The 29th International FLAIRS Conference

Key Largo, Florida, May 16 - 18, 2016

Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics

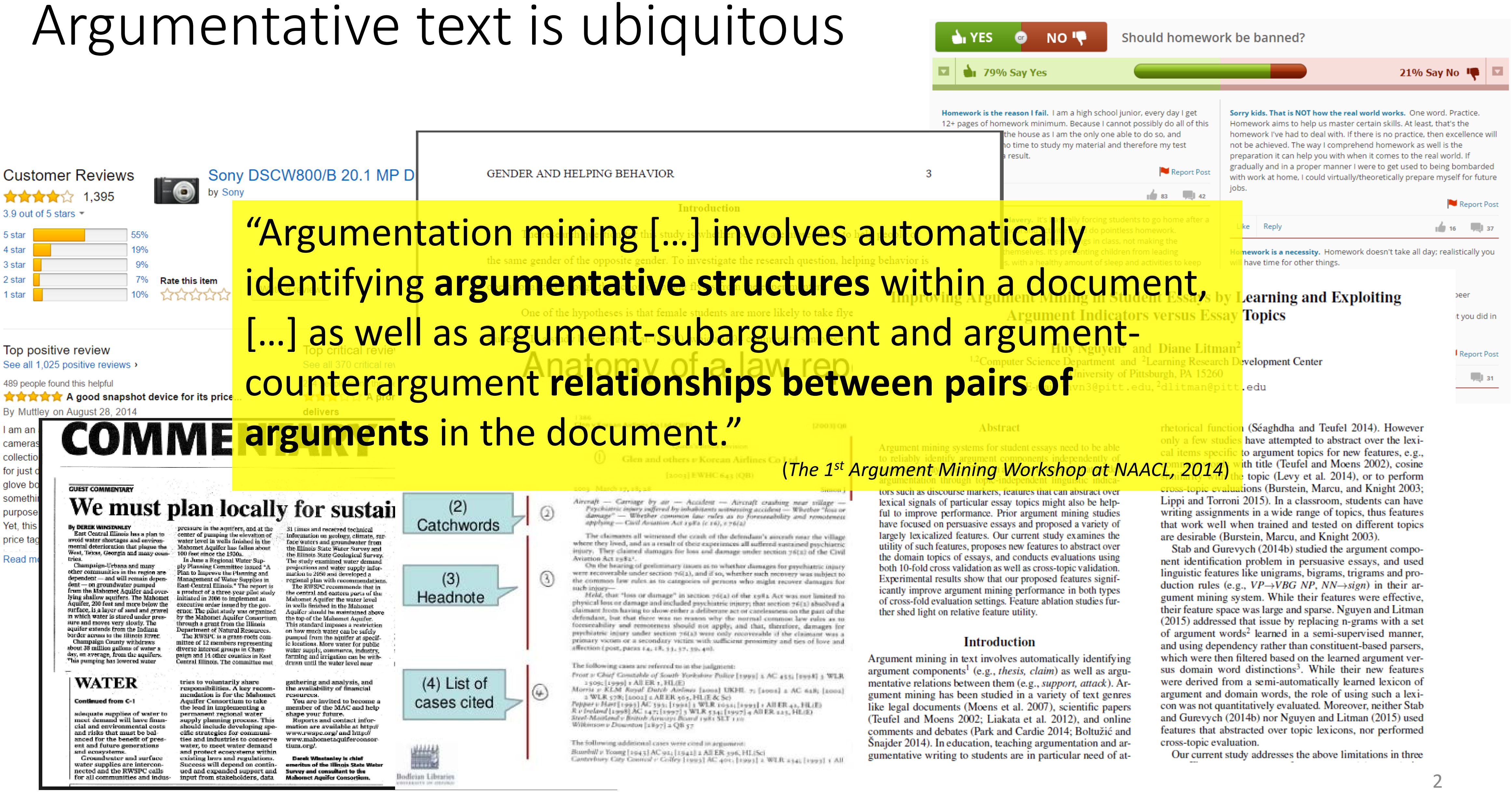
Huy Nguyen[1] Diane Litman[1,2]

[1]Computer Science Department

[2]Learning Research & Development Center

University of Pittsburgh, Pennsylvania

Argumentative text is ubiquitous



👍 YES or NO 👎 Should homework be banned?

