

Computational Argumentation — Part VI

Argument Assessment

Henning Wachsmuth

henningw@upb.de

Learning goals

▪ Concepts

- Various properties of argumentation to be assessed
- Theoretical notions of argumentation quality
- The subjective nature of argumentation properties



<https://commons.wikimedia.org>

▪ Methods

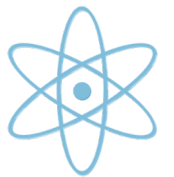
- Route kernels and more for stance and myside bias
- Feature-based and neural methods for schemes and fallacies
- Classification, regression, and graph analyses for quality



<https://pixabay.com>

▪ Associated research fields

- Argumentation theory and rhetoric
- Natural language processing



<https://pixabay.com>

▪ Within this course

- How to "understand" properties of (previously mined) arguments



Outline

- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument acquisition
- V. Argument mining
- VI. Argument assessment**
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- a) Introduction**
- b) Stance and bias
- c) Schemes and fallacies
- d) Quality in theory
- e) Absolute and relative quality assessment
- f) Objective and subjective quality assessment
- g) Conclusion

What is argument assessment?

▪ Argument(ation) assessment

- Coverage term for analysis tasks that detect, classify, rate, or otherwise judge specific properties of argumentative units, arguments, or argumentative texts

” If you wanna hear my view, I think that the EU should allow rescue boats in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats. Nothing justifies to endanger the life of innocent people.”

stance
on issue?

reasoning
scheme?

argument
quality?

framing
of issue?

author of
argument?

▪ Why argument assessment?

- Argumentative structure alone is not sufficient for many applications.
- Often, some understanding is needed of how an argument relates to an issue, how it works, and how good or important it is

What properties of argumentation to assess?

▪ What is meant by argumentation properties?

- Properties that reflect an understanding of aspects of argumentation
- Properties can be formalized as labels, scores, additional text fragments, or similar.

If you wanna hear my view, I think that the EU should allow rescue boats in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats.



4 / 5

▪ Selected properties to assess

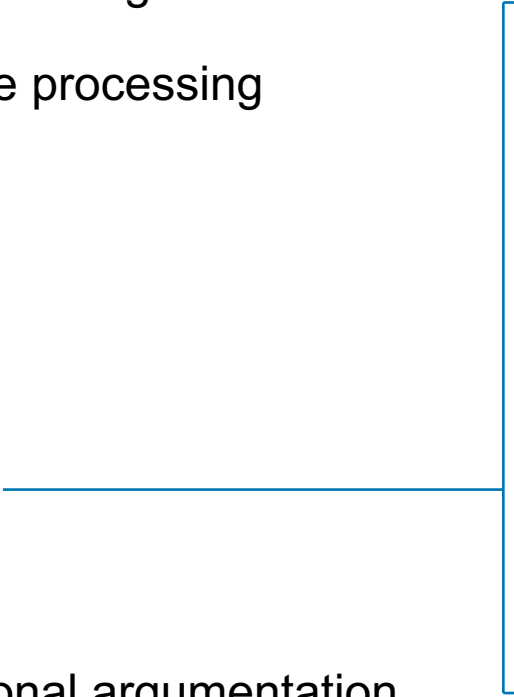
- **Subjectiveness.** Stance, myside bias, emotions, ...
- **Reasoning.** Schemes, fallacies, warrants, enthymemes, ...
- **Quality.** Logical, rhetorical, and dialectical strength, ...
- **Content.** Issues, aspects, frames, creation date, ...
- **Style.** Genre, authorship, discourse modes, rhetorical moves, ...
- **Structure.** Argumentative depth, claim centrality and divisiveness, ...

▪ Notice

- Where mining ends and assessment starts is not perfectly defined.
For example, classifying evidence types might be seen as assessment.

Next section: Stance and bias

- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument acquisition
- V. Argument mining
- VI. Argument assessment**
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- 
- a) Introduction
 - b) Stance and bias**
 - c) Schemes and fallacies
 - d) Quality in theory
 - e) Absolute and relative quality assessment
 - f) Objective and subjective quality assessment
 - g) Conclusion

Stance and myside bias

■ **Stance** (recap)

- The overall position held by a person towards some target, such as an object, statement, or issue.

Near-synonyms: Viewpoint, view, standpoint, stand, position.

- To have/take a stance on a target means to be *pro* or *con* towards it.



Con towards death penalty.

The death penalty must be abolished.

Pro towards the left claim.

It doesn't deter people from violence.

■ **Myside bias**

- Focus on information that confirms one's stance, giving disproportionately less attention to information that contradicts it

Near synonym: Confirmation bias

- An argumentative text with myside bias only supports its stance



What are stance and myside bias classification?

▪ Stance classification

- The classification of the stance of a (span of) text towards a given target.
- **Input.** An argumentative text, and a target in terms of an issue or claim
- **Output.** Whether the text is *pro* or *con*

Sometimes, also classes such as *neutral* or *not relevant* are considered.

Target: Rescue boats

” *If you wanna hear my view, I think that the EU should allow rescue boats in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats. Nothing justifies to endanger the life of innocent people.*”

**myside
bias**



<https://pixabay.com>

▪ Myside bias classification

- The classification of an argumentative text as to whether it misses opposing viewpoints or not
- **Input.** An argumentative text
- **Output.** Whether the text has *myside bias* or *no myside bias*

Not a standard task in computational argumentation, but relevant to argumentative writing support

Stance classification: Examples

- **How good are humans in stance classification?**

- What is the stance of the claims on the right to the issues on the left?

"We should ban boxing."



"Boxing remains the 8th most deadly sport."



"It is sometimes right for the government to restrict freedom of speech."



"Human rights can be limited or even pushed aside during times of national emergency."



"We should embrace multiculturalism."



"Unity is seen as an essential feature of the nation and the nation-state."



slightly modified examples of Bar-Haim et al. (2017a)

- **What makes the task challenging?**

- Stance can be expressed without mentioning the issue.
- The contrastiveness of discussed concepts needs to be accounted for.
- Positive stance can be expressed with negative sentiment, and vice versa.

But stance and sentiment polarity often correlate.

Overview of stance classification

▪ How to model stance classification computationally?

- Standard text classification trained on texts for specific issues
- Relation-like classification with the issue as one input

▪ Common features (Somasundaran and Wiebe, 2010, Hasan and Ng, 2013)

- Bag-of-words. Distribution of words or word n-grams
- Core vocab. Terms from subjectivity lexicons
- POS. Distribution of part-of-speech tags
- Discourse. Connectives and relations between units
- Sentiment. Aspect-based or topic-directed polarity
... among many others

▪ Specific stance classification approaches

- Exploit author knowledge in dialogue (Ranade et al., 2013)
- Exploit opposing views in dialogue (Hasan and Ng, 2013)
- Stance as sentiment and contrast of text and issue targets (Bar-Haim et al., 2017a)
- Route kernels for stance based on overall structure (Wachsmuth et al., 2017f)

Alice: *The EU should allow rescue boats in the Mediterranean Sea, to save the innocent refugees.*

↓ stance tend to be opposite

Bob: *So naïve... having such boats makes even more people die trying.*

↘ stance tend to be the same

Alice: *Well, I actually read that rescue boats haven't led to an increase yet.*

Stance as sentiment and contrast (Bar-Haim et al., 2017a)

▪ Task

- Given a claim relevant to a given issue, classify the claim's stance on the issue.

The issue is also supposed to have a claim-like phrasing.

Issue. *"Advertising is harmful."*

Claim. *"Marketing creates consumerism and waste."*

▪ Data

- 55 issues from iDebate, and 2394 claims from Wikipedia.
- The target of each claim and its sentiment polarity (positive or negative) were annotated manually for training.

▪ Approach in a nutshell

1. Identify the target of the issue and the claim.
2. Classify the sentiment polarity towards each target.
3. Determine whether the targets are contrastive or not.
4. Derive stance from sentiment and contrast.

Actually, Bar-Haim et al. start with the issue target and sentiment polarity given already.

$$\begin{array}{l} \text{claim target polarity} \\ \times \text{contrastiveness} \\ \times \text{issue target polarity} \\ \hline \approx \text{stance} \end{array}$$

Stance as sentiment and contrast: Approach

- **Identify targets t_c and t_i of claim and issue**
 - **Candidate targets.** Any noun phrase
 - **Features.** Position in parse tree, relation to sentiment, Wikipedia title or not, ...
 - **Supervised classifier.** Logistic regression
- **Score polarities $p(t_c)$ and $p(t_i)$ in $[-1,1]$**
 - **Lexicon-based.** Find sentiment terms and polarity shifters from lexicons
 - **Scoring.** Based on distance to targets
- **Score contrastiveness $c(t_c, t_i)$ in $[-1,1]$**
 - **Features.** Polarity shifters, relatedness measures, Wikipedia headers, ...
 - **Supervised classifier.** Random forest
- **Score stance $s = p(t_c) \cdot c(t_c, t_i) \cdot p(t_i)$**

s can be thresholded to decide when to actually classify stance.

Issue. "Advertising is harmful."

Claim. "Marketing creates consumerism and waste."

Issue. "Advertising is harmful." **-1**

Claim. "Marketing creates consumerism and waste." **-0.7**

Advertising ↔ **Marketing** **1**

$$s = -0.7 \cdot 1 \cdot -1 = 0.7$$

Stance as sentiment and contrast: Results

▪ **Evaluation** more in (Bar-Haim et al., 2017a)

- **Data.** 25 issues (1039 claims) for training, 30 issues (1355 claims) for testing
- **Baseline.** SVM with unigram and sentiment features
- **Measure.** Accuracy@coverage depending on threshold for s (here 20–100%)

Approach	20%	40%	60%	80%	100%
Baseline	0.717	0.709	0.691	0.668	0.632
Sentiment only	0.770	0.749	0.734	0.632	0.632
Sentiment + contrast	0.847	0.793	0.740	0.632	0.632

▪ **Observations**

- Reliable for confident cases, but does not beat baseline if all are classified
- The hardest cases are those where stance is expressed without sentiment.

▪ **Extended approach** (Bar-Haim et al., 2017b)

- Automatic lexicon expansion and use of sentiment in surrounding context

Bar-Haim et al. (2017b)	0.935	0.856	0.776	0.734	0.691
-------------------------	--------------	--------------	--------------	--------------	--------------

Overview of myside bias classification

- **How to model myside bias classification computationally?**
 - Conceptually, a standard text classification task
 - Argumentative structure may naturally be predictive for myside bias.



- **Approaches to myside bias classification**
 - **Supervised classification** using various features (Stab and Gurevych, 2016)
 - **Route kernels** for myside bias using overall structure (Wachsmuth et al., 2017f)

Supervised classification of myside bias (Stab and Gurevych, 2016)

▪ Task

- Given a persuasive student essay, classify it as having myside bias or not.

▪ Approach

- Polynomial SVM on six feature types:

1. **Unigrams.** Word 1-grams
2. **Dependency.** Triples from dependency tree
3. **Production.** Rules from constituency tree
4. **Opposition.** Presence of opposing words
5. **Sentiment.** Lexicon-based overall sentiment
6. **Relations.** Types of discourse relations

▪ Data

- 402 essays, 251 w/ bias, 151 w/o bias

▪ Results

- About three out of four cases correct.

Features	Accuracy
w/o Unigrams	0.733
w/o Dependency	0.765
w/o Production	0.760
w/o Opposition	0.736
w/o Sentiment	0.756
w/o Relations	0.757
All features	0.755
Best set (1+3+4)	0.770
Majority baseline	0.624

Background: Overall structure of argumentative texts

The death penalty is a legal means that as such is not practicable in Germany.

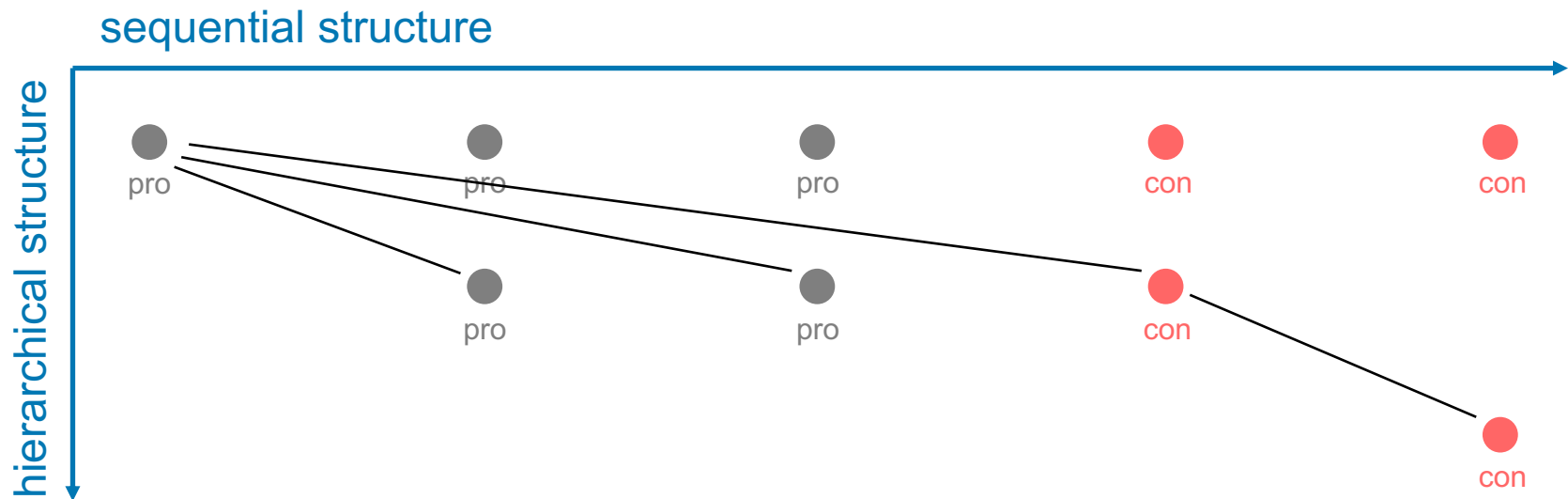
For one thing, inviolable human dignity is anchored in our constitution,

and further no one may have the right to adjudicate upon the death of another human being.

Even if many people think that a murderer has already decided on the life or death of another person,

this is precisely the crime that we should not repay with the same.

(Peldszus and Stede, 2016)



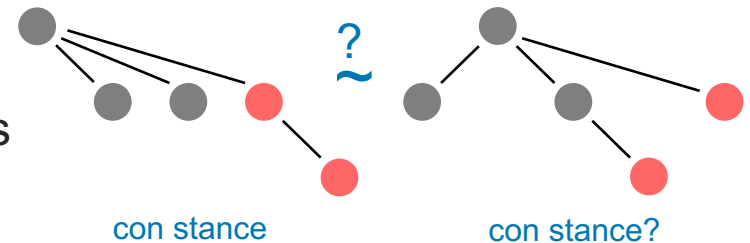
Route kernels for stance and bias (Wachsmuth et al., 2017f)

▪ Task

- Given a monological argumentative text, classify stance and myside bias without knowing the issue discussed.

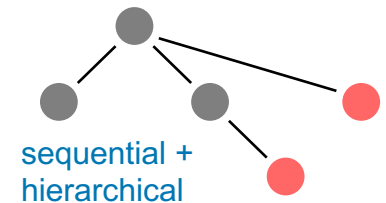
▪ Hypothesis

- The overall structure of argumentative texts is decisive for stance and myside bias.



▪ Research questions

1. How to jointly model sequential and hierarchical overall structure?
2. What model has most impact on the two tasks?



▪ Approach in a nutshell

- Start from argumentative structure of a text.
- Model overall structure with route kernels, a variation of tree kernels.
- Classify stance and myside bias based on overall structure.

