Towards Improving Current Automatic Essay Scoring and Constructive Feedback Systems

Paul Reisert

RIKEN Center for Advanced Intelligence Project @ Tohoku University, Sendai, Japan

Advisor: Kentaro Inui

Joint work with: Naoya Inoue, Farjana Sultana Mim, Keshav Singh

Informal Argumentation Working @ NII

November 18th, 2018

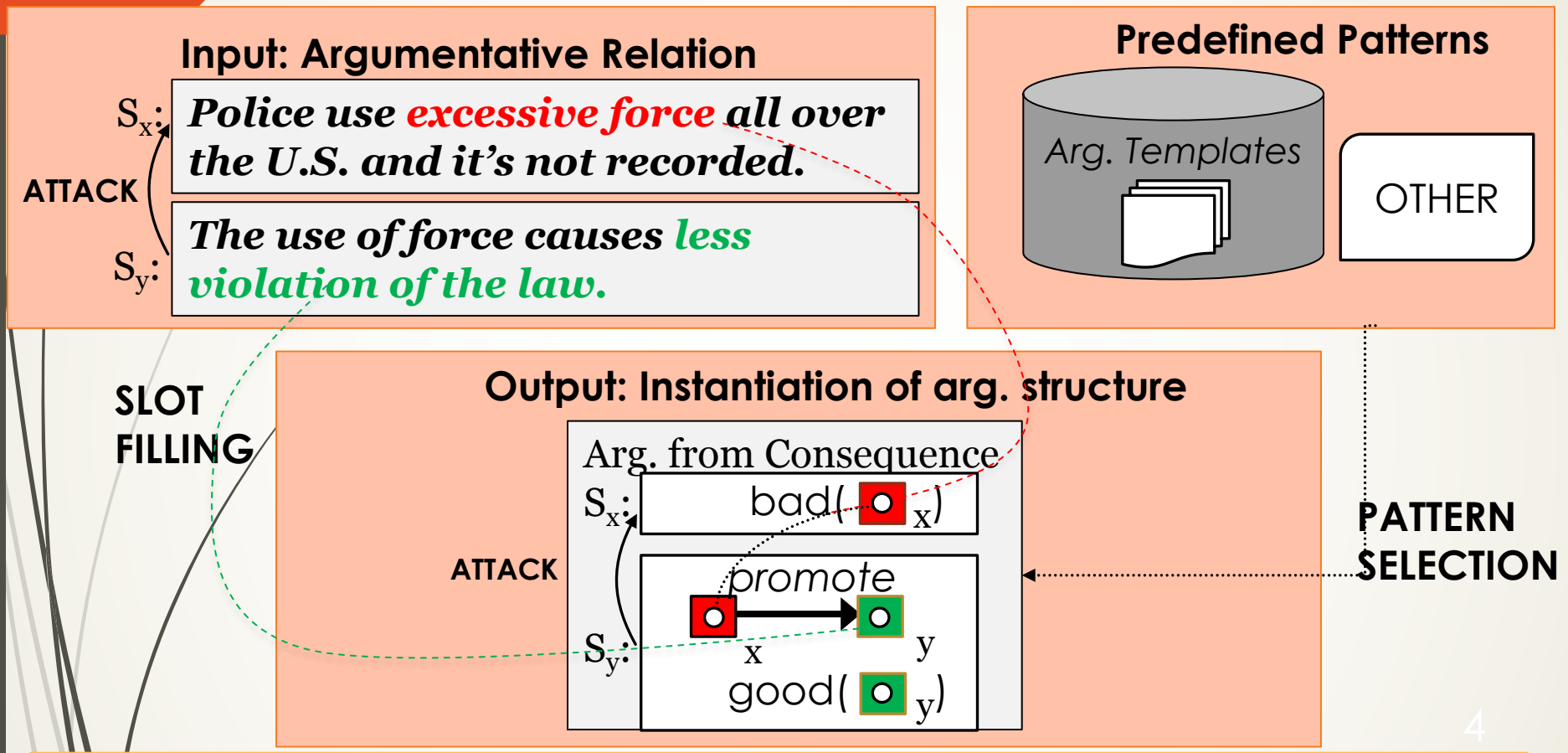
Discussion Outline

- Previous Research Topic
- Counter-Argument Generation (**Paul Reisert**)
- Improving Organization Scores for Essays (**Farjana Sultana Mim, Tohoku University, M1**)
- Implicit Warrant Identification using Background Knowledge (**Keshav Singh, Tohoku University, M1**)

Discussion Outline

- Previous Research Topic
  - Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates
- Counter-Argument Generation
- Improving Organization Scores for Essays
- Implicit Warrant Identification using Background Knowledge

Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates [5th ArgMining, EMNLP2018]



4

- Aim to capture **implicit reasoning** between argumentative components, inspired by Argumentation Schemes [Walton+, 08]
- Existing work suffers from difficult annotation guidelines [Reed+, 06]
- Created a corpus of instantiated templates on top of arg-microtexts corpus [Peldzsus+, 15] with good coverage (76%) and annotator agreement (.80 IAA)

Discussion Outline

- Previous Research Topic
- Counter-Argument Generation
- Improving Organization Scores for Essays
- Implicit Warrant Identification using Background Knowledge
- Conclusion and Future Plan

Discussion Outline

- Previous Research Topic
- Counter-Argument Generation
  - Background
  - Research Questions
  - Proposed Methodology
  - Related Work
  - Applications
  - Corpus Construction
  - Conclusion and Future Plan
- Corpus Construction
  - Crowdsourcing Interface Construction
  - Experiments and Results