👍 79% Say Yes 21% Say No 👎

Homework is the reason I fail. I am a high school junior, every day I get 12+ pages of homework minimum. Because I cannot possibly do all of this the house as I am the only one able to do so, and no time to study my material and therefore my test result.

Report Post

Sorry kids. That is NOT how the real world works. One word. Practice. Homework aims to help us master certain skills. At least, that's the homework I've had to deal with. If there is no practice, then excellence will not be achieved. The way I comprehend homework as well is the preparation it can help you with when it comes to the real world. If gradually and in a proper manner I were to get used to being bombarded with work at home, I could virtually/theoretically prepare myself for future jobs.

Report Post

Homework is a necessity. Homework doesn't take all day; realistically you will have time for other things.

Customer Reviews

★★★★☆ 1,395

3.9 out of 5 stars

5 star		55%
4 star		19%
3 star		9%
2 star		7%
1 star		10%

Sony DSCW800/B 20.1 MP D

by Sony

Rate this item ☆☆☆☆☆

Top positive review

See all 1,025 positive reviews ›

489 people found this helpful

★★★★★ A good snapshot device for its price

By Muttley on August 28, 2014

Top critical review

See all 370 critical rev

COMMENTARY

GUEST COMMENTARY

We must plan locally for sustai

By DEREK WINSTANLEY

WATER

GENDER AND HELPING BEHAVIOR 3

Introduction

Anatomy of a law rep

Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics

Huy Nguyen and Diane Litman

Computer Science Department and Learning Research Development Center, University of Pittsburgh, PA 15260

Abstract

Argument mining systems for student essays need to be able to reliably identify argument components independently of ... tors such as discourse markers, features that can abstract over lexical signals of particular essay topics might also be helpful to improve performance. Prior argument mining studies have focused on persuasive essays and proposed a variety of largely lexicalized features. Our current study examines the utility of such features, proposes new features to abstract over the domain topics of essays, and conducts evaluations using both 10-fold cross validation as well as cross-topic validation. Experimental results show that our proposed features significantly improve argument mining performance in both types of cross-fold evaluation settings. Feature ablation studies further shed light on relative feature utility.

Introduction

Argument mining in text involves automatically identifying argument components[1] (e.g., *thesis, claim*) as well as argumentative relations between them (e.g., *support, attack*). Argument mining has been studied in a variety of text genres like legal documents (Moens et al. 2007), scientific papers (Teufel and Moens 2002; Liakata et al. 2012), and online comments and debates (Park and Cardie 2014; Boltužić and Šnajder 2014). In education, teaching argumentation and argumentative writing to students are in particular need of at-

rhetorical function (Séaghdha and Teufel 2014). However only a few studies have attempted to abstract over the lexical items specific to argument topics for new features, e.g., ... with title (Teufel and Moens 2002), cosine similarity with the topic (Levy et al. 2014), or to perform cross-topic evaluations (Burstein, Marcu, and Knight 2003; Lippi and Torroni 2015). In a classroom, students can have writing assignments in a wide range of topics, thus features that work well when trained and tested on different topics are desirable (Burstein, Marcu, and Knight 2003).

