives (*JJ*), adverbs (*RB*), and prepositions or subordinating conjunctions (*IN*) were more common in claims. Meanwhile, significant n-grams in non-claims contained more concrete words, nouns (*NN*, *NNS*), cardinal numbers (*CD*), personal pronouns (*PRP*), determiners (*DT*), and inflectional forms of verbs (*VBP*, *VBN*, *VBD*) compared with the claim sub-corpus.

Table 1

Top n-grams with highest keyness values in claims and non-claims.

Significant Bigrams in Claims	Keyness	Significant Bigrams in Non-claims	Keyness	Significant Trigrams in Claims	Keyness	Significant Trigrams in Non-claims	Keyness
in conclusion	223.06	for instance	51.01	i believe that	67.08	more and more	8.38
i believe	77.18	able to	16.84	in my opinion	56.66	some people think	6.52
to sum	60.44	to go	13.11	to sum up	56.02	are able to	5.92
i think	56.57	i had	10.96	my point of	38.96	to go to	5.48
sum up	56.15	who have	10.96	point of view	35.94	in order to	5.03
in my	51.99	if you	10.85	as far as	28.50	in the past	4.41
believe that	50.85	did not	10.27	i prefer to	26.42		
my opinion	48.89	go to	10.04	first of all	23.95		
i strongly	44.68	means that	8.98	agree with the	19.66		
agree that	37.86	it was	8.85	from my point	19.30		
Significant POS Bigrams in Claims	**Keyness**	**Significant POS Bigrams in Non-claims**	**Keyness**	**Significant POS Trigrams in Claims**	**Keyness**	**Significant POS Trigrams in Non-claims**	**Keyness**
NN VBP	43.72	NN VBD	98.47	NN VBP IN	77.74	VBD TO VB	29.46
VBP IN	33.53	PRP VBD	69.91	RB VBP IN	32.32	VBD DT NN	28.74
NN MD	24.30	VBD RB	51.71	JJ VBP VBN	26.42	NN VBD RB	22.04
RBR JJ	19.58	VBD TO	40.06	VBP IN DT	19.86	IN PRP VBD	18.97
NNS MD	17.46	VBD DT	34.17	NN RB VBP	19.78	NN VBD DT	15.51
VBZ RBR	16.51	VBD VBN	25.34	TO VB RP	17.45	NN VBD VBN	15.35
JJ VBP	15.03	RB VBD	22.53	NNS MD VB	15.50	DT NNS RB	15.35
IN VBG	13.77	PRP VBP	20.96	NN MD VB	15.05	DT NN NN	13.35
NNS VBZ	12.76	VBZ VBN	19.40	VBZ RBR JJ	13.68	NN NN VBD	13.16
MD VB	11.74	VBD JJ	18.28	JJ NN VBZ	13.44	VBD IN NN	12.97

After removing features that were highly correlated ($r > .70$), 10 features remained to build the classification models: the position of the sentence in the essay, the normalized sentence position in the paragraph, the word count of the sentence, the frequency of significant bigrams and POS bigrams of claims and non-claims in the sentence, the frequency of significant trigrams and POS trigrams of claims in the sentence, and the frequency of significant POS trigrams of non-claims in the sentence.

Table 2

Performance of the multiple classifiers on claim detection in the test set.

		TP	TN	FP	FN	Precision	Recall	F1-score	Accuracy
LR	Claim	347	629	445	161	0.44	0.68	0.53	0.62
	Non-claim	629	347	161	445	0.80	0.59	0.67	
	Macro Avg			-		0.62	0.63	0.60	
	Weighted Avg					0.68	0.62	0.63	
BNB	Claim	129	965	109	379	0.54	0.25	0.35	0.69
	Non-claim	965	129	379	109	0.72	0.90	0.80	
	Macro Avg			-		0.63	0.58	0.57	
	Weighted Avg					0.66	0.69	0.65	
GNB	Claim	214	885	189	294	0.53	0.42	0.47	0.69
	Non-claim	885	214	294	189	0.75	0.82	0.79	
	Macro Avg			-		0.64	0.62	0.63	
	Weighted Avg					0.68	0.69	0.68	
LSVC	Claim	194	930	144	314	0.57	0.38	0.46	0.71
	Non-claim	930	194	314	144	0.75	0.87	0.80	
	Macro Avg			-		0.66	0.62	0.63	
	Weighted Avg					0.69	0.71	0.69	
RF	Claim	261	895	179	247	0.59	0.51	0.53	0.72
	Non-claim	895	261	247	179	0.78	0.83	0.80	
	Macro Avg			-		0.68	0.66	0.67	
	Weighted Avg					0.71	0.72	0.71	
NN	Claim	219	914	160	289	0.58	0.43	0.49	0.72
	Non-claim	914	219	289	160	0.76	0.85	0.80	
	Macro Avg			-		0.67	0.64	0.65	
	Weighted Avg					0.70	0.72	0.70	

Note: LR = Logistic Regression, BNB = Bernoulli Naive Bayes, GNB = Gaussian Naive Bayes, LSVC = Linear Support Vector Classification, RF = Random Forest, NN = Neural Network

Test performance for all classification models is reported in Table 2. A Random Forest classifier yielded the most balanced overall performance identifying claims and non-claims in argumentative essays. Therefore, we selected the Random Forest classifier as the final model to predict claims and non-claims in the validation data set.