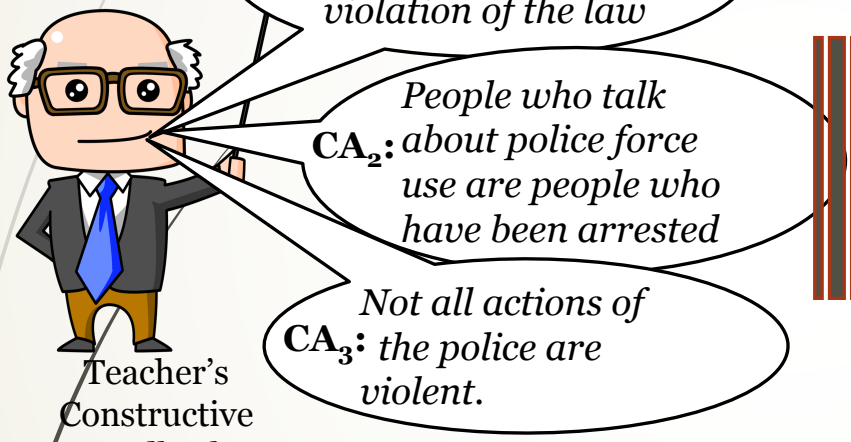
Part 1: Counter-Argument Generation (Paul Reisert)

Big Picture

Prompt P1: Are police too willing to use force?

Argument A1: *Police are too willing to use force. Police are using excessive force all over the U.S. and it's not recorded.*

Student A Essay (Input)



Teacher's Constructive Feedback **(Counter-Arguments)** **Part 1**

Inform

Student A

Part 3: Quality Scores
Organization
Content
etc.

Revision

Output:

Revised Argument R1: *Police are too willing to use force, but as a result, crime is reduced. Although many people think that arrested individuals discuss this issue, police are using excessive force all over the U.S. Granted, this force is not always violent.*

Part 2: Machine is required to understand implicit arguments (i.e. warrants)
A1 assumes "*force does not cause less violation of the law*"

Research Questions (RQs)

- **MainRQ1: How one can scale the educational process of producing counter-arguments automatically with the help of NLP technology?**
  - RQ1: Can we make a large-scale training dataset for this task which can be used for training a computational model?
  - RQ2: Even if we create the training data, how can we reasonably generate counter-arguments for prompts with limited training data?

Methodology

2. Encoder-Decoder Model

1. Corpus Construction

Prompt 1:	Arg. 1 → Counter-Argument 1 (CA1)
Prompt 1:	Arg. 2 → CA2
Prompt 2:	Arg. 1 → CA3

Analyze

Counter-Argument Typology

train

generate

useful for

refine

In-Domain (seen prompts)

Prompt 1:	Arg. 3 → CA3
Prompt 2:	Arg. 3 → CA3

Out-Domain (unseen prompts)

Prompt 3:	Arg. 1 → CA1
Prompt 4:	Arg. 1 → CA1

useful for

3. Preliminary Feedback Experiment with Actual Students

Apply to

Related Work (1/2)

- Teaching critical questions about argumentation through the revising process: effects of strategy instruction on college students' argumentative essays [Song & Ferretti, 2013]
  - Showed the importance of argumentation schemes in revising essays
  - Small sample of essays
  - Teachers manually graded the works

Argument From Consequences

Argumentation Scheme:

Use **good** consequences to **support** the implementation of a policy.
If the policy is implemented, then good consequences will (may plausibly) occur.
Therefore, the policy should be implemented.

Use **bad** consequences to **reject** the implementation of a policy.
If the policy is implemented, then bad consequences will (may plausibly) occur.
Therefore, the policy should not be implemented.

Critical Questions:

1. *Are these claimed consequences likely to happen if the policy is implemented?*
2. *If the policy is implemented, what evidence supports that these consequences will occur?*
3. *Are there other consequences of the opposite value that should be taken into account?*

Argument From Example

Argumentation Scheme:

Argument from example to support the reason:
There is an example to show that the reason is true.
Therefore the reason is true.

Critical questions:

1. *Is the example true?*
2. *Is the example typical of the kinds of the examples that illustrate the generalization?*
3. *Are there any special circumstances that might impair the generalizability of the examples?*

Related Work (2/2)

- Neural Argument Generation Augmented with Externally Retrieved Evidence [Hua and Lang, 2018]

- Retrieval of the Best Counterargument without Prior Topic Knowledge [Wachsmuth et al., 2018]
  - Focuses on finding the best counter-argument

Does not focus on local argument.
Output not too informative.

Input: Putin is trying to re-form a "Soviet Union" with his past actions in Georgia and current actions in Ukraine
Human: So far Putin has only really targeted areas where the local population is majority russian (or historically majority russian) and sympathetic to russia. That's not to say that he wouldn't go further, but I don't think there is yet any compelling reason to believe he would.
Retrieval: In 1968, the Soviet Union and Warsaw pact allies invaded Czechoslovakia to halt the Prague spring reforms. It was this significant lack of flexibility that was exploited by the US to undermine the strength of the Soviet Union and thus foster its reform. In the summer of 1991, the new union treaty, which would have turned the Soviet Union into a much looser union, was agreed upon by eight republics.
Our Model: There are two reasons you are so far beyond. There is no reason to see the military army. You can infer what they assume it, so they tend to protect up. There. If we assume it were bad, they can not be controlled. So we're talking going to ignore the wars of the world. The fact of the matter is not bad. I would also assume it. However, the government can not be reminded of the world.