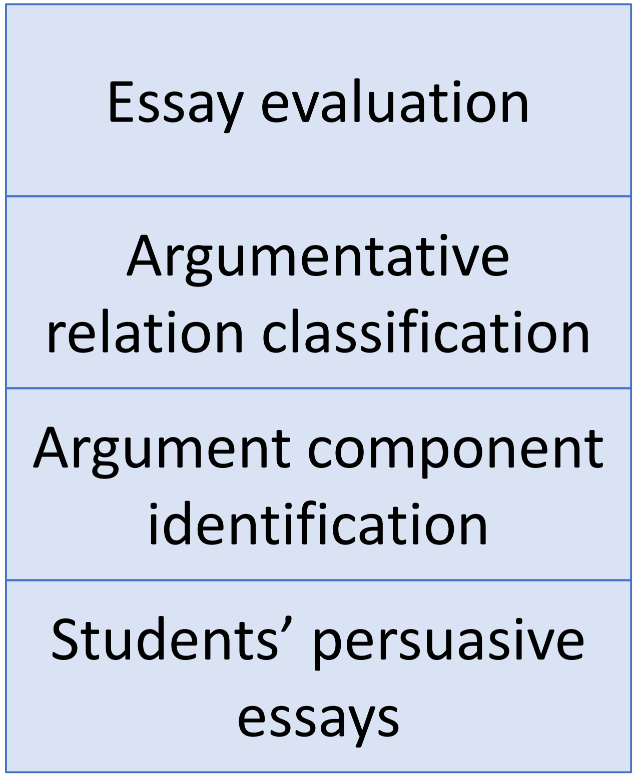
Stab and Gurevych (2014b) studied the argument component identification problem in persuasive essays, and used linguistic features like unigrams, bigrams, trigrams and production rules (e.g., *VP→VBG NP, NN→sign*) in their argument mining system. While their features were effective, their feature space was large and sparse. Nguyen and Litman (2015) addressed that issue by replacing n-grams with a set of argument words[2] learned in a semi-supervised manner, and using dependency rather than constituent-based parsers, which were then filtered based on the learned argument versus domain word distinctions[3]. While their new features were derived from a semi-automatically learned lexicon of argument and domain words, the role of using such a lexicon was not quantitatively evaluated. Moreover, neither Stab and Gurevych (2014b) nor Nguyen and Litman (2015) used features that abstracted over topic lexicons, nor performed cross-topic evaluation.

Our current study addresses the above limitations in three

> “Argumentation mining […] involves automatically identifying **argumentative structures** within a document, […] as well as argument-subargument and argument-counterargument **relationships between pairs of arguments** in the document.”

(The 1st Argument Mining Workshop at NAACL, 2014)

