Argument Mining from Text for Teaching and Assessing Writing

Diane Litman

Professor, Computer Science Department
Co-Director, Intelligent Systems Program
Senior Scientist, Learning Research & Development Center

University of Pittsburgh
Pittsburgh, PA USA

Visiting Fellow (Lent Term): Emmanuel College & Dialogue Systems Group

Outline

- ***Argumentative Writing / Argument Mining***
- Algorithms and Applications
  - Automated Writing Assessment
  - Teaching with Diagramming and Peer Review
- Summary and Current Directions

Why teach argumentative writing?

- Studies show students:
  - lack competence in argument writing (Oostdam, et al., 1994; Oostdam & Emmelot, 1991).
  - do not integrate their arguments into a high-level structure or coherent position (Keith, Weiner, & Lesgold, 1991).
  - Even if compose-aloud protocols show students mentally connect position statement & supporting details, *connections not evident in writing* (Durst, 1987).

Slide modified from Kevin D. Ashley, Ilya Goldin. 2011

Research Question

- Can ***argument mining*** be used to better teach, assess, and understand ***argumentative writing***?

- **Approach:** Technology design and evaluation
  - System enhancements that improve student learning
  - Argument analytics for teachers
  - Experimental platforms to test research predictions

Argument Mining

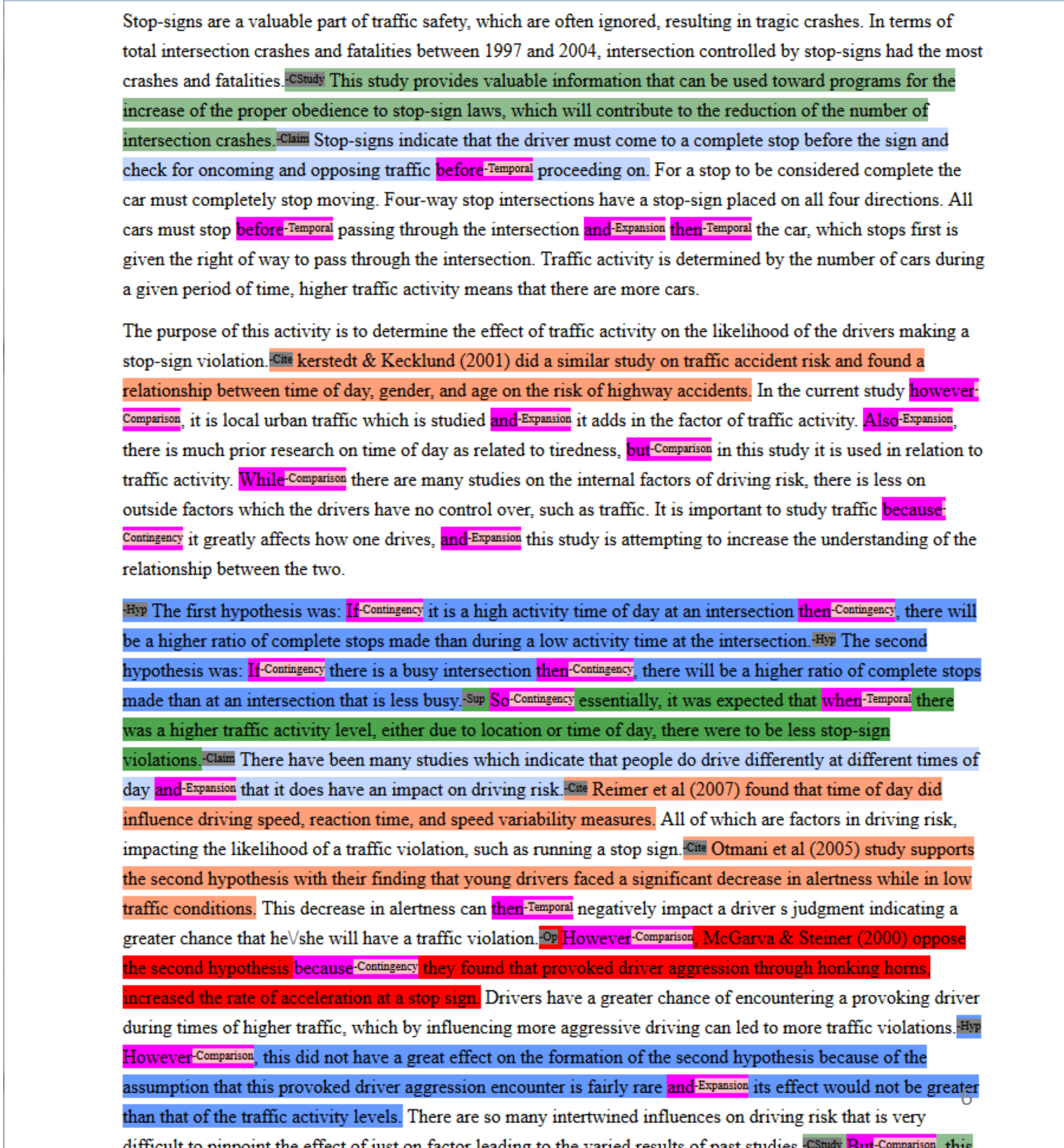
- "… exploits the techniques and methods of natural language processing … for semi-automatic and automatic recognition and extraction of structured argument data from unstructured … texts." *[SICSA Workshop on Argument Mining, July 2014]*

Mining a College Essay for *Argument Diagram Ontology*

(i.e., node and arc types)

- Current Study
- Claim
- Citation
- Hypothesis
- Supports
- Opposes

- None



Stop-signs are a valuable part of traffic safety, which are often ignored, resulting in tragic crashes. In terms of total intersection crashes and fatalities between 1997 and 2004, intersection controlled by stop-signs had the most crashes and fatalities. -CStudy This study provides valuable information that can be used toward programs for the increase of the proper obedience to stop-sign laws, which will contribute to the reduction of the number of intersection crashes. -Claim Stop-signs indicate that the driver must come to a complete stop before the sign and check for oncoming and opposing traffic before-Temporal proceeding on. For a stop to be considered complete the car must completely stop moving. Four-way stop intersections have a stop-sign placed on all four directions. All cars must stop before-Temporal passing through the intersection and-Expansion then-Temporal the car, which stops first is given the right of way to pass through the intersection. Traffic activity is determined by the number of cars during a given period of time, higher traffic activity means that there are more cars.

The purpose of this activity is to determine the effect of traffic activity on the likelihood of the drivers making a stop-sign violation. -Cite kerstedt & Kecklund (2001) did a similar study on traffic accident risk and found a relationship between time of day, gender, and age on the risk of highway accidents. In the current study however-Comparison, it is local urban traffic which is studied and-Expansion it adds in the factor of traffic activity. Also-Expansion, there is much prior research on time of day as related to tiredness, but-Comparison in this study it is used in relation to traffic activity. While-Comparison there are many studies on the internal factors of driving risk, there is less on outside factors which the drivers have no control over, such as traffic. It is important to study traffic because-Contingency it greatly affects how one drives, and-Expansion this study is attempting to increase the understanding of the relationship between the two.

-Hyp The first hypothesis was: If-Contingency it is a high activity time of day at an intersection then-Contingency, there will be a higher ratio of complete stops made than during a low activity time at the intersection. -Hyp The second hypothesis was: If-Contingency there is a busy intersection then-Contingency, there will be a higher ratio of complete stops made than at an intersection that is less busy. -Sup So-Contingency essentially, it was expected that when-Temporal there was a higher traffic activity level, either due to location or time of day, there were to be less stop-sign violations. -Claim There have been many studies which indicate that people do drive differently at different times of day and-Expansion that it does have an impact on driving risk. -Cite Reimer et al (2007) found that time of day did influence driving speed, reaction time, and speed variability measures. All of which are factors in driving risk, impacting the likelihood of a traffic violation, such as running a stop sign. -Cite Otmani et al (2005) study supports the second hypothesis with their finding that young drivers faced a significant decrease in alertness while in low traffic conditions. This decrease in alertness can then-Temporal negatively impact a driver s judgment indicating a greater chance that he\/she will have a traffic violation. -Op However-Comparison, McGarva & Steiner (2000) oppose the second hypothesis because-Contingency they found that provoked driver aggression through honking horns, increased the rate of acceleration at a stop sign. Drivers have a greater chance of encountering a provoking driver during times of higher traffic, which by influencing more aggressive driving can led to more traffic violations. -Hyp However-Comparison, this did not have a great effect on the formation of the second hypothesis because of the assumption that this provoked driver aggression encounter is fairly rare and-Expansion its effect would not be greater than that of the traffic activity levels. There are so many intertwined influences on driving risk that is very difficult to pinpoint the effect of just on factor leading to the varied results of past studies. -CStudy But-Comparison, this

Mining a Grade School Essay for *Evidence*

I was convinced that winning the fight of poverty is achievable in our lifetime. Many people couldn't afford medicine or bed nets to be treated for malaria . Many children had died from this dieseuse even though it could be treated easily. But now, bed nets are used in every sleeping site . And the medicine is free of charge. Another example is that the farmers' crops are dying because they could not afford the nessacary fertilizer and irrigation . But they are now, making progess. Farmers now have fertilizer and water to give to the crops. Also with seeds and the proper tools . Third, kids in Sauri were not well educated. Many families couldn't afford school . Even at school there was no lunch . Students were exhausted from each day of school. Now, school is free . Children excited to learn now can and they do have midday meals . Finally, Sauri is making great progress. If they keep it up that city will no longer be in poverty. Then the Millennium Village project can move on to help other countries in need.

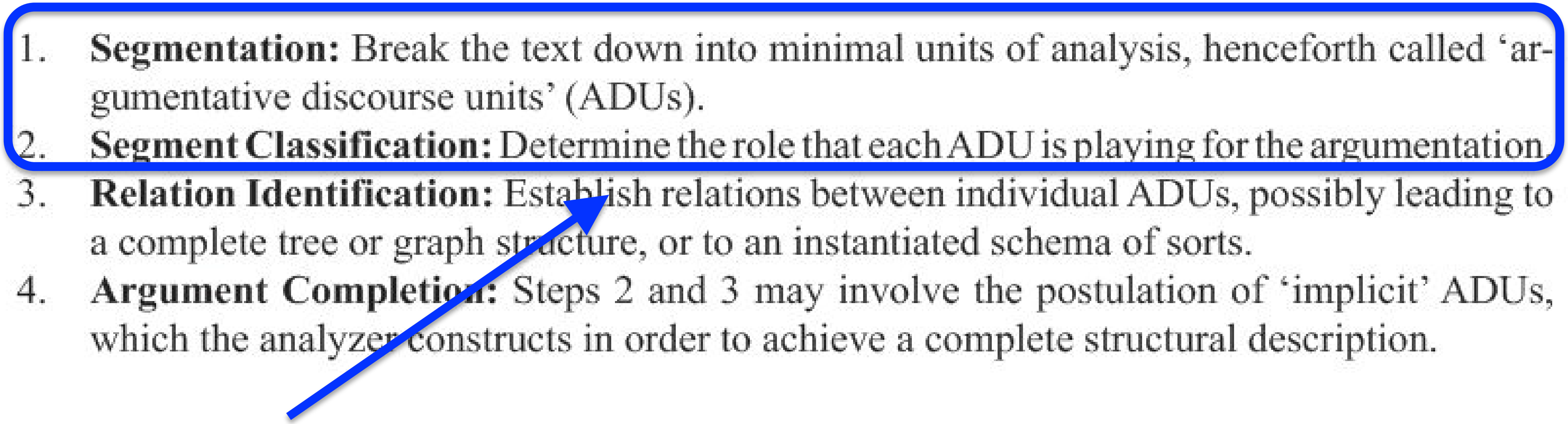
Mining a High School Essay for a *Thesis Statement*

Violence in America

In the spring of 2011, John F. Shick met a man in a New Mexico parking lot and bought two pistols for $810, along with extra clips, ammunition and a holster. This is one reason that there is so much violence in America, because people who are mentally ill, and people who aren't, can just buy a gun anywhere, there are no laws that say they cant. The two biggest factors that contribute to America's violent society are relaxed gun laws and mental illness.

There are numerous accounts on which people with mental illness don't receive the help they need and sometimes do things to things to hurt people, like the account on which was just mentioned John Shick bought two guns and shot and killed a person…

Argument Mining Subtasks [Peldszus and Stede, 2013]



1. **Segmentation:** Break the text down into minimal units of analysis, henceforth called 'argumentative discourse units' (ADUs).
2. **Segment Classification:** Determine the role that each ADU is playing for the argumentation.
3. **Relation Identification:** Establish relations between individual ADUs, possibly leading to a complete tree or graph structure, or to an instantiated schema of sorts.
4. **Argument Completion:** Steps 2 and 3 may involve the postulation of 'implicit' ADUs, which the analyzer constructs in order to achieve a complete structural description.

- Scope of today's talk
- Even partial argument mining can support useful applications
  - Mostly non-structural (e.g., *no relations, no argument schema*)
  - Application-dependent roles (e.g., *no premises*)
  - But, *real data*!!

Argument Mining for Education

- Challenges
  - Noisy data (e.g., adult learners, children)
  - Real-time algorithms; robust at scale
  - Meaningful features

- Opportunities
  - Human in the loop (e.g., peer review)
    - Errors as student reflection opportunities

Outline

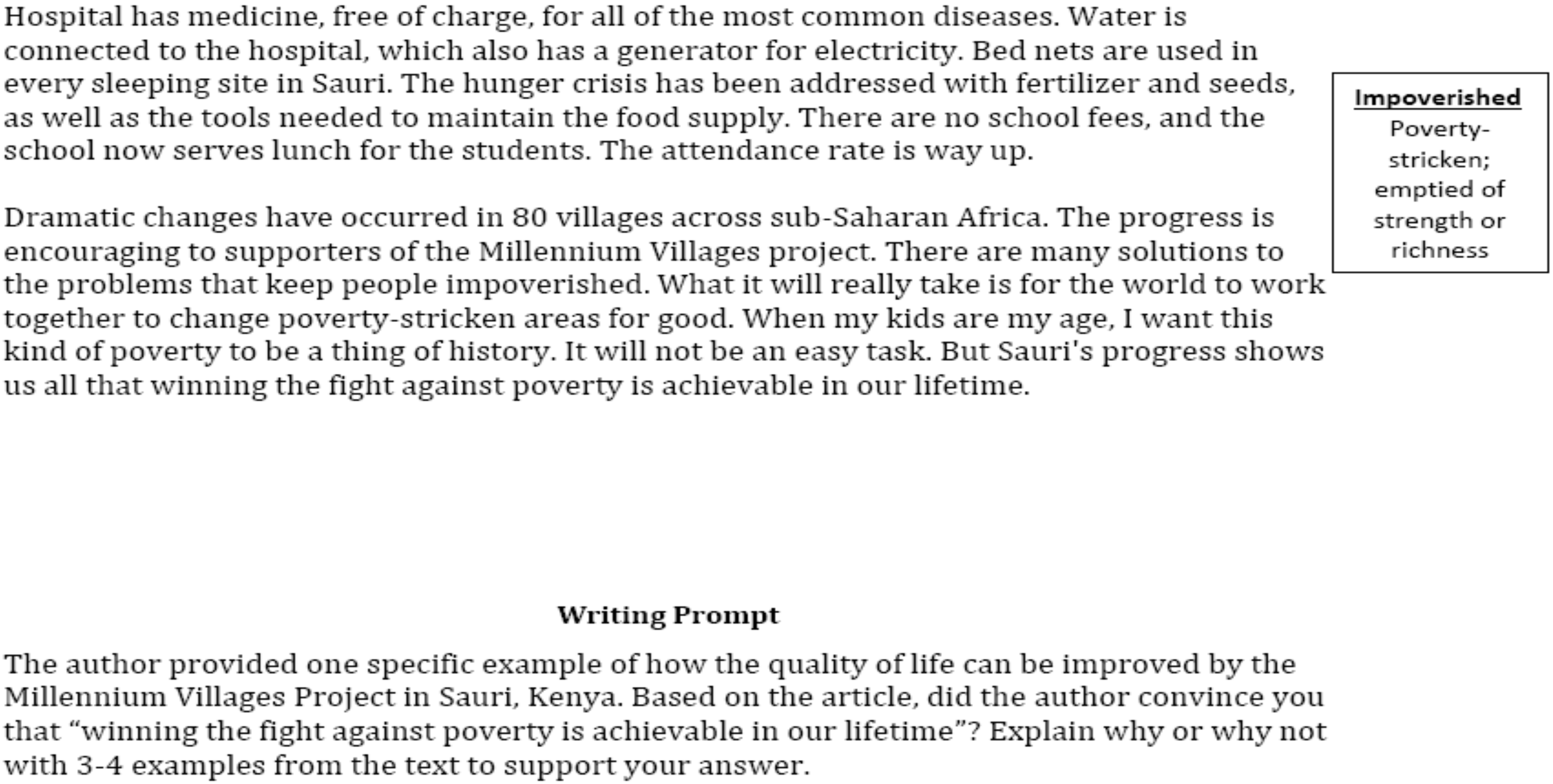
- Argumentative Writing / Argument Mining
- Algorithms and Applications
  - ***Automated Writing Assessment***
  - Teaching with Diagramming and Peer Review
- Summary and Current Directions

Why Automatic Writing Assessment?