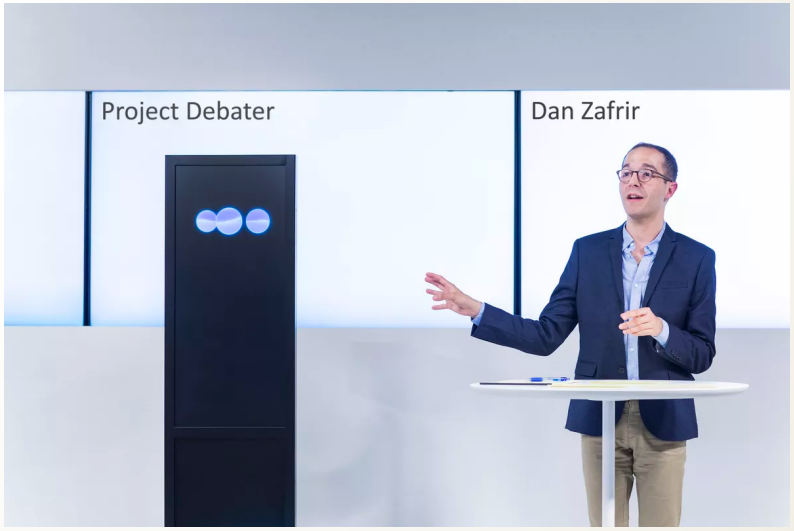
Argument *"Gun ownership is an integral aspect of the right to self defence.* (conclusion)
Law-abiding citizens deserve the right to protect their families in their own homes, especially if the police are judged incapable of dealing with the threat of attack. [...]" (premise)

While the conclusion seems well-reasoned, the web portal directly provides a counter to the argument:

Counterargument *"Burglary should not be punished by vigilante killings of the offender. No amount of property is worth a human life. Perversely, the danger of attack by homeowners may make it more likely that criminals will carry their own weapons. If a right to self-defence is granted in this way, many accidental deaths are bound to result. [...]"*

Applications

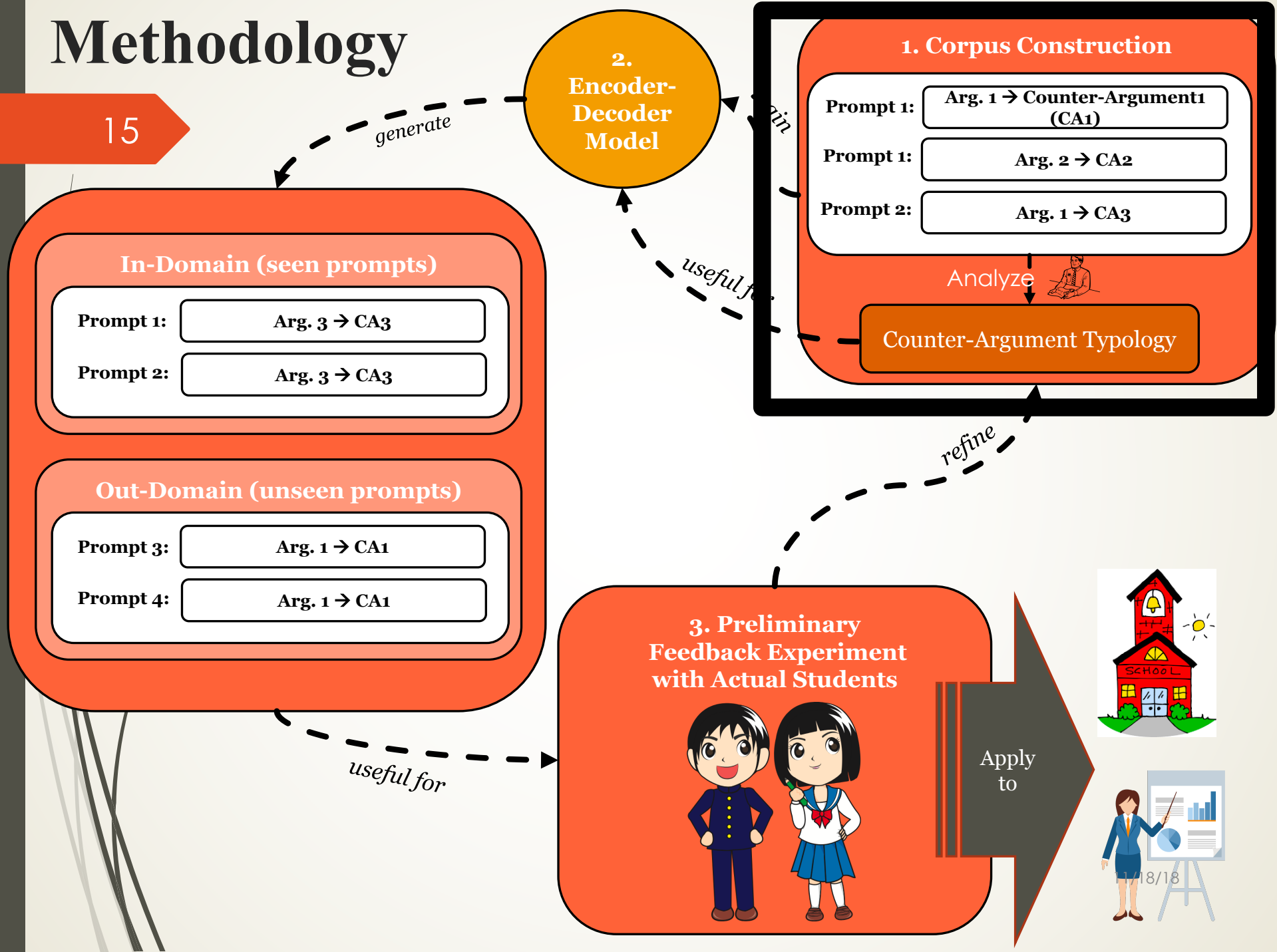
- **Essay scoring** [Persing&Ng, 2015; Ghosh+, 2016; Wachsmuth+ 2016]
- **Argumentative Writing Support** [Stab+ 2014; Stab&Gurevych, 2017]
- **AI Debating Systems** [https://www.research.ibm.com/artificial-intelligence/project-debater/]



Discussion Outline

- Research Overview
- Corpus Construction
  - Crowdsourcing Trial
  - Experiments and Results
- Conclusion and Future Plan

Methodology



Corpus Construction

- **Counter-Argument Generation (CAG) via Crowdsourcing (CS)**
  - *RQ1: Can we make a large-scale training dataset for this task which can be used for training a computational model?*
  - CS Worker must be able to identify reasoning or factual flaw in the original argument for producing counter-argument
  - Why CS?
    - Groups outperform individuals on reasoning tasks [Trouche et al., 2014]
    - Large-scale
    - Fast
- **Two CS Tasks**
  - Generation: Ask workers to generate a counter-argument.
  - Verification: Ask workers to verify the generated counter-argument.

