Computational Argumentation — Part V

Argument Assessment

Henning Wachsmuth

https://ai.uni-hannover.de

Natural Language Processing

Leibniz Universität Hannover

Learning goals

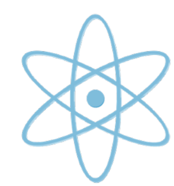
- **Concepts**
  - Various properties of argumentation to be assessed
  - Theoretical notions of argumentation quality
  - The subjective nature of certain properties

https://commons.wikimedia.org

- **Methods**
  - Route kernels and more for stance and myside bias
  - Feature-based and neural methods for schemes and fallacies
  - Multitask learning and graph analyses for argument 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 mining

V. Argument assessment

- **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

VI. Argument generation

VII. Applications of computational argumentation

VIII. Conclusion

Argument assessment

- **Argument(ation) assessment**
  - Coverage term for 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 such boats. Nothing justifies to endanger the life of innocent people.”

stance on issue? | **reasoning scheme?** | **argument quality?** | **framing of issue?** | **human Values?**

- **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

Properties of argumentation

- **What is meant by properties?**
  - Meta-information that reflects 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 such 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, frames, human values, ...
  - Style. Genre, authorship, discourse modes, rhetorical moves, ...
  - Structure. Argumentative depth, claim centrality, diviseness, ...

- **Notice**
  - Where mining ends and assessment starts is not defined exactly.

    For example, classifying evidence types might be seen as assessment.

Next section: Stance and bias

Stance and myside bias classification

- **Stance classification**
  - The classification of the stance of a 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 such 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 (i.e., whether it only supports its own stance)
  - Input. An argumentative text
  - Output. Whether the text has *myside bias* or *no myside bias*

Stance classification: Examples

- **Example: Stance classification of claims**
  - What is the stance of the claims on the right to the issues on the left?

Issue		Claim
"We should ban boxing."	←	*"Boxing remains the 8th most deadly sport."*
"We should embrace multiculturalism."	←	*"Unity is seen as an essential feature of the nation and the nation-state."*

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

- **Challenges of stance classification**
  - 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

- **Modeling stance classification**
  - Standard text classification trained for specific issues
  - Relation-like classification with the issue as input
  - Graph-based analysis over all arguments in a debate

- **Selected general approaches**
  - Feature-based classification (Somasundaran and Wiebe, 2010)
  - Exploit author knowledge in dialogue (Ranade et al., 2013)
  - Exploit opposing views in dialogue (Hasan and Ng, 2013)

  … among many others

- **Selected argument-specific approaches**
  - Stance as sentiment and contrast of text and issue targets (Bar-Haim et al., 2017a/b)
  - Route kernels based on overall structure of texts (Wachsmuth et al., 2017d)
  - Graph convolutional network on whole debate structure (Barrow et al., 2021)

Target: Rescue boats

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 supposed to have a claim-like phrasing itself.

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.

claim target polarity
× contrastiveness
× issue target polarity

≈ stance

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?, ...
  - Classification. Logistic regression

Issue. *”**Advertising** is harmful.“*

Claim. *”**Marketing** creates consumerism and waste.“*

- **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

Issue. *”Advertising is harmful.“* –1

Claim. *”Marketing creates consumerism and waste.“* –0.7

- **Score contrastiveness $c(t_c, t_i)$ in** [–1,1]
  - Features. Polarity shifters, relatedness measures, Wikipedia headers, ...
  - Classification. Random forest

Advertising ↔ ***Marketing*** 1

- **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.

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

Stance as sentiment and contrast: Results

- **Evaluation** (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

- **Modeling myside bias classification**
  - Conceptually, a standard text classification task
  - Argumentative structure may be predictive for myside bias.

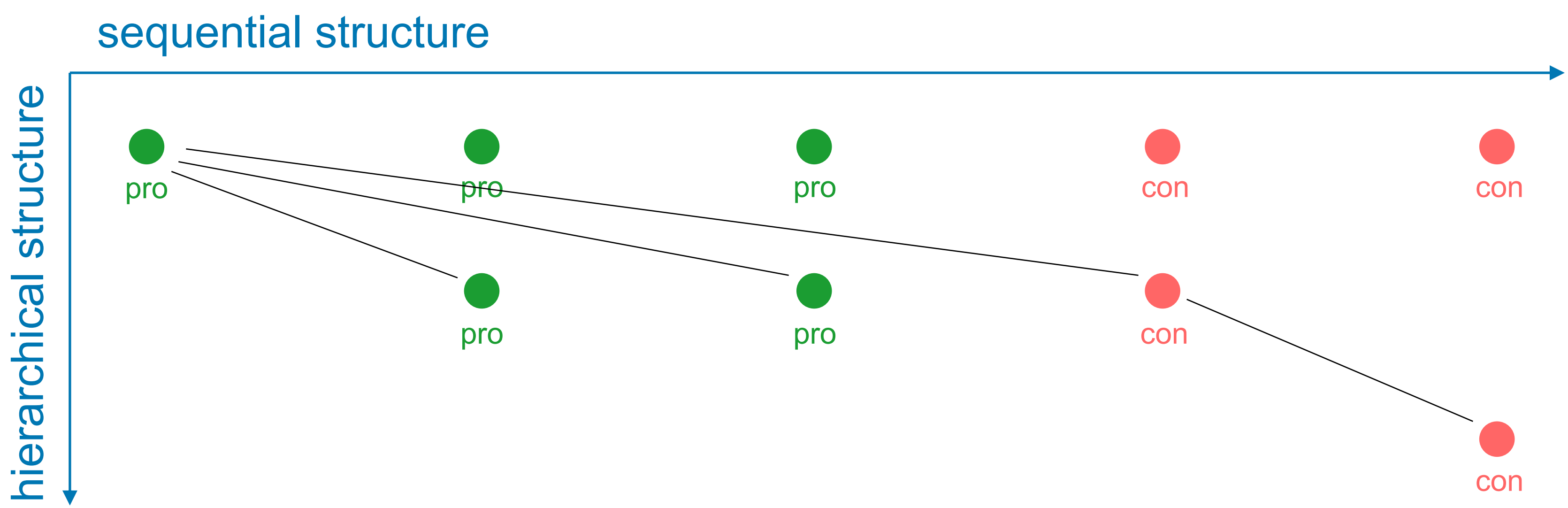
https://www.gesher-jds.org

- **Approaches to myside bias classification**
  - Supervised classification using various features (Stab and Gurevych, 2016)
  - Route kernels based on overall structure of texts (Wachsmuth et al., 2017d)

Background: Overall structure of argumentative texts

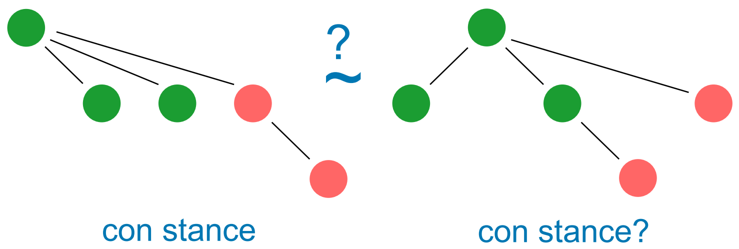
The death penalty is a legal means that as such is not practicable in Germany.	*For one thing, inviolable human dignity is anchored in our constitution,*	*and further no one may have the right to adjudicate upon the death of another human being.*	*Even if many people think that a murderer has already 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)

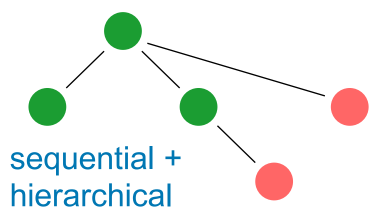


Route kernels for stance and bias (Wachsmuth et al., 2017d)

- **Task**
  - Given a monological argumentative text, classify stance and myside bias (without knowing the issue discussed).
- **Hypothesis**
  - The overall structure of the text is decisive for stance and myside bias.



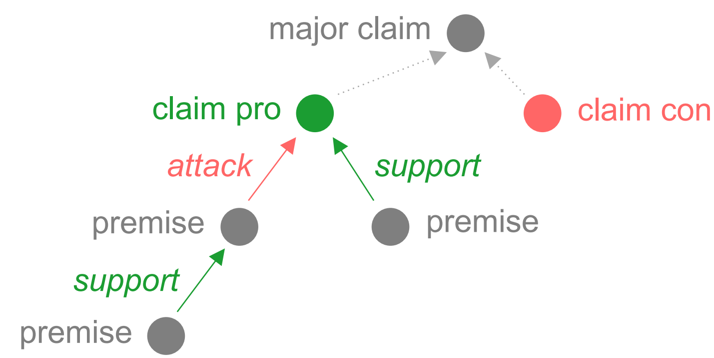
- **Research questions**
  1. How to jointly model sequential and hierarchical overall structure?
  2. What model has most impact on the two tasks?



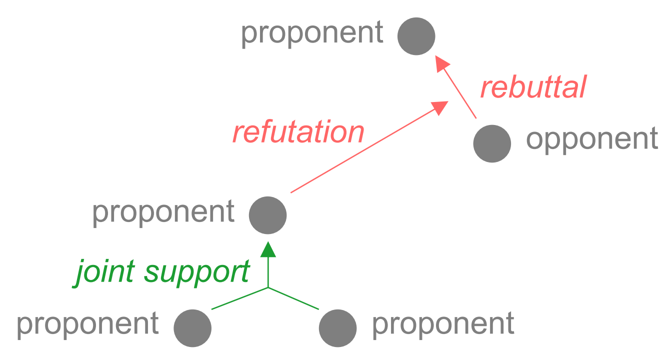
- **Approach in a nutshell**
  - Start from argumentative structure of a text.
  - Model overall structure with so called *route kernels*.
  - Classify stance and myside bias based on kernels.