- Essential for Massive Open Online Courses (MOOCs) and tutoring systems
- Even in traditional classes, frequent assignments can limit the amount of teacher feedback

An Example Writing Assessment Task: Response to Text (RTA)

- *MVP*, Time for Kids – informational text



Hospital has medicine, free of charge, for all of the most common diseases. Water is connected to the hospital, which also has a generator for electricity. Bed nets are used in every sleeping site in Sauri. The hunger crisis has been addressed with fertilizer and seeds, as well as the tools needed to maintain the food supply. There are no school fees, and the school now serves lunch for the students. The attendance rate is way up.

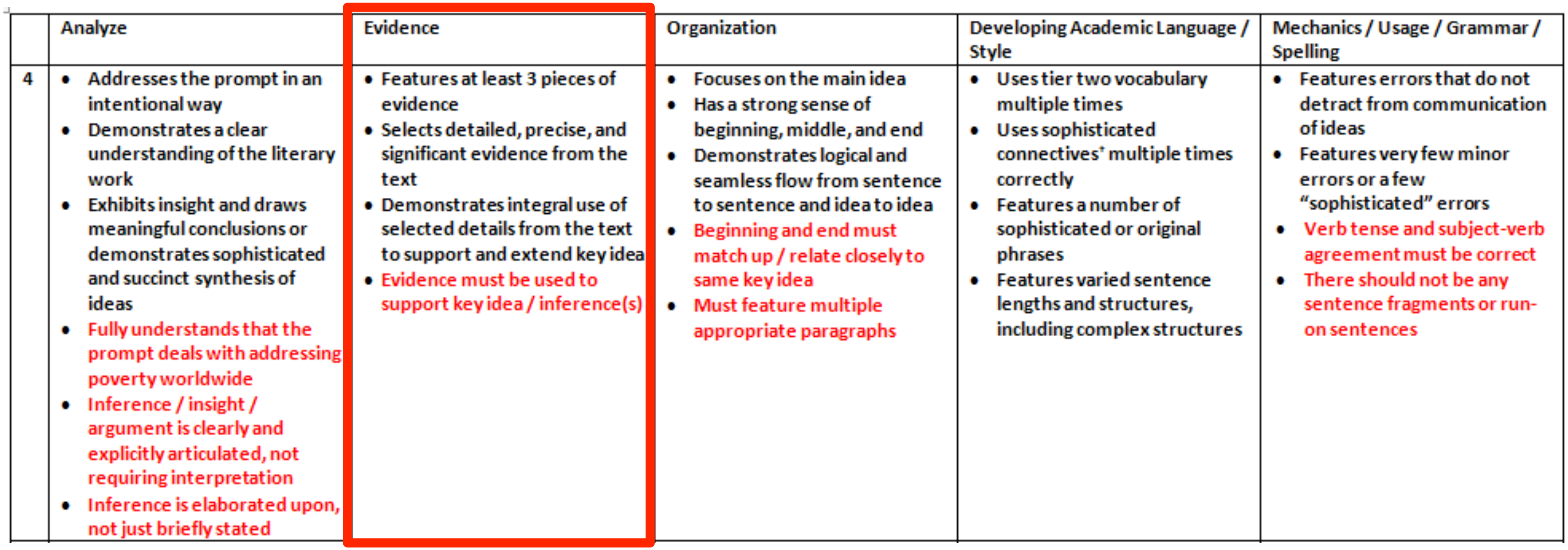
Dramatic changes have occurred in 80 villages across sub-Saharan Africa. The progress is encouraging to supporters of the Millennium Villages project. There are many solutions to the problems that keep people impoverished. What it will really take is for the world to work together to change poverty-stricken areas for good. When my kids are my age, I want this kind of poverty to be a thing of history. It will not be an easy task. But Sauri's progress shows us all that winning the fight against poverty is achievable in our lifetime.

Impoverished
Poverty-stricken; emptied of strength or richness

Writing Prompt

The author provided one specific example of how the quality of life can be improved by the Millennium Villages Project in Sauri, Kenya. Based on the article, did the author convince you that "winning the fight against poverty is achievable in our lifetime"? Explain why or why not with 3-4 examples from the text to support your answer.

Excerpt from the Scoring Rubric



	Analyze	Evidence	Organization	Developing Academic Language / Style	Mechanics / Usage / Grammar / Spelling
4	• Addresses the prompt in an intentional way • Demonstrates a clear understanding of the literary work • Exhibits insight and draws meaningful conclusions or demonstrates sophisticated and succinct synthesis of ideas • Fully understands that the prompt deals with addressing poverty worldwide • Inference / insight / argument is clearly and explicitly articulated, not requiring interpretation • Inference is elaborated upon, not just briefly stated	• Features at least 3 pieces of evidence • Selects detailed, precise, and significant evidence from the text • Demonstrates integral use of selected details from the text to support and extend key idea • Evidence must be used to support key idea / inference(s)	• Focuses on the main idea • Has a strong sense of beginning, middle, and end • Demonstrates logical and seamless flow from sentence to sentence and idea to idea • Beginning and end must match up / relate closely to same key idea • Must feature multiple appropriate paragraphs	• Uses tier two vocabulary multiple times • Uses sophisticated connectives* multiple times correctly • Features a number of sophisticated or original phrases • Features varied sentence lengths and structures, including complex structures	• Features errors that do not detract from communication of ideas • Features very few minor errors or a few "sophisticated" errors • Verb tense and subject-verb agreement must be correct • There should not be any sentence fragments or run-on sentences

Audience Participation: Evidence Scoring (1 or 4?)

Student 1: Yes, because even though proverty is still going on now it does not mean that it can not be stop. Hannah thinks that proverty will end by 2015 but you never know. The world is going to increase more stores and schools. But if everyone really tries to end proverty I believe it can be done. Maybe starting with recycling and taking shorter showers, but no really short that you don't get clean. Then maybe if we make more money or earn it we can donate it to any charity in the world. Proverty is not on in Africa, it's practiclly every where! Even though Africa got better it didn't end proverty. Maybe they should make a law or something that says and declare that proverty needs to need. There's no specic date when it will end but it will. When it does I am going to be so proud, wheather I'm alive or not**.**

Student 2: I was convinced that winning the fight of poverty is achievable in our lifetime. Many people couldn't afford medicine or bed nets to be treated for malaria . Many children had died from this dieseuse even though it could be treated easily. But now, bed nets are used in every sleeping site . And the medicine is free of charge. Another example is that the farmers' crops are dying because they could not afford the nessacary fertilizer and irrigation . But they are now, making progess. Farmers now have fertilizer and water to give to the crops. Also with seeds and the proper tools . Third, kids in Sauri were not well educated. Many families couldn't afford school . Even at school there was no lunch . Students were exhausted from each day of school. Now, school is free . Children excited to learn now can and they do have midday meals . Finally, Sauri is making great progress. If they keep it up that city will no longer be in poverty. Then the Millennium Village project can move on to help other countries in need.

Automatic Scoring via Argument Mining

Student 1: Yes, because even though proverty is still going on now it does not mean that it can not be stop. Hannah thinks that proverty will end by 2015 but you never know. The world is going to increase more stores and schools. But if everyone really tries to end proverty I believe it can be done. Maybe starting with recycling and taking shorter showers, but no really short that you don't get clean. Then maybe if we make more money or earn it we can donate it to any charity in the world. Proverty is not on in Africa, it's practiclly every where! Even though Africa got better it didn't end proverty. Maybe they should make a law or something that says and declare that proverty needs to need. There's no specic date when it will end but it will. When it does I am going to be so proud, wheather I'm alive or not. **(SCORE=1)**

Student 2: I was convinced that winning the fight of poverty is achievable in our lifetime. Many people couldn't afford medicine or bed nets to be treated for malaria . Many children had died from this dieseuse even though it could be treated easily. But now, bed nets are used in every sleeping site . And the medicine is free of charge. Another example is that the farmers' crops are dying because they could not afford the nessacary fertilizer and irrigation . But they are now, making progess. Farmers now have fertilizer and water to give to the crops. Also with seeds and the proper tools . Third, kids in Sauri were not well educated. Many families couldn't afford school . Even at school there was no lunch . Students were exhausted from each day of school. Now, school is free . Children excited to learn now can and they do have midday meals . Finally, Sauri is making great progress. If they keep it up that city will no longer be in poverty. Then the Millennium Village project can move on to help other countries in need. **(SCORE=4)**

Automatic Scoring of an Analytical Response-To-Text Assessment (RTA)

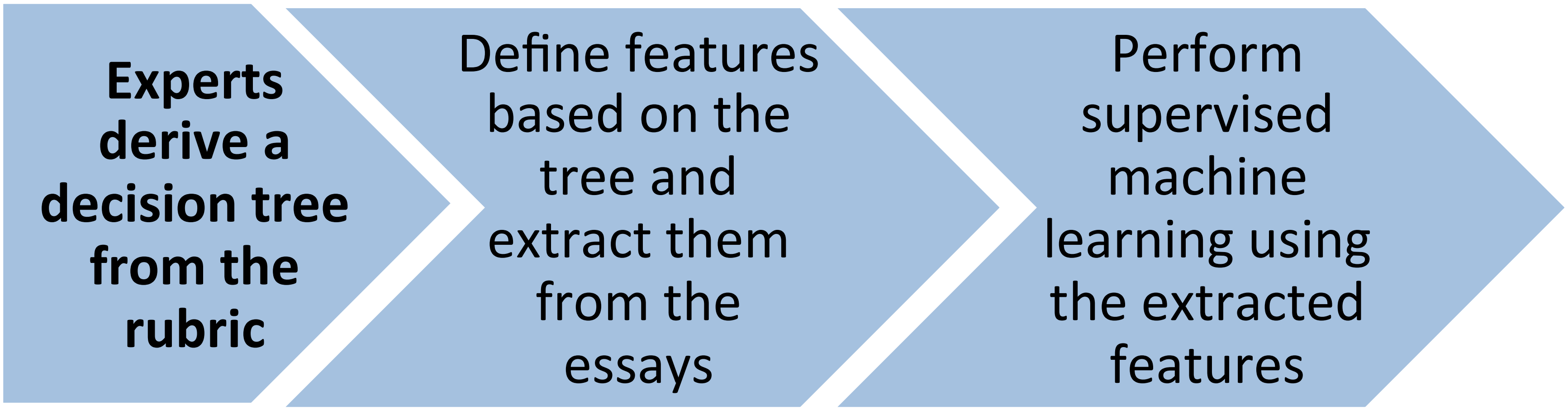
[Rahimi, **Litman**, Correnti, Matsumura, Wang & Kisa, 2014]

- Long-term goals
  - informative feedback for students and teachers
  - large-scale research on the impact of instruction and policies

- Current work
  - writing assessment via meaningful features that operationalize the ***Evidence*** rubric of RTA

- Argument mining subtasks
  - **segmentation**: spans of text
  - **segment classification**: evidence from text (or not)

Automatic Scoring Approach

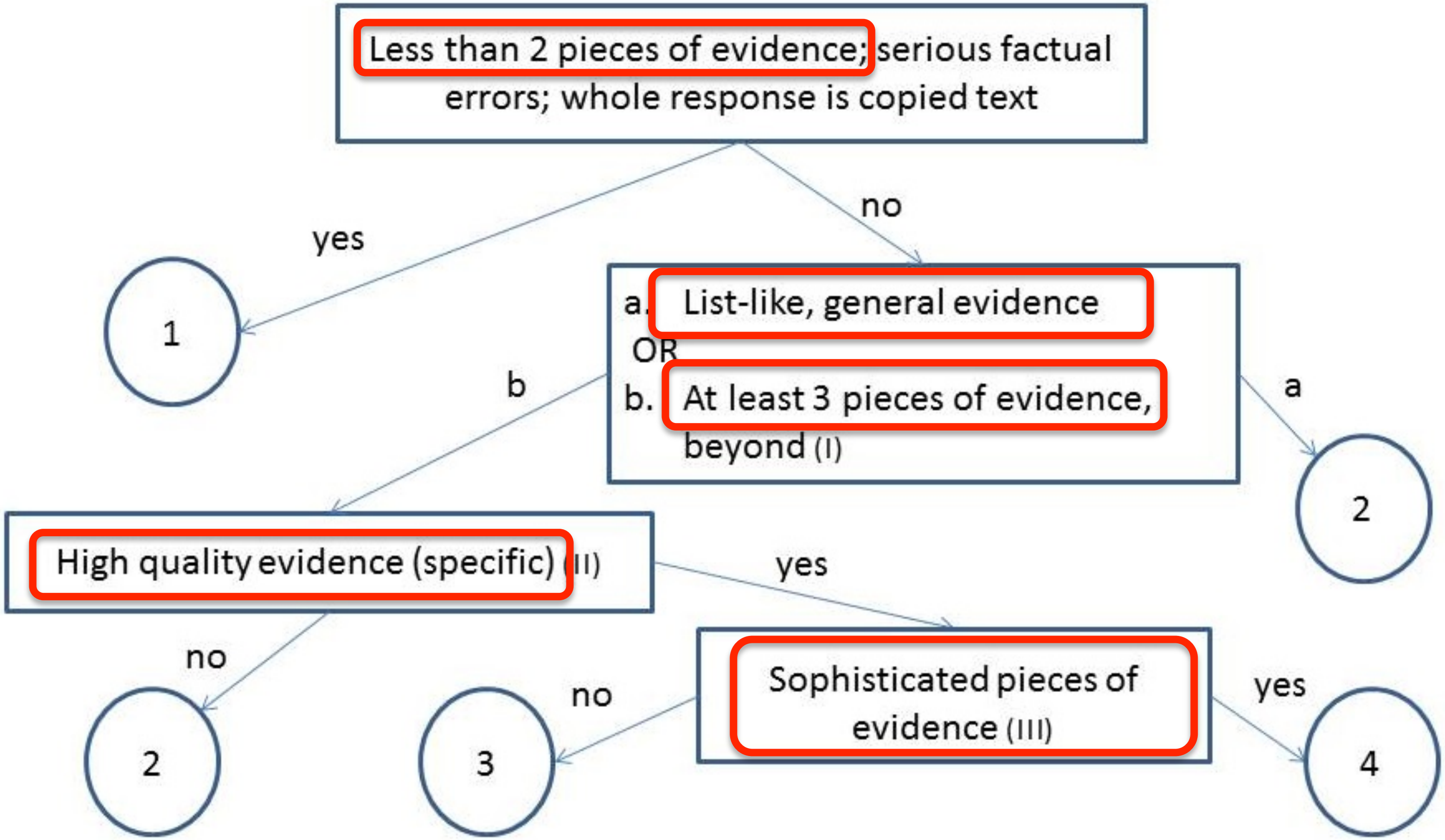


Experts derive a decision tree from the rubric

Define features based on the tree and extract them from the essays

Perform supervised machine learning using the extracted features

Scoring Essays for Evidence



Less than 2 pieces of evidence; serious factual errors; whole response is copied text

yes

1

no

a. List-like, general evidence

OR

b. At least 3 pieces of evidence, beyond (I)

a

2

b

High quality evidence (specific) (II)

no

2

yes

Sophisticated pieces of evidence (III)