CS Trial Experiment

- **Dataset**
  - Persuasive Essay Corpus [Stab+ 2014]
  - Claim-Premise pairs
- **Platform**
  - Figure Eight (Crowdflower)
- **Settings**
  - Default settings
  - Level 1 reliability (quick, less reliable workers)
  - No time limit
- **Number of workers**
  - 25 counter arguments
  - Judged by 3 annotators each

CA Generation Interface

Generation Interface

topic : There She Is, Miss America
claim : Miss America is good for women
premise : Miss America gives honors and education scholarships.

Please write a counter-argument that attacks the claim, premise, or both. (required)

Enter the text here.

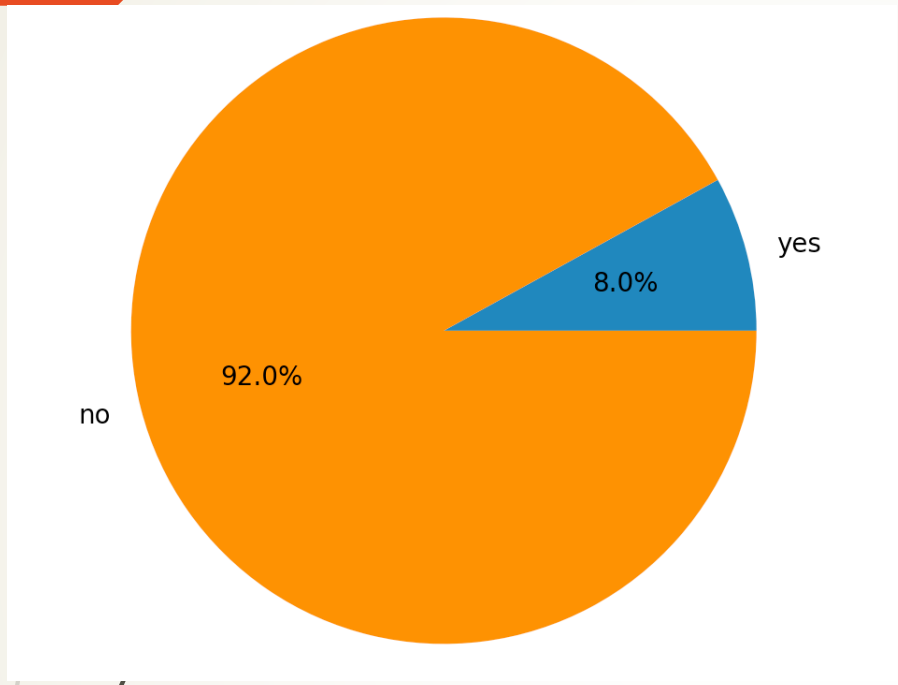
Verification Interface

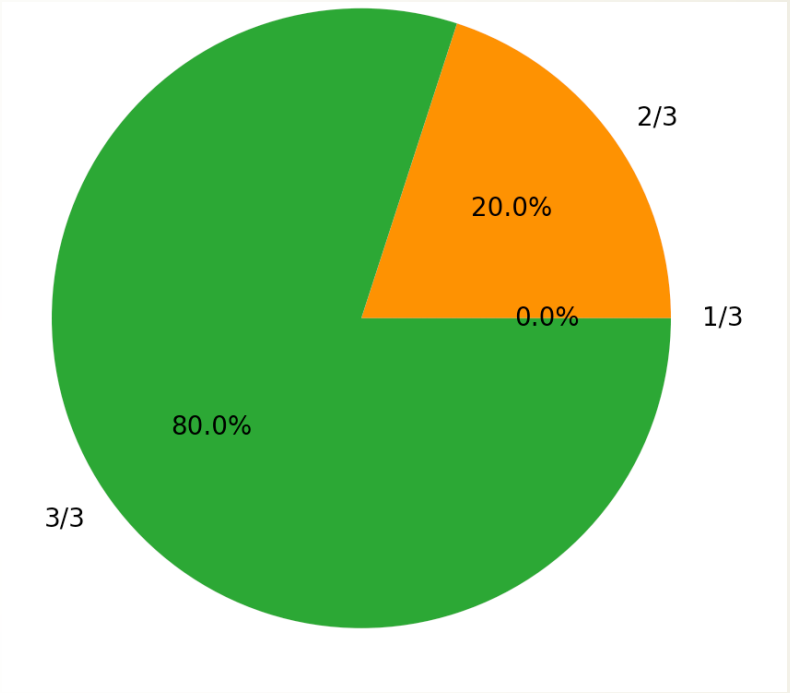
Topic : There She Is, Miss America
Claim : Miss America is good for women
Premise : Miss America gives honors and education scholarships.
Counter-Argument : Miss America is very bed specialy for women, married and with kids

Does the counter-argument attack the claim, premise, or both? (required)

- ✓ Select one
- Yes
- No
- Unsure

CAG Verification for First Trial (T1)





- **Results**
  - Almost 92% of the counter-arguments were bad
  - Analyzed the results →

Good/Bad CAs for T1

20

Topic	Target	Source	Good Counter-Arguments
The Internet is an adequate source of academic information	the Internet is an adequate source of academic information, which will potentially fulfill the needs of university pupils	the Internet offers a more effective and practical method of studying	The internet is also offering some misleading and harmful method of studying.
Living in smal towns	another advantage of small towns is living costs	we can save time and money	Life is not cheaper in all small towns.
Children engagement in paid work	when children take jobs, they tend to be more responsibl	whether they can earn money or not will depend on their effectiveness and attitudes in working	Children working means they have the money to get in the wrong direction.

Topic	Target	Source	Counter-Argument
The Internet is an adequate source of academic information	the Internet is an adequate source of academic information, which will potentially fulfill the needs of university pupils	the Internet offers a more effective and practical method of studying	the Internet offers a more effective and practical method of studying
Establishing a new university in your community	building the university may lead to some social problems	These social problems may impair the quality of life in the community	yes I agree
Is it necessary for children or not?	they would be able to develop their personalities and sense of reliance	Having knowledge about other countries and their languages lead to extend the child's vision	Is it necessary for children or not?