Overview of our research

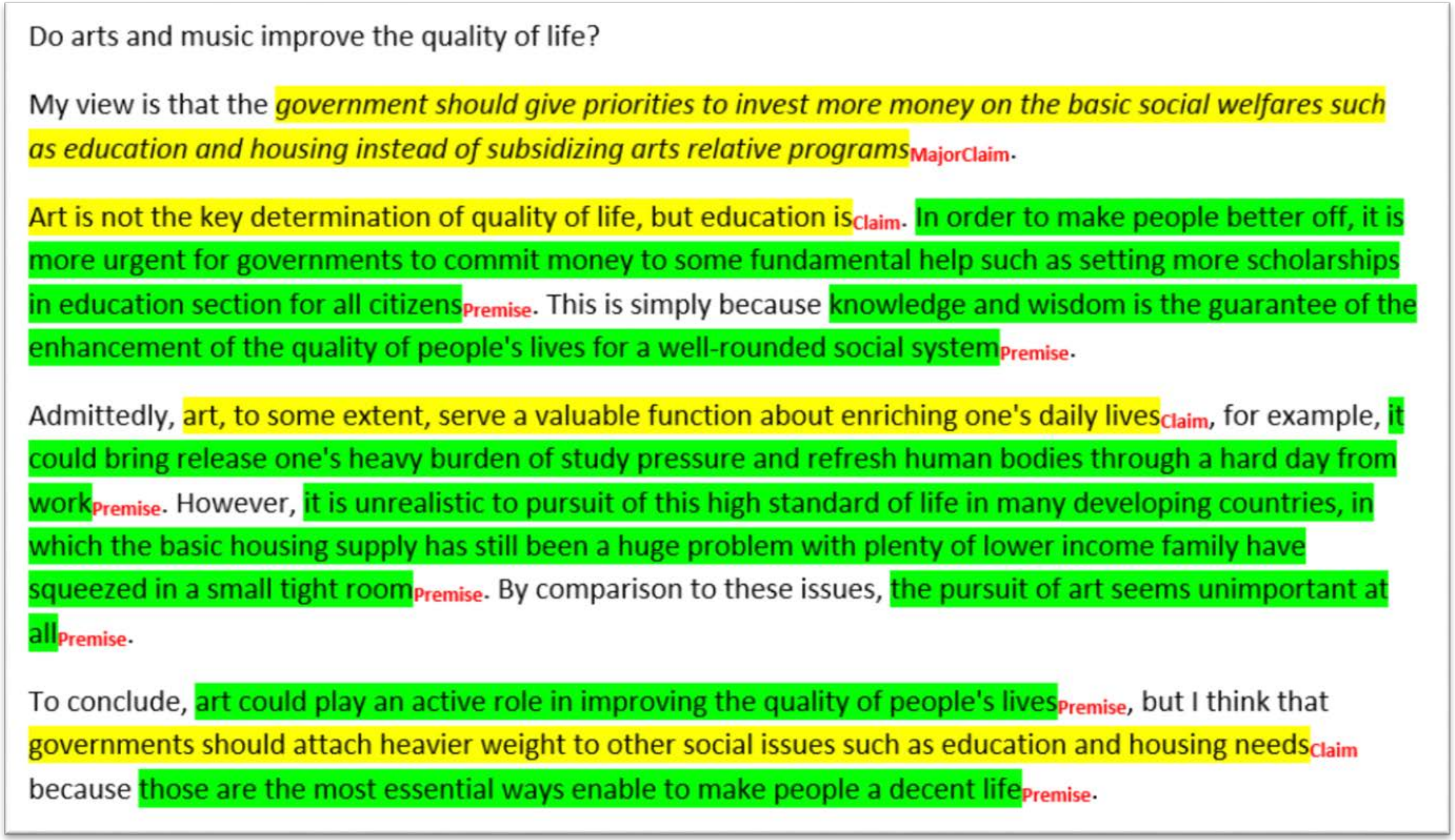


Essay evaluation

Argumentative relation classification

Argument component identification

Students’ persuasive essays



Do arts and music improve the quality of life?

My view is that the *government should give priorities to invest more money on the basic social welfares such as education and housing instead of subsidizing arts relative programs*$_{\text{MajorClaim}}$.

Art is not the key determination of quality of life, but education is$_{\text{Claim}}$. In order to make people better off, it is more urgent for governments to commit money to some fundamental help such as setting more scholarships in education section for all citizens$_{\text{Premise}}$. This is simply because knowledge and wisdom is the guarantee of the enhancement of the quality of people's lives for a well-rounded social system$_{\text{Premise}}$.

Admittedly, art, to some extent, serve a valuable function about enriching one's daily lives$_{\text{Claim}}$, for example, it could bring release one's heavy burden of study pressure and refresh human bodies through a hard day from work$_{\text{Premise}}$. However, it is unrealistic to pursuit of this high standard of life in many developing countries, in which the basic housing supply has still been a huge problem with plenty of lower income family have squeezed in a small tight room$_{\text{Premise}}$. By comparison to these issues, the pursuit of art seems unimportant at all$_{\text{Premise}}$.