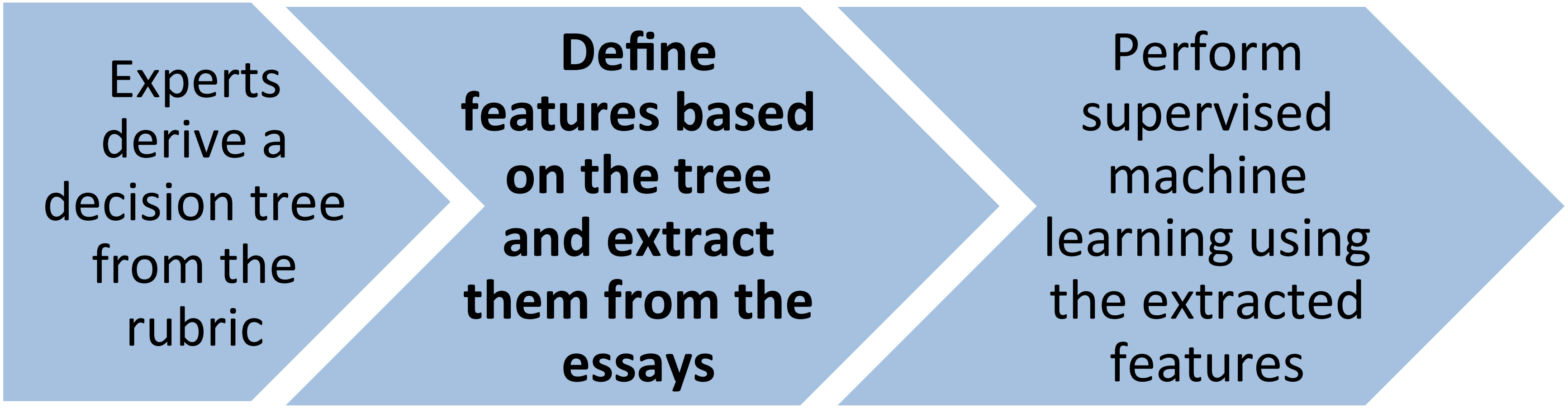
no

3

yes

4

Automatic Scoring Approach

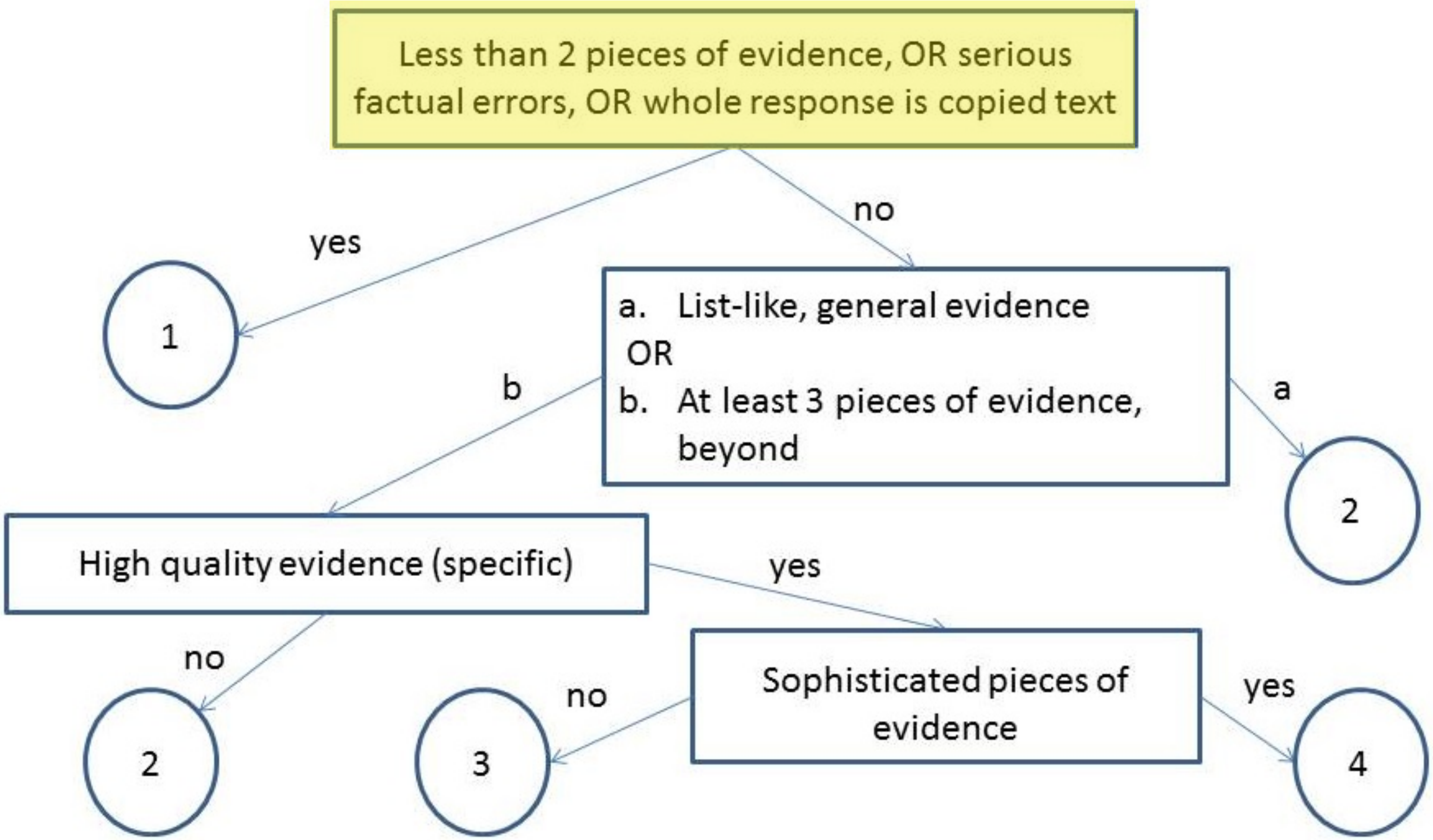


Experts derive a decision tree from the rubric

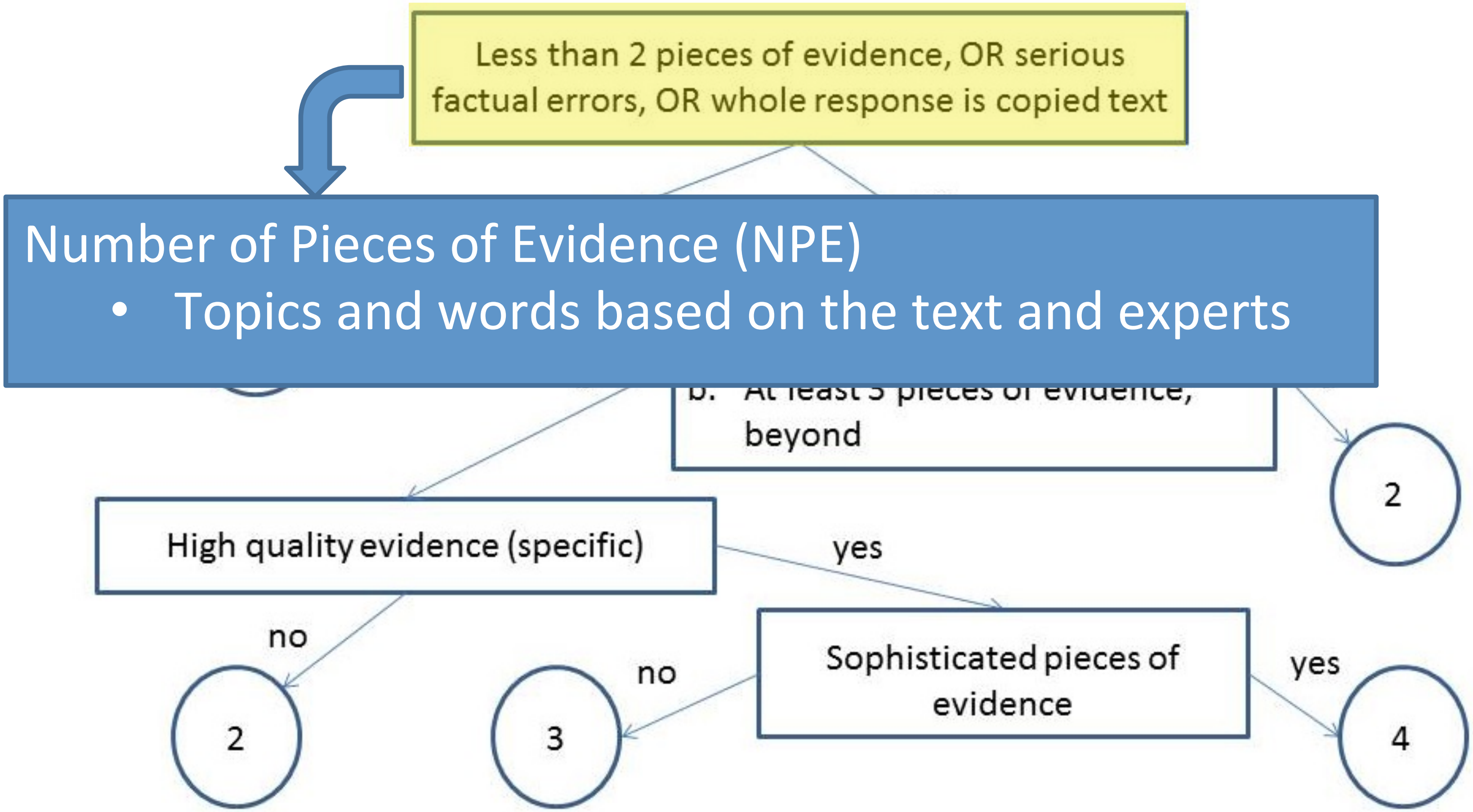
Define features based on the tree and extract them from the essays

Perform supervised machine learning using the extracted features

Our Features


Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text
yes
no
1
a. List-like, general evidence
OR
b. At least 3 pieces of evidence, beyond
b
a
2
High quality evidence (specific)
yes
no
2
no
3
Sophisticated pieces of evidence
yes
4

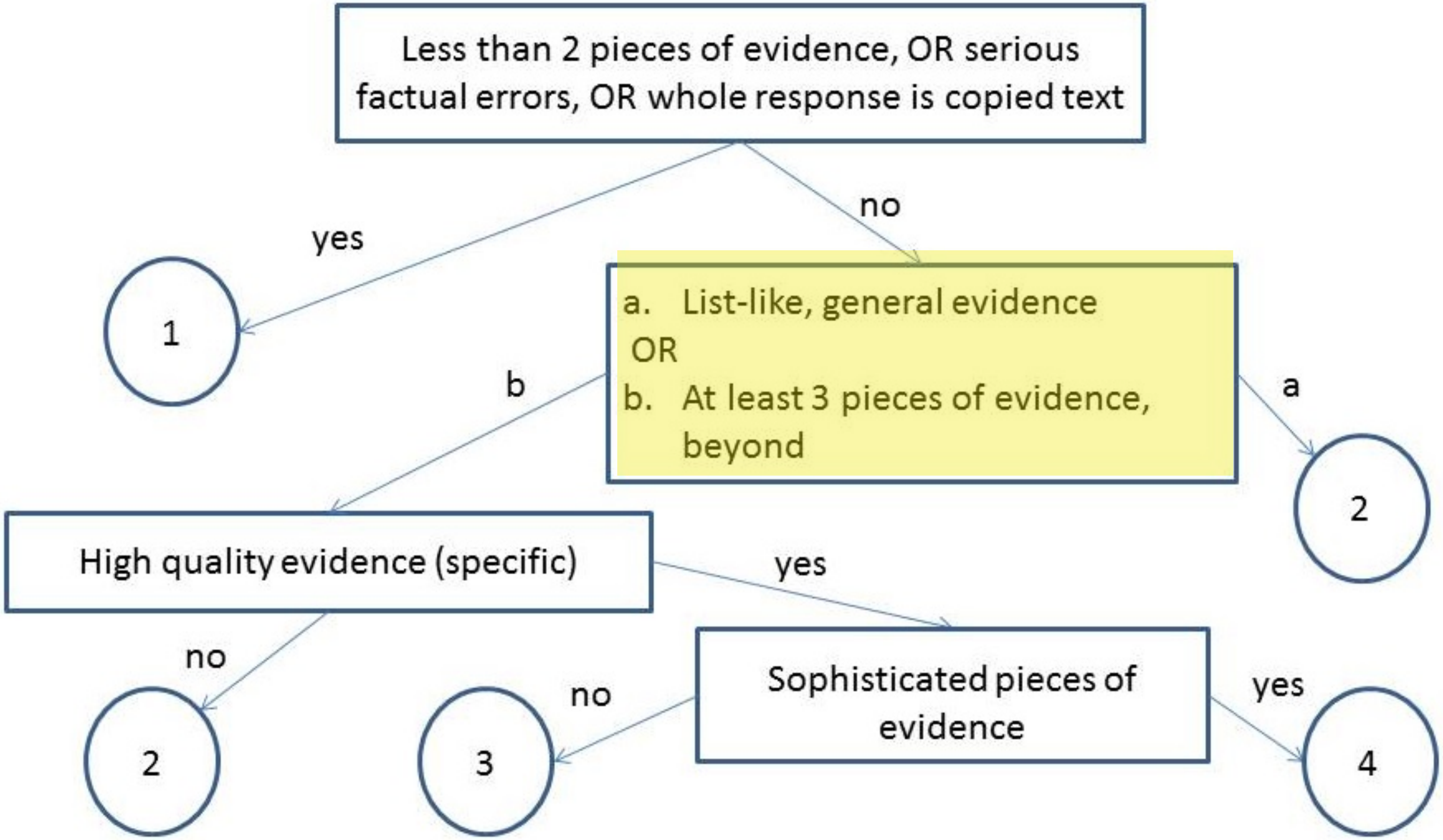
Our Features


Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text
b. At least 3 pieces of evidence, beyond
2
High quality evidence (specific)
yes
no
2
no
3
Sophisticated pieces of evidence
yes
4

Number of Pieces of Evidence (NPE)