Route kernels for stance and bias: Tasks and data

▪ Myside bias on AAE-v2

(Stab and Gurevych, 2016)

- 402 persuasive student essays
- Proprietary argument model
- 251 myside bias, 151 no myside bias

▪ Stance on Arg-Microtexts

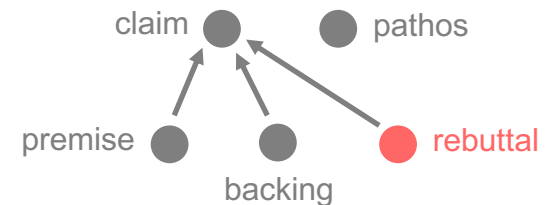
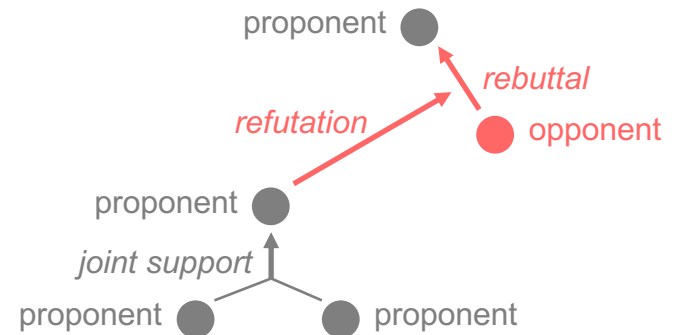
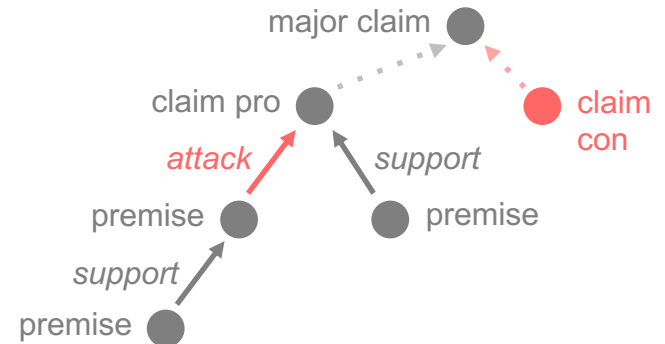
(Peldszus and Stede, 2016)

- 112 short argumentative texts
- Freeman model (Freeman, 2011)
- 46 pro stance, 42 con stance, 24 unlabeled

▪ Genre on Web Discourse (for comparison)

(Habernal and Gurevych, 2015)

- 340 argumentative web texts
- Modified Toulmin model (Toulmin, 1958)
- 216 comments, 46 blog posts, 73 forum posts, 5 articles



Route kernels for stance and bias: A unified model

Map specific models to unified model

- Order nodes according to position.
- Encode stance towards parent as node label.
- Model relations between node *pairs* only.
- The root implicitly defines main claim.

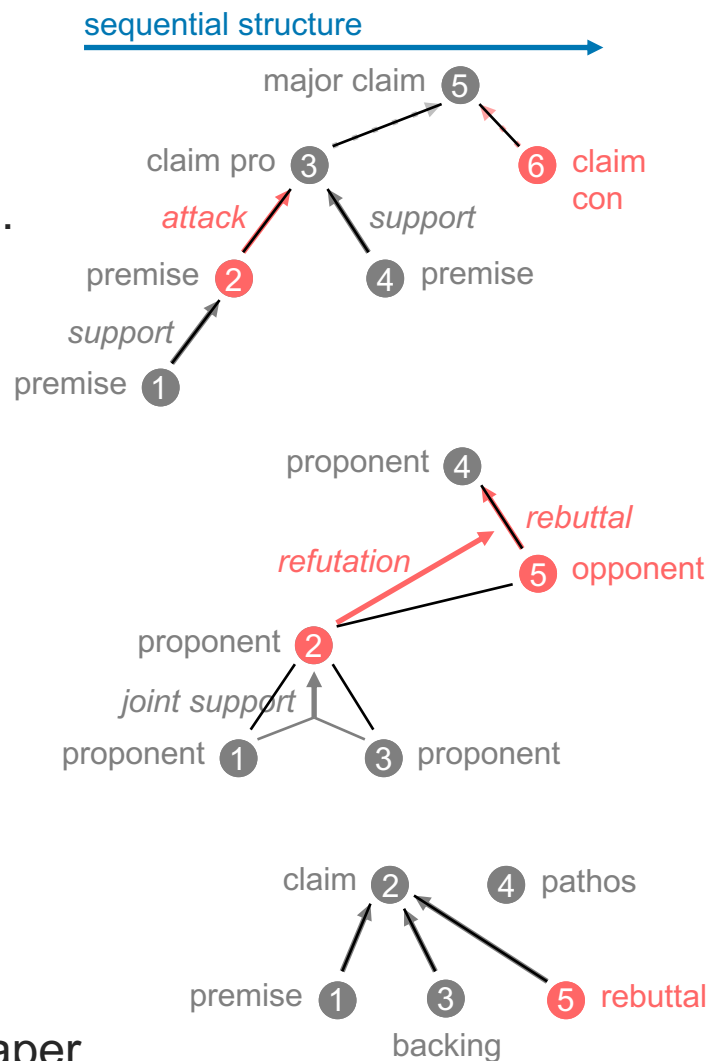
Pros and cons

- + Sequential structure captured
- + Same analyses on all corpora
- + Comparisons across corpora
- + Simpler argument mining (hypothesized)
- Partly less expressive

In this lecture, only unified model

- For experiments with specific models, see paper.

(Wachsmuth et al., 2017f)



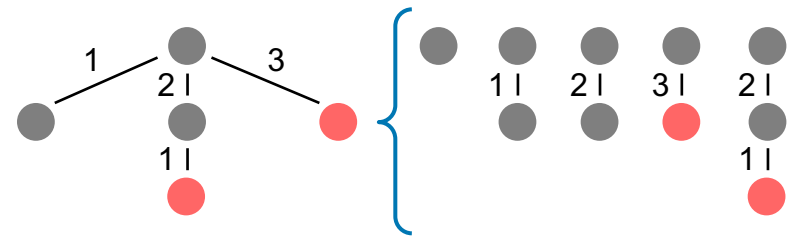
Background: Route kernels (see also lecture part V)

Kernel methods in machine learning (recap)

- Kernels represent instances in a task-specific implicit feature space.
- Kernel functions compute similarities used by classifiers, such as SVMs.
- Tree kernels capture hierarchical structures.

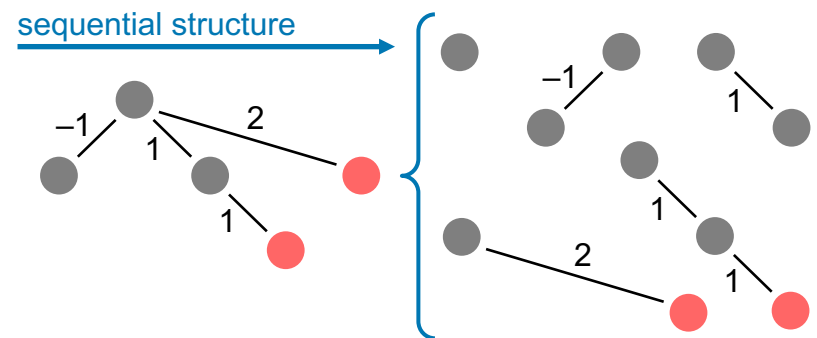
Route kernels (Aiolli et al., 2009)

- Capture both sequential and hierarchical structure
- Tree kernel with edge labels, indicating node positions relative to siblings



Adapted route kernel for arguments

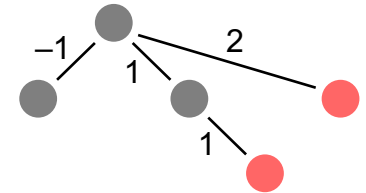
- Model all paths starting from the root of a tree
- A polynomial kernel "combines" paths to capture full overall structure.
- Positions are relative to parent node



Route kernels for stance and bias: Approach

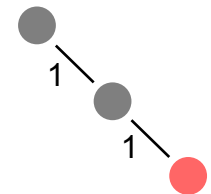
Overall structure as a positional tree

- A tree $T = (V, E)$ where nodes in V represent argumentative units and edges in E relations between two units
- Node labels.** Each node is labeled as *pro* or *con*.
- Edge labels.** Node position in a text relative to parent node



Kernel function for overall structure

- Let two trees $T = (V, E)$ and $T' = (V', E')$ be given.
- The similarity of the trees is defined as:

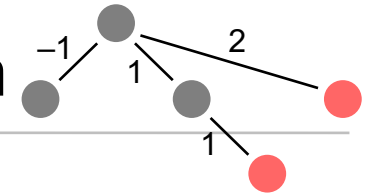


$$K_{\xi\pi}(T, T') = \left(\sum_{v \in V} \sum_{v' \in V'} \frac{\delta(\xi(v), \xi(v')) \cdot \delta(\pi(v), \pi(v'))}{(|V| \cdot |V'|)^2} \right)^d$$

1 for identical paths, 0 otherwise
 Node label path from root to v
 Edge label path from root to v
 Degree of polynomial ($d = 2$ best in experiments)

Sum over all pairs of paths of the two trees
 Normalization over maximum possible score

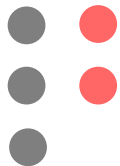
Route kernels for stance and bias: Evaluation



Overall structure approaches

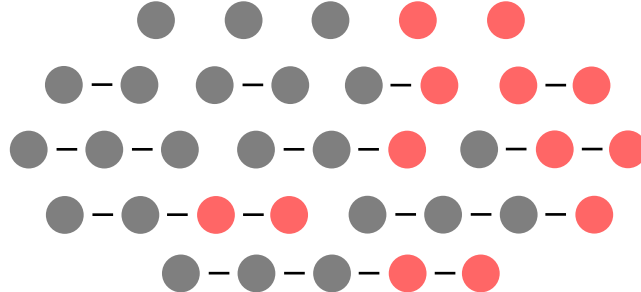
frequencies

linear kernel



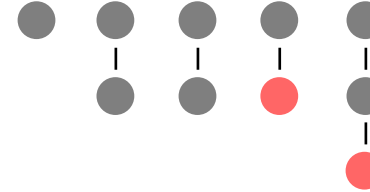
sequences

subsequence kernel



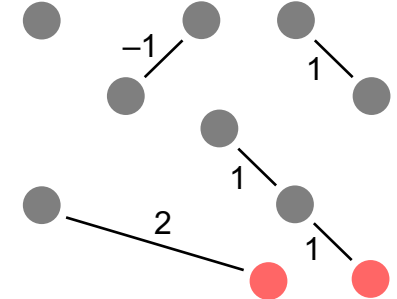
hierarchies

tree path kernel



routes

adapted route kernel



Baseline approaches

majority

always majority class

46 **pro stance**
42 ~~con stance~~

pos

linear kernel

part-of-speech
1-, 2-, and 3-grams

tokens

linear kernel

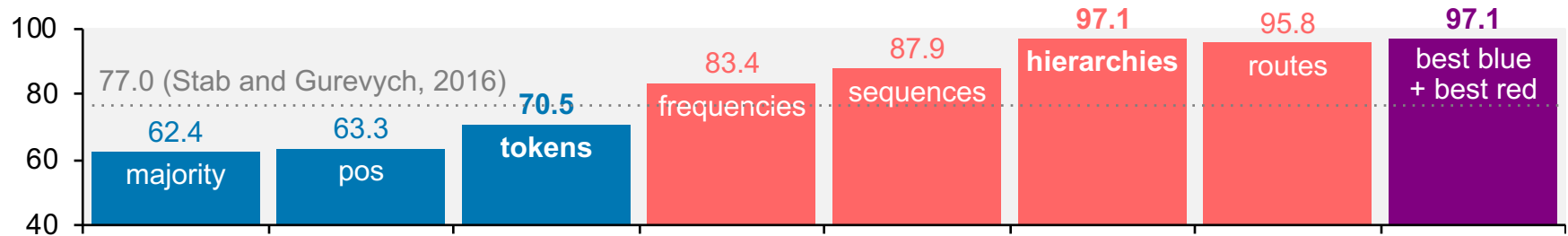
token
1-, 2-, and 3-grams

Experiments on ground-truth argument corpora

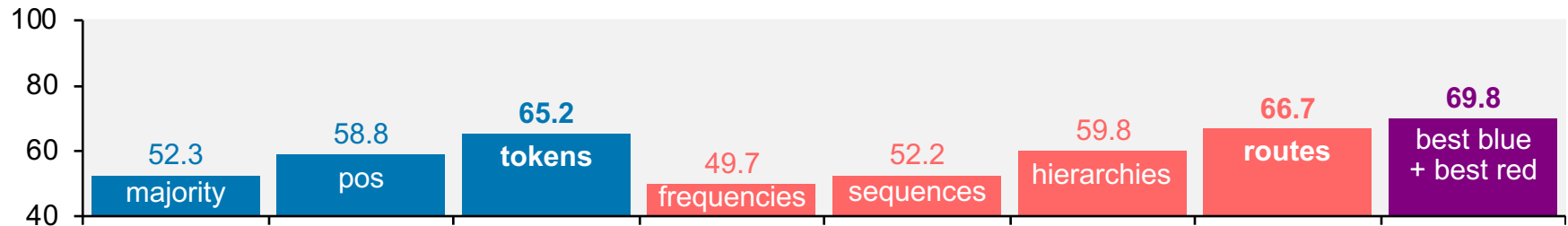
- SVM for each kernel evaluated in repeated 10-fold cross-validation
- Hyperparameters of SVM tuned on training set with balanced class weights

Route kernels for stance and bias: Results

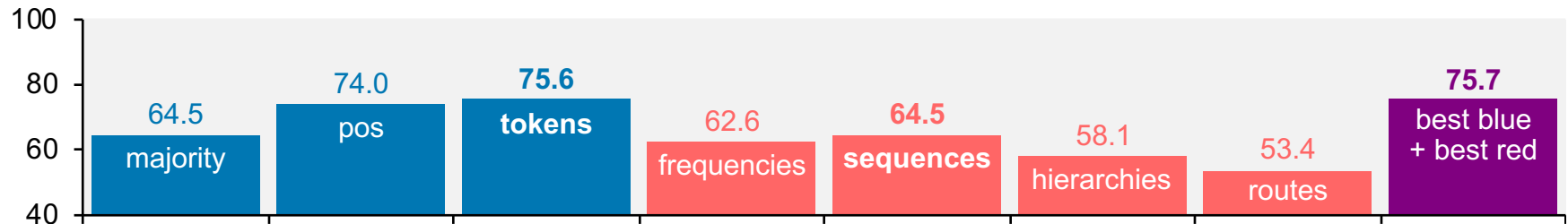
■ Myside bias accuracy on AAE-v2



■ Stance accuracy on Arg-Microtexts



■ Genre accuracy on Web Discourse



Stance and bias: Discussion

- **Effective stance and myside bias classification**
 - Approaches to stance achieve an accuracy < 0.8 in most settings.
 - Stance is subjective, so a notably higher accuracy may not be feasible.
 - Too few approaches to myside bias exist to make a conclusive statement.
- **Impact of argumentative structure**
 - At least for entire argumentative texts, modeling overall structure is important.
 - Theoretically, modeling hierarchical structure “solves” myside bias.
 - Practically, the impact depends on the effectiveness of argument mining.
- **Stance classification, an independent task**
 - Stance classification is also studied apart from computational argumentation.
 - Not in all literature on the topic, arguments are considered explicitly.
 - Still, the notion of stance implies an argumentative context.

Next section: Schemes and fallacies

- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument acquisition
- V. Argument mining
- VI. Argument assessment**
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- a) Introduction
- b) Stance and bias
- c) Schemes and fallacies**
- d) Quality in theory
- e) Absolute and relative quality assessment
- f) Objective and subjective quality assessment
- g) Conclusion

Argumentation schemes and fallacies (recap)

▪ Argumentation scheme

- The form of inference from an argument's premises to its conclusion.
- Around 60 deductive, inductive, and especially abductive schemes exist.

▪ Example schemes

- Argument from example
- Argument from consequence
- **Argument from position to know**

▪ Fallacy

- An argument with some (often hidden) flaw in its reasoning, i.e., it has a failed or deceptive scheme.