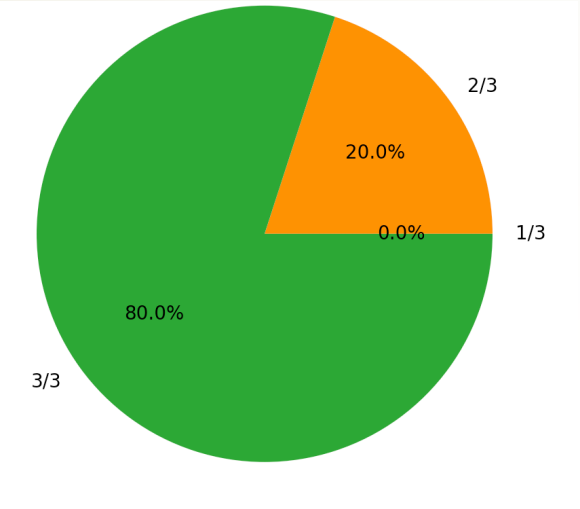
Copy-paste

Second Trial (T2)

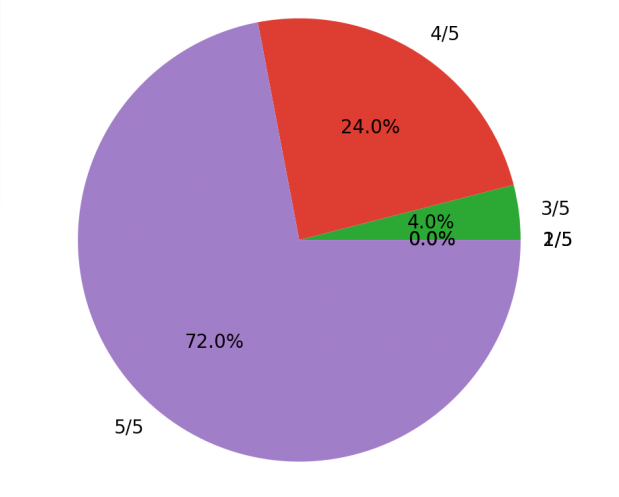
- Generation of text has difficulties in crowdsourcing [Budzianowski+, EMNLP2018]
- Experimented with settings for reducing erroneous input
  - **minimum time for 5 instances to 50 seconds (10 seconds per instance)**
    - Removes worker from task if they complete in less than 50 seconds
    - Prevents copy and paste
  - **level 3**
    - Guarantees FigureEight's most reliable annotators
    - Slower than level 1,but more reliable
  - **10円 per question**
    - Motivates the worker to try harder
- Workers
  - 25 instances, judged by 5 workers each

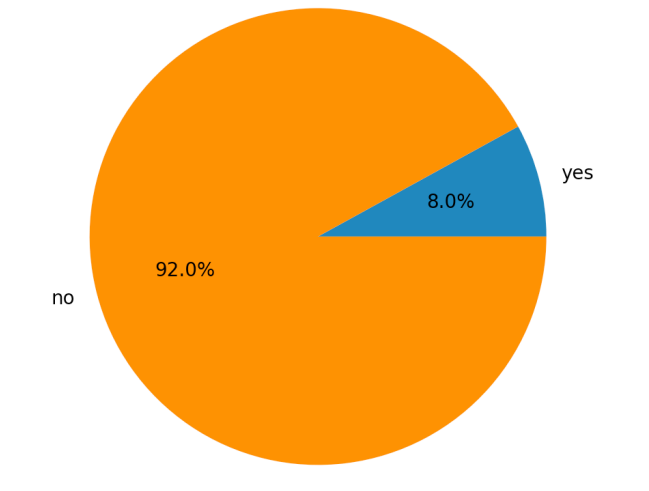
Comparison of Results

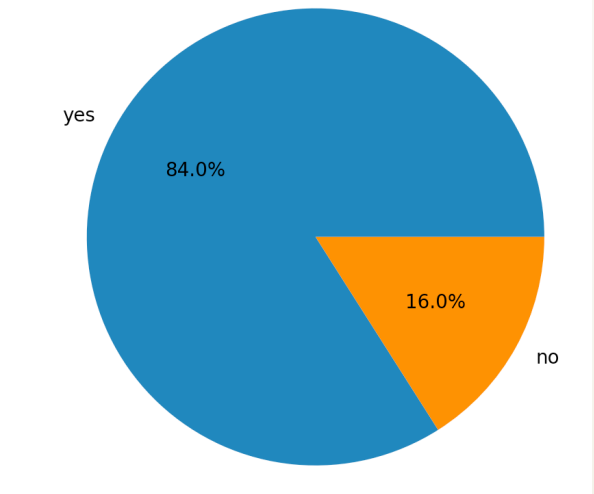
T1



T2







- 92% 'not counter-argument' to '84% yes'!
- Minimum time setting prevents copy-paste

Guidelines

Can You Write A Counter-Argument?

Instructions ▲

Overview

Greetings! We really appreciate you being a worker for this crowdsourcing task. The job is as follows. For a given **topic**, someone has stated two texts (**claim** and **premise**). In this work, we would like for you to write a **counter-argument** against the **claim, premise,** or **both** in your OWN WORDS. Please make sure the **counter-argument** is in English and is only one sentence long.

Steps

- 1. Carefully read the **topic**, **claim**, and **premise**.
- 2. Write, in your own words, a **counter-argument** to attack the **claim, premise,** or **both.** Please use the list of examples below as a hint. Please only write one sentence and use English only.
  - For this part, please do not copy and paste anything. Unfortunately, such work will be rejected.

Task Benefits

- This task will help your thinking skills and understanding of arguments improve. If you like to debate, your debating skills will significantly improve.