- Topics and words based on the text and experts

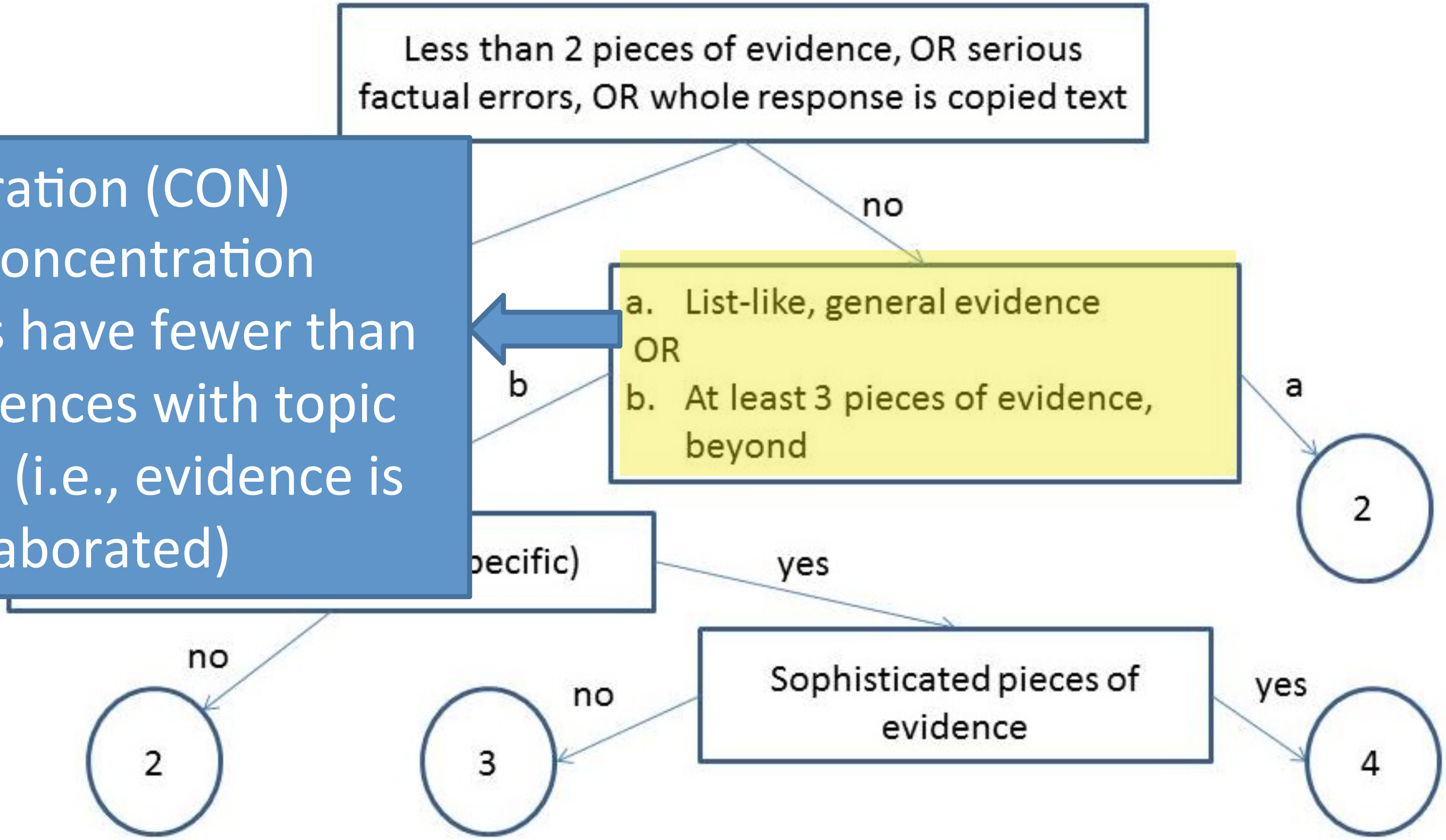
Our Features


Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text
yes
no
1
a. List-like, general evidence
OR
b. At least 3 pieces of evidence, beyond
b
a
2
High quality evidence (specific)
yes
no
2
no
3
Sophisticated pieces of evidence
yes
4

Our Features

Concentration (CON)

- High concentration essays have fewer than 3 sentences with topic words (i.e., evidence is not elaborated)



Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text

no

a. List-like, general evidence
OR
b. At least 3 pieces of evidence, beyond

a

b

2

pecific)

yes

no

2

Sophisticated pieces of evidence

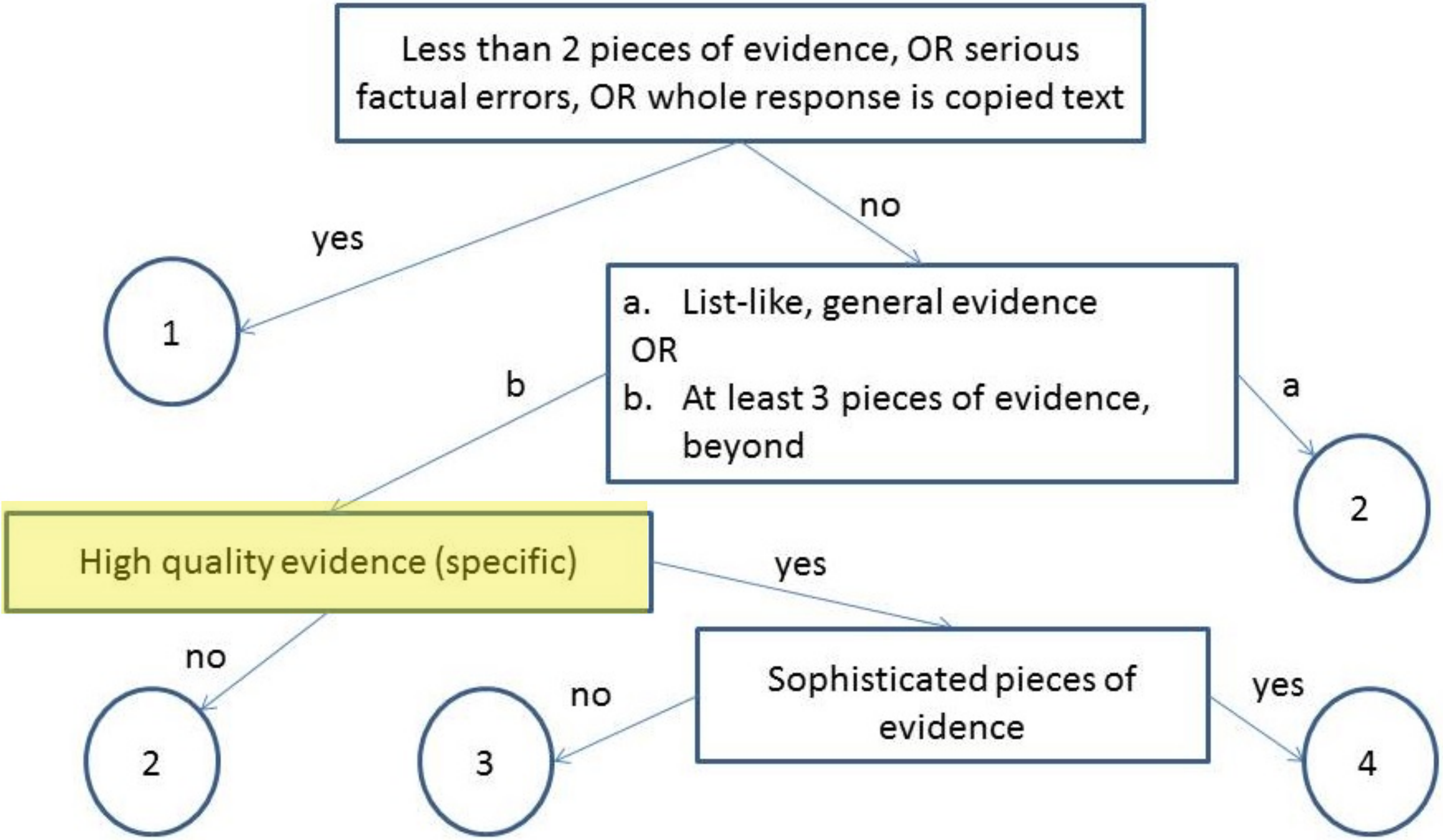
no

yes

3

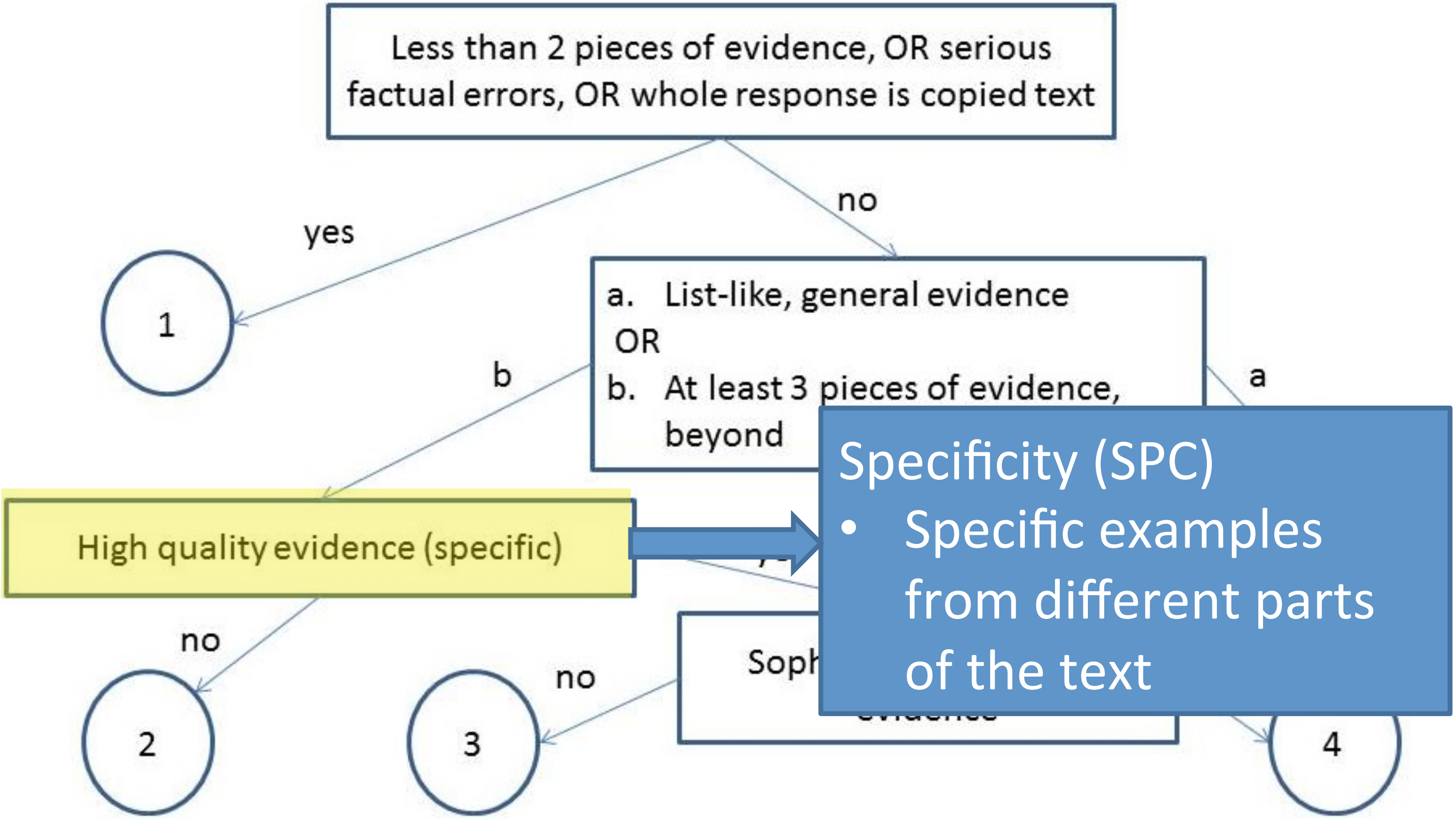
4

Our Features



Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text
yes
no
1
a. List-like, general evidence
OR
b. At least 3 pieces of evidence, beyond
b
a
2
High quality evidence (specific)
yes
no
2
Sophisticated pieces of evidence
no
yes
3
4

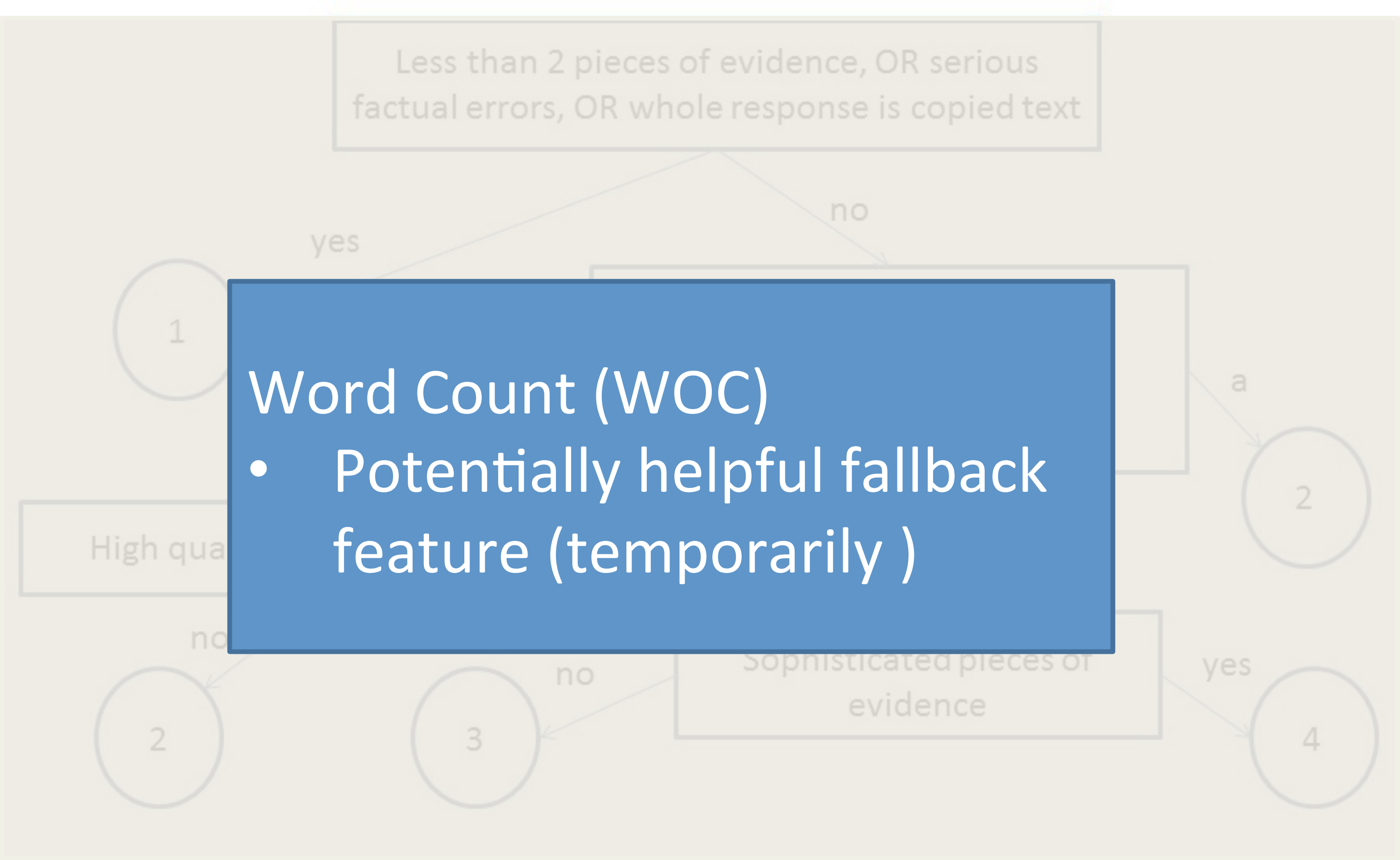
Our Features


Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text
yes
no
1
a. List-like, general evidence
OR
b. At least 3 pieces of evidence, beyond
b
a
High quality evidence (specific)
no
2
no
3
Soph
4