▪ Example types of fallacies

- **Ad-hominem.** Attacking the opponent instead of attacking their arguments
- **Appeal to ignorance.** Taking lack of evidence as proof for the opposite

Conclusion *A is true.*

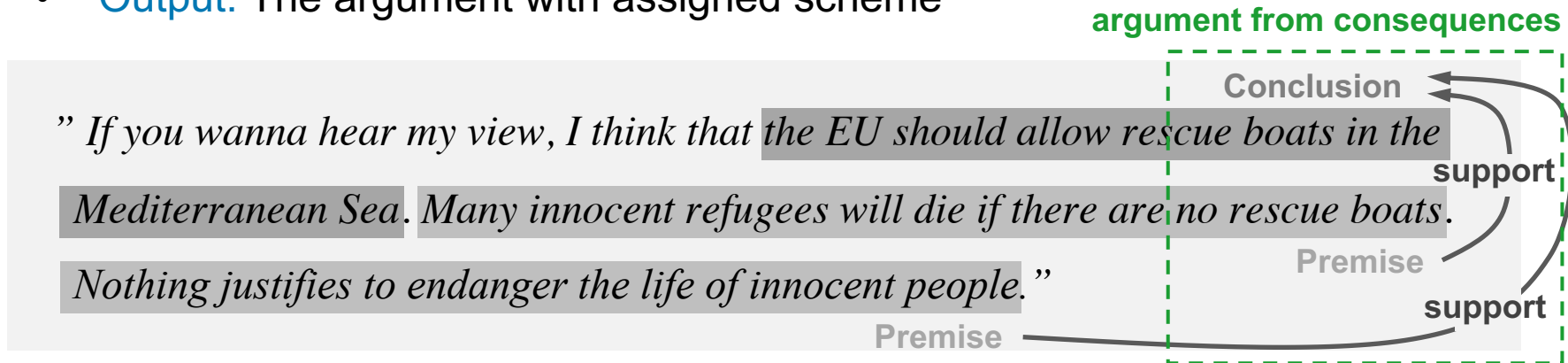
Major premise *Source E is in a position to know about things in a subject domain S with proposition A.*

Minor premise *E asserts that A is true (in domain S).*

What are scheme classification and fallacy detection?

▪ Scheme classification

- The assignment of an argumentation scheme to an argument from a given scheme set
- **Input.** An argument, usually with annotated structure
- **Output.** The argument with assigned scheme



fallacy?

▪ Fallacy detection

- The identification of arguments being a fallacy of a type from a set of types
- **Input.** An argument, possibly with annotated structure
- **Output.** Whether or not the argument is a fallacy of a certain type

Example: Correct or fallacious argumentation scheme?

■ How good are humans in analyzing schemes?

- Is the following example a correct instance of *argument from position to know*?
- Check the critical questions below.

Conclusion *A is true.*

Major premise *Source E is in a position to know about things in a subject domain S with proposition A.*

Minor premise *E asserts that A is true (in domain S).*

Conclusion *Cigarettes are not addictive.*

Major premise *James W. Johnston (the CEO of RJ Reynolds Tobacco Company) is an expert on tobacco.*

Minor premise *Johnston testified before Congress that tobacco is not an addictive substance.*

(thanks to Jonas Bülling for this example)

■ Critical questions

- Is Johnston in a position to know about cigarette addictiveness? **yes**
- Did Johnston assert that it's true that cigarettes are not addictive? **(yes)**
- Is Johnston a reliable source? **no!**

Overview of scheme and fallacy detection

▪ Schemes and fallacies in argumentation

- Describe how the reasoning in an argument works or is flawed, respectively

▪ How to model scheme classification?

- Conceptually, a text classification task
- The few existing approaches realize it as a one-vs.-all or one-vs.-one task.

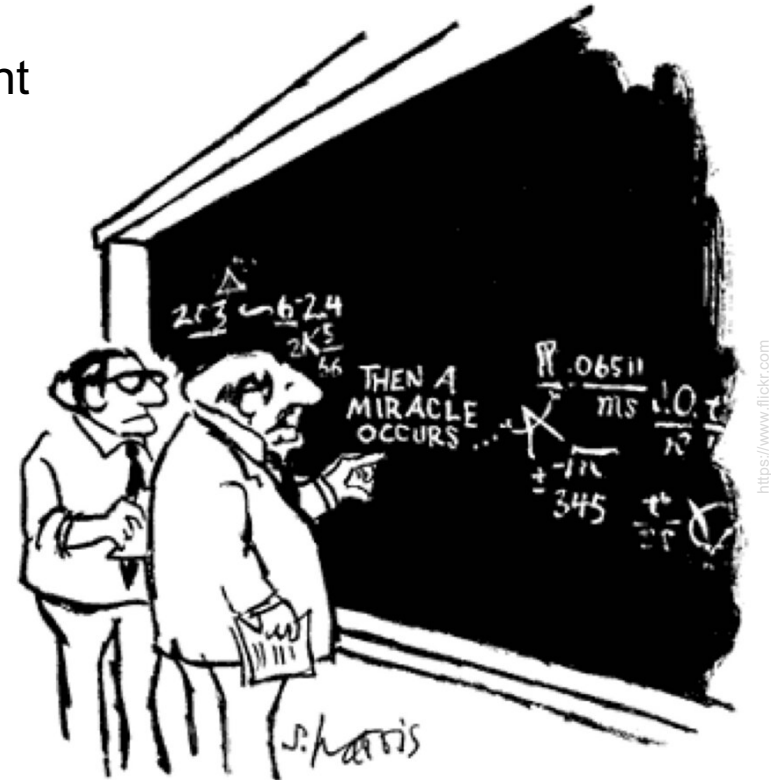
▪ How to model fallacy detection?

- Conceptually, the same
- The few existing approaches consider only specific types of fallacies.

▪ Selected approaches

- [Scheme classification](#) with tailored features (Feng and Hirst, 2011; Lawrence and Reed, 2016)
- [Ad-hominem argument detection](#) on the web (Habernal et al., 2018)

(thanks to Natalie Lüke for this illustration)



"I think you should be more explicit here in step two."

Classifying schemes with tailored features (Feng and Hirst, 2011)

▪ Task

- Given the premises and conclusion of an argument, assign one scheme from a set of given schemes.

▪ Research question

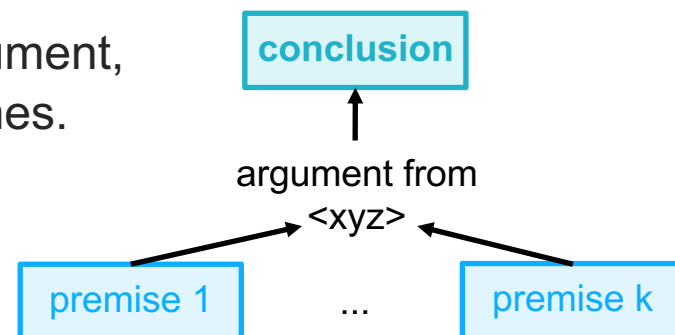
- How visible is the scheme of an argument in its text and its structure?

▪ Data

- The Araucaria corpus with 658 mixed argumentative texts, annotated for Walton's argumentation schemes (Walton et al., 2008)
- Only the five most frequent schemes considered (see next slide)

▪ Approach in a nutshell

- Compute features tailored to argumentation schemes.
- Classify schemes with standard supervised learning.



Classifying schemes with tailored features: Scheme set

▪ Argument from verbal classification

Minor pr. *a has property F.*

Major pr. *For all x, if x has property F, then x can be classified as having property G.*

Conclusion *a has a property G.*

▪ Argument from example

Minor pr. *In this particular case, the individual a has property F and also property G.*

Conclusion *If x has property F, then it also has property G.*

▪ Argument from cause to effect

Minor pr. *In this case, A occurs.*

Major pr. *Generally, if A occurs then B will occur.*

Conclusion *B will occur.*

▪ Practical reasoning

Minor pr. *I have a goal G.*

Major pr. *Carrying out this action A is a means to realize G.*

Conclusion *I ought to carry out A.*

▪ Argument from consequences

Major pr. *If A is done, good (bad) consequences will occur.*

Conclusion *A should (not) be done.*

Classifying schemes with tailored features: Examples

“Censorship is [...] the hallmark of an authoritarian regime. For example, one of Nazi Germany’s first acts was to burn all the books [...] which offended their sensibilities, beliefs, and values.”

from
example

“[You shouldn’t build a] road into the heart of the Amazon. [This] will likely result in commercialization and destruction of the valuable Amazon habitat.”

from
conseq.

“If we want to stop the counterfeit products, we have to make new products more unique.”

practical
reasoning

“[The] Iraq war [is] illegal. There is no law [...] that sanctions attacks on guys because you have good reason to believe they are bad, and could threaten you.”

verbal
classific.

“The crisis [of a party] is likely to have an effect on other opposition parties. The public’s disappointment with the [party] will lead to an erosion of confidence in the opposition.”

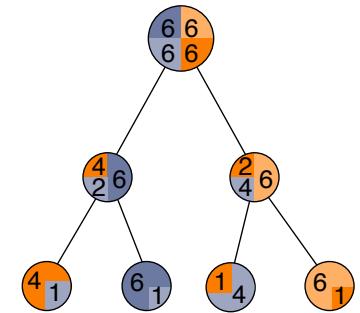
cause
to effect

(thanks to Jonas Bülling for these examples)

Classifying schemes with tailored features: Approach

▪ Approach

- C4.5 decision tree for supervised classification
- Feature engineering for all five argumentation schemes



▪ Features tailored to all schemes

- **Location.** Relative positions and distances of premises and conclusion
- **Statistics.** Premise/conclusion length ratio, number of premises
- **Structure.** Linked or convergent (given in ground truth!)

▪ Features tailored to specific schemes

- **Cue phrases**, e.g., "for example", "result", "want"
- **Indicating patterns**, e.g., causal WordNet relations
- **Sentiment.** Positive and negative words
- **Word similarity** between central words in premise and conclusion

from
example

cause
to effect

practical
reasoning

from
conseq.

verbal
classific.

Classifying schemes with tailored features: Results

- **10-fold cross-validation**

- **One-against-all.** 50% target scheme, 50% all others (once for all schemes)
- **One-against-one.** 50% scheme A, 50% scheme B (once for all scheme pairs)

- **Results (accuracy)**

Features	Acc.	Example	Practical reas.	Cause to effect	Consequ.
Verbal classific.	0.632	0.860	0.983	0.856	0.642
From consequ.	0.629	0.869	0.979	0.867	
Cause to effect	0.704	0.806	0.942		
Practical reas.	0.908	0.931			
From example	0.906				

- **Observations**

- High effectiveness for some schemes, but two schemes were confused often.
Both less training data and less clear linguistic indicators may be reasons.
- Ultimately, focusing on five schemes limits the applicability of the approach.

Ad-hominem arguments on the web (Habernal et al., 2018)

That's an ad hominem fallacy
Calvin!!

"YOU'RE FACE IS AN
AD HOMINEM!!!"



<https://yrepacademy.com>



Ad-hominem arguments on the web: Task and data

▪ What is an ad-hominem argument?

- An argument that attacks the author of an argument, not the argument itself
- According to a study, 20% of all news comments are *uncivil*. (Coe et al., 2014)

▪ Research questions

- How well can ad-hominem be identified automatically?
- What triggers ad-hominem in discussions?

▪ Data

- 2M posts from Reddit ChangeMyView
- 3866 posts (0.2%) contain ad-hominem arguments

Ad-hominem is deleted by moderators, but was made available to Habernal et al. (2018).



▪ Reddit ChangeMyView (CMV)

- An opinion poster (OP) states a view.
- Others argue for the opposite.
- OP gives Δ to convincing posts.

Deltas(s) from OP **CMV: Trump has done nothing of substance since being elected to office.**

This is kind of a counter to the other post made recently about Trump being a great president.

He pointed out things like the economy, which was growing

Ad-hominem arguments on the web: Identification

Examples

"Possible lie any harder?"

"You're making the claims, it's your job to prove it. Don't you know how debating works?"

"You're too dishonest to actually quote the verse because you know it's bullshit"

"little buddy"

"Your just an asshole"

"How can you explain that? You can't because it will hurt your feelings to face reality"

"Thank you so much for all your pretentious explanations"

"Wow. Someone sounds like a bit of an anti-semite"

"You have no capability to understand why"

"boy"

"Did you even read this?"

"Read what I posted before acting like a pompous ass"

"Do you even know what you're saying?"

"You're obviously just Nobody with enough brains to operate a computer could possibly believe something this stupid"

"Can you also use Google?"

"Reading comprehension is your friend"

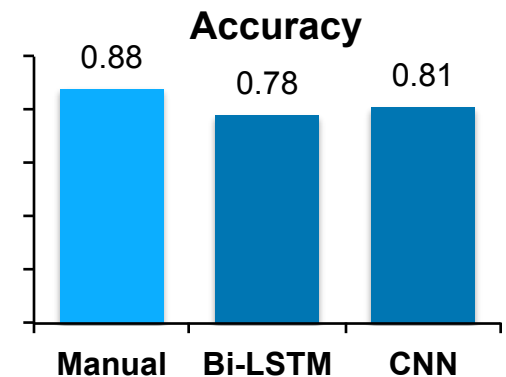
"You're using troll tactics"

"Again, how old are you?"

"You are just a liar."

Identification of ad-hominem

- **Manual.** 100 balanced arguments (50 ad-hominem) were classified by 6 workers.
- **Computational.** 7242 balanced arguments were classified by two neural classifiers (Bi-LSTM, CNN).

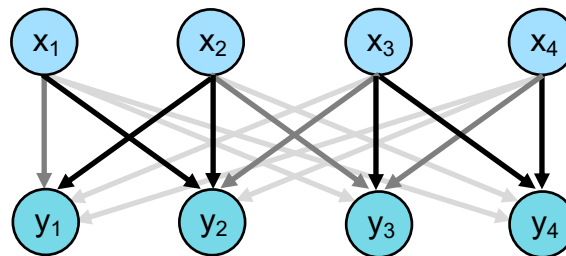


Background: Attention in neural networks (see also lecture part V)

▪ Attention

- A mechanism of RNNs that quantifies interdependencies between different parts of input and output.
- The key idea is to retain all hidden states of an input while creating the output.
- This allows learning to focus on input parts relevant to the output.

Transformer-based language models entirely rely on attention (see lecture part VII).



Edge width indicates importance

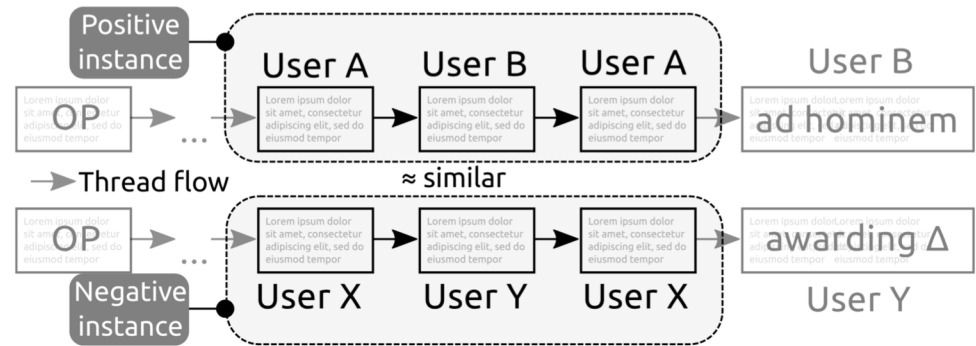
▪ Self-attention

- Quantification of interdependencies within the input only
In NLP, usually this means between the words of a sentence.
- An RNN with self-attention can provide weight values that represent the relevance it gives to different parts of an input.