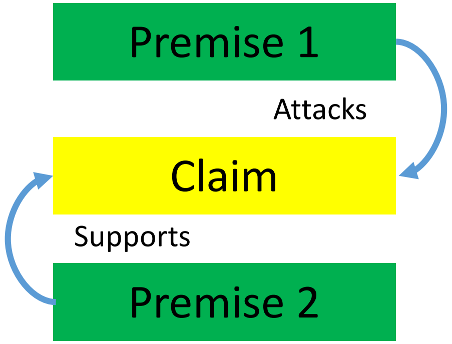
To conclude, art could play an active role in improving the quality of people's lives$_{\text{Premise}}$, but I think that governments should attach heavier weight to other social issues such as education and housing needs$_{\text{Claim}}$ because those are the most essential ways enable to make people a decent life$_{\text{Premise}}$.

Argument component identification

- Argument component: text portion with a specific role in forming the argument*

[…] To conclude, art could play an active role in improving the quality of people's lives, but I think that governments should attach heavier weight to other social issues such as education and housing needs because those are the most essential ways enable to make people a decent life.

(Persuasive Essay Corpus, Stab & Gurevych 2014)



- The step before argumentative relation mining

- This study focuses on argument component identification in student essays

* Peldszus & Stede. From Argument Diagrams to Argumentation Mining in Texts: A Survey. IJCINI 2013.

Prior argument component identification studies

- N-gram and production rule features (VP→VBG NP) [Stab & Gurevych 2014]
  - ✗ Large and sparse feature space
  - ✗ Have not considered abstraction of argument topic
- Lexicons of argument and domain words [Nguyen & Litman 2015]
  - ✗ Lacked a quantitative evaluation
- Cross-fold validation
  - ✗ Have not evaluated topic-independence of the models (e.g., train and test essays are of different topics)

Argument and domain word extraction [Nguyen & Litman 2015]

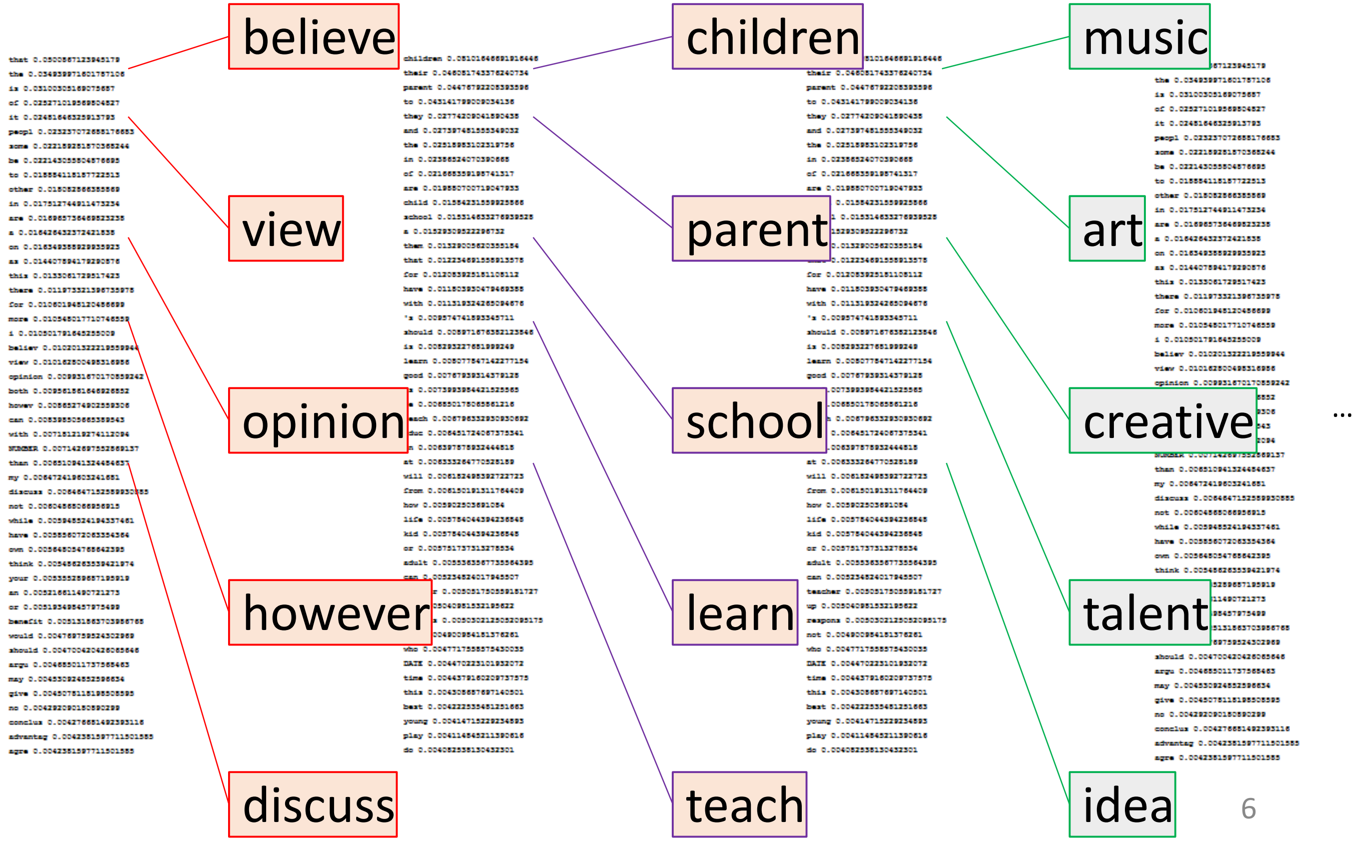
- 6794 un-annotated persuasive essays*
- Process Latent Dirichlet Allocation [Blei et al. 2003] topic model output