Specificity (SPC)

- Specific examples from different parts of the text

Our Features


Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text
yes
no
1
a
2
High qua
no
no
2
3
Sophisticated pieces of evidence
yes
4

Word Count (WOC)

- Potentially helpful fallback feature (temporarily)

Example Feature Vectors

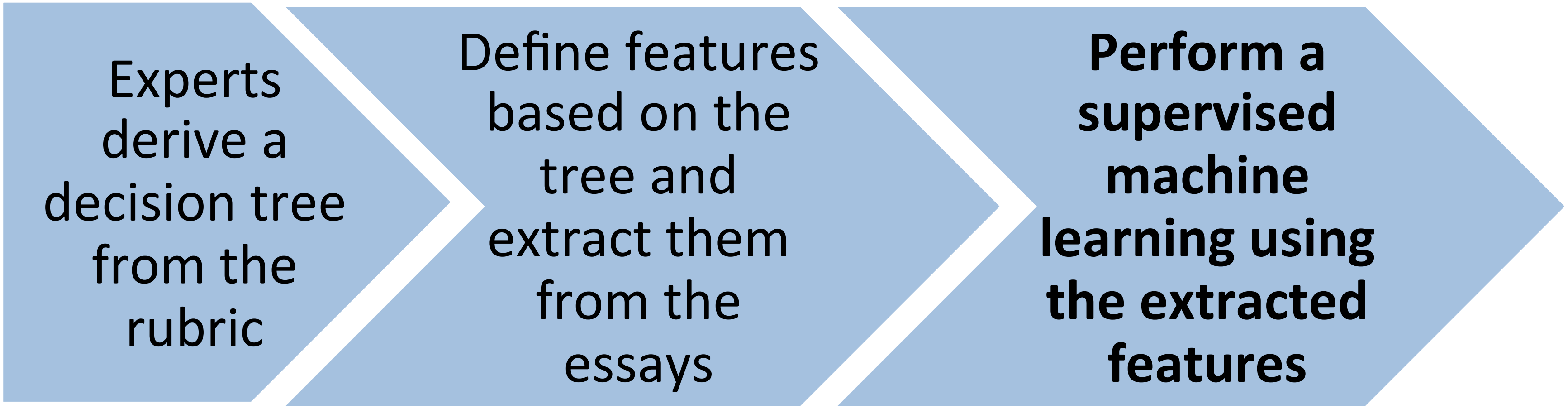
- Essay with Score=1 (from audience participation)

NPE	CON	WOC	SPC							
1	1	166	0	0	0	0	0	1	1	0

- Essay with Score=4 (from audience participation)

NPE	CON	WOC	SPC							
4	0	187	0	0	1	4	3	3	5	1

Automatic Scoring Approach



Experts derive a decision tree from the rubric

Define features based on the tree and extract them from the essays

Perform a supervised machine learning using the extracted features

Supervised Machine Learning

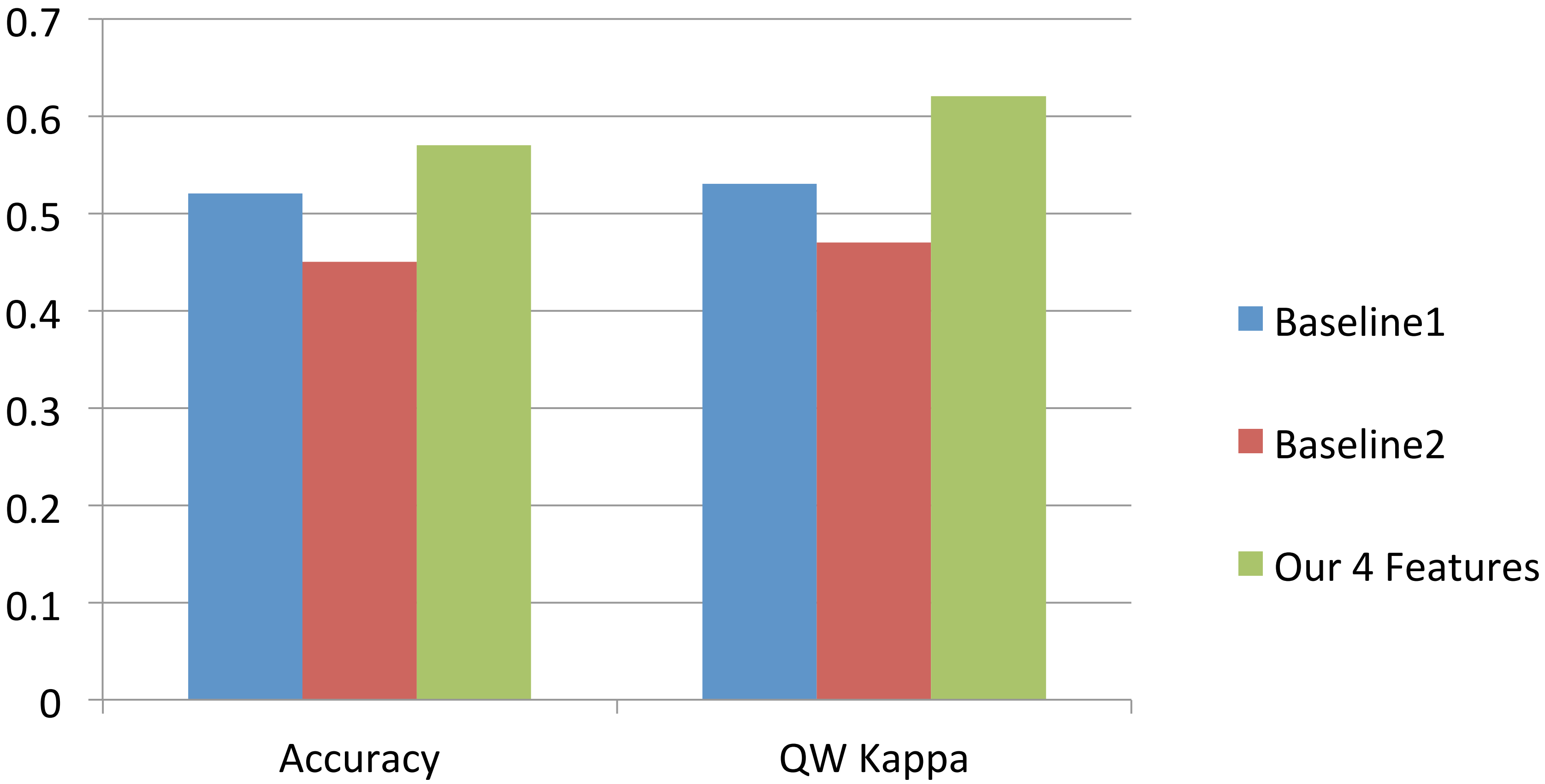
- Data [Correnti et al., 2013]
  - 1560 essays written by students in grades 4-6
    - Short, many spelling and grammatical errors

Experimental Evaluation

- Baseline1 [Mayfield 13]: one of the best methods from the Hewlett Foundation competition [Shermis and Hamner, 2012]
  - Features: primarily bag of words (top 500)

- Baseline2: Latent Semantic Analysis
  - Based on the scores of the 10 most similar essays, weighted by semantic similarity [Miller 03]

Results: Can we Automate?



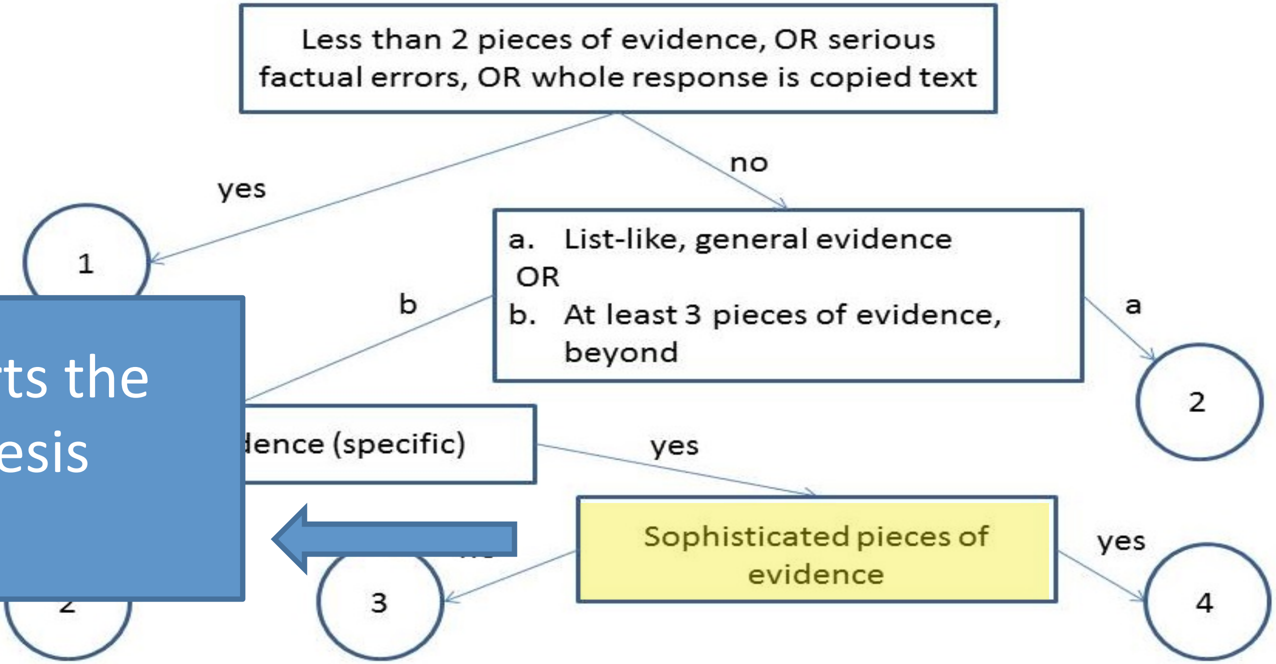
- **Proposed features** outperform both baselines

Other Results

- Removing features significantly degrades performance
- Wordcount is only useful for discriminating score 4 (where no rubric features were defined)
- Features also outperform baselines on essays from grades 6-8

Current Directions

- New types of argument mining
  - **segmentation**: sentences
  - **segment classification**: thesis statement
  - **relation identification**: supports (evidence, thesis)



Less than 2 pieces of evidence, OR serious factual errors, OR whole response is copied text
yes
no
1
a. List-like, general evidence
OR
b. At least 3 pieces of evidence, beyond
b
a
2
dence (specific)
yes
Sophisticated pieces of evidence
yes
3
4
2

Evidence supports the key idea (the thesis statement)

Current Directions (continued)

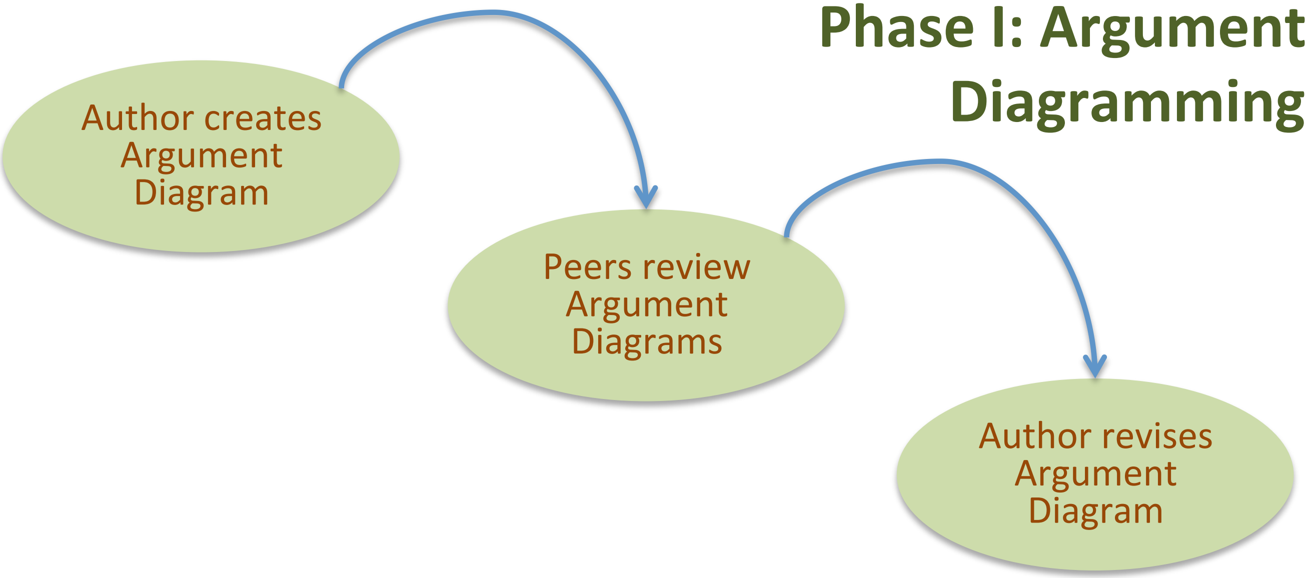
- Additional argument-related rubric dimensions
  - **Organization**

Outline

- Argumentative Writing / Argument Mining
- Algorithms and Applications
  - Automated Writing Assessment
  - ***Teaching with Diagramming and Peer Review***
- Summary and Current Directions

ArgumentPeer Project

Phase I: Argument Diagramming



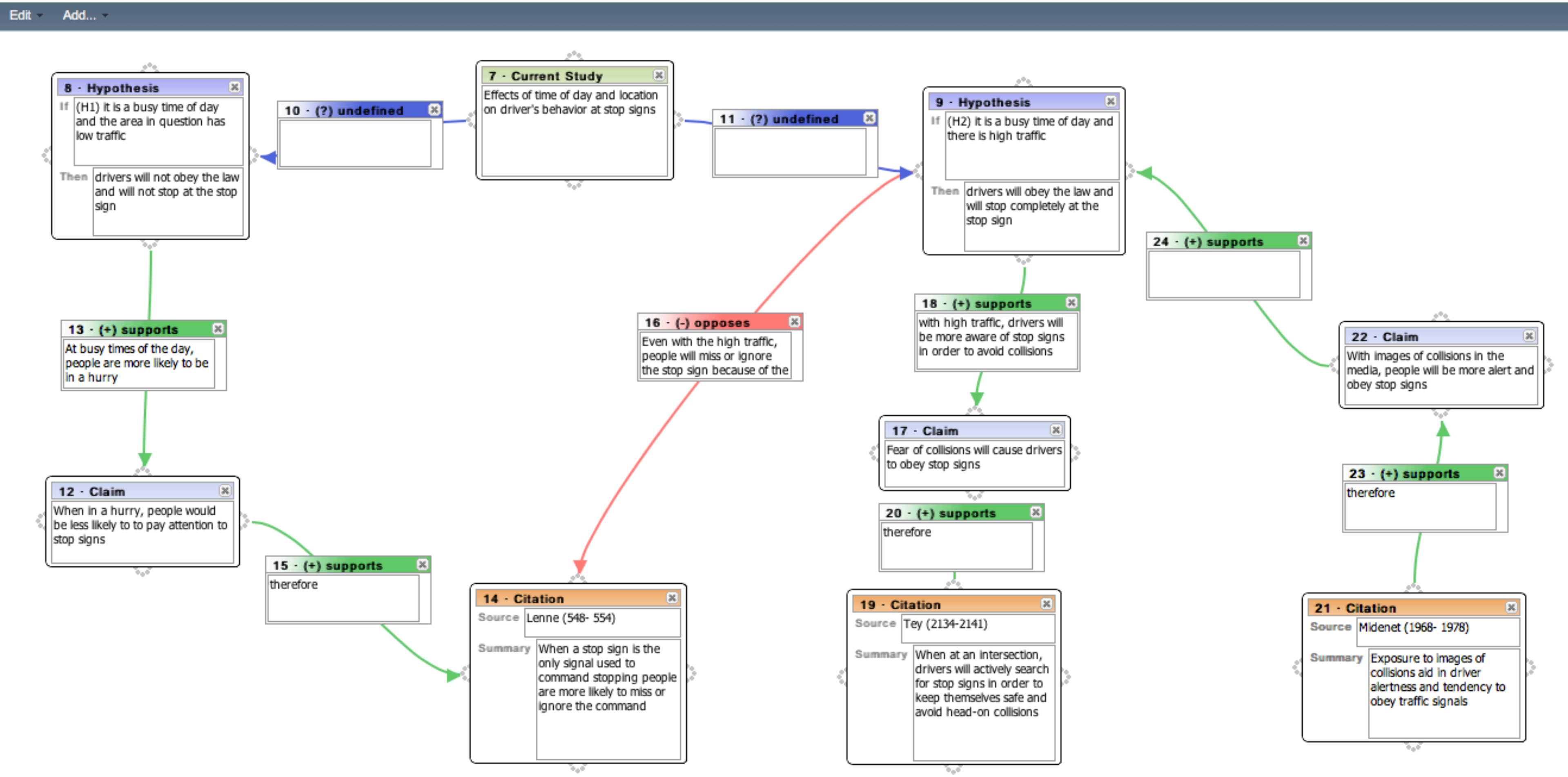
Author creates Argument Diagram

Peers review Argument Diagrams

Author revises Argument Diagram

Joint work with Kevin Ashley and Chris Schunn

Example Student Argument Diagram
(input via the LASAD system [Pinkwart et al.])