Route kernels for stance and bias: Tasks and data

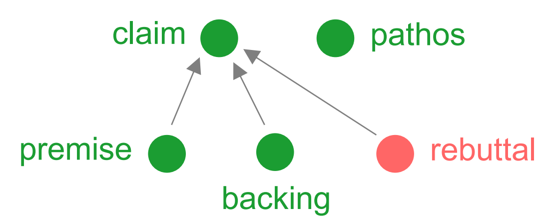
- **Myside bias on AAE-v2**
  (Stab and Gurevych, 2016)
  - 402 persuasive student essays
  - Essay-specific argument model
  - 251 myside bias, 151 no myside bias



- **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

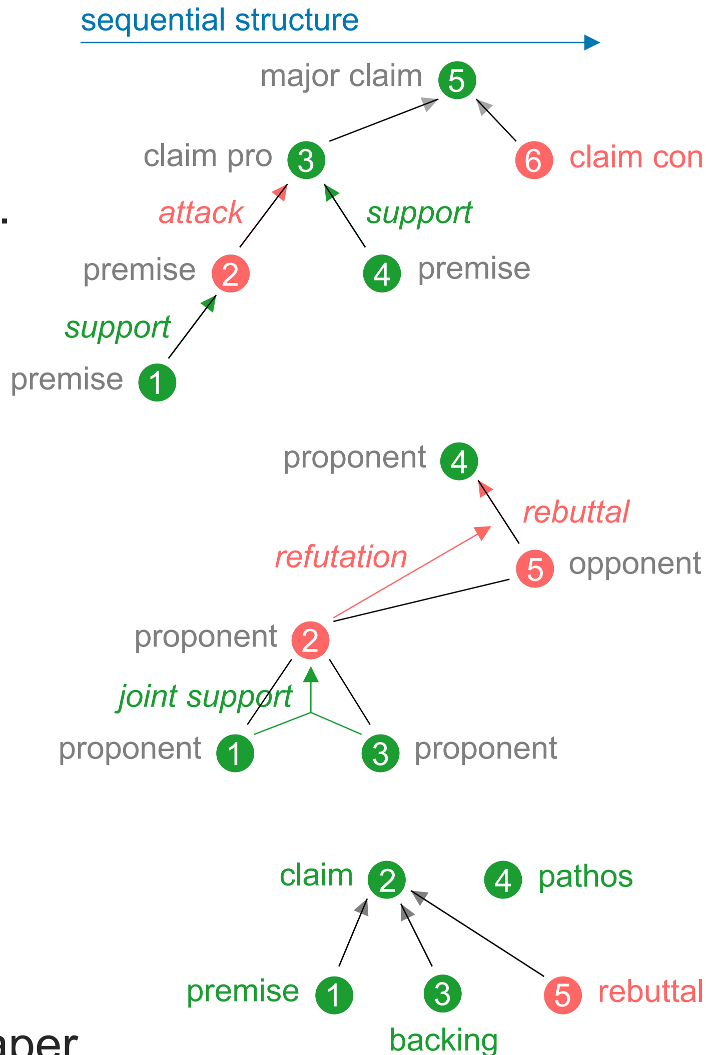


Route kernels for stance and bias: Unification

- **A unified model**
  - Order nodes according to position.
  - Encode stance towards parent as node label.
  - Model relations between node *pairs* only.
  - The root implicitly defines the thesis.

- **Pros and cons**
  - \+ Sequential structure captured
  - \+ Same analyses on all corpora
  - \+ Comparisons across corpora
  - \+ Simpler argument mining (presumably)
  - − Partly less expressive

- **In this lecture, only unified model**
  - For experiments with specific models, see paper.
    (Wachsmuth et al., 2017d)

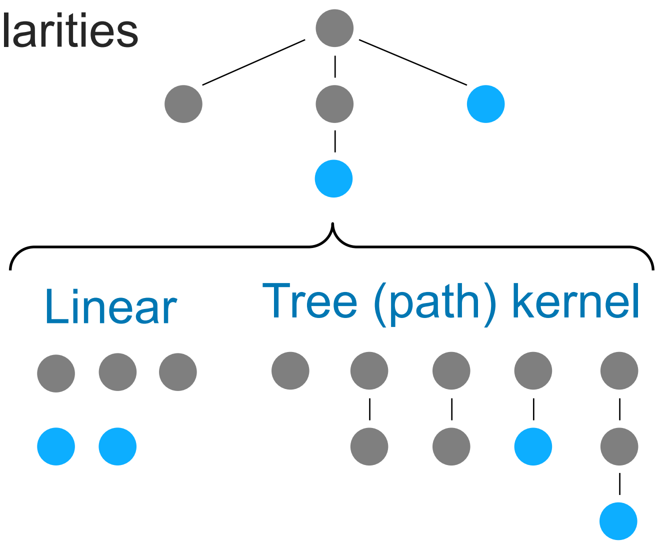


Background: Kernel methods

- **Kernel methods in machine learning**
  - Kernel methods classify instances by comparing them to known instances.
  - Strong when good features are unknown and/or when data is limited
    Often used for structured input data, such as trees

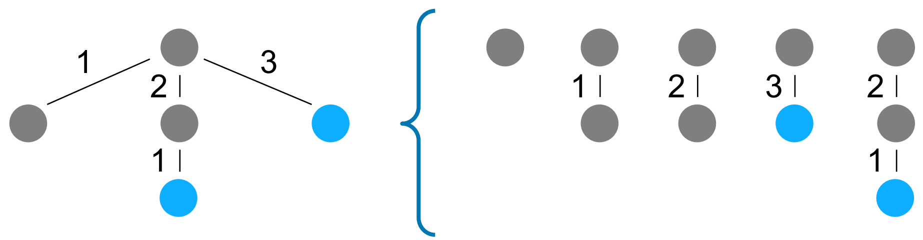
- **Kernel method in a nutshell**
  - Kernel. Represents an instance in a task-specific implicit feature space
    Different kernels can be combined mathematically.
  - Similarity function. Quantifies the similarity of any two kernels
  - Classifier. Distinguishes classes based on similarities
    A typical kernel-based classifier is the support vector machine (SVM).

- **Selected kernels for structured data**
  - Linear kernels capture distributions only
    The correspondent of standard feature vectors
  - Subsequence kernels for sequential structure
    (Mooney and Bunescu, 2006)
  - Tree kernels for hierarchical structure
    (Collins and Duffy, 2001)

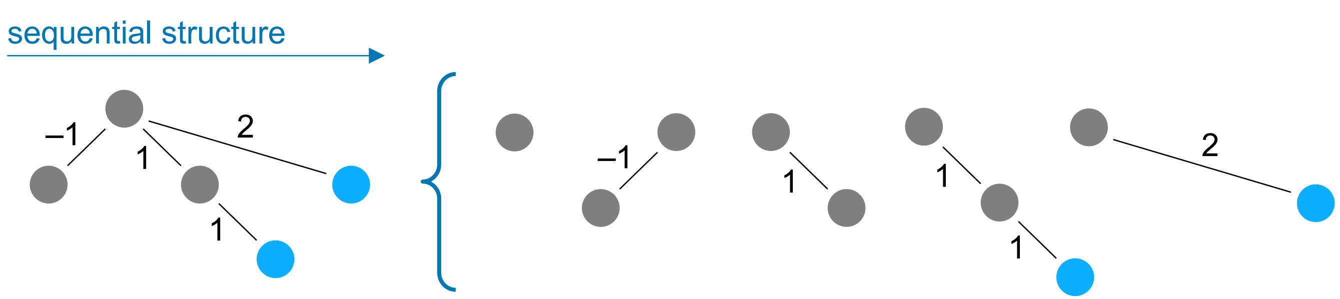


Background: Route kernels

- **Route kernel** (Aiolli et al., 2009)
  - Captures both sequential and hierarchical structure
  - Tree kernel with edge labels, indicating node positions relative to siblings
  - Models all paths starting from the root of a tree

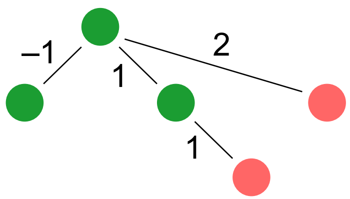


- **Adapted route kernel for arguments**
  - Positions are relative to parent node.
  - A polynomial kernel ”combines“ paths to capture full overall structure.

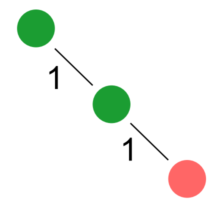


Route kernels for stance and bias: Approach

- **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 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:



1 for identical paths, 0 otherwise

Node label path from root to *v*

Edge label path from root to *v*

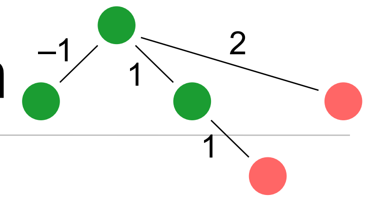
Degree of polynomial (*d* = 2 best in experiments)