Argument seeds: agree, disagree, reason, support, advantage, disadvantage, think, conclusion, result, opinion

Domain seeds: *title words that are not argument seeds or stop words*

Scoring algorithm: *looks for the most argumentative LDA topic, i.e., high argument weight and low domain weight*

Result: *263 argument words and 1806 domain words (stemmed)*



* www.essayforum.com

Example argument and domain words

Argument seeds: agree, disagree, reason, support, advantage, disadvantage, think, conclusion, result, opinion

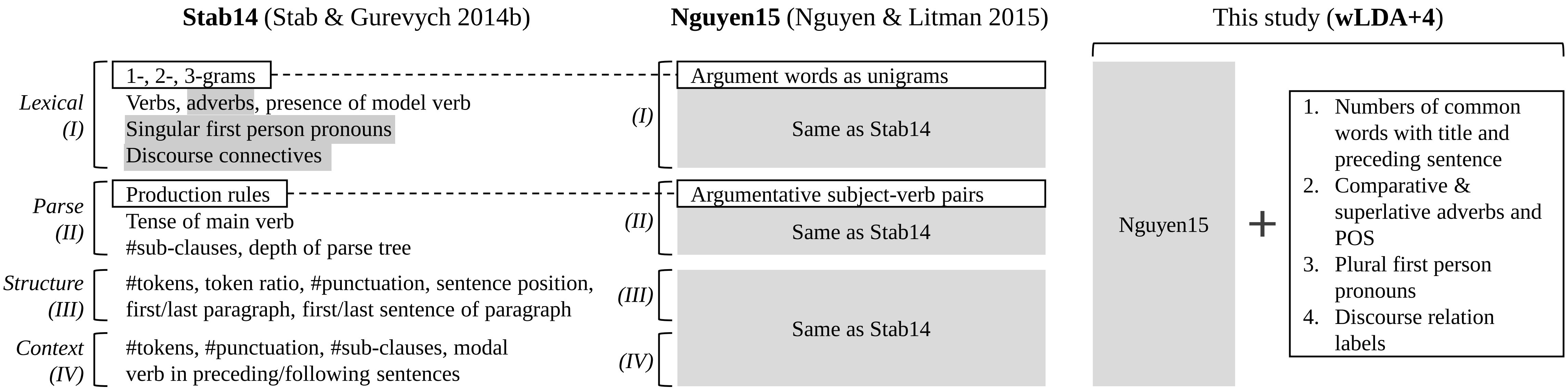
LDA topic 1: reason example support agree think because disagree statement opinion believe therefor idea conclusion