Important Definitions

- **claim:** controversial statement that requires additional information to be accepted
- **premise**: statement that acts as evidence to support the acceptability of the **claim**
- **counter-argument:** contradiction or way to attack/challenge the acceptability of the **claim, premise,** or **both**

Guideline Examples

Acceptable Examples

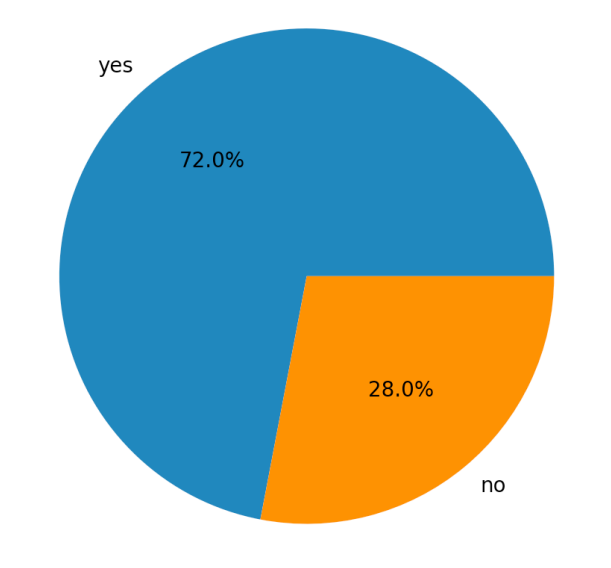
topic	claim	premise	counter-argument
Nowadays human activities are influenced by computer use	Many humans use computers everyday.	Computers help to communicate more easily.	Only some computers help humans communicate more easily.
Improve roads or public transports	Public transportation is great.	It is much safer than private transportation.	Not all public transportation is safe.
Violence in video games	Video games cause violence in young children.	When children see violence in video games, they will act it out.	Other factors influence whether children become violent or not.

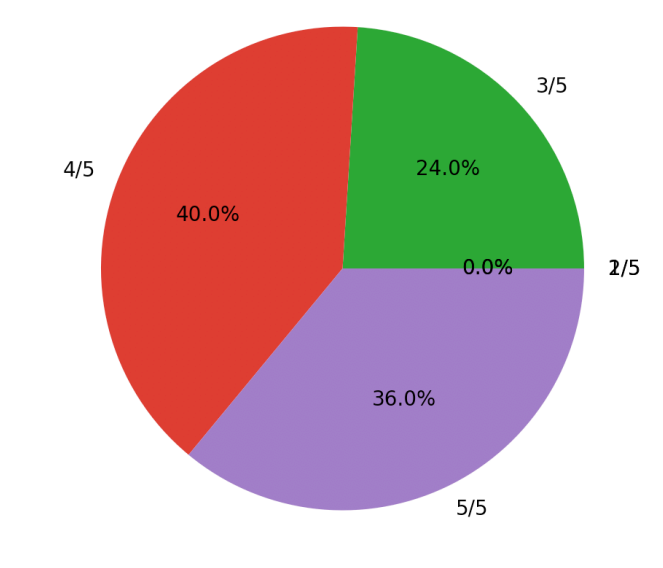
Unacceptable Examples

topic	claim	premise	counter-argument	Unacceptable Reason
Books are an adequate source of academic information	Books are an adequate source of academic information, which will potentially fulfill the needs of university pupils	books offers a more effective and practical method of studying	Books are an effective tool.	This example is not a counter-argument. It simply restates the premise.
Improve roads or public transports	Public transportation is great.	It is much safer than private transportation.	counter-argument	The word "counter-argument" only is not an acceptable answer.
Is it necessary for children or not?	they would be able to develop their personalities and sense of reliance	Having knowledge about other countries and their languages lead to extend the child's vision	la cultura es importante para los niños	This example is not in English.

Arg. Reasoning Comprehension (ARC) Task

- Sem-Eval 2018 Task [Habernal et al., NAACL2018]
  - + 2477 claim-premise-warrant pairs
  - + No context required
  - + Well-known in the Arg. Mining community
- CS Trial using ARC data (results below)
  - Can reasonably use the corpus for CA generation





Discussion Outline

- Research Overview
- Corpus Construction
- Conclusion and Future Plan

Conclusion and Future Plan

- **Conclusion**
  - Created methodology for addressing task of constructive feedback generation
  - Developed a crowdsourcing method for generating reasonable CAs
- **Future Plan**
  - **Short-term**
    - Currently conducting a mid-size corpus construction
    - Conduct crowdsourcing task for identifying type of counter-argument
  - **Long-term**
    - Extension of corpus to large-scale
    - Implementation of seq2seq model
    - Improving existing attack relation identification models using generated counter-arguments

Short-term

- Currently conducting a mid-size corpus construction
  - 500 generated counter-arguments
  - Each judged by 5 workers
- Conduct crowdsourcing task for identifying type of counter-argument

Argument A1: *Police are too willing to use force. Police are using excessive force all over the U.S. and it's not recorded.*

Not all actions of the police are violent.

Targets 'hasty generalization' fallacy

- How to typologize the remaining fallacies?

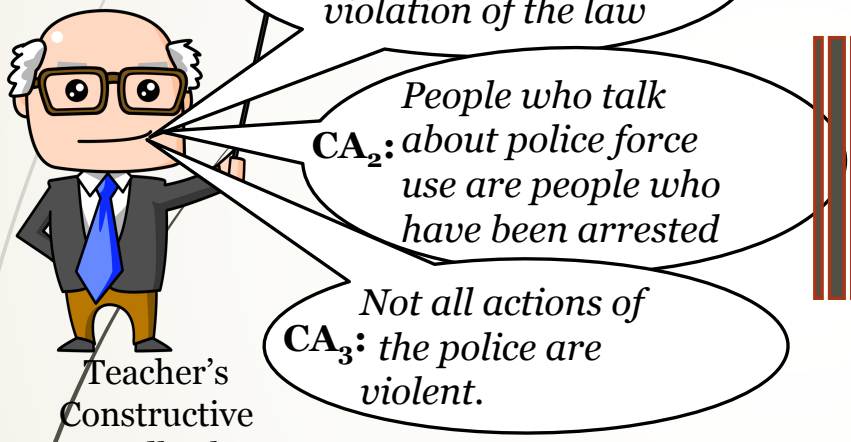
Part 2: Incorporating Background Knowledge for Warrant Identification (Keshav Singh)

Big Picture

Prompt P1: Are police too willing to use force?