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

Sum over all pairs of paths of the two trees

Normalization over maximum possible score

Route kernels for stance and bias: Evaluation



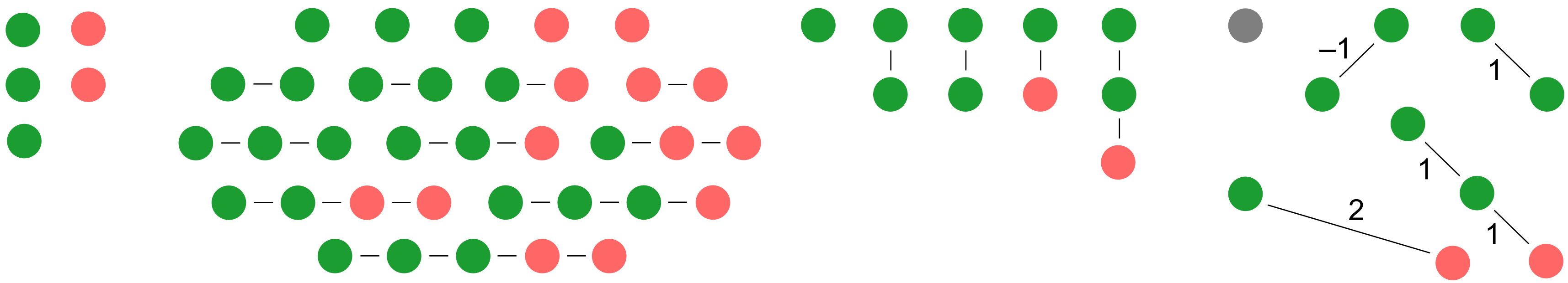
- **Overall structure approaches**

frequencies
linear kernel

sequences
subsequence kernel

hierarchies
tree path kernel

routes
adapted route kernel



- **Baselines**

majority
always majority class

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

pos
linear kernel

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

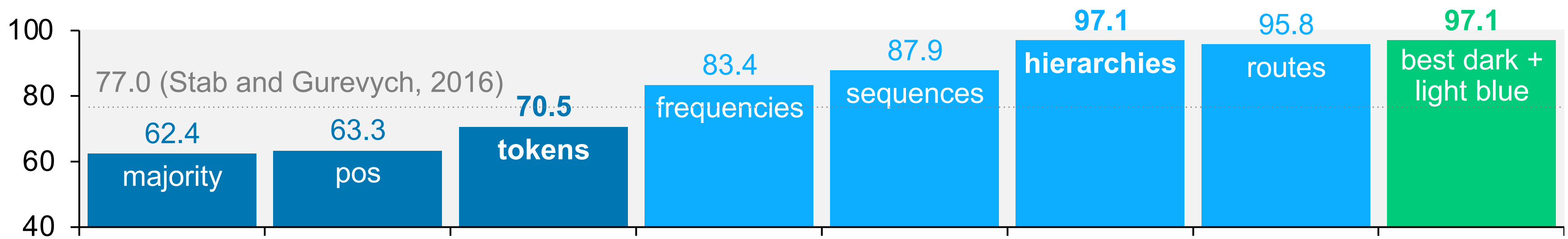
tokens
linear kernel

token
1-, 2-, and 3-grams

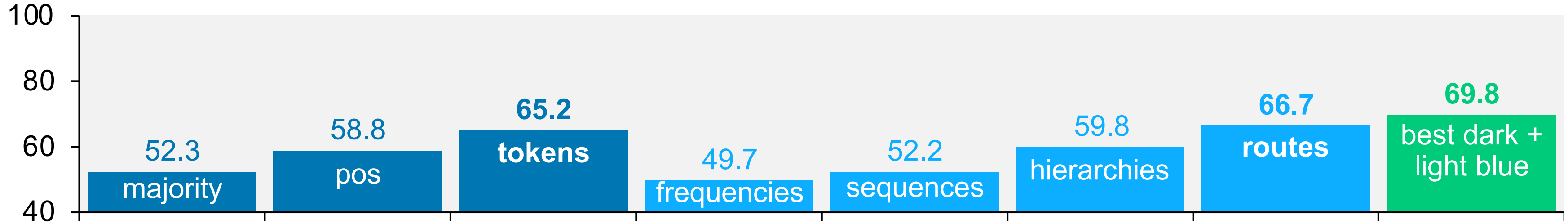
- **Experiments on ground-truth argument corpora**
  - SVM for each kernel evaluated in repeated 10-fold cross-validation
  - Hyperparameters of SVM tuned on training set with balanced class weights

Route kernels for stance and bias: Results

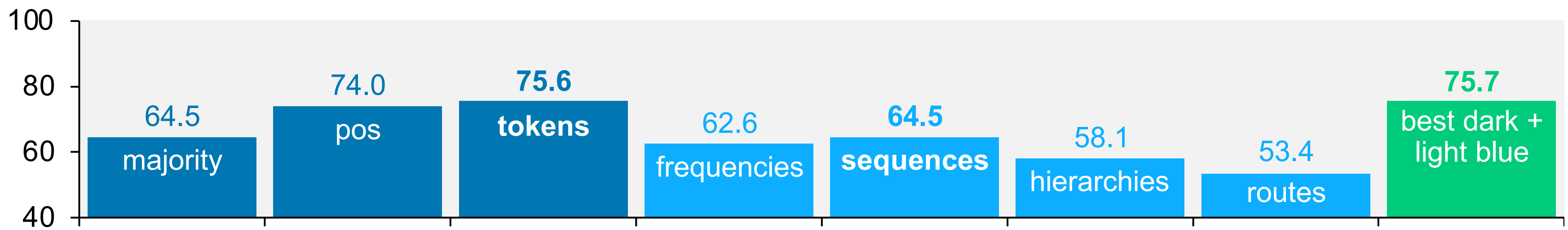
- **Myside bias accuracy** on AAE-v2



- **Stance accuracy** on Arg-Microtexts



- **Genre accuracy** on Web Discourse



Stance and myside 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 mining

V. Argument assessment

VI. Argument generation

VII. Applications of computational argumentation

VIII. 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

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

argument from consequences

Conclusion

support

Premise

support

" 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."

Premise

fallacy?

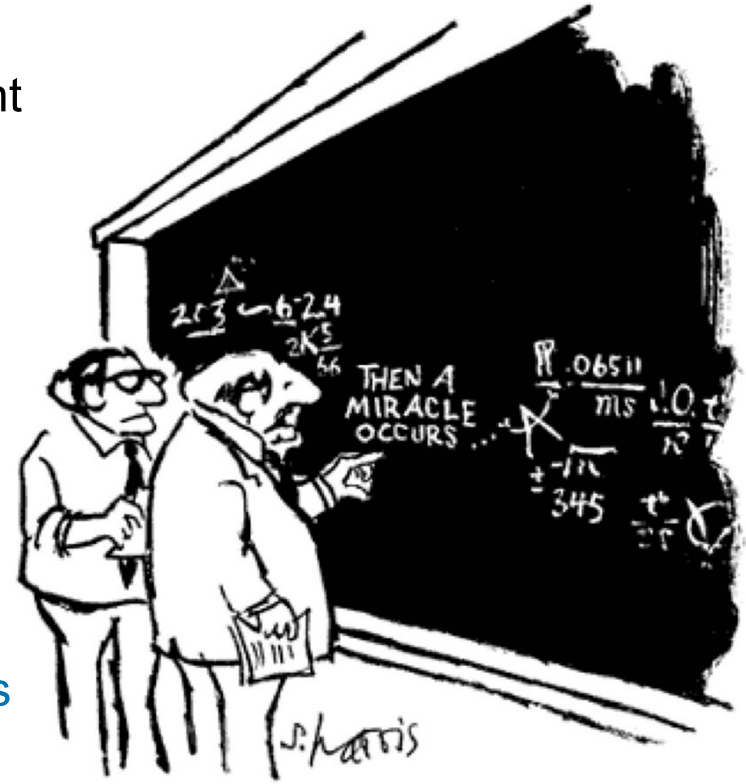
- **Fallacy detection**
  - The identification of arguments being a fallacy of a type from a set of types
  - Input. An argument, possibly with annotated structure
  - Output. Whether or not the argument is a fallacy of a certain type

Overview of scheme classification and fallacy detection

- **Schemes vs. fallacies**
  - Describe how the reasoning in an argument works or is flawed, respectively

- **Modeling the tasks**
  - Both are multiclass classification tasks.
  - Existing approaches realize them as a one-vs.-all or one-vs.-one task.

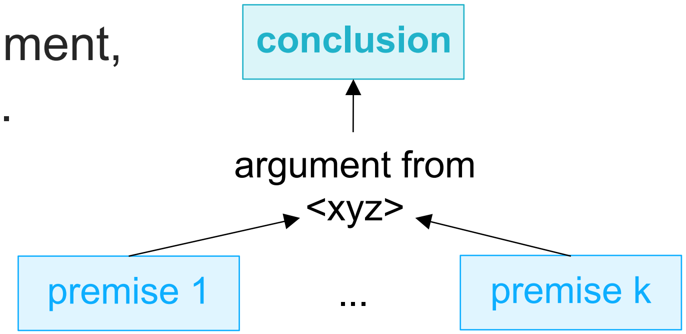
- **Selected approaches**
  - Scheme classification with tailored features (Feng and Hirst, 2011; Lawrence and Reed, 2016)
  - Ad-hominem fallacy detection using CNNs and BiLSTMs with self-attention (Habernal et al., 2018)
  - Logical fallacy detection using natural language inference (Jin et al., 2022)
  - Fallacy detection using transformers on text and audio (Goffredo et al., 2022)



"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 a scheme from a given set of schemes.



- **Research question**
  - How visible is the scheme of an argument in its text and its structure?
- **Data**
  - 658 mixed argumentative texts, annotated for 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 poperty 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.*

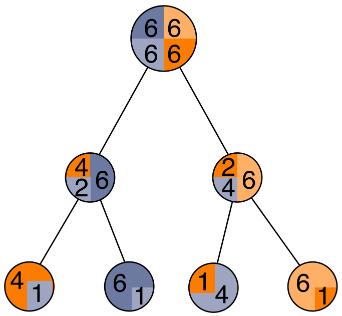
- **Practical reasoning**

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: Approach



- **Approach**
  - C4.5 decision tree for supervised classification
  - Feature engineering for all five argumentation schemes
- **Features for 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 for argument from consequence**
  - Sentiment. Positive and negative words
- **Features for argument from verbal classification**
  - Word similarity between central words in premise and conclusion
- **Features for the other three schemes**
  - Cue phrases, such as “for example”, “result”, “want”
  - Indicative patterns, such as causal WordNet relations

Classifying schemes with tailored features: Results

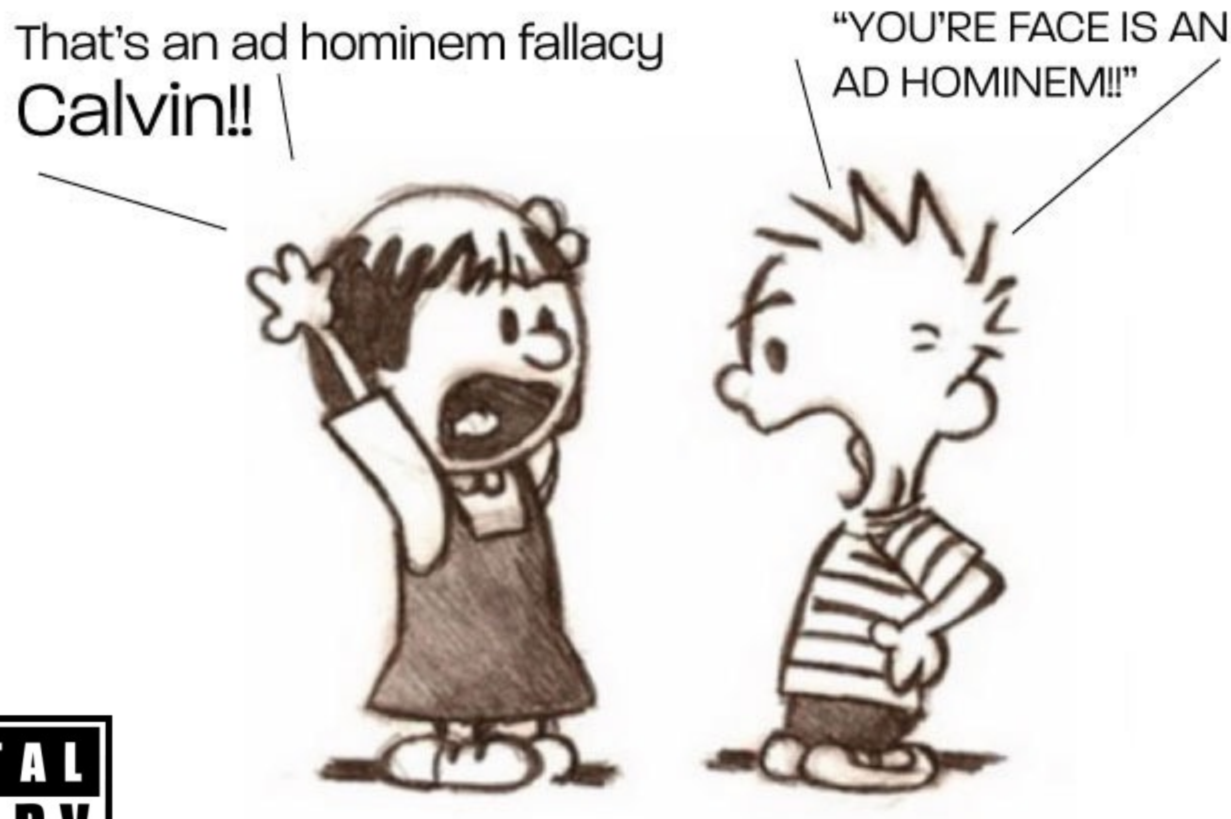
- **10-fold cross-validation**
  - One-vs.-all. 50% target scheme, 50% all others (once for all schemes)
  - One-vs.-one. 50% scheme A, 50% scheme B (once for all scheme pairs)