Ad-hominem arguments on the web: Triggers

■ Prediction of ad-hominem

- Self-attentive LSTM trained on 2852 argument 3-tuples
- Accuracy: 0.72
- Manual attention analysis:



(OOV_comment_begin) If only you would n't rely on [fallacious] (http : OOV) [arguments] (http : OOV) to make your point. So no , I do n't realize how stupid and naive I am. All I 've realized is that you are n't actually prepared to have an actual discussion .

(OOV_comment_begin) What god do you believe in ? And it 's not a fallacy when it 's very comparable to the most popular gods .

(OOV means out-of-vocabulary)

■ Terms with much attention

- Mostly topic-independent rhetorical devices
- A few loaded keywords (e.g., "rape" or "racist")
- Partly argumentation-specific

vulgar intensifiers
"... the fuck..."

direct imperatives
"You should..."

bad argumentation
"You're grasping at straws"

missing evidence
"unsupported claims!"

...

Discussion: Scheme and fallacy detection

▪ **Effective scheme and fallacy classification**

- Some schemes are reflected in words, others require deeper understanding.
- Many schemes have never been approached so far.
- Finding ad-hominem seems doable, but this may not hold for other fallacies.

▪ **Few computational approaches**

- While extensively studied in theory, computational research on schemes and fallacies is rare so far.
- For schemes, one reason lies in the complexity of getting ground-truth data.
The high number of less frequent schemes is a particular problem in this regard.
- For fallacies, their detection is often just hard, even for humans.

▪ **Why studying schemes and fallacies?**

- Knowing the scheme means to understand how an argument reasons.
- Schemes clarify what is left implicit, allowing to find *enthymemes*.
- A way to judge quality: a good argument is usually not fallacious. (Hamblin, 1970)

Next section: Quality in theory

- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument acquisition
- V. Argument mining
- VI. Argument assessment**
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- a) Introduction
- b) Stance and bias
- c) Schemes and fallacies
- d) Quality in theory**
- e) Absolute and relative quality assessment
- f) Objective and subjective quality assessment
- g) Conclusion

Argumentation quality

▪ Argumentation quality

- Natural language argumentation is rarely logically *correct* or *complete*.
- Quality reflects how *good* a unit, an argument, or argumentation is.

premises
acceptable?

linguistically
clear?

relevant to
discussion?

” If you wanna hear my view, I think that the EU should allow rescue boats in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats. Nothing justifies to endanger the life of innocent people.”

argument
cogent?

effective in
persuading?

reasonably
argued?

▪ Observations

- **Goal orientation.** What is important depends on the goal of argumentation.
- **Granularity.** Quality may be addressed at different levels of text granularity.
- **Dimensions.** Several dimensions of quality may be considered.

Argumentation quality: Theory and in practice

▪ Quality in theory

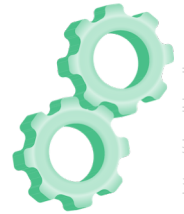
- The normative view of quality in terms of cogency, reasonableness, or similar.
- Suggests to use *absolute* quality ratings.



<https://commons.wikimedia.org>

▪ Quality in practice

- Quality is decided by the effectiveness on (some group of) people.
- *Relative* comparisons are often more suitable.



<https://de.wikipedia.org>

” Is a strong argument an effective argument which gains the adherence of the audience, or is it a valid argument, which ought to gain it? “

(Perelman and Olbrechts-Tyteca, 1969)

▪ Unresolved questions

- Should quality be aligned with how we *should* or how with we *do* argue?
- Is this actually so different? (more on this below)

Survey of existing research

argumentation
theory

assessment
approaches

Toulmin (1958)

Walton et al. (2008)

Cabrio and Villata (2012)

Braunstein et al. (2016)

van Eemeren and Grootendorst (2004)

Tindale (2007)

Hamblin (1970)

Walton (2006)

Boltužić and Šnajder (2015)

Logic

Damer (2009)

Dialectic

Cohen (2011)

Rahimi et al. (2014)

Johnson and Blair (2006)

Wachsmuth et al. (2017a)

Stab and Gurevych (2017)

**Argumentation
quality**

Mercier and Sperber (2011)

Govier (2010)

Blair (2012)

van Eemeren (2015)

Freeman (2011)

Persing and Ng (2015)

Rahimi et al. (2015)

Persing and Ng (2013)

Perelman and Olbrecht-Tyteca (1969)

Persing et al. (2010)

Feng et al. (2014)

Hoeken (2001)

Rhetoric

Tan et al. (2016)

Wei et al. (2016)

Persing and Ng (2014)

O'Keefe and Jackson (1995)

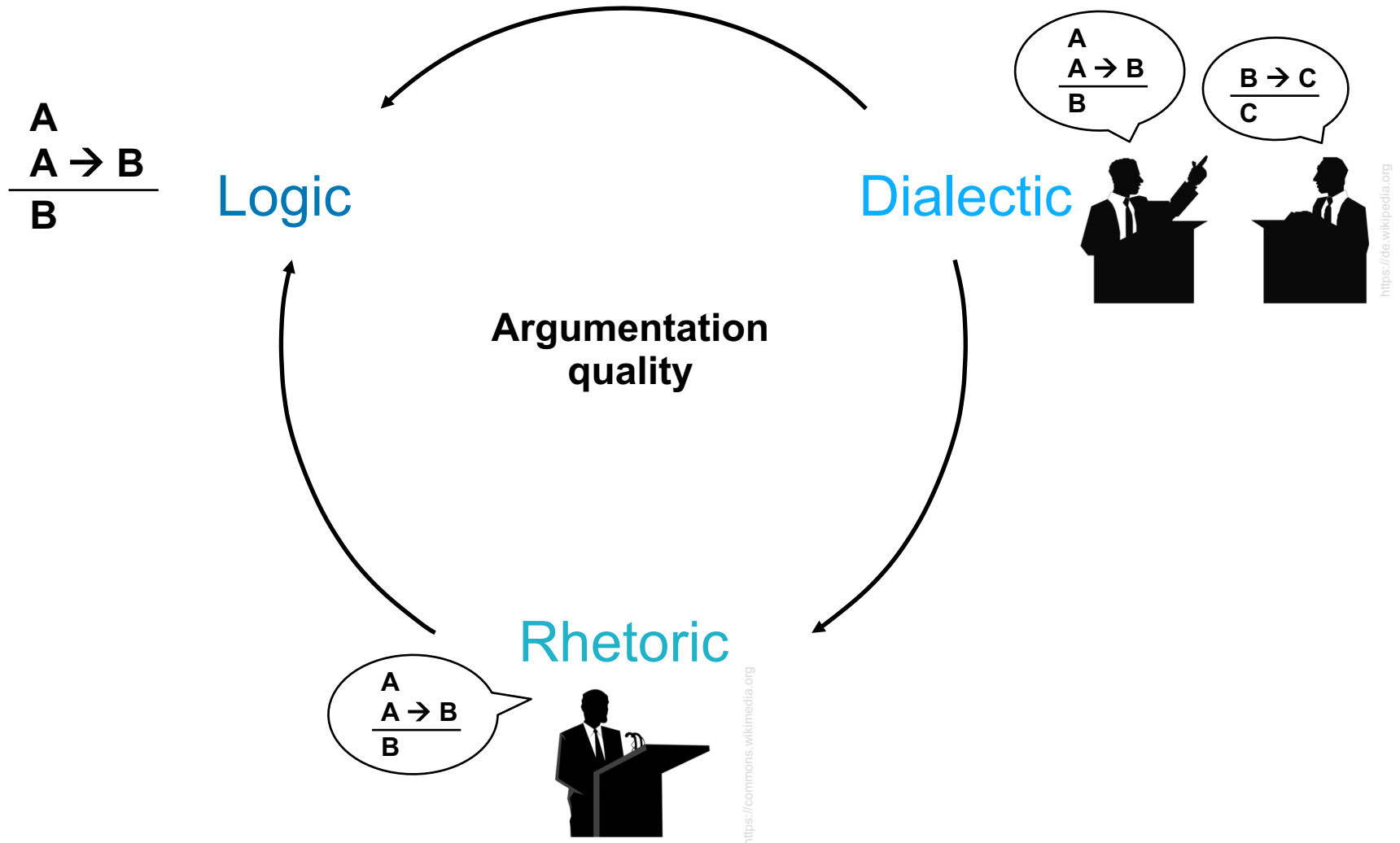
Zhang et al. (2016)

Park et al. (2015)

Aristotle (2007)

Habernal and Gurevych (2016)

Three main quality aspects (recap)



Unification of views

focus on theory

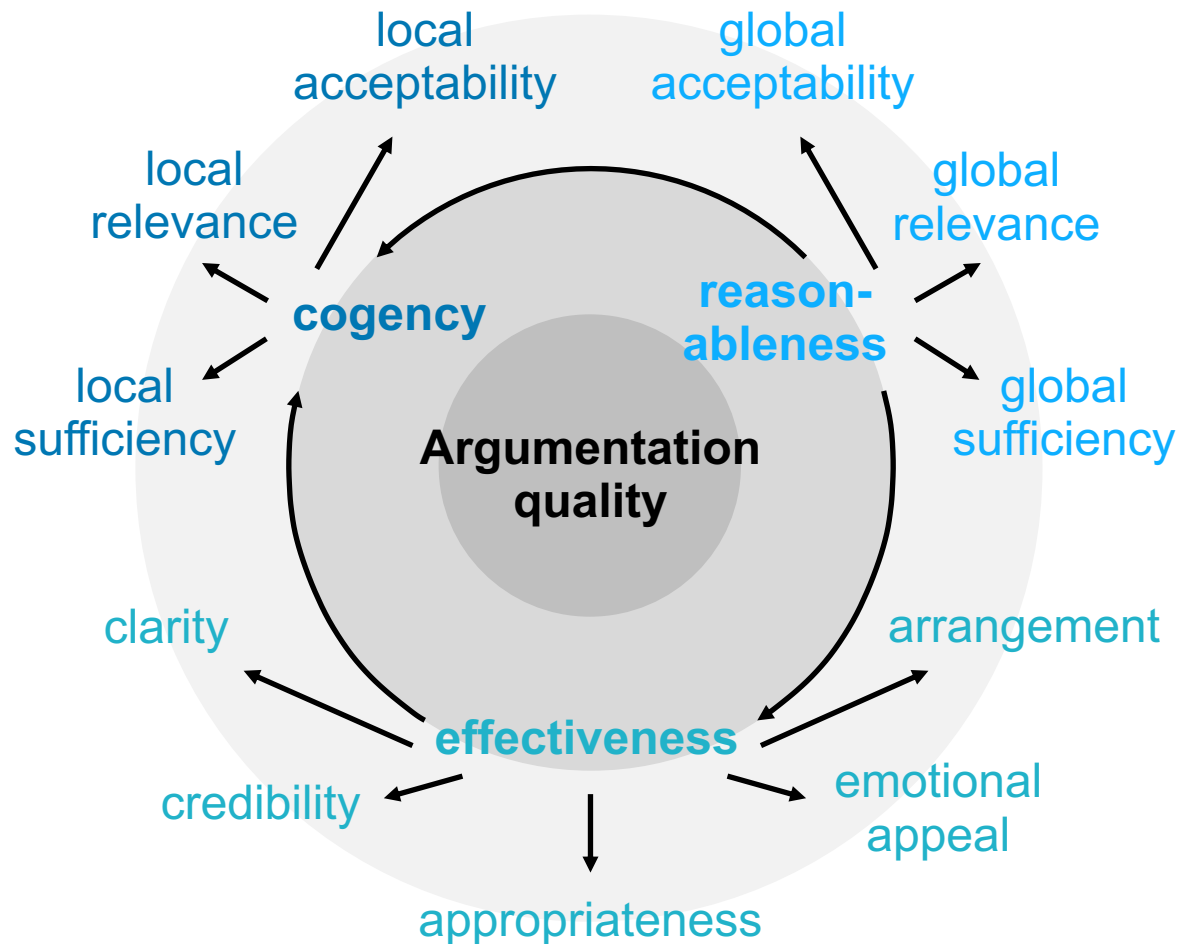
focus on accepted

prefer general

unify names



A taxonomy of argumentation quality



Quality dimensions in the taxonomy

- **A cogent argument.** Acceptable, relevant, and sufficient premises
 - **Local acceptability.** The premises are worthy being believed as true.
 - **Local relevance.** The premises are relevant to the conclusion.
 - **Local sufficiency.** The premises are sufficient to draw the conclusion.

- **Effective argumentation.** Persuades the target audience
 - **Credibility.** Make the author worthy of credence
 - **Emotional appeal.** Makes the audience open to be persuaded
 - **Clarity.** Linguistically clear and as simple as possible
 - **Appropriateness.** Linguistically matches the audience and issue
 - **Arrangement.** Presents content in the right order

- **Reasonable argumentation.** Acceptable, relevant, and sufficient
 - **Global acceptability.** Worthy being considered in the way stated
 - **Global relevance.** Contributes to resolution of issue
 - **Global sufficiency.** Adequately rebuts potential counterarguments

Logic

Rhetoric

Dialectic

Notice: cogency also adds to effectiveness, and cogency and effectiveness also add to reasonableness.

Next section: Absolute and relative quality assessment

- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument acquisition
- V. Argument mining
- VI. Argument assessment**
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- a) Introduction
- b) Stance and bias
- c) Schemes and fallacies
- d) Quality in theory
- e) Absolute and relative quality assessment**
- f) Objective and subjective quality assessment
- g) Conclusion

What is argumentation quality assessment?

▪ Argumentation quality assessment

- Identification of indisputable flaws or requirements of argumentation
- Judgment about a specific quality dimension
- Determination whether argumentation successfully achieves its goal

linguistically
clear?

effective in
persuading?

▪ Observations

- **Choice of comparison.** Dimensions can be assessed *absolutely* or *relatively*.
- **Subjectivity.** Perceived quality depends on the view of the reader/audience.
(and maybe also on the author/speaker)

▪ How to approach quality assessment?

- **Input.** Argumentative text, metadata (e.g., author), external knowledge, ...
- **Techniques.** Supervised classification/regression, graph-based analyses, ...

Importance of quality assessment

▪ Why assessing argumentation quality?

- Mining arguments and understanding the reasoning is not enough in practice.
- For successful argumentation, we need to choose the "best" arguments.
- Critical for any application of computational argumentation

"In some sense, the question about the quality of an argument is the 'ultimate' one for argumentation mining."

(Stede and Schneider, 2018)

▪ Example applications

- **Argument search.** What argument to rank highest?
- **Writing support.** How good is an argumentative text, what flaws does it have?
- **Automatic decision making.** Which arguments outweigh others?



Absolute vs. relative assessment

▪ How to assess a quality dimension computationally?

- **Absolute rating.** Assignment of a score from a predefined scale
Typical scales: Integers (possibly with half-points): 1–3, 1–4, 1–5, 1–10, -2–2, ... Real valued: [0,1], [-1,1]
- **Relative comparison.** Given two instances, which of them is better.

” If you wanna hear my view, I think that the EU should allow rescue boats in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats. Nothing justifies to endanger the life of innocent people.”

4/5

better
than

▪ Observations

- Both allow for ranking assessed instances.
- Absolute ratings entail relative comparisons and they imply a maximum and minimum.

”It’s the main job of the EU to save people’s lives, no matter whether they belong here.“

▪ Absolute vs. relative assessment

- A relative assessment is often much easier.
- Still, absolute ratings are widely spread and often work well.

Absolute quality rating: Overview

▪ Problem

- Can we predict *whether* an argument is good (cogent, effective, ...)?
- Can we rate *how* good it is?

▪ Main idea

- See quality assessment as a standard classification or regression task.
- Learn what feature or metadata speaks for quality.