Argument A1: *Police are too willing to use force. Police are using excessive force all over the U.S. and it's not recorded.*

Student A Essay (Input)



Teacher's Constructive Feedback **(Counter Argument) Part 1**

Inform

Student A

Part 3: Quality Scores
Organization
Content
etc.

Revision

Output:

Revised Argument R1: *Police are too willing to use force, but as a result, crime is reduced. Although many people think that arrested individuals discuss this issue, police are using excessive force all over the U.S. Granted, this force is not always violent.*

Part 2: Machine is required to understand implicit arguments (i.e. warrants)
A1 assumes "*force does not cause less violation of the law*"

Existing Work (Data + State of the art Model)

Topic: Is Google a Harmful Monopoly?

Additional Information: European regulators say the company's Android phone blocks rival services.

Premise (Reason): People can choose not to use Google.

And since

✔ **Warrant 0:** they can opt-out from being indexed by their search engine

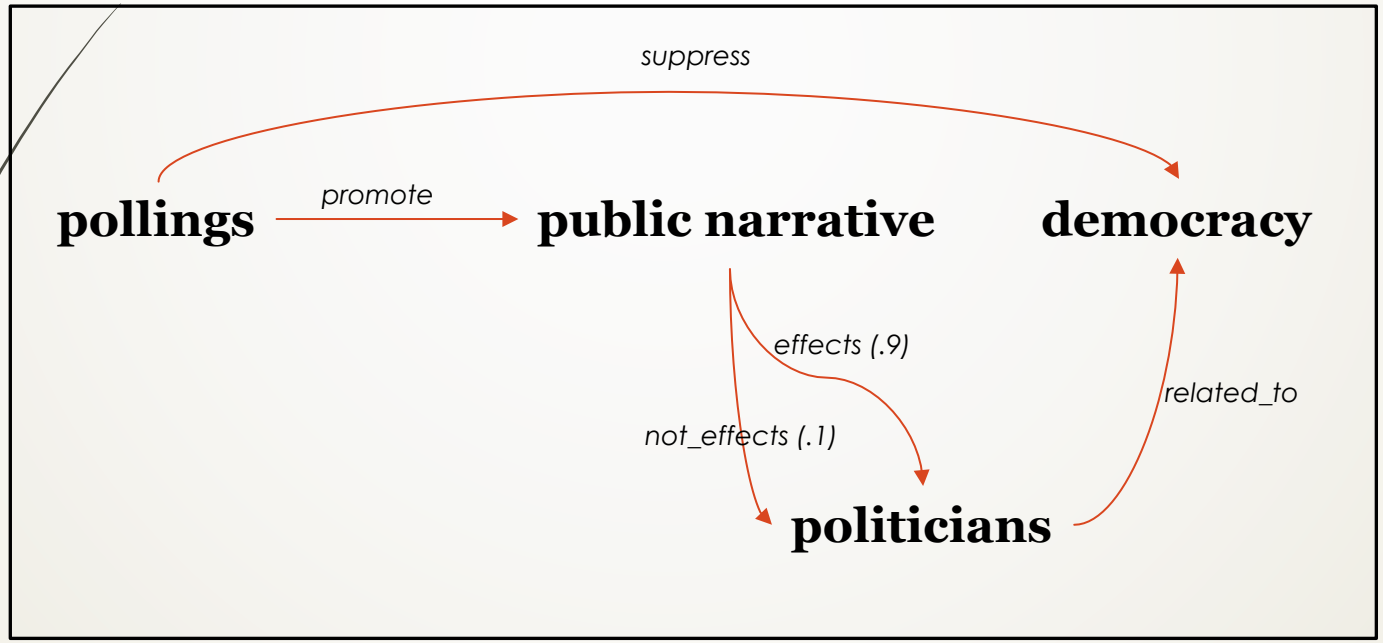
✗ **Warrant 1:** they cannot opt-out from being indexed by their search engine

Claim: Google is not a harmful monopoly

- **The Argument Reasoning Comprehension task[Habernal et al., 2018] -** Identify the correct warrant. : Given a debate title, claim and reason.
  - Dataset: 2477 claim-premise-warrant pairs
  - + Topic and additional information
- **GIST model** - Transfers inference knowledge to this task. [Choi and Lee, 2018]

Motivation

- **Claim**: *Pollings undermine democracy.*
- **Premise**: *Poll results create a public narrative rather than reality.*
- **Correct Warrant**: *Public narrative has effect on politicians.*
- **Incorrect Warrant**: *Public narrative has virtually no effect on politicians*



Plan

- Utilize existing, large-scale corpora for knowledge extraction (e.g. Wikipedia, Gigaword, etc.)
- Utilize existing relation extraction technologies for building KB
- Use the created KB to incorporate logic-based analysis of the chain of reasoning
- Devise methodology to use of this with respect to the Argument Reasoning Comprehension task

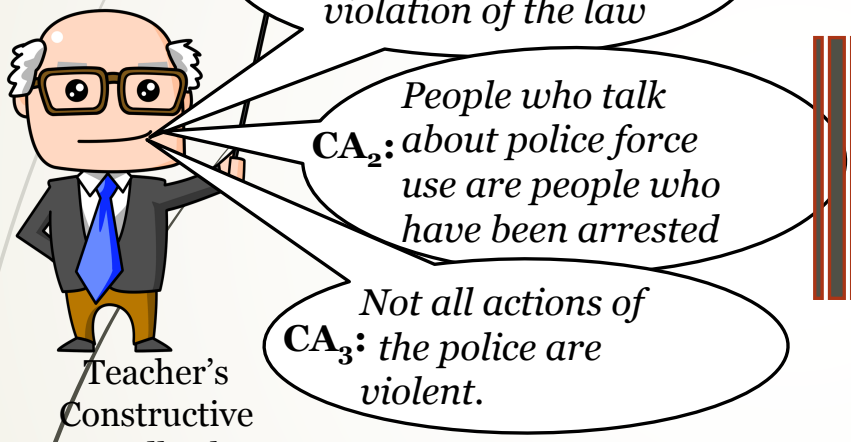
Part 3: Improving Modeling of Student Essay Organization Scoring (Farjana Sultana Mim)

Big Picture

Prompt P1: Are police too willing to use force?