- **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)



PARENTAL ADVISORY EXPLICIT CONTENT

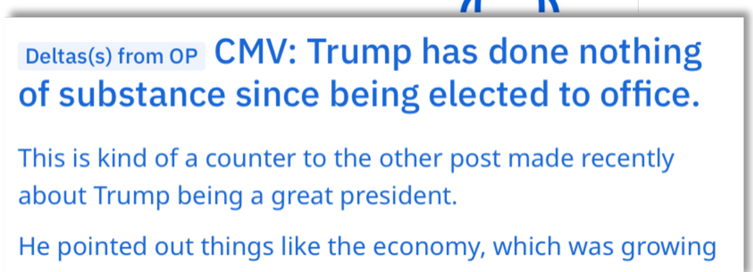
Ad-hominem arguments on the web: Task and data

- **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.



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"

"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"

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

"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?"

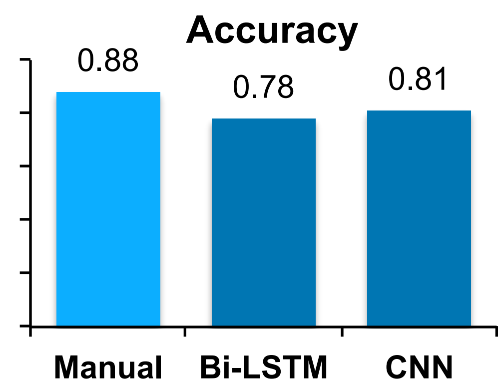
"You are just a liar."

"Reading comprehension is your friend"

"You're using troll tactics"

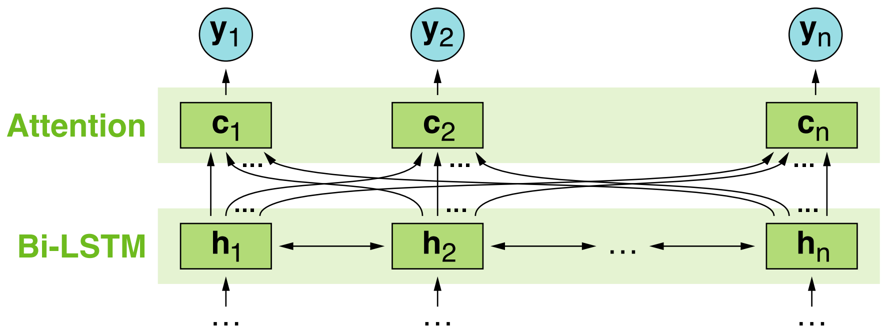
"Again, how old are you?"

- **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)

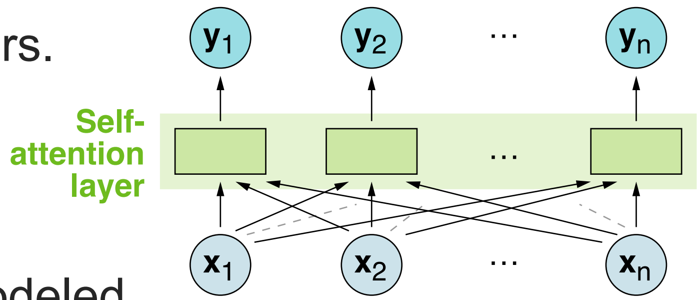


Background: Self-attention in neural networks

- **Attention** (recap)
  - Learn which inputs are relevant to which outputs.
  - Condition context **c** on *all* outputs of layer **h**: $\mathbf{c} := f(\mathbf{h}_1, \dots, \mathbf{h}_n)$.



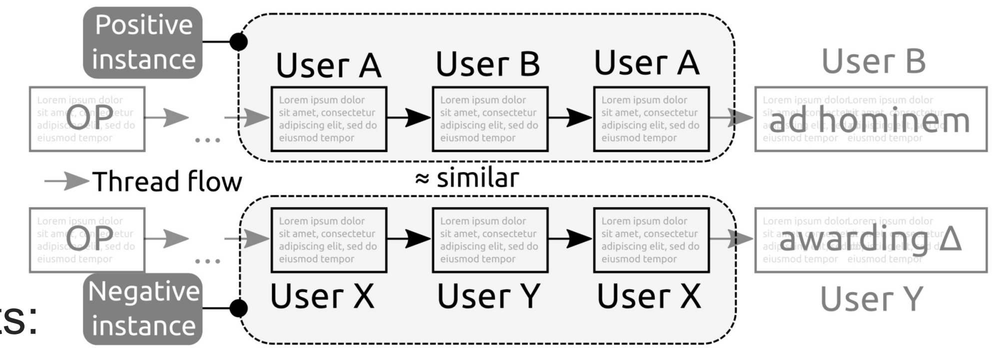
- **Self-attention**
  - Learn which inputs are relevant to which others.
  - Condition representation of input $\mathbf{x}_i$ on all inputs: $\mathbf{h}_i := g(\mathbf{x}_1, \dots, \mathbf{x}_n)$.
    For generation, only on previous inputs: $\mathbf{h}_i := g(\mathbf{x}_1, \dots, \mathbf{x}_i)$.
  - This way, any input interdepencies can be modeled.
    Transformer-based language models entirely rely on self-attention (see lecture part VI).



- **Deriving input weights from self-attention**
  - By summing up all weights an input $\mathbf{x}_i$ gets when computing output $\mathbf{y}$, its importance for $\mathbf{y}$ can be approximated.

Ad-hominem arguments on the web: Triggers

- **Prediction of ad-hominem**
  - Self-attentive LSTM trained on 2852 argument 3-tuples
  - Accuracy. 0.72
  - Visualizing self-attention weights:



(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 high weights for prediction**
  - 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...”

missing evidence
”unsupported claims!”

bad argumentation
”You're grasping at straws”

...

Discussion: Scheme and fallacy detection

- **Effective classification**
  - Some schemes are reflected in words, others require deeper understanding.
  - Many schemes have never been approached so far.
  - Ad-hominem shows linguistic patterns, but this does not hold for all fallacies.

- **Few computational approaches**
  - While extensively studied in theory, computational research on schemes and fallacies is still limited.
  - 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

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 such boats. Nothing justifies to endanger the life of innocent people."*

argument cogent?

effective in persuading?

reasonably argued?

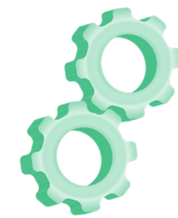
- **Observations**
  - Goal orientation. What is important depends on the goal of argumentation.
  - Granularity. Quality may be addressed at different levels of text granularity.
  - Dimensions. Several dimensions of quality may be considered.

Argumentation quality: Theory and 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.



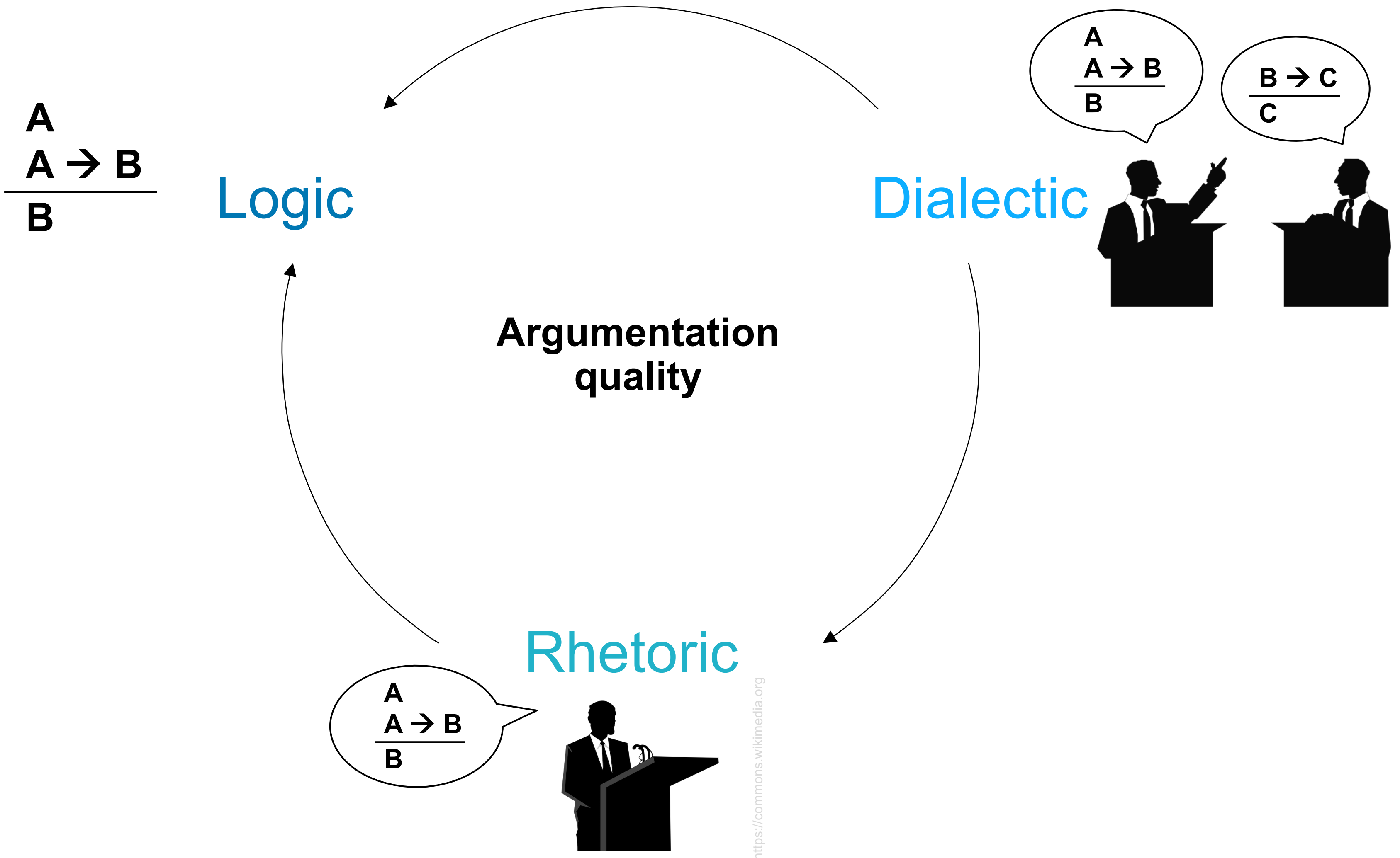
https://de.wikipedia.org

> *" 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)

- **Arising questions**
  - Should we align quality with how we *should* argue or with how we *do* argue?
  - Is this actually so different?