Conclusion
Premises

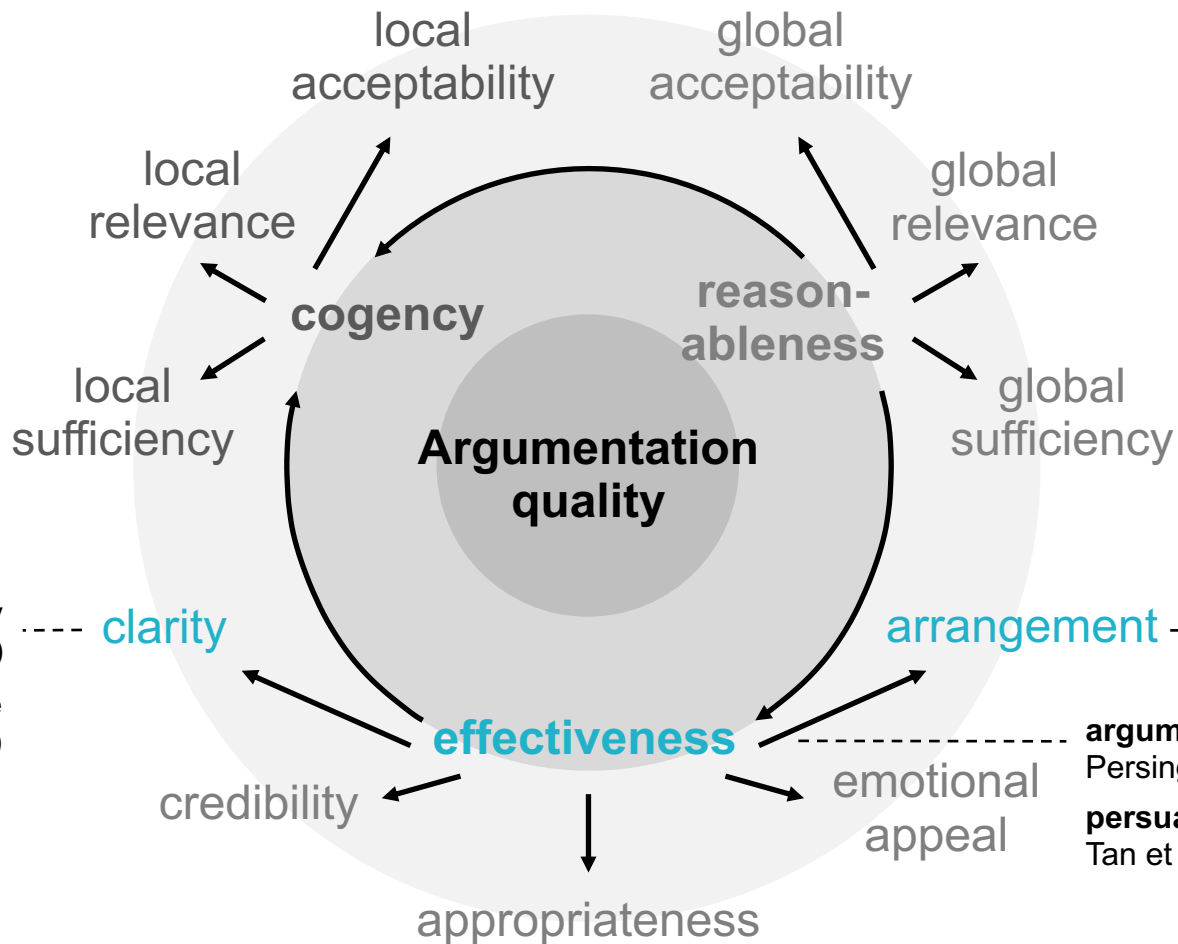
4/5

▪ Selected approaches

- **Level of support.** Count of evidence supporting conclusion (Rahimi et al., 2014)
- **Persuasiveness.** Prediction based on interaction of participants (Tan et al., 2016)
- **Organization and strength.** Assessment based on structure (Wachsmuth et al., 2016)
- **Sufficiency.** Classification with convolutional neural network (Stab and Gurevych, 2017)
- **Sufficiency.** Classification based on generated conclusion (Gurcke et al., 2021)

The last one will be discussed in lecture part VII.

Absolute quality rating: Dimensions covered here



Rating quality based on interaction (Tan et al., 2016)

▪ Task

- In a discussion, what will persuade someone who is open to be persuaded?

▪ Approach

- Analyze correlations of linguistic and interaction features with persuasion.
- Predict based on features as to whether persuasion will happen.

▪ Data

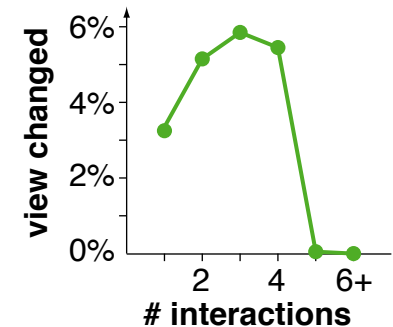
- 20k+ discussions from Reddit ChangeMyView
- **Discussion.** An opinion poster (OP) states a view, others argue against, OP gives Δ to convincing arguments



<https://de.wikipedia.org>

▪ Selected results

- **Accuracy.** 69% in balanced setting
- **Insights.** Some interactions and many participants help; style not too similar to OP most persuasive



Rating quality based on mining (Wachsmuth et al., 2016)

▪ Task

- Given a persuasive essay, score argumentation-related quality dimensions.

▪ Dimensions (Persing et al., 2010; Persing and Ng, 2013–2015)

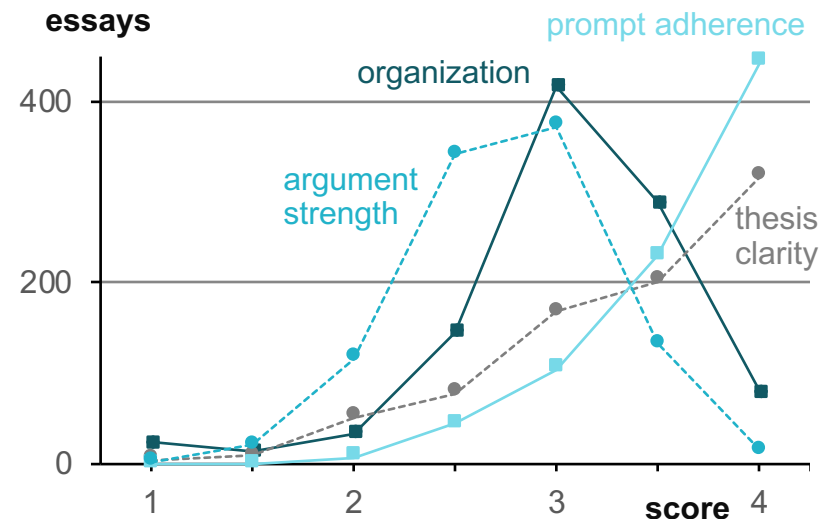
- **Organization.** How well is the argumentation arranged?
- **Thesis clarity.** How easy to understand is the thesis?
- **Prompt adherence.** How close does the essay stay to the issue?
- **Argument strength.** How strong is the argument made for the thesis?

▪ Research question

- Can we leverage argument mining to assess the argumentation quality of persuasive essays?

▪ Data

- 800–1003 essays with scores in [1,4] annotated for each dimension



Rating quality based on mining: Mining and analysis

▪ Mining

- **Task.** Classify sentence-level units as thesis, conclusion, premise, or none
- **Data.** AAE corpus (Stab and Gurevych, 2014a)
- **Approach.** SVM with different standard features

Approach	Accuracy	F ₁
Majority baseline	52.5	36.1
Stab and Gurevych (2014b)	77.3	72.6
Mining approach	74.5	74.5

▪ Analysis

- **Task.** Compute most common unit role flows
- **Data.** All paragraphs of all 6085 essays in ICLE corpus (Granger et al., 2009)

Unit role flows	Average	First	Last
Conclusion, Premises	25.1%	–	13.1%
Conclusion, Premises, Conclusion	17.0%	–	27.2%
None, Thesis	3.4%	25.9%	–
Premises, Conclusion	2.9%	–	2.7%

Rating quality based on mining: Example essay

■ Prompt

"Some people say that in our modern world, dominated by science and technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?"

Organization 3.0
Thesis clarity 2.0
Prompt adherence 4.0
Argument strength 2.0

■ Essay

None

"If we take a look back in time we are in a position to see man dreaming, philosophizing and using his imagination of whatever comes his way. We see man transcending his ego I a way and thus becoming a God - like figure. And by putting down these sacred words, what is taking shape in my mind is the fact that using his imagination Man is no longer this organic and material substance like his contemporary counterpart who is putting his trump card on science, technology and industrialization but Man is a way transcends himself through his imagination.

Introduction

Conclusion

For instance, if we take into account the Renaissance or Romantic periods of mankind and close our eyes we could see Shakespeare applying his imagination in the fancy world of his comedies: elf and nymphs circling the stage making it a dream that will lost forever in our minds. We could even hear their high-pitched weird chuckle piercing with a gentle touch our ears, but "open those eyes that must eclipse the day" and you'll see the high-tech wiping out every trace of the human elevated spirit that have dominated over the previous centuries. What we see now is "deux aux machina" or the fake "God from the machine" who with the touch of a button could unleash Armageddon.

Premise

Body

For poets and literate people of yore it was a common idea to transcend reality or to go beyond it by using their imagination not by using reason as we the homosapiens of our time do. For example, if we indulge in entertaining the idea of the film "The matrix" it has a lot to do with the period of Romanticism. But the difference is that a poet from that time could transcend reality, become one with Nature, and cruise wherever he wants using his imagination. Whereas now in the 21st century and in "The matrix" in particular the scientific type of Man thinks that at last he has succeeded in making travelling without boundaries via the virtual reality of his PC.

Body

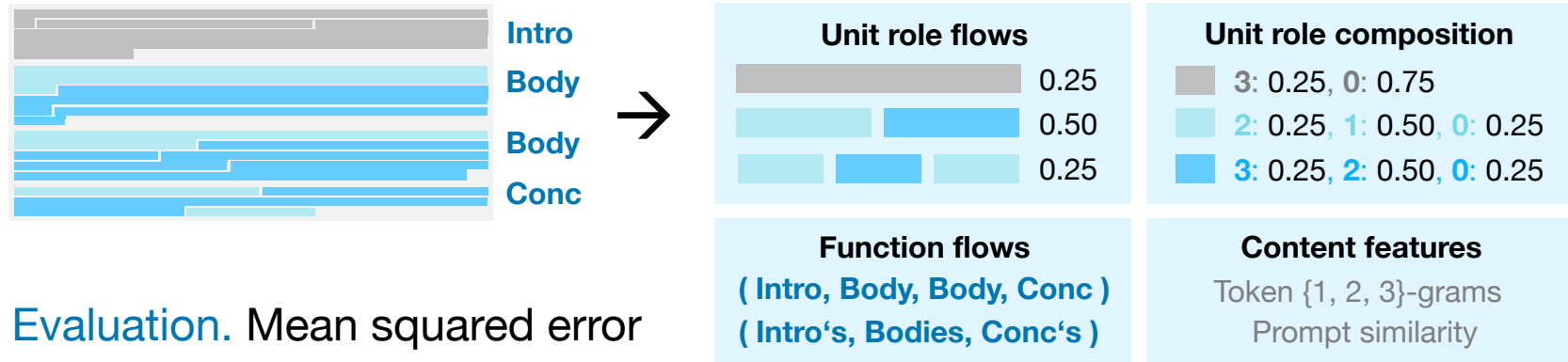
As a logical conclusion to my essay I would like to put only one thing. 'Wouldn't it be better if imagination makes the world go round'. If I was to answer this question, the answer would be positive, but given the aquisitive or consumer society conditions we live in let's make a match between imagination and science. It would be somewhat more realistic."

Conclusion

Rating quality based on mining: Approach and results

Assessment

- Approach.** SVM based on argument-specific and standard features



- Evaluation.** Mean squared error for each quality dimension

Approach	Organization	Clarity	Adherence	Strength
Average baseline	0.349	0.469	0.291	0.266
Persing et al. (2010–2015)	0.175	0.369	0.197	0.244
Assessment approach	0.164	0.425	0.216	0.226
— Unit role flows	0.234	0.461	0.247	0.242
— Unit role composition	0.194	0.457	0.239	0.239
— Function flows	0.220	0.478	0.255	0.251
— Content features	0.336	0.425	0.231	0.236

Relative quality comparison: Overview

▪ Problem

- Rating the quality of an argument in isolation may be hard or even doubtful.
- Is there an easier or more realistic way to assess quality?

▪ Main idea

- Often, we are only interested in the best available argument.
- Then, it's enough to compare the quality of an argument to others.
- **Dilemma.** Unclear in the end whether the best argument is good

Conclusion
Premises

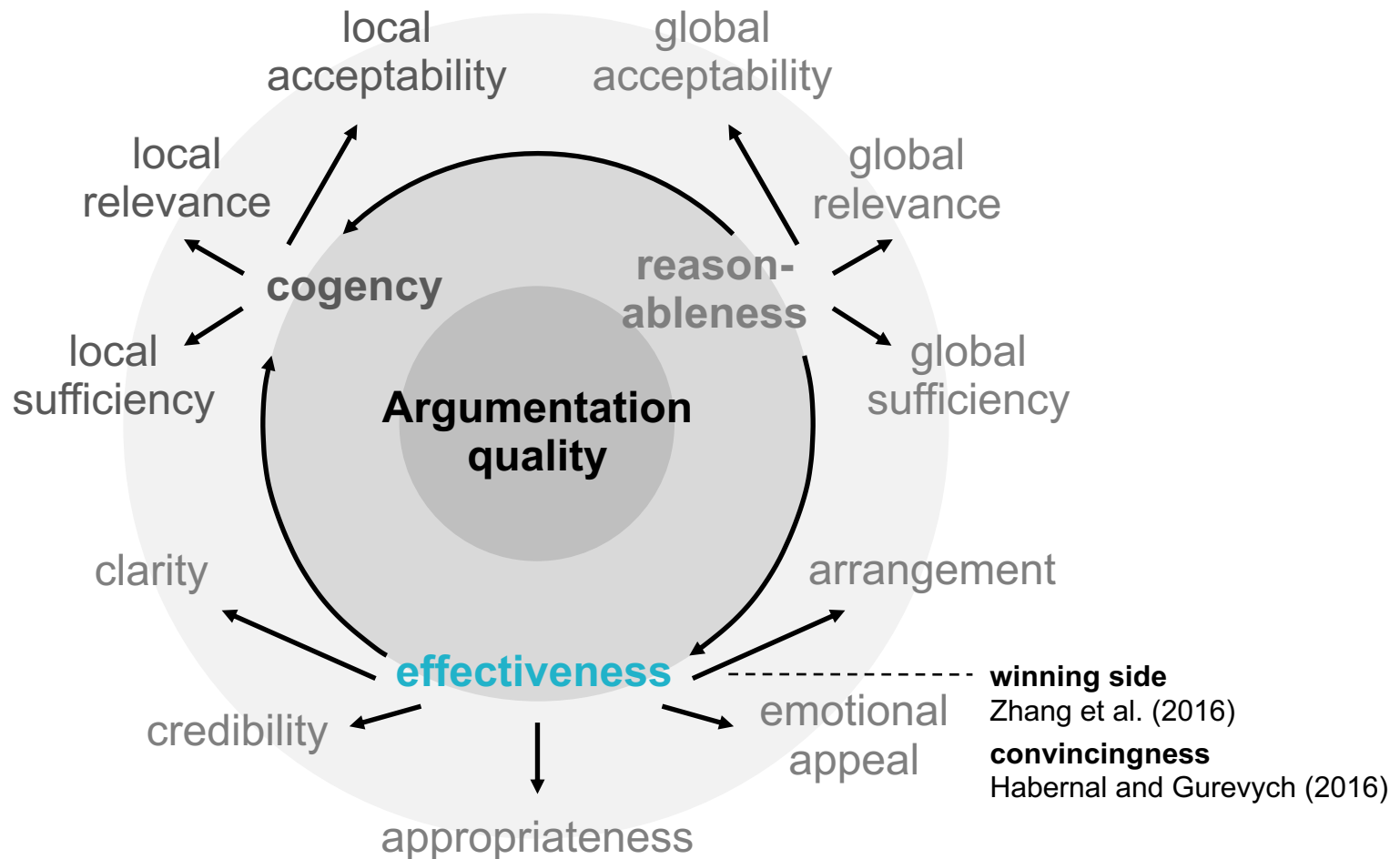
VS

Conclusion
Premises

▪ Existing approaches

- **Winning side.** Prediction of the debate winner from debate flow (Zhang et al., 2016)
- **Winning side.** Prediction of the winner from content and style (Wang et al., 2016)
- **Convincingness.** Argument quality comparison with SVM and Bi-LSTM (Habernal and Gurevych, 2016)
- **Level of support.** Ranking of arguments by support of claim (Braunstein et al., 2016)

Relative quality comparison: Dimensions covered here



Comparing quality based on debate flow (Zhang et al., 2016)

▪ Task

- Given a full Oxford-style debate, which side wins?