8 - Hypothesis
If (H1) it is a busy time of day and the area in question has low traffic
Then drivers will not obey the law and will not stop at the stop sign

10 - (?) undefined

7 - Current Study
Effects of time of day and location on driver's behavior at stop signs

11 - (?) undefined

9 - Hypothesis
If (H2) it is a busy time of day and there is high traffic
Then drivers will obey the law and will stop completely at the stop sign

24 - (+) supports

13 - (+) supports
At busy times of the day, people are more likely to be in a hurry

16 - (-) opposes
Even with the high traffic, people will miss or ignore the stop sign because of the

18 - (+) supports
with high traffic, drivers will be more aware of stop signs in order to avoid collisions

22 - Claim
With images of collisions in the media, people will be more alert and obey stop signs

17 - Claim
Fear of collisions will cause drivers to obey stop signs

12 - Claim
When in a hurry, people would be less likely to to pay attention to stop signs

23 - (+) supports
therefore

20 - (+) supports
therefore

15 - (+) supports
therefore

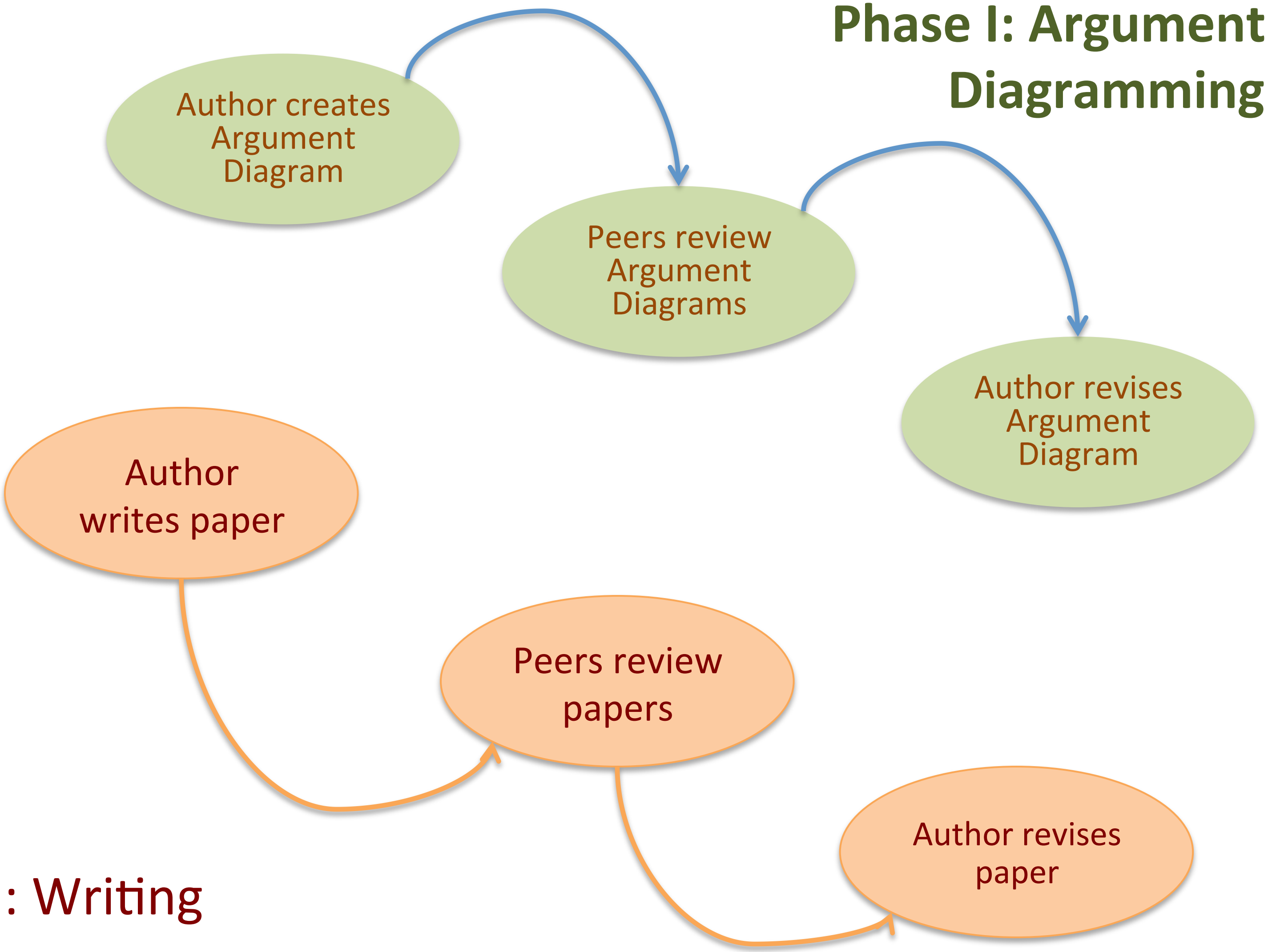
14 - Citation
Source Lenne (548- 554)
Summary When a stop sign is the only signal used to command stopping people are more likely to miss or ignore the command

19 - Citation
Source Tey (2134-2141)
Summary When at an intersection, drivers will actively search for stop signs in order to keep themselves safe and avoid head-on collisions

21 - Citation
Source Midenet (1968- 1978)
Summary Exposure to images of collisions aid in driver alertness and tendency to obey traffic signals

ArgumentPeer Project

Phase I: Argument Diagramming



Author creates Argument Diagram

Peers review Argument Diagrams

Author revises Argument Diagram

Author writes paper

Peers review papers

Author revises paper

Phase II: Writing

Joint work with Kevin Ashley and Chris Schunn

Why Diagram before Writing?

- Argument diagramming can improve essays
  - more support and opposition; increased relevance of citations [Ashley et al., 2014]
  - more supporting premises, evidence, explicit relations; increased accuracy [Harrell & Wetzel, 2013]

- Features extracted from argument diagrams can predict essays grades [Lynch et al., 2014]

- **Our research:** Can argument diagram constructs be *mined* directly from essays?

Audience Participation: Argument Mining

Written Essay Excerpt:

There have been many studies which indicate that people do drive differently at different times of day and that it does have an impact on driving risk. Reimer et al (2007) found that time of day did influence driving speed, reaction time, and speed variability measures.

Argument Diagram Ontology:

Node Types

- Current Study
- Hypothesis
- Claim
- Citation

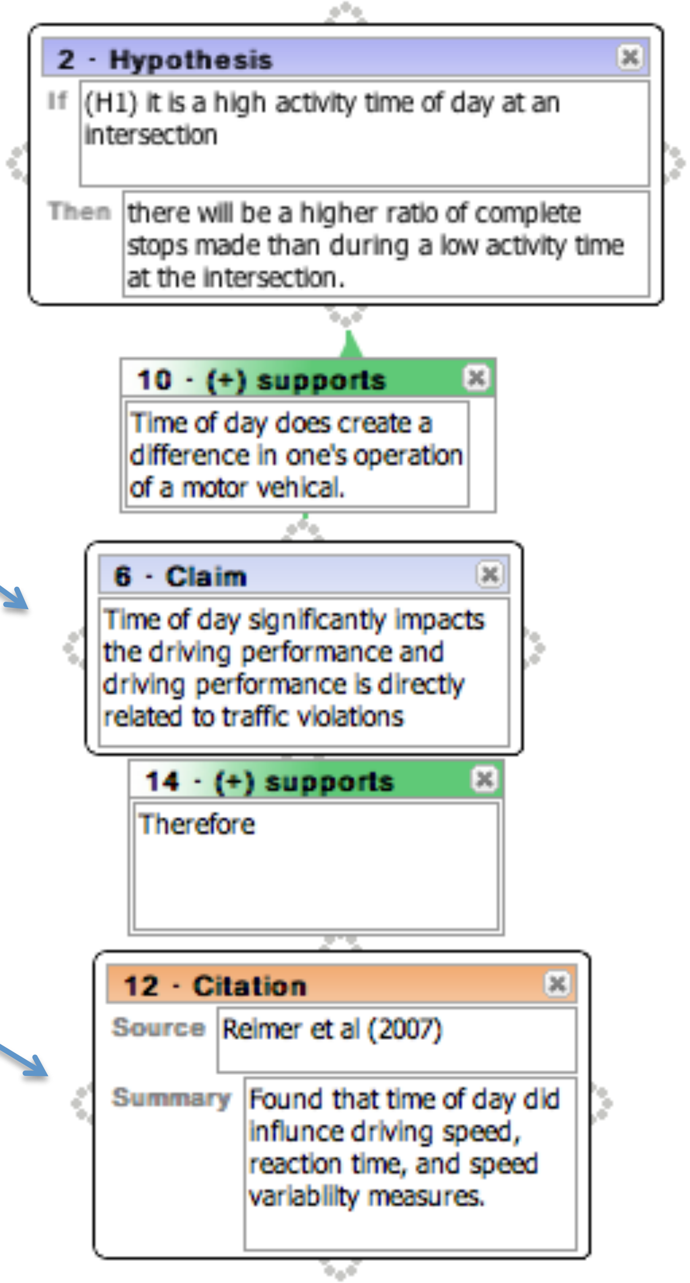
Arc Types

- Supports
- Opposes

Example Argument Mining Output

There have been many studies which indicate that people do drive differently at different times of day and that it does have an impact on driving risk. [Claim]

Reimer et al (2007) found that time of day did influence driving speed, reaction time, and speed variability measures… [Citation]



2 · Hypothesis

If (H1) It is a high activity time of day at an intersection

Then there will be a higher ratio of complete stops made than during a low activity time at the intersection.

10 · (+) supports

Time of day does create a difference in one's operation of a motor vehical.

6 · Claim

Time of day significantly impacts the driving performance and driving performance is directly related to traffic violations

14 · (+) supports

Therefore

12 · Citation

Source Reimer et al (2007)

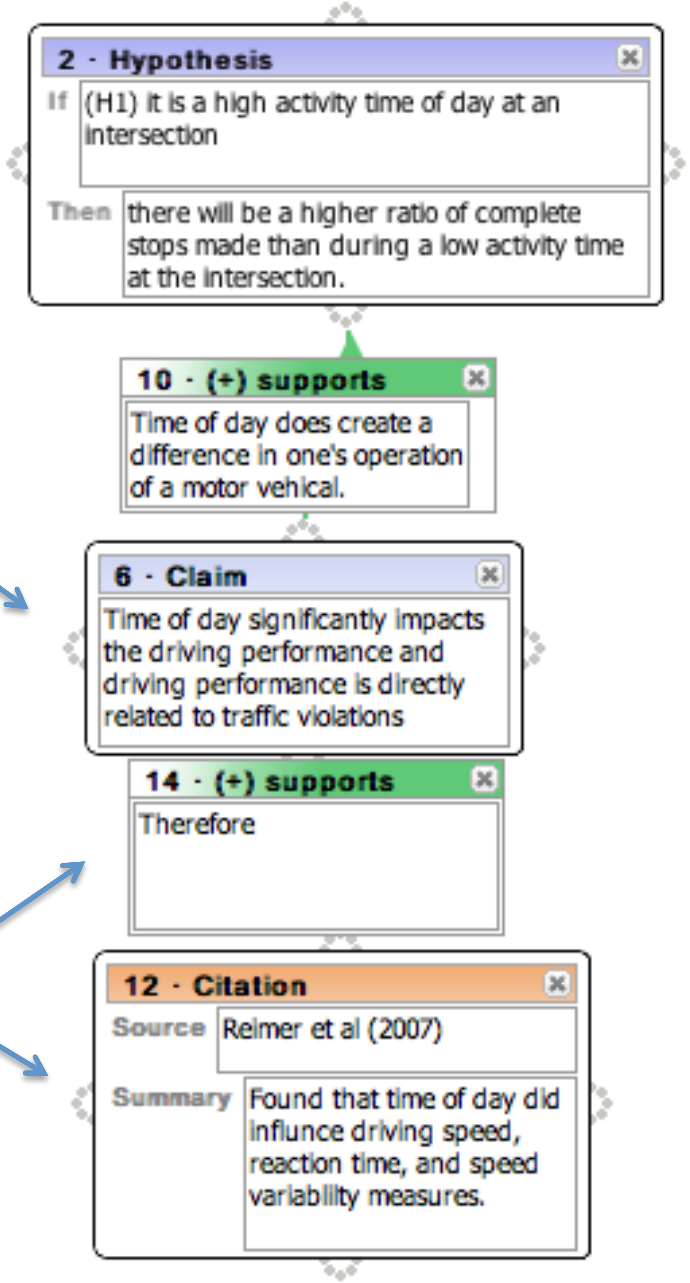
Summary Found that time of day did influnce driving speed, reaction time, and speed variablility measures.

Example Argument Mining Output

There have been many studies which indicate that people do drive differently at different times of day and that it does have an impact on driving risk. [Claim]

Reimer et al (2007) found that time of day did influence driving speed, reaction time, and speed variability measures… [Citation]

Citation supports Claim



2 - Hypothesis

If (H1) it is a high activity time of day at an intersection

Then there will be a higher ratio of complete stops made than during a low activity time at the intersection.

10 - (+) supports

Time of day does create a difference in one's operation of a motor vehical.

6 - Claim

Time of day significantly impacts the driving performance and driving performance is directly related to traffic violations

14 - (+) supports

Therefore

12 - Citation

Source Reimer et al (2007)

Summary Found that time of day did influnce driving speed, reaction time, and speed variability measures.

Audience Participation: Essay Scoring (1 or 4?)