Correlational analysis (see Table 3) indicated the number of predicted claims (r = .35, p < .001) and the average position of non-claims in essay (r = -.19, p < .001) showed at least a small effect size (r > .099) with essay quality and were not strongly correlated with text length (r < .70). These variables were selected for inclusion in our regression analysis to predict essay quality.

Table 3

Correlations between essay quality and predicted data.

	1	2	3	4	5	6	7	8	9	10	11	12
Holistic score	1	0.37	0.35	0.08	-0.08	0.03	-0.04	-0.19	0.04	0.01	0.07	0.43
Number of non-claims	0.37	1	0.43	-0.27	0.27	-0.27	-0.21	-0.26	-0.01	-0.34	-0.13	0.76
Number of claims	0.35	0.43	1	0.70	-0.70	0.06	0.00	-0.33	-0.01	-0.21	0.05	0.52
Percentage of claims	0.08	-0.27	0.70	1	-1.00	0.25	0.19	-0.14	0.03	0.04	0.19	0.00
Percentage of non-claims	-0.08	0.27	-0.70	-1.00	1	-0.25	-0.19	0.14	-0.03	-0.04	-0.19	0.00
Average position of non-claims in paragraph	0.03	-0.27	0.06	0.25	-0.25	1	-0.32	0.17	-0.14	0.12	0.05	-0.15
Average position of claims in paragraph	-0.04	-0.21	0.00	0.19	-0.19	-0.32	1	-0.31	0.48	0.09	0.27	-0.11
Average position of non-claims in essay	-0.19	-0.26	-0.33	-0.14	0.14	0.17	-0.31	1	-0.83	0.07	-0.08	-0.29
Average position of claims in essay	0.04	-0.01	-0.01	0.03	-0.03	-0.14	0.48	-0.83	1	0.05	0.18	0.02
Average word count of non-claims	0.01	-0.34	-0.21	0.04	-0.04	0.12	0.09	0.07	0.05	1	0.43	0.21
Average word count of claims	0.07	-0.13	0.05	0.19	-0.19	0.05	0.27	-0.08	0.18	0.43	1	0.26
Word count of essay	0.43	0.76	0.52	0.00	0.00	-0.15	-0.11	-0.29	0.02	0.21	0.26	1

Note: 1 = Holistic score, 2 = Number of non-claims, 3 = Number of claims, 4 = Percentage of claims, 5 = Percentage of non-claims, 6 = Average position of non-claims in paragraph, 7 = Average position of claims in paragraph, 8 = Average position of non-claims in essay, 9 = Average position of claims in essay, 10 = Average word count of non-claims, 11 = Average word count of claims, 12 = Word count of essay

A significant regression equation was reported (R^2 = .132, F(2,2266) = 172.3.4, p < .001). The model explained 13.2% of the variance of the human scores. Two significant predictors of essay quality were included in the model: number of predicted claims (β = .132, p < .001) and the average position of predicted non-claims in essay (β = -2.829, p < .001).

Discussion

In this study, we extracted content-based linguistic features and structure-based features to train and test six machine-learning models on predicting discourse type (claims and non-claims) in argumentative essays. The best-performing model (the Random Forest model) was used to make predictions for the number, percentage, and positionality of claims and non-claims in our validation set. We then examined links between these predicted features and essay quality. The correlation analysis indicated that the predicted number of claims and the average position of predicted non-claims in essay were indicators of essay quality, predicting 13.2 percent of the variance.

In future work, we intend to investigate the relations between argumentation elements from a broader perspective. We will include more argumentation elements such as major claims, primary claims, counterarguments, and rebuttals, and the relations between these discourse elements. We will also include more diverse linguistic features in our models including discourse markers, cohesion features, and syntactic indices. Our objective is to construct complete models of essay quality, including the complex array of discourse elements and their functional relationships. Such models will enhance feedback algorithms in automated tutoring systems and contribute to our theoretical understandings of writing, a fundamental aspect of discourse processes.

References

Burstein, J., & Marcu, D. (2003). A machine learning approach for identification thesis and conclusion statements in student essays. *Computers and the Humanities*, *37*(4), 455-467.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (2001). Enriching Automated Essay Scoring Using Discourse Marking.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998, August). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 206-210). Association for Computational Linguistics.

Nguyen, H., & Litman, D. (2015, June). Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining* (pp. 22-28).

Nguyen, H., & Litman, D. (2016, August). Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1127-1137).

Persing, I., & Ng, V. (2015, July). Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 543-552).

Stab, C., & Gurevych, I. (2014, August). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 1501-1510).

Stab, C., & Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, *43*(3), 619-659.