Argument A1: *Police are too willing to use force. Police are using excessive force all over the U.S. and it's not recorded.*

Student A Essay (Input)



Teacher's Constructive Feedback **(Counter Argument) Part 1**

Inform

Student A

Part 3: Quality Scores
Organization
Content
etc.

Revision

Output:

Revised Argument R1: *Police are too willing to use force, but as a result, crime is reduced. Although many people think that arrested individuals discuss this issue, police are using excessive force all over the U.S. Granted, this force is not always violent.*

Part 2: Machine is required to understand implicit arguments (i.e. warrants)
A1 assumes "*force does not cause less violation of the law*"

Existing Work

- Motivation: Incorporate structured information into textual information
- Previous work does not incorporate the existing structure, e.g:

- **Heuristic rules** for sentence and paragraph labels to represent [Ng&Persing, 2010]

For example: **Introduction**, **Body**, **conclusion** etc. (paragraph label) and Rebuttal, Elaboration, Thesis etc. (sentence label)

presence of however, but , argue → Rebuttal sentence

Main Idea, Support, Conclusion sentence → Body paragraph

- **Argumentative features** (i.e. claim, premise, etc.) on top of Ng's heuristic rules [Wachsmuth et al., 2016]

3 types of ADU features:

1/ ADU flows (e.g: (claim, premise, claim))
2/ ADU n-grams
3/ ADU compositions

Ongoing Work

- ICLE corpus introduction
  - 91% of the ICLE text are argumentative
  - Average Essay length 617 (tokens)
  - Total 6086 essays.
  - 1003 essays are annotated with organization score (Score range: 0-4)
- Baseline model 1:
  - Neural AES model (Taghipur & Ng, 2016) + Persing rules (Persing et. al, 2010)
- Results (Organization):

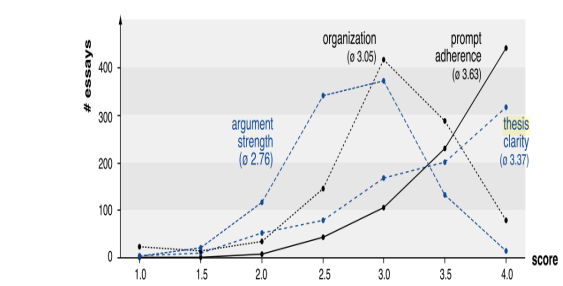
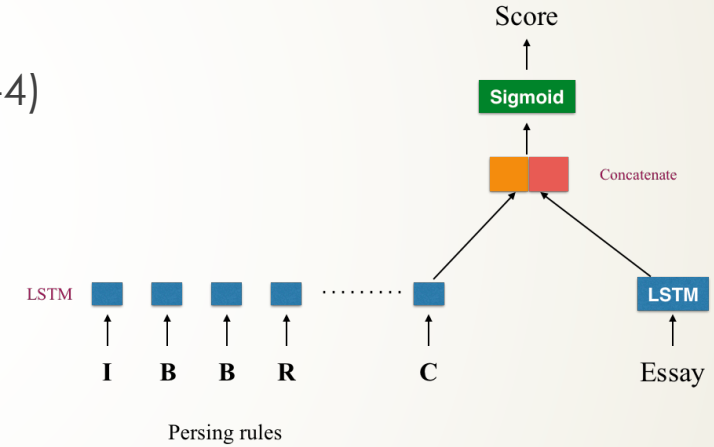


Figure 5: Distribution of essays over the possible scores from [1.0, 4.0] in the datasets of the four tasks.

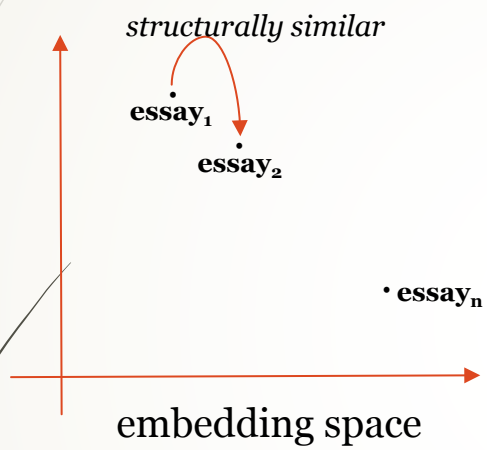


Persing rules

	Persing et. al., 2010	Wachsmuth et. al., 2016	Baseline 1
MSE	0.175	0.164	0.162
MAE	0.323	0.314	0.314

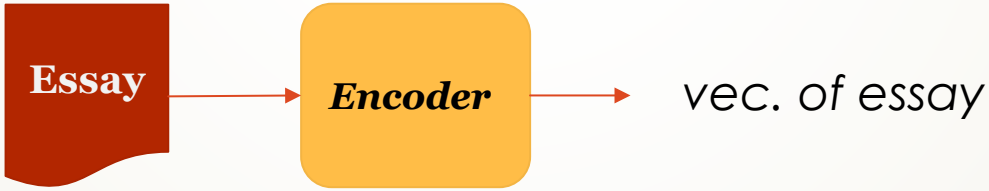
Plan

Unsupervised Learning of Discourse Structure-aware Text Representation for Essay Scoring



(1) Discourse markers
(2) Arg. Components

PDTB
S&G17



ICLE: 6000 texts