Human Argument Mining

- **Hypothesis: 1, Supports: 2, Opposes: 1, Citation: 7 (6 unique)**

Frequently, people encounter situations in their environment where they can either choose to be prosocial or antisocial. It is therefore of great importance to study what makes people choose either response. Previous research has indicated that as group size increases, prosocial responses decrease, a phenomenon known as the bystander effect **(Fischer, Krueger, Greitemeyer, Vogrincic, Kastenmuller, Frey, Heene, Wicher, Kainbacher, 2011).** In our particular study, a participant sneezed on an elevator and noted whether or not somebody had a nice response such as "God bless you." The nice response would be equivalent to a prosocial behavior. The results of this study can give insight into what makes people more or less likely to engage in prosocial behavior in different situations, so this study has good external validity. This research topic is worth of study because the results may show how likely people are to help others in a more critical situation, such as somebody getting mugged, etc. It is important to realize that sneezing on an elevator is not necessarily a dangerous situation to act prosocially in, but the result will lead to a better understanding of whether or not gender and group size contribute to prosocial responses.

Previous research has shown that females seem to be more likely to engage in prosocial behavior as distance decreases, and this correlates to the hypothesis of this particular study **(McCullough, Tabak, 2010).** From this same previous research, it seems that females are more likely to do this because females tend to behave in more altruistic ways than males. In other words, females should give a nice response to a sneeze more often than males, and this was part of our hypothesis. Previous research has also indicated that people who have the knowledge of how to behave prosocially in a particular situation are more likely to do so **(Anker, Feeley, 2011).** If somebody in the elevator is from a certain religion or has recently moved and is not aware of what to say after somebody sneezes, then they most likely will not give a response at all, and this was not considered when putting together the results of the study. This should have been important to consider because knowledge of what to say after a sneeze could be a confounding variable in the results. Lastly, previous studies have shown that moral frames often determine whether or not people will give a prosocial response **(Thornberg, 2010)**. Overall, the better somebody's moral frames are, the more likely they are to give a prosocial response. This correlates to our particular study because the people who didn't give a response could possibly have negative moral frames. A future experiment could study this and the results would be interesting and could give additional information about what makes people behave prosocially.

In our study, it was hypothesized that as group size increased, the likelihood of a prosocial response would increase, and that females were more likely to give a prosocial response than males. Previous research from McCullough, etc (2011) conflicted with the hypothesis that as group size increases, prosocial responses should increase. McCullough's study showed this to be the opposite – as group size increases, prosocial behavior decreases, a phenomenon known as the bystander effect. Previous research about gender, however, showed that females are more likely to behave prosocially due to their nature and altruistic characteristics; this evidence supports our **hypothesis (Levine, Cassidy, Jentzsch, 2010).**

Audience Participation: Essay Scoring (1 or 4?)

Human Argument Mining

- **Hypothesis: 1, Supports: 5, Opposes: 3, Citation: 15 (6 unique)**

In today's tough economic climate, any factors that may contribute to a successful business or enterprise have become increasingly important. One prevalently researched factor is prosocial behavior which is comparable to helping behaviors. Research has shown that prosocial behaviors have been associated with job satisfaction, productivity, and client satisfaction **(Danzis & Stone---Romero, 2009)**. Therefore the understanding of what situations support and encourage prosocial behaviors could be an asset to any manager or business (**Danzis & Stone--- Romero, 2009)**.

Previous research has focused on relationships of prosocial behavior between gender and the effect of group size. With regards to gender, there has been much debate over whether females or males are inherently more prosocial. Research has been done on gender by implementing a training program on empathy with children in day care centers between the ages of 5---6 **Kallipuska, 1991). The results of this study concluded that females were more prosocial than males based on teacher appraisals (Kallipuska, 1991)**.**Other research supports that females are more prosocial but only to familiar individuals or when following a social norm (Danzis & Stone---Romero 2009)**. **In contrast, males were considered to be more "chivalrous"(p.722) which would promote them to exhibit more prosocial behavior toward strangers relative to females(Danzis & Stone---Romero, 2009)**.

In addition to studies on the relationship between prosocial behavior and gender, there is also research between group sizes and prosocial behavior. **One of the most well--known psychological phenomena is the bystander effect which describes a diffusion of responsibility as group size increases** (**Garcia, Weaver, Darley, & Moskowitz, 2002)**.More recent studies have described an implicit bystander effect which claims that it is even possible to create diffusion of responsibility even when no one else is around by priming participants about group mentality **(Garcia, Weaver, Darley, & Moskowitz, 2002).** There has even been debate over the implications of the implicit identity effect. One research study claimed that the implicit identity can only be used to inhibit prosocial behavior such as through the bystander effect **(Garcia, Weaver, Darley, & Moskowitz, 2002)**. **Yet, conflicting research has found evidence to support a theory that the implicit identity effect can be used to encourage prosocial behavior if it follows the group's social norms (Levine & Crowther, 2008). Also, a study by Krupka and Weber (2009) found that even when people observe others acting selfishly, on average this produces prosocial behavior due to the focusing effecting of** norms (2009).Therefore the focusing effect of norms can be defined as being aware of the behavior of others around**Krupka & Weber, 2009).**

The current study sought to find relationships between prosocial behavior, gender and group size. Prosocial behavior was defined as someone verbally responding to the observer's sneeze. In order to create a controlled environment, the study took place on elevators in on---campus buildings only. Unlike previous research, our study did not prime the participants before they entered the elevator so it was comparable to the control group of the Krupka & Weber study(2009). **I predicted that females would respond more than males and that responses would be more likely in a group of more than two other people on the elevator.According to the Kallipuska study, females appear to be more prosocial and polite than males and would be more likely to respond to a sneeze(1991)**. With regards to group size, research shows that the focusing effect of norms becomes more apparent when others are present**(Krupka & Weber, 2009).One contradiction to this hypothesis would be the Garcia, Weaver, Darley, & Moskowitz study that the implicit bystander effect can cause a diffusion of responsibility in large groups which could cause no one to respond to the sneeze(2002)**.

Actual Expert Essay Scores (via rubric)

SCORE = 1

- Sentences/Words: 22/564
- Hypothesis: 1
- Supports: 2
- Opposes: 1
- Citation: 7 (6 unique)

SCORE = 4

- Sentences/Words: 26/610
- Hypothesis: 1
- Supports: 5
- Opposes: 3
- Citation: 15 (6 unique)

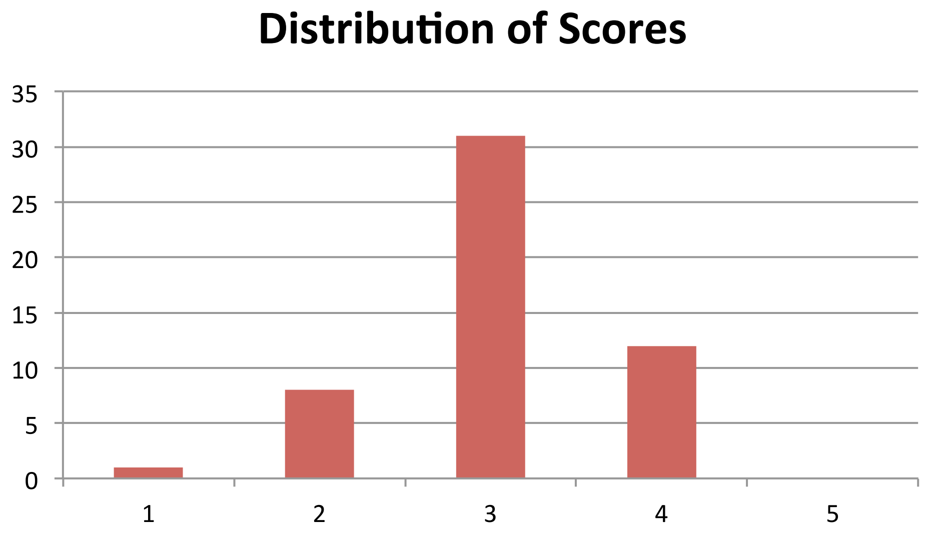
Ontology-Based Argument Mining and Automatic Essay Scoring

[Ong, **Litman** & Brusilovsky, 2014; Nguyen & Litman, submitted]

- System recognizes diagram ontology in essays
- System scores essays using recognized ontology
- Argument mining subtasks
  - **Segmentation**: sentences
  - **Segment classification:** diagram ontology tags
    - E.g., Study, Hypothesis, Opposes, Supports, Claim, Citation *(instructor-defined)*
    - None

Initial Data

- 52 first-draft essays from two undergraduate psychology courses
  - Written after diagramming and peer-feedback
  - Average length: 5.2 paragraphs, 28.6 sentences
  - Expert scores: Average = 3.03



Essay Processing Pipeline

1. Discourse Processing
   – Tag essays with discourse connective senses
     - *Expansion, Contingency, Comparison, Temporal, Non-discourse*
     - Tagger from Upenn

2. Argument Ontology Mining
   – Tag essays with diagram ontology elements
     - Rule-based algorithm

3. Ontology-Based Scoring
   – Use the mined argument to score the essays
     - Rule-based algorithm

Experimental Results: Extrinsic Evaluation
[Ong, Litman & Brusilovsky, 2014]

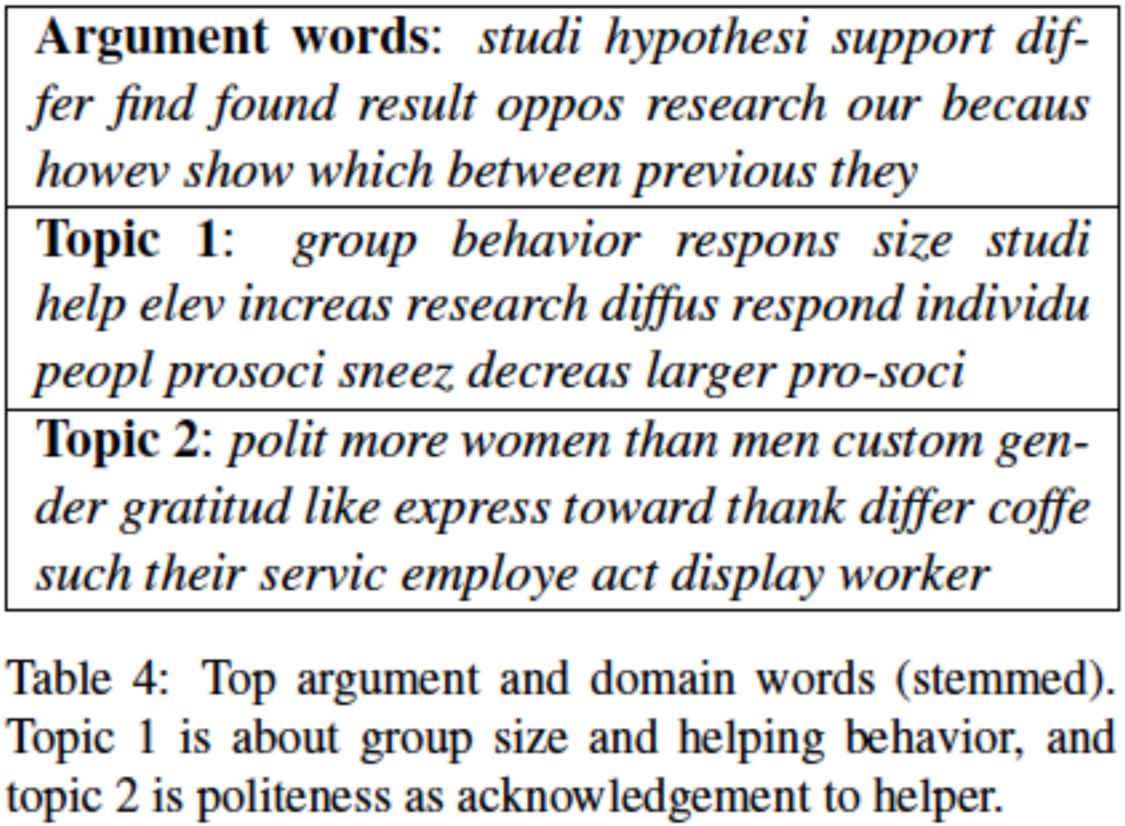
- Spearman Correlation trend between automatically generated and expert scores
  - number of automatically generated tags for diagram elements are positively correlated with score

- These ranking results can focus teacher or peer reviewer attention on particular types of papers

Current Directions: Intrinsic Evaluation (with H. Nguyen)

- Argument annotation of 116 first-draft essays (Kappa > .6)
  - *Hypothesis*: 186 sentences
  - *Support/Opposition*: 460 sentences
  - *None*: 2208 sentences
- Features for supervised learning
  - Persuasive essays [Stab & Gurevych, 2014]
  - Academic writing style
  - LDA-enabled features and constraints
    - Ontology seeding to extract argument vs. domain words

Post-Processed LDA Output



Argument words: *studi hypothesi support differ find found result oppos research our becaus howev show which between previous they*
Topic 1: *group behavior respons size studi help elev increas research diffus respond individu peopl prosoci sneez decreas larger pro-soci*
Topic 2: *polit more women than men custom gender gratitud like express toward thank differ coffe such their servic employe act display worker*

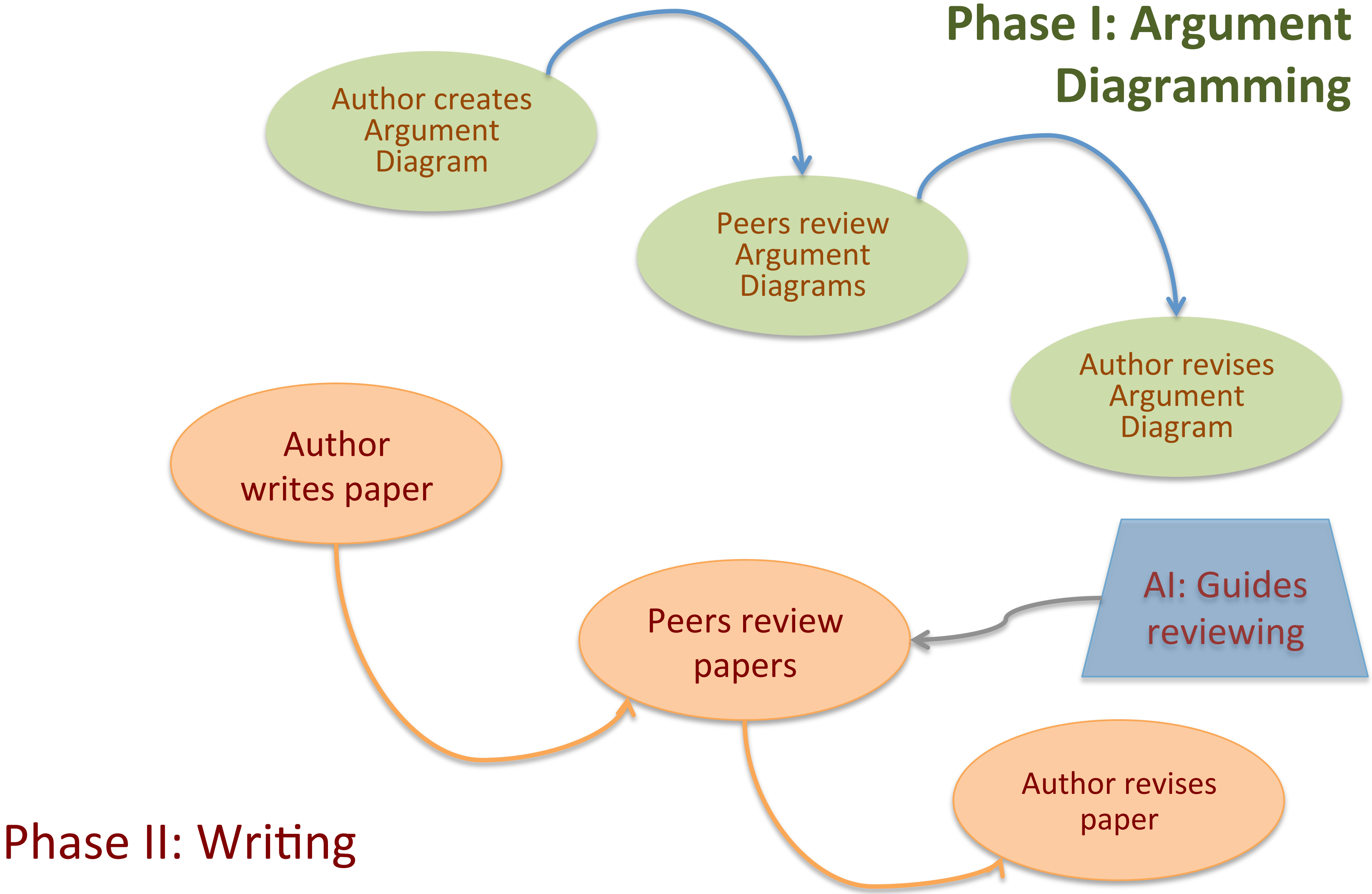
Table 4: Top argument and domain words (stemmed). Topic 1 is about group size and helping behavior, and topic 2 is politeness as acknowledgement to helper.