Three main quality aspects (recap)



Quality dimensions

focus on theory | **focus on** accepted | **prefer** general | **unify** names

validity
soundness
premise acceptability
intersubjective acceptability
argument acceptability

level of support
local/probative relevance
fallaciousness
global/dialectical relevance
argument relevance

amount of evidence
Logic
cogency
reason-ableness
Dialectic
prominence

premise sufficiency
amount of rebuttal
dialectical sufficiency
sufficiency
satisfac-toriness

Argumentation quality

well-formedness
strength
convincingness

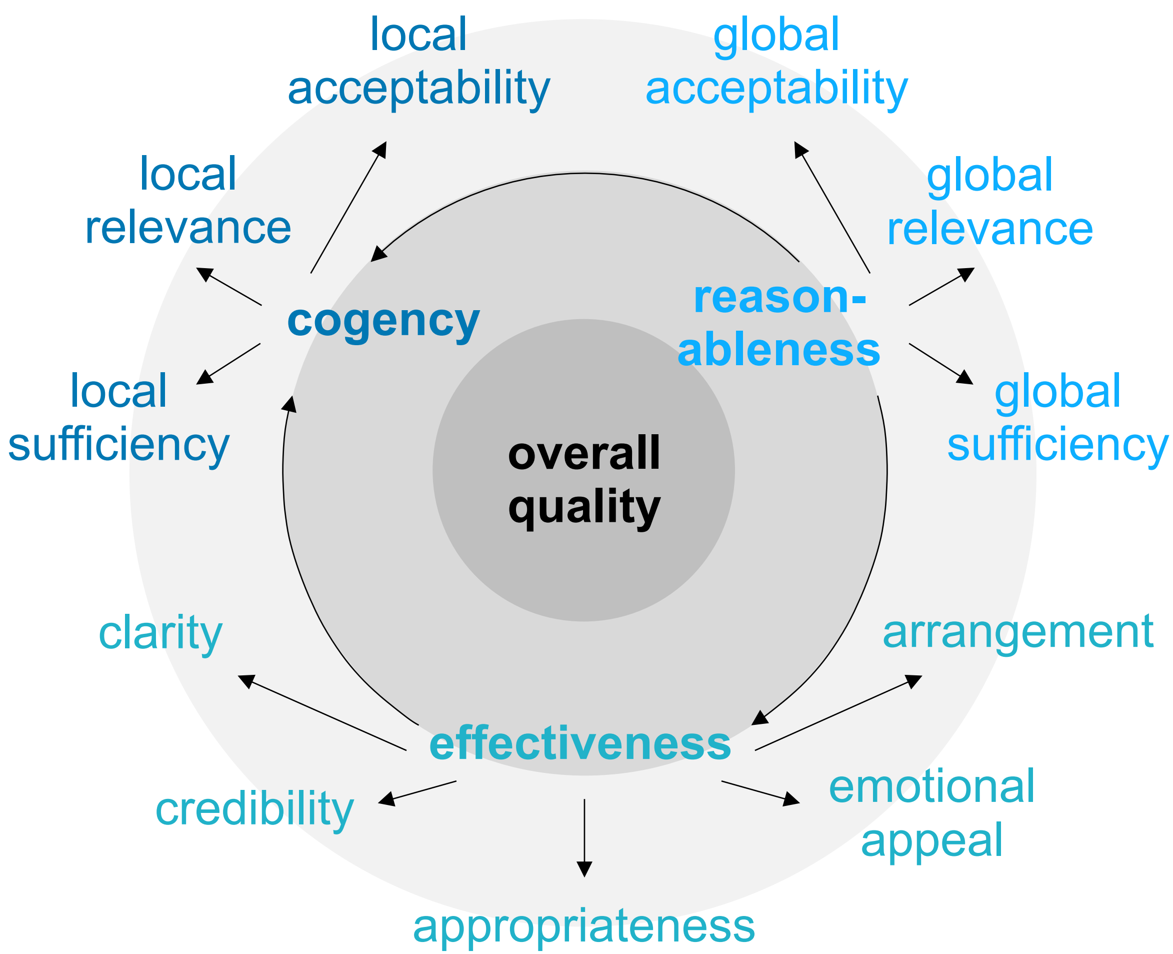
global coherence
clarity of style
argument strength
winning side
arrangement

thesis clarity
effectiveness
organization

prompt adherence
credibility
Rhetoric
emotional appeal

evaluability
appropriateness of style
persuasiveness

A taxonomy of argumentation quality (Wachsmuth et al., 2017b)



Quality dimensions in the taxonomy

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

Logic

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

Rhetoric

- **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

Dialectic

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

Next section: Absolute and relative quality assessment

Argument quality assessment

- **Argument(ation) 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)

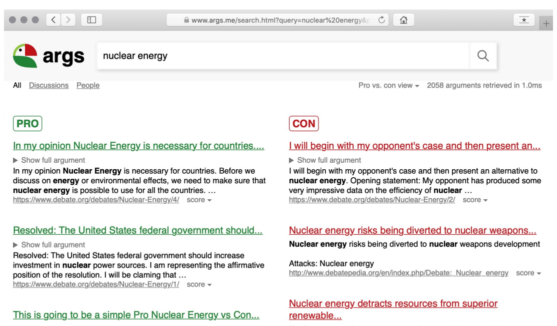
- **How to assess quality?**
  - Input. Argumentative text, metadata (e.g., author), external knowledge, ...
  - Techniques. Supervised classification/regression, graph-based analyses, ...

Importance of quality assessment

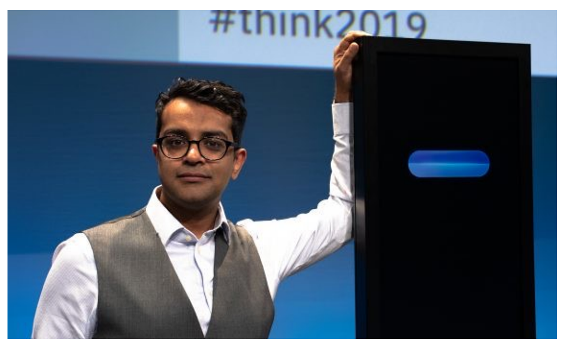
- **Why assessing quality?**
  - Mining arguments and understanding the reasoning is not enough in practice.
  - For successful argumentation, we need to choose the "best" arguments.
  - Critical for any application of computational argumentation

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

(Stede and Schneider, 2018)

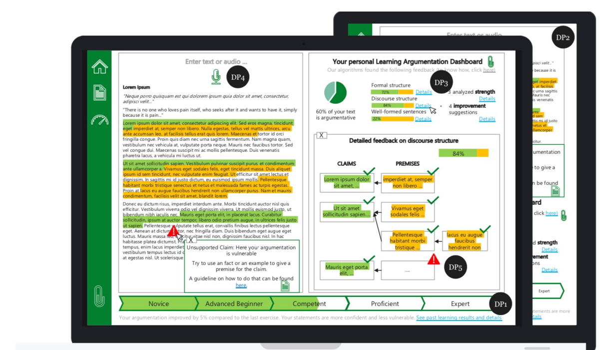


Argument search



https://flickr.com

Debating technology



(Wambsganss and Rietsche, 2019)

Writing assistance

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 such boats. Nothing justifies to endanger the life of innocent people."

4 / 5

better than

"It's the main job of the EU to save people's lives, no matter whether they belong here."

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

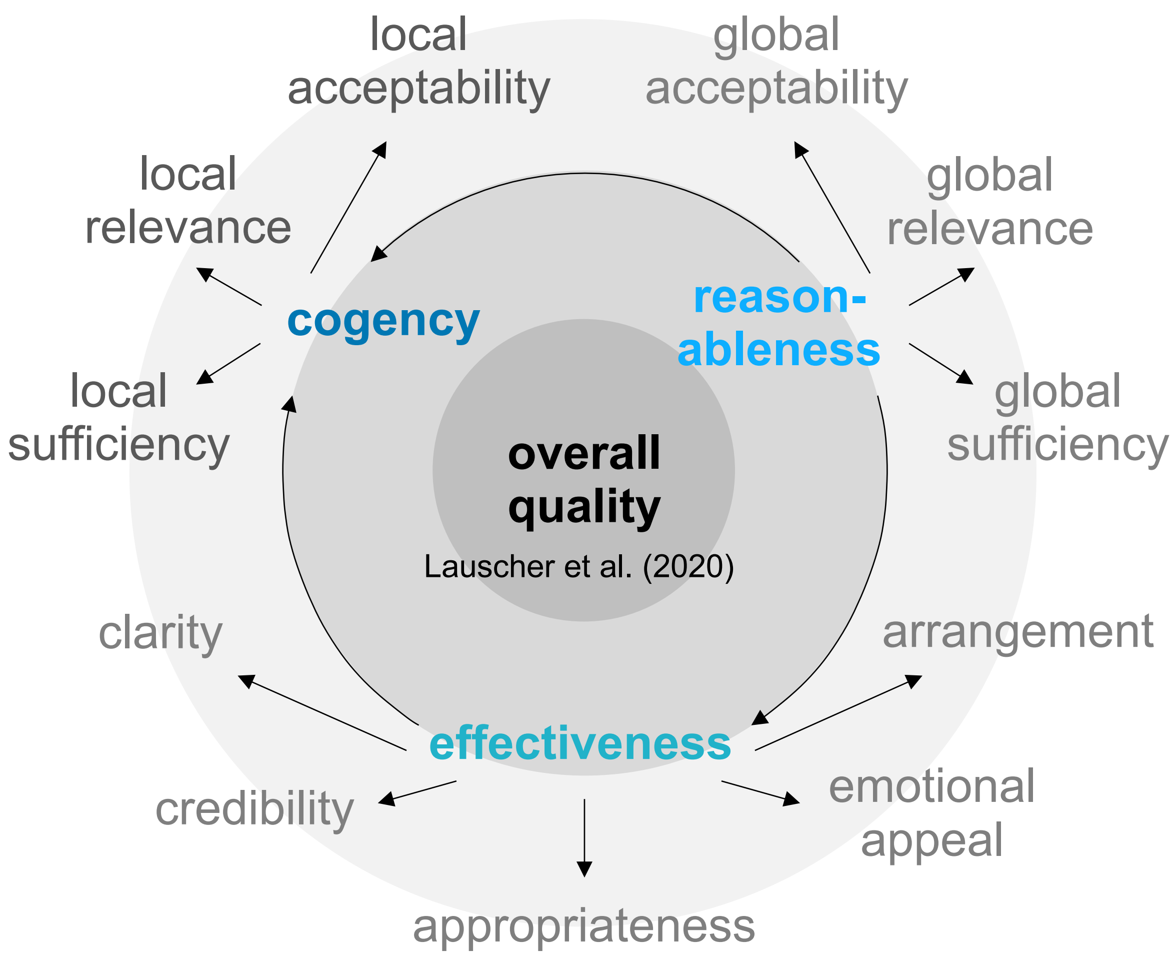
- **Main idea**
  - See quality assessment as a standard classification or regression task.
  - Learn what feature or metadata speaks for quality.