▪ Approach

- Mining of supporting points for each side
- Modeling of the "conversational flow":
When does a side put forward own points, when does it attack opponent points?
- Logistic regression classifier with features capturing the flow

"Millennials don't stand a chance"

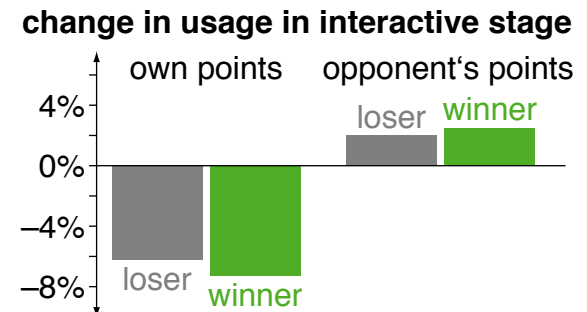
debt college
boomer **pro** reality
economy volunteer
home **con** engage

▪ Data

- 108 Intelligence² debates (117 turns on average)
- Winning side and audience feedback given

▪ Results

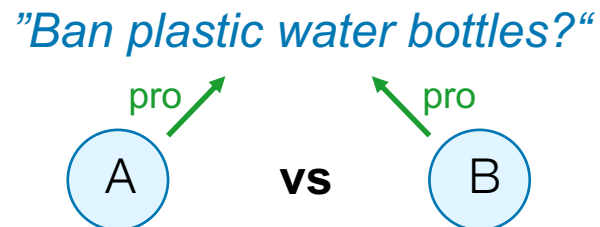
- **Accuracy.** Approach (0.65) beats audience feedback (0.60)
- **Insights.** Attacking the opponent's points better than focus on own points



Comparing quality with SVM and Bi-LSTM (Habernal and Gurevych, 2016)

▪ Task

- Given two arguments with the same topic and stance, which one is more convincing?



▪ Supervised learning approaches

- **SVM.** SVM with RBF kernel using various linguistic features
- **Bi-LSTM.** Bi-directional long short-term memory neural network

Notice: The focus of the paper was not the approaches but the data construction.

▪ Crowdsourced data

- 16,927 pairs of 1052 debate portal arguments for 32 topic-stance pairs
- Each annotated 5 times for convincingness (most reliable annotation taken)

Reliability can be estimated with MACE (Hovy et al., 2013). Annotators also had to give reasons.

▪ Results in 32-fold cross-validation

- **Accuracy.** SVM (0.78) beats Bi-LSTM (0.76); human performance 0.93
- **Insights.** Surface features like capitalization easy, "inverted" sentiment hard

Absolute vs. relative assessment ~ Theory vs. practice

▪ Data representing theory

(Wachsmuth et al., 2017b)

- Absolute expert ratings
- Normative guidelines
- 15 predefined quality dimensions

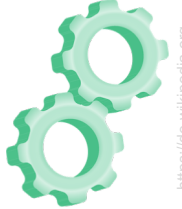


<https://commons.wikimedia.org>

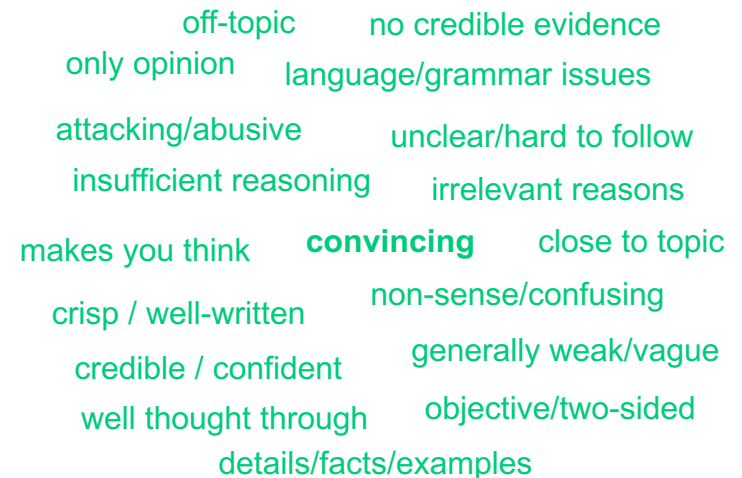
▪ Data representing practice

(Habernal and Gurevych, 2016)

- Relative lay comparisons
- No guidelines
- 17+1 resulting reason labels



<https://de.wikipedia.org>



▪ Empirical comparison of theory and practice

(Wachsmuth et al., 2017d)

- 736 argument pairs are available with ratings *and* labels.
- Compute Kendall's τ correlations of all dimensions and reasons.

How different is assessment in theory and in practice?

▪ Selected insights

- **Convincing** correlates most with **overall quality** (0.64)
- Generally high "correlations" between 0.3 and 1.0
- Perfect: **Global acceptability** + **attacking/abusive** (1.0)
- Mostly very intuitive, such as **clarity** + **unclear** (0.91)
- Top **overall quality** for **well thought through** (mean score 1.8 of 3)
- Lowest **overall quality** for **off-topic** (mean score 1.1 of 3)
- Few unintuitive results, e.g., "only" 0.52 for **credibility** + **no credible evidence**
- **Local sufficiency** + **global sufficiency** hard to separate

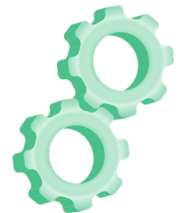
▪ Conclusions

- Theory and practice match more than expected.
- Theory can guide quality assessment in practice.
- Practice indicates what to focus on to simplify theory.



<https://commons.wikimedia.org>

VS



<https://de.wikipedia.org>

Next section: Objective & subjective quality assessment

- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument acquisition
- V. Argument mining
- VI. Argument assessment**
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- a) Introduction
- b) Stance and bias
- c) Schemes and fallacies
- d) Quality in theory
- e) Absolute and relative quality assessment
- f) Objective and subjective quality assessment**
- g) Conclusion

The role of participants in argumentation (recap)

▪ Author (or speaker)

- Argumentation is connected to the person who argues.
- The same argument is perceived differently depending on the author.

” The EU should allow rescue boats. Many innocent refugees will die if there are no rescue boats. “



<https://pixabay.com>



<https://commons.wikimedia.org>

▪ Reader (or audience)

- Argumentation often targets a particular audience.
- Different arguments and ways of arguing work for different readers.

” According to a recent UN study, the number of rescue boats had no effect on the number of refugees who try. “



<https://pixabay.com>

▪ Questions

- May the assessment ignore the author/speaker? And the reader/audience?
The author/speaker is unknown in some application scenarios, but rarely the reader/audience is.

Objective and subjective quality assessment

■ Subjectiveness of quality assessment

- Many dimensions are inherently subjective.
- Quality depends on the subjective weighting of different aspects of an issue.
- Also, it depends on preconceived opinions.

"Should we buy a Chesterfield armchair?"



(thanks to Christian Kock for this example)

■ Example: Which argument is more relevant?

"The death penalty legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."

"The death penalty doesn't deter people from committing serious violent crimes. The thing that deters is the likelihood of being caught and punished."

■ Two ways to approach this problem (both detailed below)

- **Either**, focus on properties that can be assessed "objectively".
- **Or**, include a model of the reader/audience in the quality assessment.

Objective quality assessment: Overview

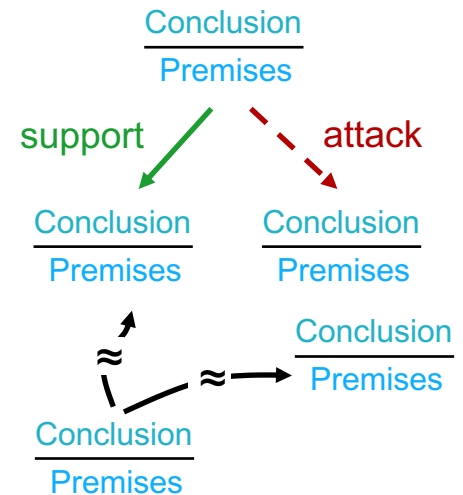
■ Problem

- How to assess quality without learning from subjective annotations?
- What are objective quality indicators?

■ Main idea

- Assess quality based on the structure induced by the set of all arguments.
- Works for both for absolute and relative assessment
- **Dilemma.** Evaluation on subjective annotations?

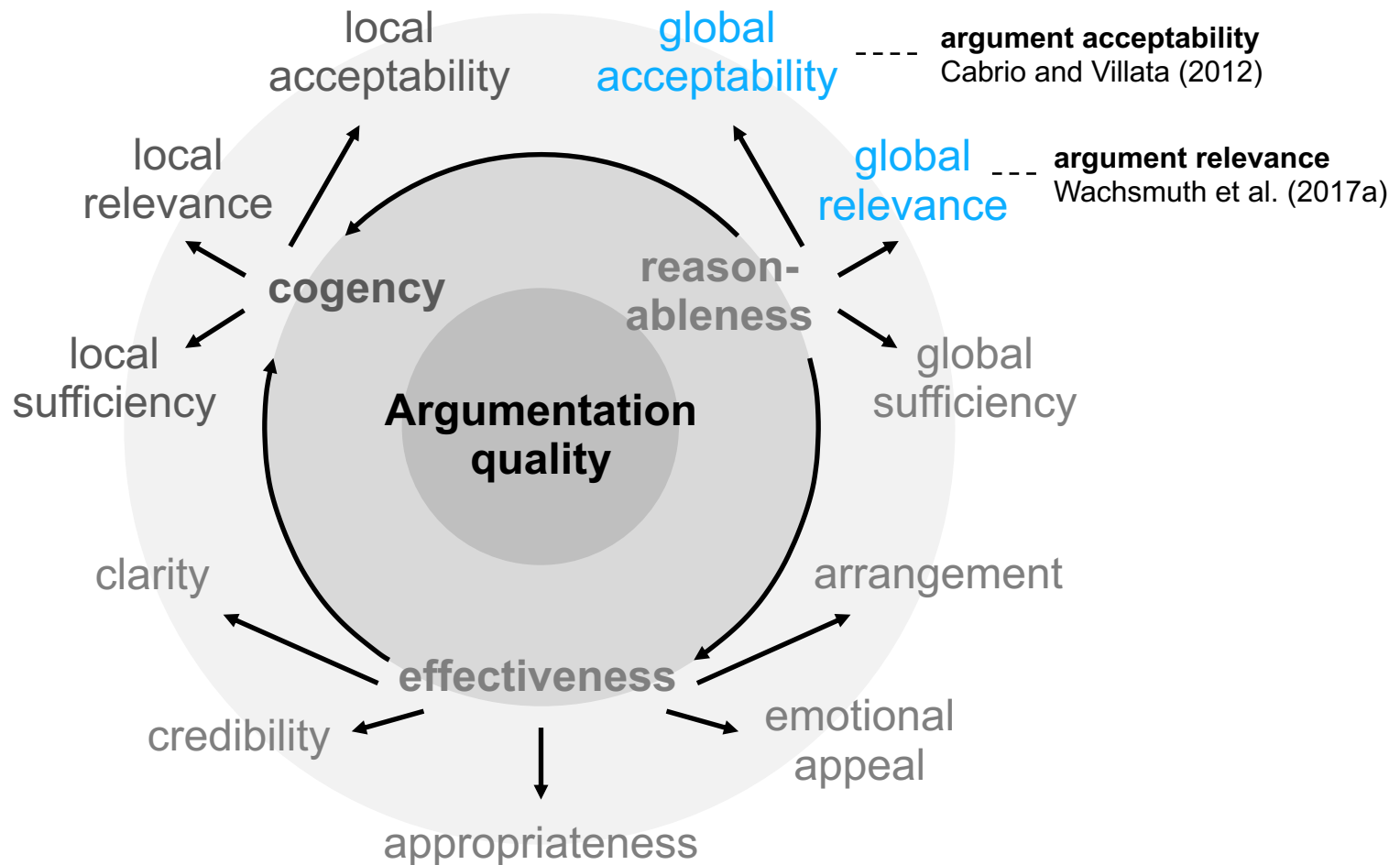
A possible solution is to rely on majority assessments of many annotators.



■ Existing approaches

- **Acceptability.** Assessment based on attack relations (Cabrio and Villata, 2012)
- **Relevance.** Assessment based on reuse of units (Wachsmuth et al., 2017a)
- **Prominence.** Assessment based on argument frequency (Boltužic and Šnajder, 2015)

Objective quality assessment: Dimensions covered here

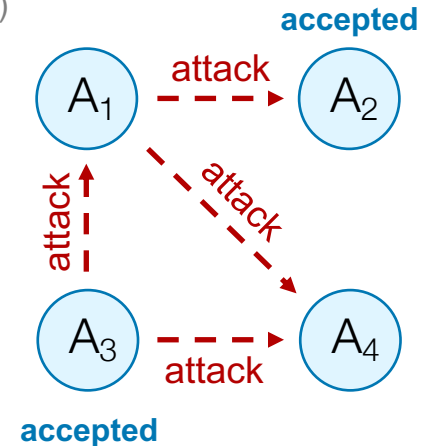


Objective assessment based on attacks (Cabrio and Villata, 2012)

▪ **Background: Abstract argumentation framework** (Dung, 1995)

- A directed graph where nodes represent arguments and edges attack relations between arguments
- Graph analysis reveals whether to accept an argument.
- **Accepted.** If all arguments attacking it are rejected
- **Not accepted.** If an accepted argument attacks it

Extensions with weightings and with support+attack exist.



▪ **Approach**

- Given a set of arguments, use textual entailment algorithm to classify attacks.
- Assess acceptability of arguments following Dung's framework.

▪ **Evaluation**

- Tested on 100 argument pairs from idebate.org, 45 attacking each other
- **Attack classification.** Accuracy 0.67
- **Acceptability assessment.** Accuracy 0.75

Objective assessment based on reuse (Wachsmuth et al., 2017a)

▪ Task

- Given a set of arguments, which one is most relevant to some issue?
- **Problem.** Relevance is highly subjective

"The death penalty legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."

"The death penalty doesn't deter people from committing serious violent crimes. The thing that deters is the likelihood of being caught and punished."

▪ Research question

- Can we develop an "objective" measure of relevance?

▪ Key hypothesis

- The relevance of a conclusion depends on what other arguments across the web use it as a premise.
- **Rationale.** Author cannot control who "cites" a conclusion in this way.

▪ Approach

- Ignore content and reasoning of arguments (for now).
- Derive relevance structurally from the reuse of conclusions at web scale.