= Argument seeds & variants, discourse connectives, stop words

LDA topic 2: city live big house place area small apart town build community factory urban

LDA topic 3: children parent school education teach kid adult grow childhood behavior taught

Baseline vs. Proposed models



Stab14 (Stab & Gurevych 2014b)

Lexical (I)
- 1-, 2-, 3-grams
- Verbs, adverbs, presence of model verb
- Singular first person pronouns
- Discourse connectives

Parse (II)
- Production rules
- Tense of main verb
- #sub-clauses, depth of parse tree

Structure (III)
- #tokens, token ratio, #punctuation, sentence position, first/last paragraph, first/last sentence of paragraph

Context (IV)
- #tokens, #punctuation, #sub-clauses, modal verb in preceding/following sentences

Nguyen15 (Nguyen & Litman 2015)

(I)
- Argument words as unigrams
- Same as Stab14

(II)
- Argumentative subject-verb pairs
- Same as Stab14

(III) *(IV)*
- Same as Stab14

This study (**wLDA+4**)

Nguyen15 +

1. Numbers of common words with title and preceding sentence
2. Comparative & superlative adverbs and POS
3. Plural first person pronouns
4. Discourse relation labels

10-fold cross validation (data was randomly split into training and test sets)

Cross writing-prompt validation (training and test essays are of different prompts)

Ablated models

- Replace argument and domain lexicons in wLDA+4 model
  - SEED model: uses only argument and domain seeds

    Extracted lexicons vs. Seed words

  - woLDA model: does not use seed words or the two lexicons
    - Removes argument word features
    - Uses all subject-verb pairs

    With lexicons vs. Without lexicon

Persuasive Essay Corpus [Stab & Gurevych 2014]

- 90 persuasive essays*
  - MajorClaim
  - Claim
  - Premise

government should give priorities to invest more money on the basic social welfares [...]

I think that governments should attach heavier weight to other social issues such as education and housing needs

those are the most essential ways enable to make people a decent life

- 3 expert annotators
  - Accuracy 0.88
  - Krippendorff's α_U 0.72

MajorClaim	Claim	Premise	Non-argumentative
90	429	1033	327

* www.essayforum.com

Evaluation method

- Cross-fold validation
    - Randomly: 10-fold cross validation
    - By-prompt: cross writing prompt validation

- In each folding
    - Select top 100 features in training folds (InfoGain + Ranking)
    - Train prediction model with top 100 features
    - Record prediction output on the test fold

Experimental results: cross validation

	10-fold cross validation					Cross-prompt validation				
	Stab14	**Nguyen15**	**woLDA**	**SEED**	**wLDA+4**	**Stab14**	**Nguyen15**	**woLDA**	**SEED**	**wLDA+4**
Accuracy	0.787*	0.792*	0.780*	0.781*	**0.805**	0.780*	0.796	0.774*	0.776*	**0.807**
Kappa	0.639*	0.649*	0.629*	0.632*	**0.673**	0.623*	0.654+	0.618*	0.623*	**0.675**
Precision	0.741*	0.745*	0.746*	0.740*	**0.763**	0.722*	0.757*	0.751	0.734	**0.771**
Recall	0.694*	0.698*	0.695*	0.695*	**0.720**	0.670*	0.695*	0.681*	0.686*	**0.722**

Best values in bold. +: p < 0:1, *: p < 0:05 by T-test when comparing with wLDA+4

Obtains comparable performances between two experiment settings

Proposed model (wLDA+4) performs the best in 10-fold cross validation

12 groups:

- 11 single-prompt groups (73 essays)
- 1 mixed group of minor prompts (17 essays)