Conclusion
Premises

4 / 5

- **Selected approaches**
  - Rating quality based on argument mining using SVR (Wachsmuth et al., 2016)
  - Rating quality using multitask learning (Lauscher et al., 2020)
  - Sufficiency classification using transformer-based generation (Gurcke et al., 2021)
  - Classifying 14 appropriateness flaws with transformers (Ziegenbein et al., 2023)
  - Learning interactions of quality dimensions with adapters (Falk and Lapesa, 2023)
  - Rating all taxonomy dimensions by prompting LLMs (Mirzakhmedova et al., 2024)

Absolute quality rating: Dimensions covered here

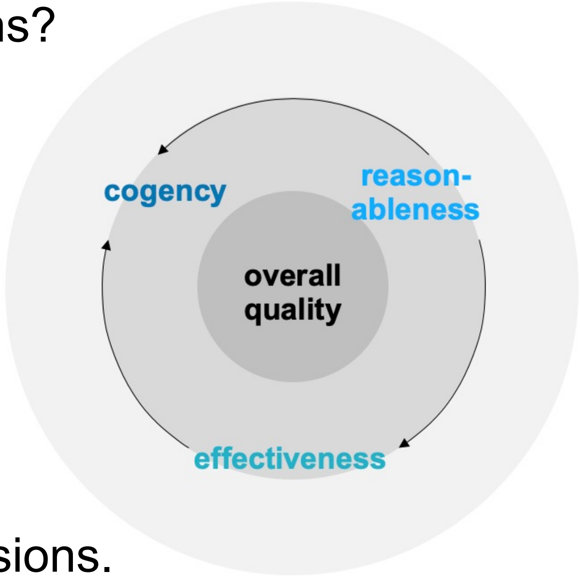


Rating quality using multitask learning

- **Task**
  - Given an argument, rate all four main taxonomy dimensions from 1 to 5
  - Dimensions. Cogency, effectiveness, reasonableness, overall quality

- **Research questions**
  - How to exploit the interdependencies between the dimensions?
  - How well can they be assessed in different domains?

- **Hypothesis**
  - Modeling all four dimensions jointly enables a more effective quality assessment.



- **Presented approach** (Lauscher et al., 2020)
  - Encode input argument with a BERT transformer.
  - Use *multitask learning* to jointly rate all four dimensions.
  - Also explore potential impact of subdimensions on parent dimension.

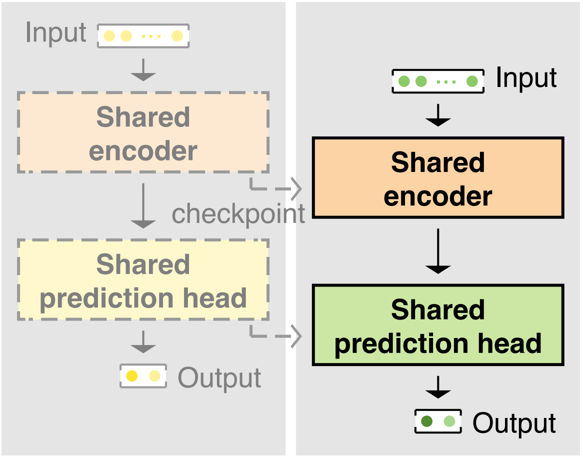
Background: Multitask learning

- **Multitask learning**
  - Many tasks are not independent from each other.
  - By tackling them jointly, all can be solved better.
  - Also, reduces need for training data per task
  - Techniques. Continual learning, joint learning, …
- **Continual multitask learning**
  - One shared model across all tasks
  - First train on auxiliary tasks, then on primary task.
  - Useful when tasks similar and one is in focus

    Example: Toxicity detection auxiliary task for social bias detection
- **Joint multitask learning**
  - Encoding shared, but different prediction heads
  - Train all tasks simultaneously or round-robin.
  - Useful when tasks are somehow related

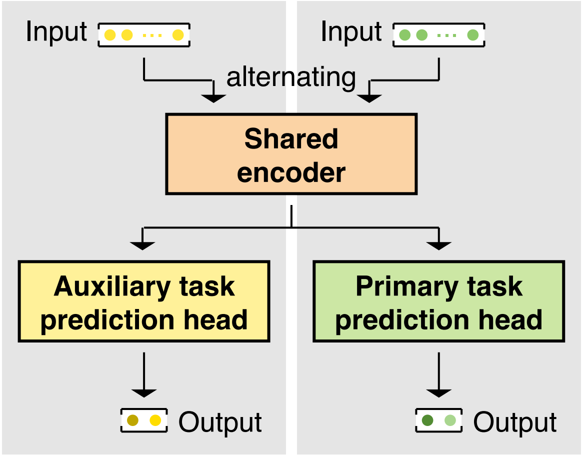
    This is the technique used in the following.

Continual multitask learning



Training #1 only | Training #2 + Inference

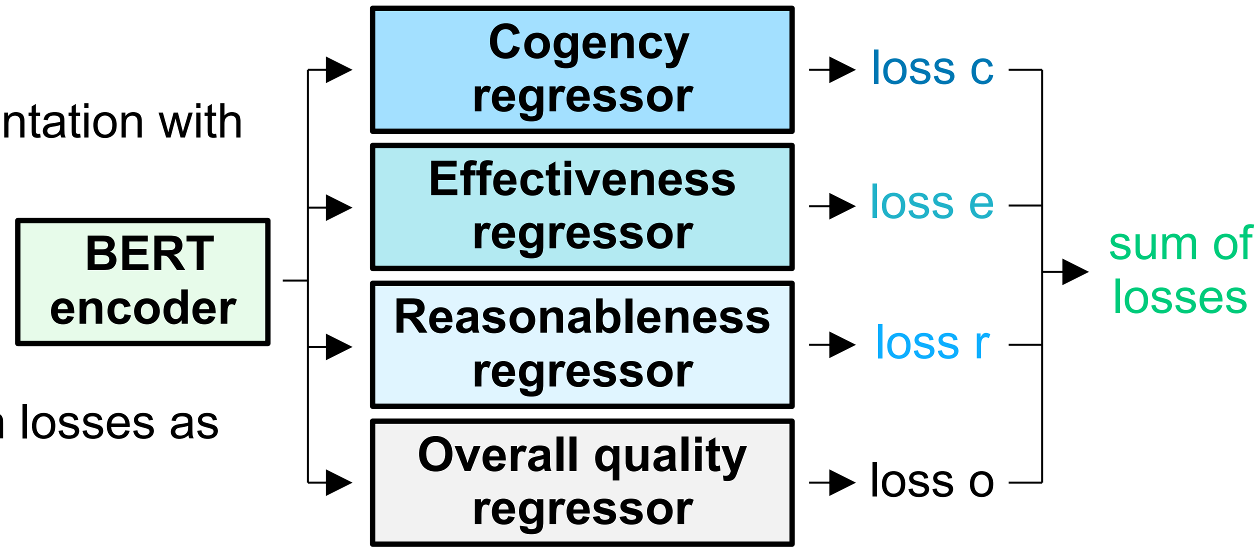
Joint multitask learning



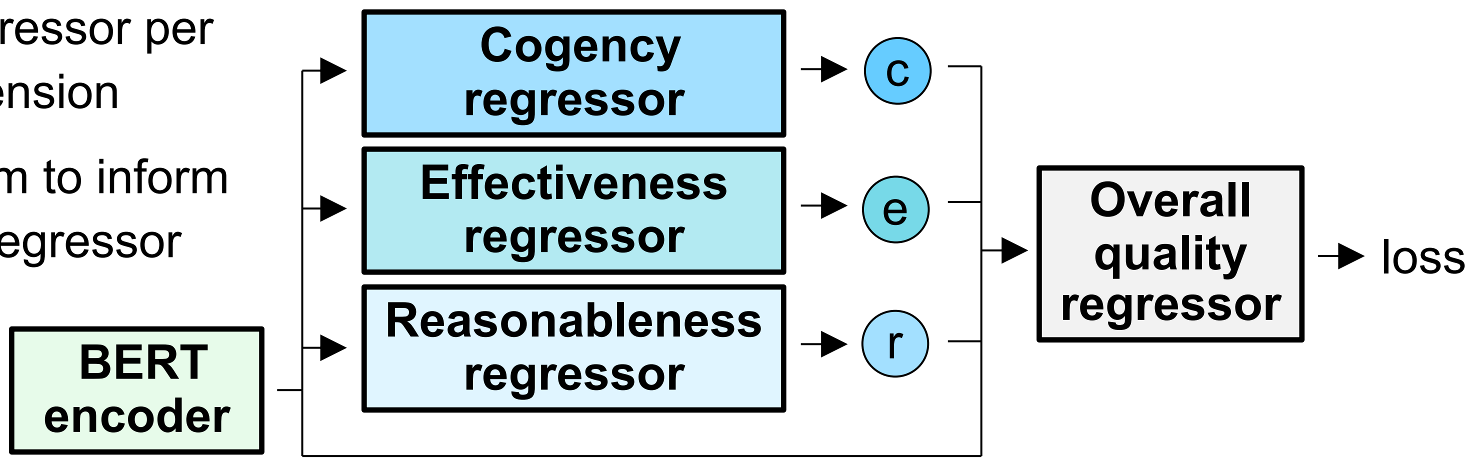
Training only | Training + Inference

Rating quality using multitask learning: Approaches

- **Flat multitask BERT**
  - Argument representation with shared encoder
  - One regressor per dimension
  - Sum of regression losses as overall loss