Conclusion

Premises

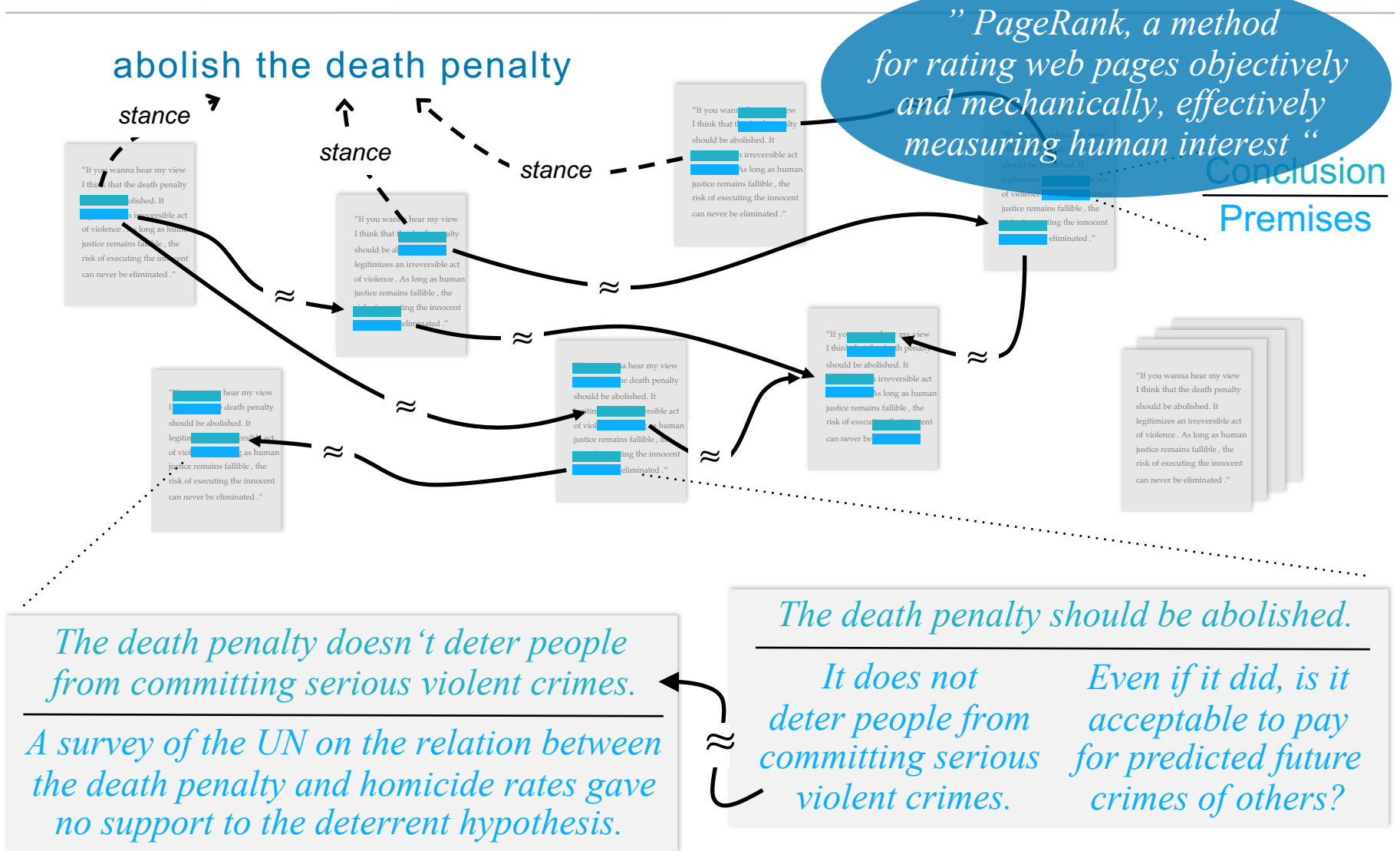


≈

Conclusion

Premises

Objective assessment based on reuse: Argument graph



Objective assessment based on reuse: Approach

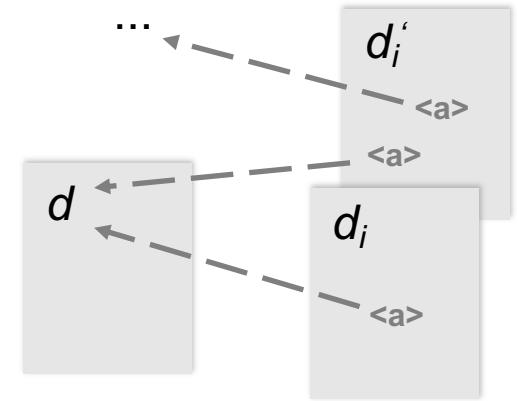
- Original PageRank score of a web page d (Page et al., 1999)

same score for each page

$$p(d) = (1 - \alpha) \cdot \underbrace{\frac{1}{|D|}}_{\text{ground relevance}} + \alpha \cdot \underbrace{\sum_i \frac{p(d_i)}{|D_i|}}_{\text{recursive relevance}}$$

page d_i links to d
page d_i links to d

pages d_i links to
pages d_i links to



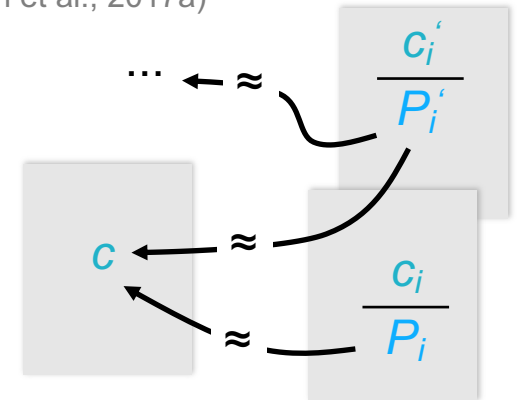
- Adapted PageRank score of an argument unit c (Wachsmuth et al., 2017a)

PageRank of page d containing c

$$\hat{p}(c) = (1 - \alpha) \cdot \underbrace{\frac{p(d) \cdot |D|}{|A|}}_{\text{ground relevance}} + \alpha \cdot \underbrace{\sum_i \frac{\hat{p}(c_i)}{|P_i|}}_{\text{recursive relevance}}$$

conclusion c_i uses c as premise
conclusion c_i uses c as premise

premises of c_i
premises of c_i



- Argument relevance is aggregation of premise scores.
 - Minimum, average, maximum, or sum

Objective assessment based on reuse: Results

▪ Evaluation of unsupervised ranking approaches

PageRank
of premises

\hat{p}

Frequency
of premises

Σ

Similarity
of units

$c \sim P$

Sentiment
of premises



Number
of premises

$|P|$

Random
ranking



each for minimum, average, maximum, and sum aggregation

▪ Experiment on graph with 18k arguments

57 argument corpora from www.aifdb.org

- Rank with each approach
- Correlate with benchmark rankings

▪ Results

- PageRank with sum aggregation best
- Notable correlation despite ignorance of content and inference
- Other quality assessment should follow

best rank correlation (higher is better)

#	Approach	Kendall's τ
1	PageRank	0.28
2	Number	0.19
3	Sentiment	0.12
4	Frequency	0.10
5	Similarity	0.02
6	Random	0.00

Objective assessment based on reuse: Examples



<https://de.wikipedia.org>

” Technology has enhanced the daily life of humans. ”

#3 *” The use of technology has revolutionized business. ”*

#1 *” The internet has enabled us to widen our knowledge. ”*

#2 *” Technology has given us a means of social interaction that wasn't possible before. ”*



<https://pixabay.com>

” Strawberries are the best choice for your breakfast meal. ”

#1 *” Berries are superfoods because they're so high in antioxidants without being high in calories, says Giovinazzo MS, RD, a nutritionist at Clay health club and spa, in New York City. ”*

#3 *” Strawberries are good for your ticker. ”*

#2 *” One cup of strawberries, for instance, contains your full recommended daily intake of vitamin C, along with high quantities of folic acid and fiber. ”*

Inclusion of Subjectivity: Overview

▪ Problem

- Ultimately, effective argumentation requires considering the target audience.
- Humans would barely argue without doing so.

▪ Main idea

- Model the target audience within quality assessment.
- This also includes to have audience-specific ground-truth annotations.



▪ Missing approaches

- Audience models have rarely been included explicitly so far.
- Implicitly, some annotated corpora may actually represent specific audiences.
- Recent studies analyze the quality perception of different audiences.

▪ Studies

- **Different personalities.** Effectiveness of emotional vs. rational arguments (Lukin et al., 2017)
- **Different ideologies.** Effectiveness of news editorials (El Baff et al., 2018)

Effectiveness based on personality (Lukin et al., 2017)

▪ Hypothesis

- People with different personalities are open to different types of arguments.

▪ Study

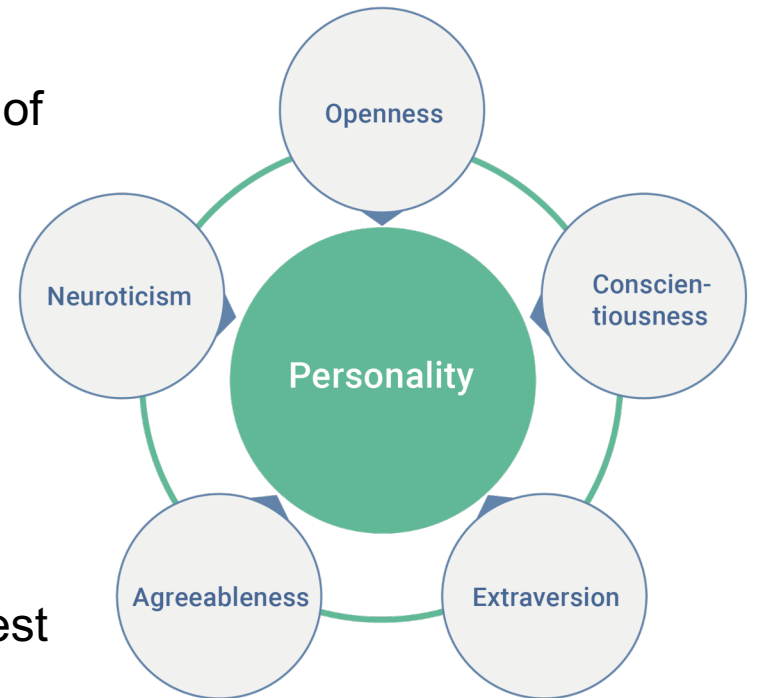
- Impact of personality on the effectiveness of emotional and factual arguments
- **Personality.** Here, the "Big Five"

▪ Data

- 5185 arguments from online dialogues
- **Quality.** Each annotated for whether it changed the belief (to pro, to con, neither)
- **Personality.** Each annotator did Big Five test

▪ Selected insights

- Agreeable people easiest to predict (F_1 0.48), extroverted hardest (F_1 0.44)
- Factual arguments best for agreeable people, emotional best for open people

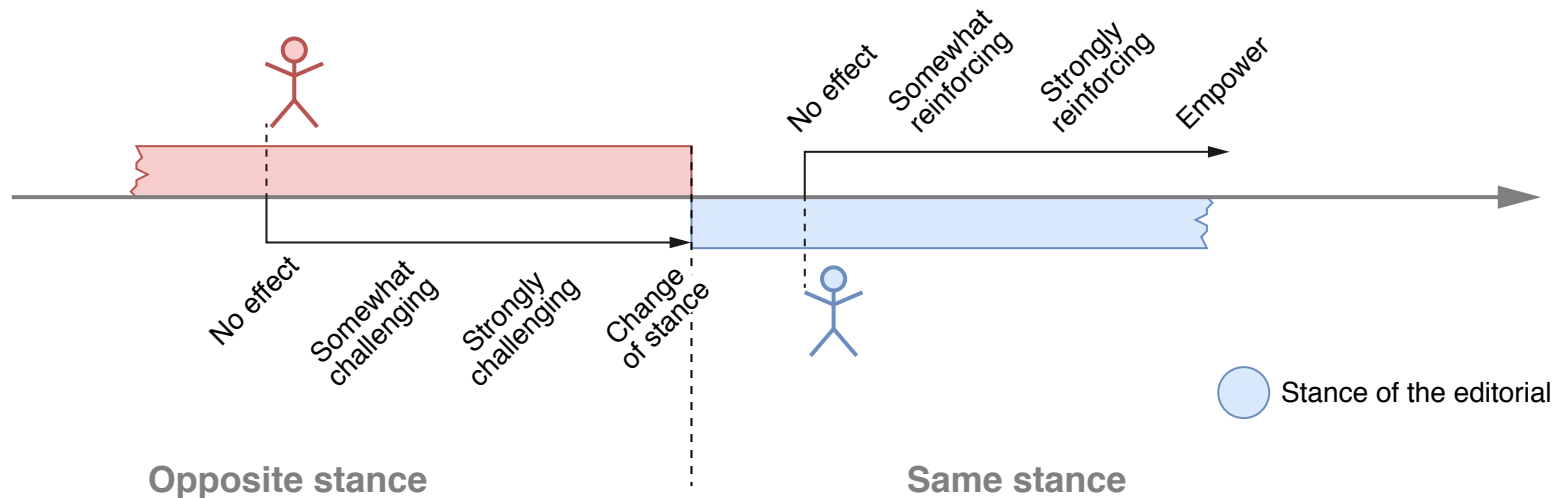


<https://commons.wikimedia.org>

Effectiveness based on ideology (El Baff et al., 2018)

▪ Effects of news editorials

- News editorials are said to shape public opinion, but they rarely *change* a reader's prior stance.
- Rather, they challenge or reinforce stance — or neither.



▪ Dialectical notion of argumentation quality

- A good editorial reinforces one side and challenges the other.
- Or it challenges both sides.

Effectiveness based on ideology: Data

▪ Hypothesis

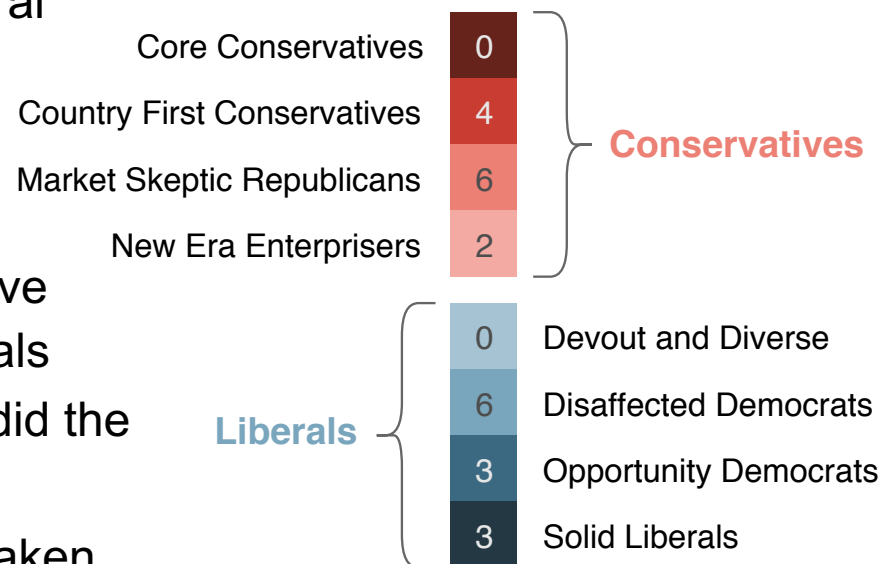
- Prior stance depends on political ideology (and personality).
- Ideology needs to be known to assess the effectiveness of news editorials.