Prompts: school, technologies, prepared food …

Experimental results: holdout test sets

	Stab’s test set		Nguyen’s test set	
	Stab’s reported	wLDA+4	Nguyen’s reported	wLDA+4
Accuracy	0.77	**0.82**	0.83	**0.84**
Kappa	–	**0.68**	0.69	**0.71**
F1	0.73	**0.75**	0.76	**0.78**
Precision	0.77	**0.79**	0.79	**0.81**
Recall	0.68	**0.73**	0.74	**0.76**

Feature evaluation

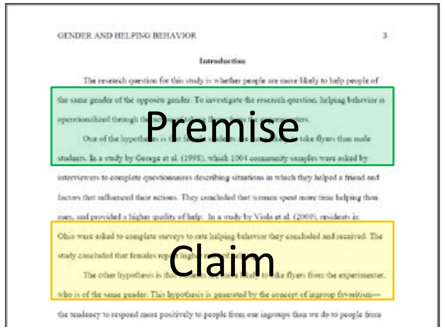
- Among all top features used to train the models
  - 49% are argument words
  - 8% are argumentative subject-verb pairs

LDA-enabled features in Nguyen15

- In the top-50
  - Common word counts
  - Comparative adverbs, and *RBR* part-of-speech
  - Person pronouns *WE*, *OUR*
  - Discourse labels *Expansion*, *Contingency*

Proposed features in this study

Conclusions and future work



GENDER AND HELPING BEHAVIOR 3

Introduction

The research question for this study is whether people are more likely to help people of the same gender of the opposite gender. To investigate the research question, helping behavior is operationalized through the [illegible] experimenters.

One of the hypotheses is that [illegible] take flyers than male students. In a study by George et al. (1998), which 1004 community samples were asked by interviewers to complete questionnaires describing situations in which they helped a friend and factors that influenced their actions. They concluded that women spent more time helping than men, and provided a higher quality of help. In a study by Viola et al. (2009), residents in Ohio were asked to complete surveys to rate helping behavior they concluded and received. The study concluded that females rep[illegible]

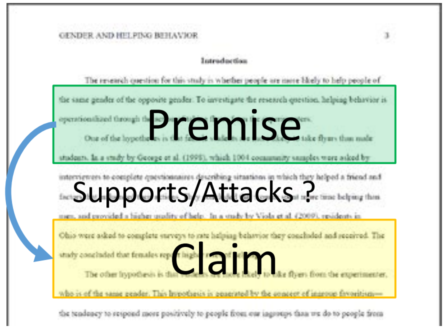
The other hypothesis is that [illegible] likely to take flyers from the experimenter, who is of the same gender. This hypothesis is generated by the concept of ingroup favoritism—the tendency to respond more positively to people from our ingroups than we do to people from

- A study on argument component identification

- New features that model argument indicators and abstract over essay topics
  - A necessary supplement to the learned and noisy argument and domain words

- Cross-topic and 10-fold cross validations
  - Proposed model obtained comparable performances

- Our next study focuses on argumentative relation classification



GENDER AND HELPING BEHAVIOR 3

Introduction

The research question for this study is whether people are more likely to help people of the same gender of the opposite gender. To investigate the research question, helping behavior is operationalized through the [illegible] experimenters.

One of the hypotheses is that [illegible] take flyers than male students. In a study by George et al. (1998), which 1004 community samples were asked by interviewers to complete questionnaires describing situations in which they helped a friend and factors that influenced their actions. They concluded that women spent more time helping than men, and provided a higher quality of help. In a study by Viola et al. (2009), residents in Ohio were asked to complete surveys to rate helping behavior they concluded and received. The study concluded that females rep[illegible]

The other hypothesis is that [illegible] likely to take flyers from the experimenter, who is of the same gender. This hypothesis is generated by the concept of ingroup favoritism—the tendency to respond more positively to people from our ingroups than we do to people from

Thank you!

This research is supported by NSF Grant 1122504