- **Hierarchical multitask BERT**
  - One regressor per subdimension
  - Use them to inform overall regressor



Rating quality using multitask learning : Evaluation

- **Baselines**
  - SVR tf-idf. Support vector regression with TF-IDF features
  - SVR arg. Support vector regression with argument features (Wachsmuth et al., 2016)
  - BERT single-task. BERT regressors trained individually on each dimension

    All manually annotated for the four quality dimensions

- **Data**

CQA. 2088 arguments from Yahoo! Answers

https://commons.wikimedia.org

Debate. 2103 arguments from debate portals

https://en.wikipedia.org

Review. 1104 restaurant reviews from Yelp

https://en.wikipedia.org

- **Experiments**
  - Here, focus on quality assessment with training and test in single domains
  - In the paper, also joint-domain and out-of-domain assessment

Rating quality using multitask learning: Results

- **Results** (pearson correlation)

Approach	Cogency CQA	Cogency Debate	Cogency Review	Effectiveness CQA	Effectiveness Debate	Effectiveness Review
SVR tf-idf	.444	.257	.384	.411	.120	.340
SVR arg (Wachsmuth et al., 2016)	.503	.429	.464	.523	.450	.432
BERT single-task	.587	.503	.554	.612	.542	**.555**
BERT flat multitask	.633	**.541**	**.561**	**.671**	**.570**	.514
BERT hierarchical multitask	**.638**	.474	.541	.670	.532	.486

Approach	Reasonableness CQA	Reasonableness Debate	Reasonableness Review	Overall quality CQA	Overall quality Debate	Overall quality Review
SVR tf-idf	.457	.247	.452	.389	.265	.450
SVR arg (Wachsmuth et al., 2016)	.476	.399	.432	.492	.432	.533
BERT single-task	**.665**	.418	.609	.652	.511	**.605**
BERT flat multitask	.664	**.473**	.610	**.667**	**.537**	.588
BERT hierarchical multitask	.626	.408	**.611**	.661	.494	.593

- BERT Flat multitask is best on average, but with some deviations

Relative quality comparison: Overview

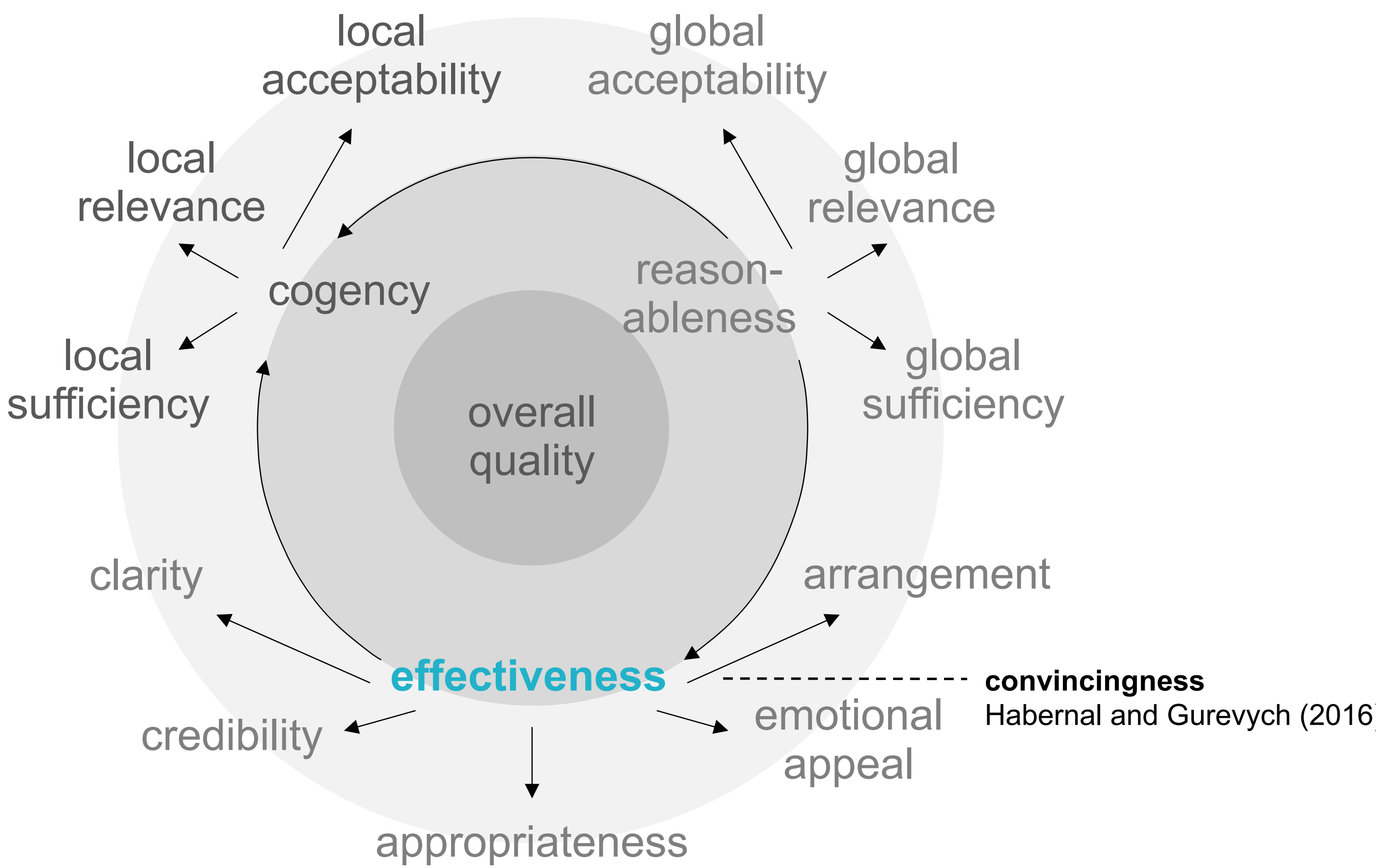
- **Problem**
  - Rating the quality of an argument in isolation may be hard or even doubtful.
  - Is there an easier or more realistic way to assess quality?

- **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.
  - Downside: Unclear whether the best argument is good

$$\frac{\text{Conclusion}}{\text{Premises}} \quad \textbf{vs} \quad \frac{\text{Conclusion}}{\text{Premises}}$$

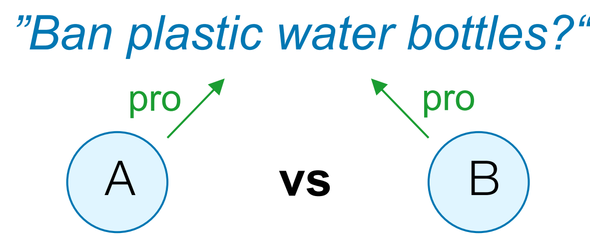
- **Selected approaches**
  - Debate winner prediction from the flow of talking points (Zhang et al., 2016)
  - Convincingness comparison with SVM and Bi-LSTM (Habernal and Gurevych, 2016)
  - Selection of the preferable argument with BERT (Toledo-Ronen et al., 2019)
  - Ranking claim revisions by clarity with SBERT+SVMRank (Skitalinskaya et al., 2021)

Relative quality comparison: Dimension covered here



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

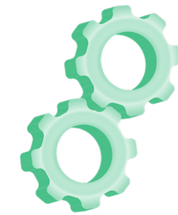
Absolute vs. relative assessment ~ Theory vs. practice

- **Data representing theory**
  (Wachsmuth et al., 2017b)
  - Absolute expert ratings
  - Normative guidelines
  - 15 predefined quality dimensions

https://commons.wikimedia.org

local acceptability
global acceptability
local relevance
cogency
reason-ableness
global relevance
local sufficiency
overall quality
global sufficiency
clarity
arrangement
effectiveness
credibility
emotional appeal
appropriateness

- **Data representing practice**
  (Habernal and Gurevych, 2016)
  - Relative lay comparisons
  - No guidelines
  - 17+1 resulting reason labels



https://de.wikipedia.org

off-topic
no credible evidence
only opinion
language/grammar issues
attacking/abusive
unclear/hard to follow
insufficient reasoning
irrelevant reasons
makes you think
convincing
close to topic
crisp / well-written
non-sense/confusing
credible / confident
generally weak/vague
well thought through
objective/two-sided
details/facts/examples

- **Empirical comparison of theory and practice** (Wachsmuth et al., 2017c)
  - 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)

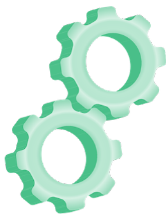
  - Top **overall quality** for well thought through (mean score 1.8 of 3)
  - Lowest **overall quality** for off-topic (mean score 1.1 of 3)

  - 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 mining

V. Argument assessment

VI. Argument generation

VII. Applications of computational argumentation

VIII. 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

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.

- **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."*

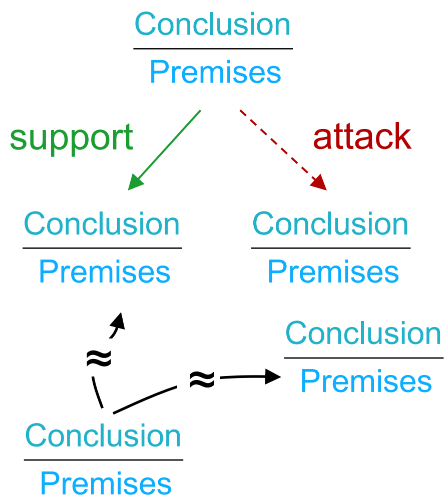
> *" The death penalty doesn't deter people from committing serious violent crimes. The thing that deters is the likelihood of being caught and punished."*

- **Two ways to approach this problem**
  - 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