▪ Study

- Impact of ideology (and personality) on the effectiveness of news editorials
- **Ideology.** Here, conservative vs. liberal

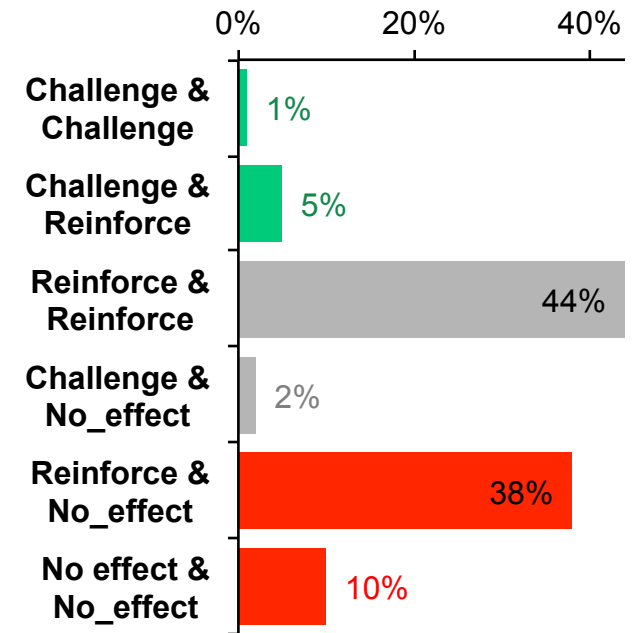
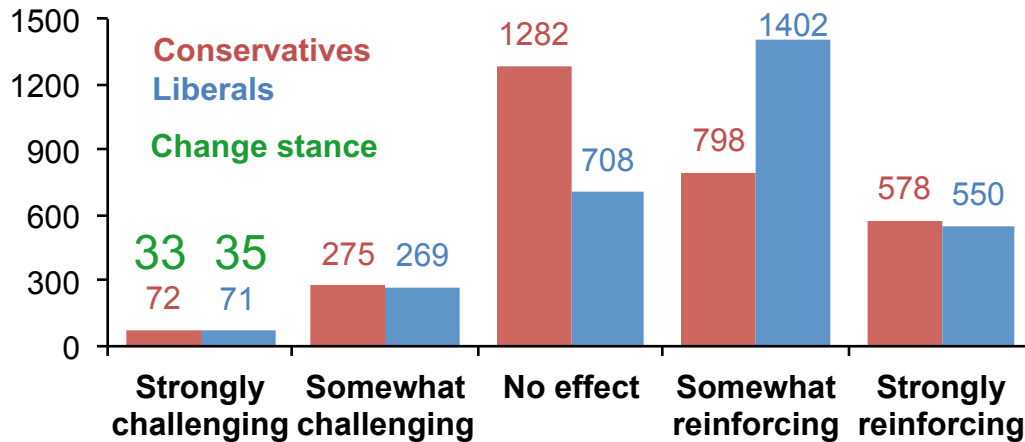
▪ Data

- 1000 editorials from NYTimes
- **Quality.** Each annotated for persuasive effect by 3 conservatives and 3 liberals
- **Ideology.** All 24 annotators (in total) did the Political Typology Quiz.
- **Personality.** Also, Big Five test was taken.



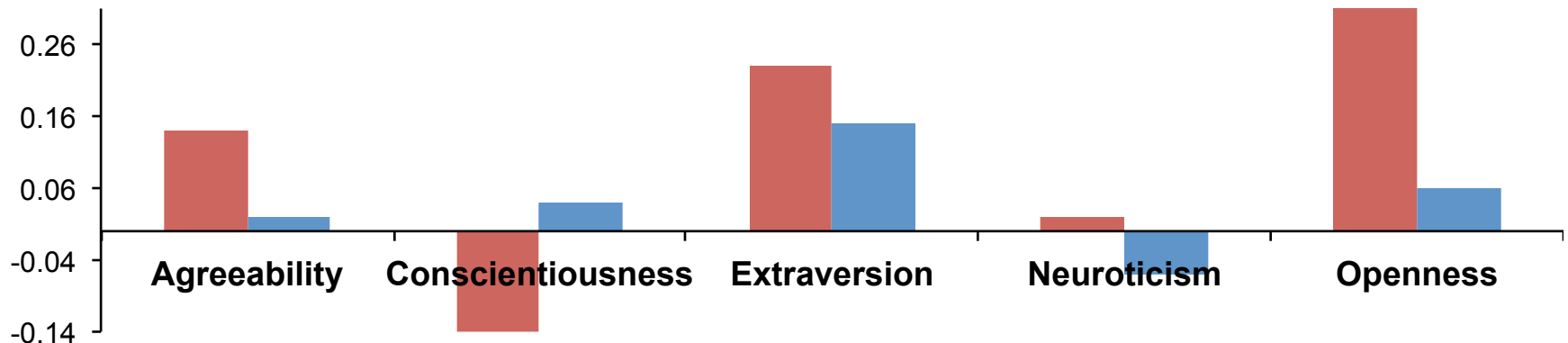
Effectiveness based on ideology: Results

Majority effect distribution in the corpus



Effect depending on ideology and personality

Kendall's τ correlation with challenge/reinforce



Effect assessment depending on ideology (El Baff et al., 2020)

▪ Task

- Given a news editorial and a reader's ideology, predict the persuasive effect.

▪ Approach

- SVM using five style feature types:
- **LIWC**. Psyche-related words
- **NRC**. Emotion/Sentiment words
- **MPQA-S**. Subjective words
- **MPQA-A**. Argumentative words
- **ADUs**. Distribution of ADU types

+ Lemma n-grams for comparison to content

▪ Data

- As above, 80% training, 20% test

▪ Results

- Only for liberals, significant micro-F₁ gains over random baseline achieved

For liberals, style seems at least as discriminative as content.

Features	Conserv.	Liberals
LIWC	0.26	0.40
NRC	0.29	0.39
MPQA-S	0.28	0.38
MPQA-A	0.29	0.41
ADUs	0.31	0.36
Best style set	0.37	*0.49
Lemma n-grams	0.38	*0.49
Best overall	0.36	**0.54
Random baseline	0.34	0.26

Next section: Conclusion

- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument acquisition
- V. Argument mining
- VI. Argument assessment**
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- a) Introduction
- b) Stance and bias
- c) Schemes and fallacies
- d) Quality in theory
- e) Absolute and relative quality assessment
- f) Objective and subjective quality assessment
- g) Conclusion**

Conclusion

Argument assessment

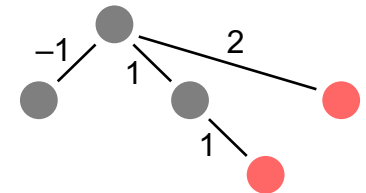
- Classification of issue-related subjectiveness properties
- Interpretation of the reasoning of an argument
- Judgment of several quality dimensions of an argument



argument from consequences

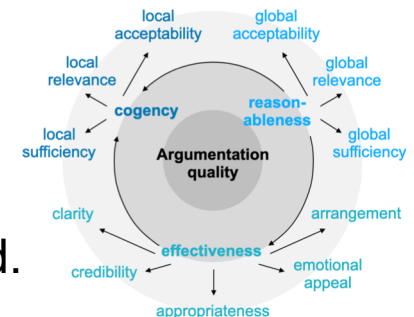
Subjectiveness and reasoning

- Stance, bias, argumentation schemes, fallacies, and more
- Stance classification is a major and extensively-studied task.
- Reasoning-related methods are still limited.



Argumentation quality

- Several dimensions are considered in theory and practice.
- Absolute rating and relative comparison approaches exist.
- Subjectiveness may be included or somehow circumvented.



References

- **Aioli et al. (2009).** Fabio Aioli, Giovanni Da San Martino, and Alessandro Sperduti. 2009. Route kernels for trees. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 17–24.
- **Aristotle (2007).** Aristotle (George A. Kennedy, Translator). *On Rhetoric: A Theory of Civic Discourse*. Clarendon Aristotle series. Oxford University Press, 2007.
- **Bar-Haim et al. (2017a).** Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance Classification of Context-Dependent Claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, 2017.
- **Bar-Haim et al. (2017b).** Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. Improving Claim Stance Classification with Lexical Knowledge Expansion and Context Utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, 2017.
- **Blair (2012).** J. Anthony Blair. *Groundwork in the Theory of Argumentation*. Springer Netherlands, 2012.
- **Boltužic and Šnajder (2015).** Filip Boltužic and Jan Šnajder. Identifying Prominent Arguments in Online Debates using Semantic Textual Similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, 2015.
- **Braunstain et al. (2016).** Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. Supporting Human Answers for Advice-seeking Questions in CQA Sites. In *Proceedings of the 38th European Conference on IR Research*, pages 129–141, 2016.
- **Cabrio and Villata (2012).** Elena Cabrio and Serena Villata. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, 2012.
- **Coe et al. (2014).** Kevin Coe, Kate Kenski, and Stephen A. Rains. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64(4):658–679, 2014.

References

- **Cohen (2001)**. Daniel H. Cohen. Evaluating Arguments and Making Meta-Arguments. *Informal Logic*, 21(2):73–84, 2001.
- **Damer (2009)**. T. Edward Damer. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Wadsworth, Cengage Learning, Belmont, CA, 6th edition, 2009.
- **Dung (1995)**: Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321–357, 1995.
- **El Baff et al. (2018)**. Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, 2018.
- **El Baff et al. (2020)**. Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, to appear 2020.
- **Feng and Hirst (2011)**. Vanessa Wei Feng and Graeme Hirst. Classifying Arguments by Scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 987–996, 2011.
- **Feng et al. (2014)**. Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949. Dublin City University and Association for Computational Linguistics, 2014.
- **Freeley and Steinberg (2009)**. Austin J. Freeley and David L. Steinberg. *Argumentation and Debate*. Cengage Learning, 12th edition, 2008.
- **Freeman (2011)**. *Argument Structure: Representation and Theory*. Springer, 2011.

References

- **Govier (2010).** Trudy Govier. *A Practical Study of Argument*. Wadsworth, Cengage Learning, Belmont, CA, 7th edition, 2010.
- **Granger et al. (2009).** Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. *International Corpus of Learner English (version 2)*, 2009.
- **Gurcke et al. (2021).** Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. *Assessing the Sufficiency of Arguments through Conclusion Generation*. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, 2021.
- **Habernal and Gurevych (2016).** Ivan Habernal and Iryna Gurevych. 2016. *Which Argument is More Convincing? Analyzing and Predicting Convincingness of Web Arguments using Bidirectional LSTM*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.
- **Habernal et al. (2018).** Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. *Before Name-calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation*. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 386–396, 2018.
- **Hamblin (1970).** Charles L. Hamblin. *Fallacies*. Methuen, London, UK, 1970.
- **Hasan and Ng (2013).** Kazi Saidul Hasan and Vincent Ng. *Stance Classification of Ideological Debates: Data, Models, Features, and Constraints*. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, 2013.
- **Hoeken (2001).** Hans Hoeken. *Anecdotal, Statistical, and Causal evidence: Their Perceived and Actual Persuasiveness*. *Argumentation*, 15(4):425–437, 2001.
- **Hovy et al. (2013).** Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. *Learning Whom to Trust with MACE*. In *Proceedings of NAACL-HLT 2013*, pages 1120–1130.

References

- **Johnson and Blair (2006).** Ralph H. Johnson and J. Anthony Blair. 2006. Logical Self-defense. International Debate Education Association.
- **Lawrence and Reed (2016).** John Lawrence and Chris Reed. Argument Mining Using Argumentation Scheme Structures. In Proceedings of the Sixth International Conference on Computational Models of Argument, pages 379–390, 2016.
- **Lukin et al. (2017).** Stephanie Lukin, Pranav Anand, Marilyn Walker and Steve Whittaker. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 741–752, 2017.
- **O’Keefe and Jackson (1995).** Daniel J. O’Keefe and Sally Jackson. Argument Quality and Persuasive Effects: A Review of Current Approaches. In *Argumentation and Values: Proceedings of the Ninth Alta Conference on Argumentation*, pages 88–92, 1995.
- **Mercier and Sperber (2011).** Hugo Mercier and Dan Sperber. 2011. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 34:57–111.
- **Page et al. (1999).** Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120, 1999.
- **Park et al. (2015).** Joonsuk Park, Cheryl Blake, and Claire Cardie. Toward Machine-assisted Participation in eRulemaking: An Argumentation Model of Evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210, 2015.
- **Peldszus and Stede (2016).** Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: 1st European Conference on Argumentation*.
- **Perelman and Olbrecht-Tyteca (1969).** Chaïm Perelman and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric: A Treatise on Argumentation* (John Wilkinson and Purcell Weaver, translator). University of Notre Dame Press.

References

- **Persing and Ng (2013):** Isaac Persing and Vincent Ng. Modeling Thesis Clarity in Student Essays. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 260–269, 2013.
- **Persing and Ng (2014):** Isaac Persing and Vincent Ng. Modeling Prompt Adherence in Student Essays. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1534–1543, 2014.
- **Persing and Ng (2015):** Isaac Persing and V. Ng. Modeling Argument Strength in Student Essays. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 543–552, 2015.
- **Persing et al. (2010).** Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 229–239, 2010.
- **Rahimi et al. (2014).** Zahra Rahimi, Diane J. Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. Automatic Scoring of an Analytical Response-to-Text Assessment. In Proceedings of the 12th International Conference on Intelligent Tutoring Systems, pages 601–610, 2014.
- **Rahimi et al. (2015).** Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. Incorporating Coherence of Topics as a Criterion in Automatic Response-to-Text Assessment of the Organization of Writing. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 20–30, 2015.
- **Somasundaran and Wiebe (2010):** Swapna Somasundaran and Janyce Wiebe. Recognizing Stances in Ideological On-Line Debates. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 116–124, 2010.
- **Stab and Gurevych (2014a).** Christian Stab and Iryna Gurevych. Annotating Argument Components and Relations in Persuasive Essays. In Proceedings of the 25th Conference on Computational Linguistics, pages 1501–1510, 2014.

References

- **Stab and Gurevych (2014b)**. Christian Stab and Iryna Gurevych. Identifying Argumentative Discourse Structures in Persuasive Essays. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pages 46–56, 2014.
- **Stab and Gurevych (2016)**. Christian Stab and Iryna Gurevych. Recognizing the Absence of Opposing Arguments in Persuasive Essays. In Proceedings of the Third Workshop on Argument Mining (ArgMining2016), pages 113–118, 2016.
- **Stab and Gurevych (2017)**. Christian Stab and Iryna Gurevych. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 980–990, 2017.
- **Stede and Schneider (2018)**. Manfred Stede and Jodi Schneider. Argumentation Mining. Synthesis Lectures on Human Language Technologies 40, Morgan & Claypool, 2018.
- **Ranade et al. (2013)**: Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. Stance Classification in Online Debates by Recognizing Users' Intentions. In: Proc. of the SIGDIAL 2013, 61–69, 2013.
- **Tindale (2007)**. Christopher W. Tindale. 2007. Fallacies and Argument Appraisal. Critical Reasoning and Argumentation. Cambridge University Press.
- **Toulmin (1958)**. Stephen E. Toulmin. The Uses of Argument. Cambridge University Press, 1958.
- **van Eemeren (2015)**. Frans H. van Eemeren. Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics. Argumentation Library. Springer International Publishing, 2015.
- **van Eemeren and Grootendorst (2004)**. Frans H. van Eemeren and Rob Grootendorst. 2004. A Systematic Theory of Argumentation: The Pragma-Dialectical Approach. Cambridge University Press, Cambridge, UK.

References

- **Wachsmuth et al. (2016).** Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Using Argument Mining to Assess the Argumentation Quality of Essays. In: Proceedings of the 26th International Conference on Computational Linguistics, pages 1680–1692, 2016.
- **Wachsmuth et al. (2017a).** Henning Wachsmuth, Benno Stein, and Yamen Ajjour. "PageRank" for Argument Relevance. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 1116–1126, 2017.
- **Wachsmuth et al. (2017b).** Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational Argumentation Quality Assessment in Natural Language. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 176–187, 2017.
- **Wachsmuth et al. (2017d).** Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation Quality Assessment: Theory vs. Practice. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Vancouver, Canada, pages 250–255, 2017.
- **Wachsmuth et al. (2017f).** Henning Wachsmuth, Giovanni Da San Martino, Dora Kiesel, and Benno Stein. The Impact of Modeling Overall Argumentation with Tree Kernels. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2369–2379, 2017.
- **Walton (2006).** Douglas Walton. Fundamentals of Critical Argumentation. Cambridge University Press, 2006.
- **Walton et al. (2008).** Douglas Walton, Christopher Reed, and Fabrizio Macagno. Argumentation Schemes. Cambridge University Press, 2008.
- **Wang et al. (2017).** Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. Transactions of the Association for Computational Linguistics 5:219–232, 2017.

References

- **Wei et al. (2016).** Zhongyu Wei, Yang Liu, and Yi Li. Is this Post Persuasive? Ranking Argumentative Comments in Online Forum. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 195–200, 2016.
- **Zhang et al. (2016).** Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational Flow in Oxford-style Debates. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 136–141, 2016.