- New Features: # domain words, # argument words, argument unigrams
- New Constraints: e.g. selected dependency relations

Evaluation Results

- 10-fold cross-essay validation
  - .56 Kappa (versus .51 for Stab14 baseline)
- Holdout validation with unseen topics
  - .58 Kappa (versus .50 for Stab14 baseline)
- New features significantly improve performance, generalize better, and compact the feature space
- Current work: argument relations; essay scoring; use of diagrams

ArgumentPeer Project



Phase I: Argument Diagramming

Author creates Argument Diagram

Peers review Argument Diagrams

Author revises Argument Diagram

Author writes paper

Peers review papers

AI: Guides reviewing

Author revises paper

Phase II: Writing

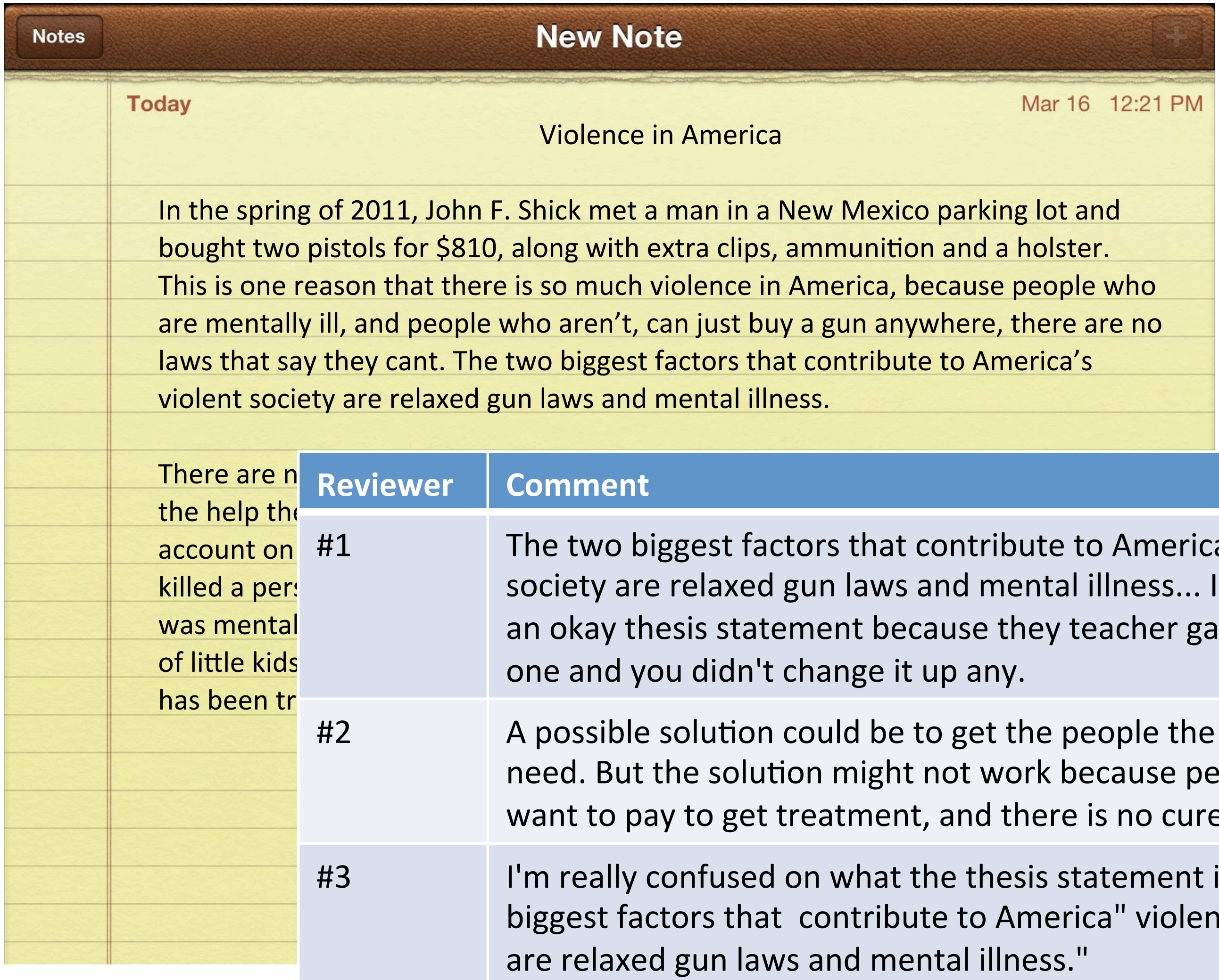
Joint work with Kevin Ashley and Chris Schunn

Why Peer Review?

- Pros
  - Quantity and diversity of review feedback
  - Students learn by reviewing

- Cons
  - Reviews are often not stated in effective ways
  - Reviews and papers do not focus on core aspects
    - E.g., no discussion of thesis statements

Audience Participation: Argument Mining (Detect the Thesis)



Violence in America

In the spring of 2011, John F. Shick met a man in a New Mexico parking lot and bought two pistols for $810, along with extra clips, ammunition and a holster. This is one reason that there is so much violence in America, because people who are mentally ill, and people who aren't, can just buy a gun anywhere, there are no laws that say they cant. The two biggest factors that contribute to America's violent society are relaxed gun laws and mental illness.

There are n
the help the
account on
killed a pers
was mental
of little kids
has been tr

Reviewer	Comment	Back Evaluation
#1	The two biggest factors that contribute to America violent society are relaxed gun laws and mental illness... I think it's an okay thesis statement because they teacher gave of us this one and you didn't change it up any.	We didnt have to
#2	A possible solution could be to get the people the help they need. But the solution might not work because people don't want to pay to get treatment, and there is no cure	This is not my thesis
#3	I'm really confused on what the thesis statement is, "The two biggest factors that contribute to America" violent society are relaxed gun laws and mental illness."	That's obviously my thesis because its the last sentence
#4	Good thesis.	thanks

Identifying Thesis and Conclusion Statements in Student Essays to Scaffold Peer Review [Falakmasir, Ashley, Schunn, **Litman** 2014]

- How well can machine learning distinguish thesis or conclusion sentences from other sentences in student essays?
  - Feature extraction via natural language processing
- Argument mining subtasks
  - **Segmentation**: sentences
  - **Segment classification**: probability of being thesis, conclusion, …

Data

- 432 high school essays from 8 assignments
  - 18,081 sentences
- Thesis/conclusion manually identified by experts
  - AP English Language & Composition

Features for Machine Learning

- 3 feature sets inspired by [Burstein et al. 2003]
  - **Positional:** paragraph number, sentence number in the paragraph, type of paragraph (first, body, last)
  - **Sentence Level**: syntactic, semantic, frequent words
  - **Essay Level**: number of keywords among the most frequent words of the essay, number of words overlapping with the assignment prompt, and a sentence importance score based on Rhetorical Structure Theory

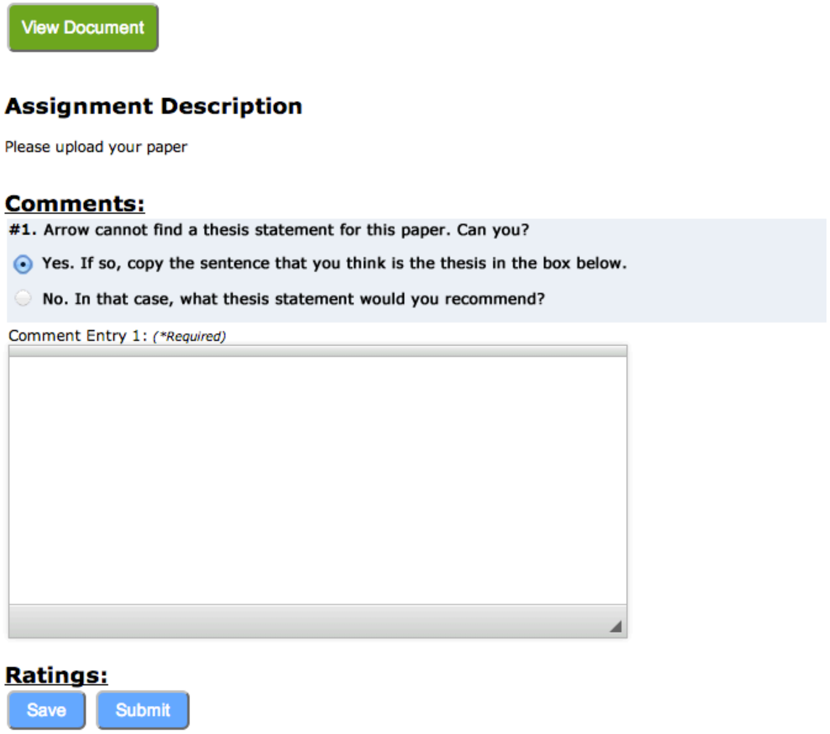
Argument Mining Results

- F-Measure on unseen test set
  - Thesis Sentences
    - Positional Baseline: .57
    - Machine Learning: .74
  - Conclusion Sentences
    - Positional Baseline: .55
    - Machine Learning: .67

- Even with a small training corpus, NLP features can detect core argument sentences in essays better than the baseline

Scaffolding During Peer Review [Falakmasir, Jabbari et al., in progress]

When argument mining cannot find a thesis sentence, the first peer review prompt is:



Assignment Description

Please upload your paper

Comments:

#1. Arrow cannot find a thesis statement for this paper. Can you?

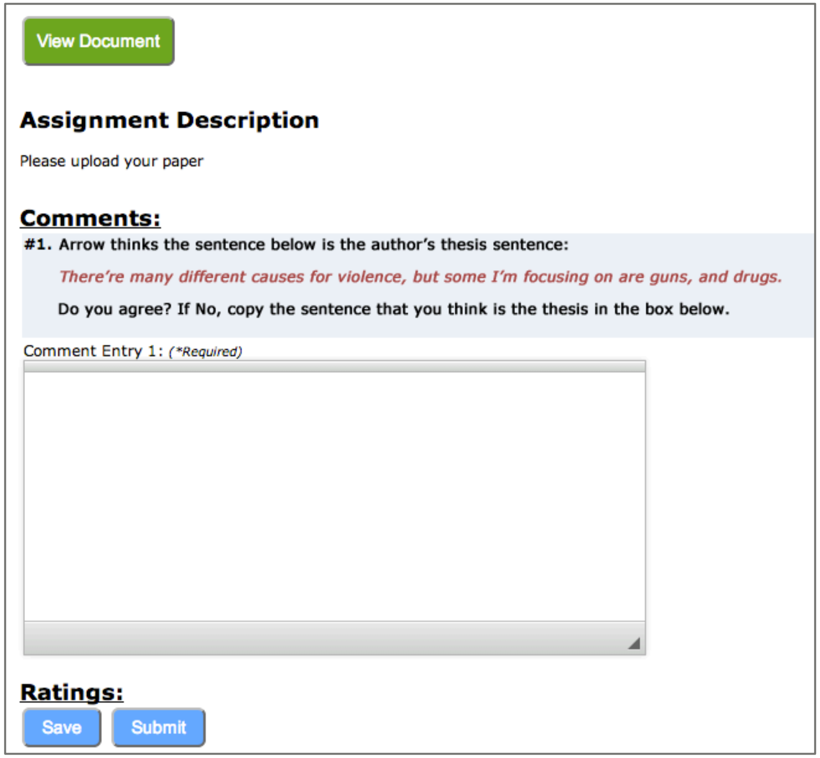
● Yes. If so, copy the sentence that you think is the thesis in the box below.

○ No. In that case, what thesis statement would you recommend?

Comment Entry 1: *(*Required)*

Ratings:

When argument mining identifies a thesis sentence, the first peer review prompt is:



Assignment Description

Please upload your paper

Comments:

#1. Arrow thinks the sentence below is the author's thesis sentence:

There're many different causes for violence, but some I'm focusing on are guns, and drugs.

Do you agree? If No, copy the sentence that you think is the thesis in the box below.

Comment Entry 1: *(*Required)*

Ratings:

Evaluation in High-School Classrooms

Argument mining performance

- E.g., how many actual theses had highest score?

Educational performance

- impact of scaffolding for peer review comments

Outline

- Argumentative Writing / Argument Mining
- Algorithms and Applications
  - Automated Writing Assessment
  - Teaching with Diagramming and Peer Review
- ***Summary and Current Directions***

Summary

- NLP-supported argument mining for teaching and assessing writing
  - Feature / Algorithm Development
    - Noisy and diverse data
    - Meaningful features
    - Real-time performance
  - Experimental Evaluations
    - Response-to-Text Assessment (grade school)
    - Ontology-Guided Essay Ranking (university undergraduates)
    - Intelligent Scaffolding for Peer Review (high school)
- Even non-structural and application-dependent argument mining can support useful applications!

New Work: *Temporal* Argument Mining [Zhang and Litman, 2014]

- How are *arguments changed* during paper revision?
  - **1st Draft:** The fifth level of Hell "contains the Wrathful" (110). Wrathful people want to intentionally harm others. **Saddam Hussein and Osama Bin Laden come** to mind when mentioning wrathful people.
  - **2nd Draft:** The fifth level of Hell "contains the Wrathful" (110). Wrathful people want to intentionally harm others. **Fidel Castro comes** to mind when mentioning wrathful people
  - **Revision modifies *evidence***
    - Rather than *come* -> *comes*
- Argument mining subtasks
  - **Segmentation**: sentences
    - Alignment needed between 1st and 2nd paper drafts
  - **Segment classification**: argumentation purpose of revision (if any)

Acknowledgements

- Response to Text Assessment
  - **Faculty:** Richard Correnti, Lindsay Clare Matsumura
  - **PhD Students:** Zahid Kisa, Zahra Rahimi, Elaine Wang

- SWoRD and ArgumentPeer
  - **Faculty:** Kevin Ashley, Amanda Godley, Chris Schunn
  - **Postdocs:** Alok Baikadi, Amanda Crowell, Lisa Fazio, Jordan Lippman
  - **PhD Students:** Sara DeMartino**,** Mohammad Falakmasir**,** Fataneh Jabari**,** Adam Loretto**,** Collin Lynch**,** Huy Nguyen, Wenting Xiong, Fan Zhang
  - **Undergraduates:** Alexandra Brusilovsky, Chris Clark, Nathan Ong
  - **Staff:** Alexsandar Ivetic, Carmela Rizzo

Thank You!

- Questions?

- Further Information (and downloads!)
  - http://www.cs.pitt.edu/~litman