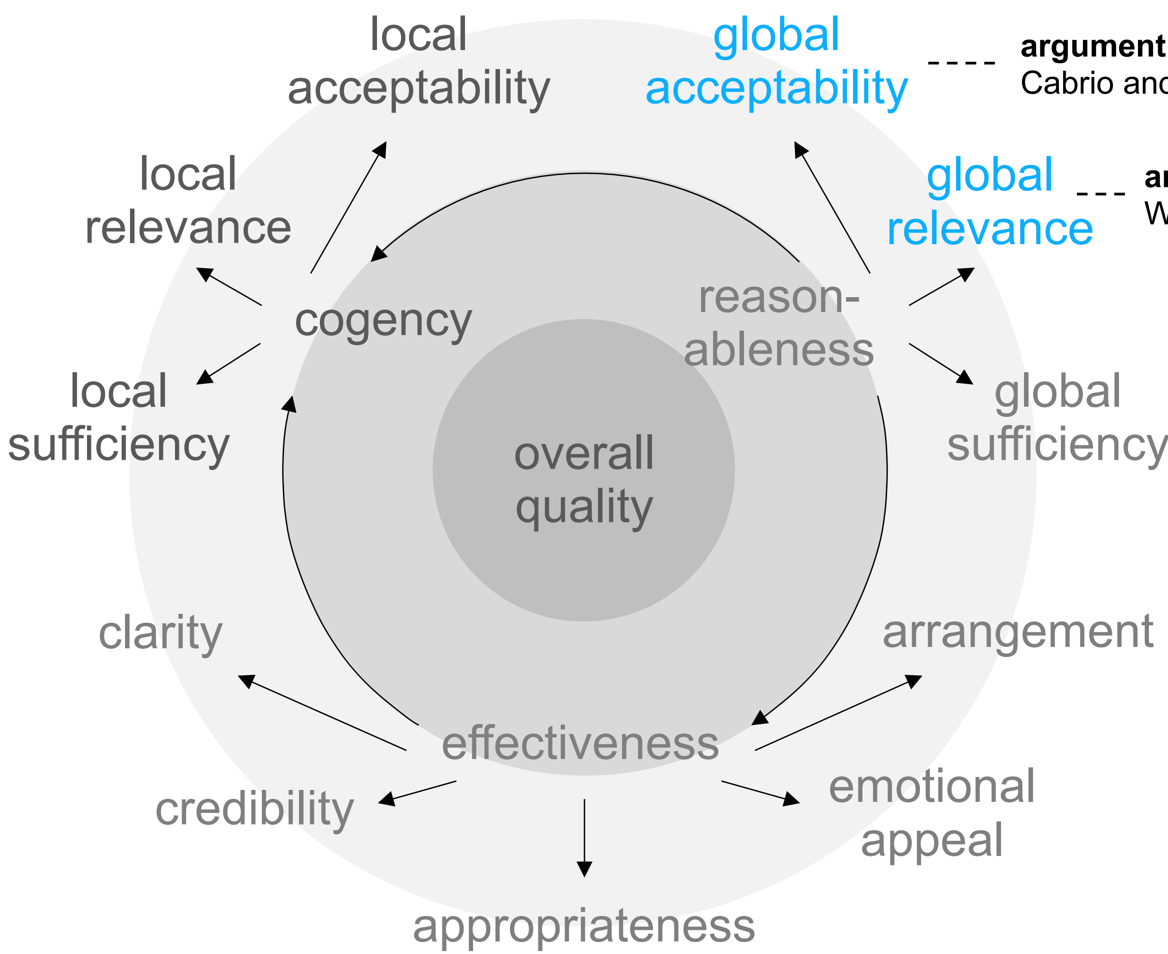
- **Problem**
  - How to assess quality without learning from subjective annotations?
  - What are objective quality indicators?

- **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 way out is to rely on majority assessments of many annotators.



- **Selected approaches**
  - Assessment of acceptability based on attack relations (Cabrio and Villata, 2012)
  - Assessment of prominence based on argument frequency (Boltužic and Šnajder, 2015)
  - Assessment of relevance based on reuse of units (Wachsmuth et al., 2017a)

Objective quality assessment: Dimensions covered here



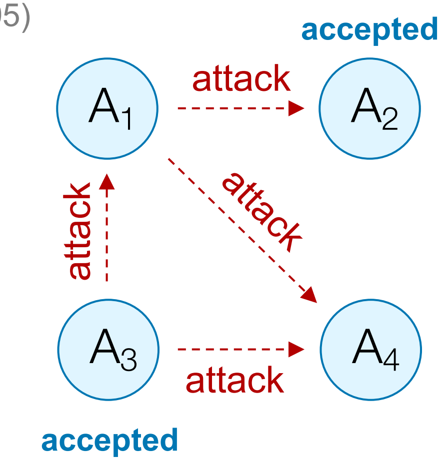
argument acceptability
Cabrio and Villata (2012)

argument relevance
Wachsmuth et al. (2017a)

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 a debate portal, 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

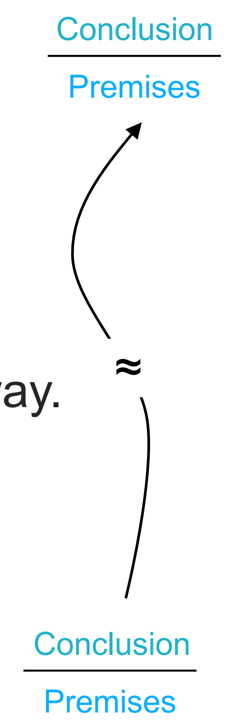
> *" The death penalty legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."*

> *" 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?

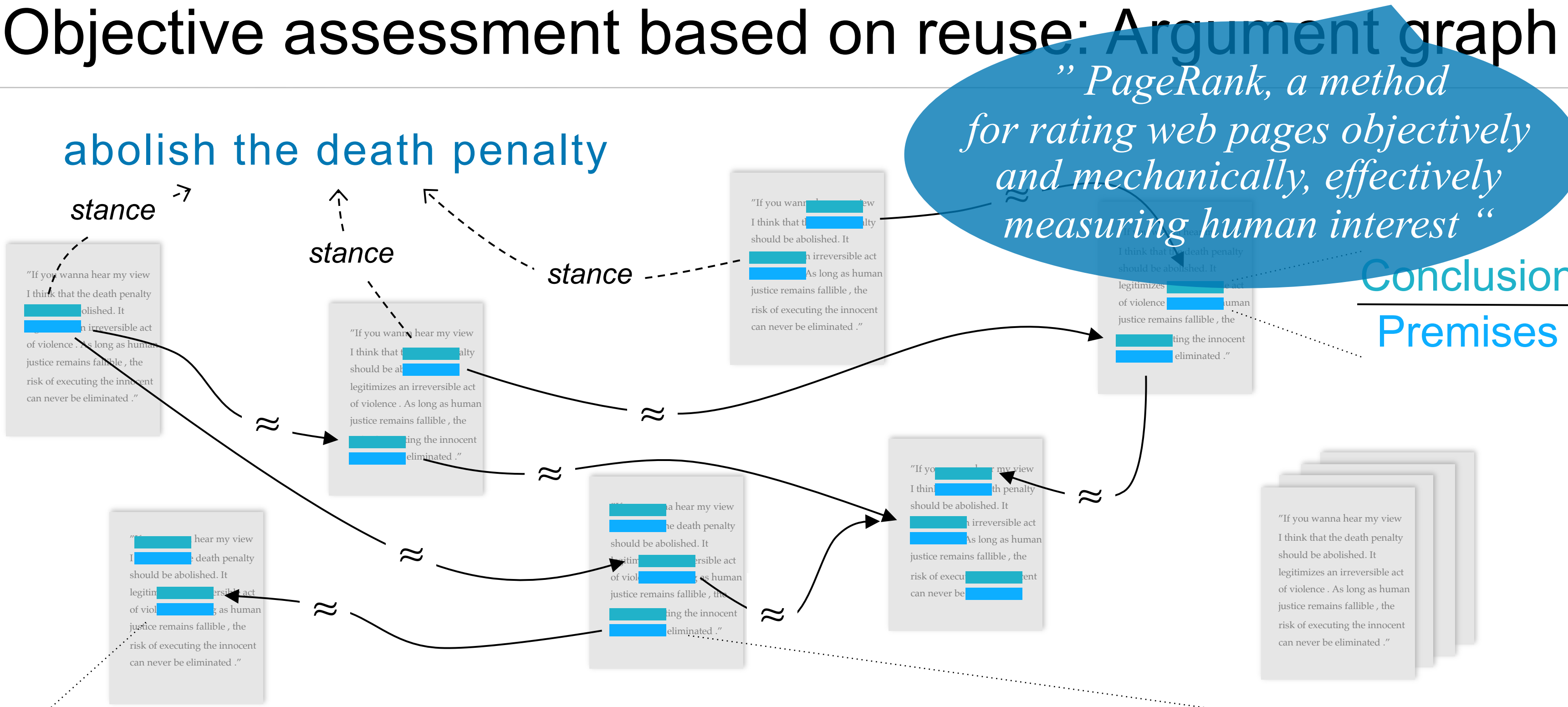
- **Key hypothesis**
  - The relevance of a conclusion depends on what other arguments across the web use it as a premise.
  - Rationale. Author cannot control who "cites" a conclusion in this way.

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



Page et al. (1999)

Objective assessment based on reuse: Argument graph



The death penalty doesn‘t deter people from committing serious violent crimes.

A survey of the UN on the relation between the death penalty and homicide rates gave no support to the deterrent hypothesis.

≈

The death penalty should be abolished.

It does not deter people from committing serious violent crimes.

Even if it did, is it acceptable to pay for predicted future crimes of others?

Objective assessment based on reuse: Approach

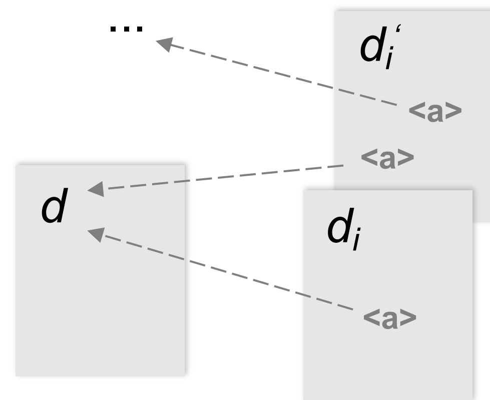
- **Original PageRank score** of a web page *d* (Page et al., 1999)

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

\# pages d_i links to



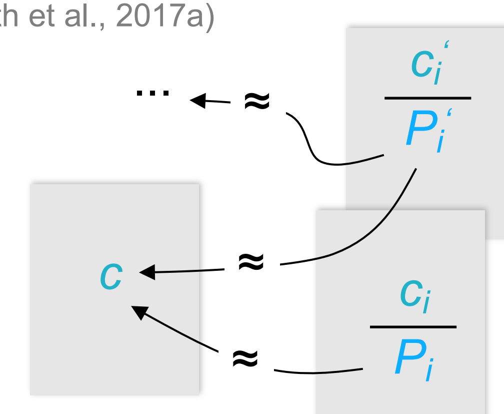
- **Adapted PageRank score** of an argument unit *c* (Wachsmuth et al., 2017a)

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

\# premises of c_i



- **Argument relevance** derived from aggregated premise scores
  - Minimum, average, maximum, or sum

Objective assessment based on reuse: Results

- **Evaluation of unsupervised ranking approaches**

PageRank of premises	Frequency of premises	Similarity of units	Sentiment of premises	Number of premises	Random ranking
$\hat{p}$	$\sum$	$c \sim P$		$\lvert P \rvert$	

each for minimum, average, maximum, and sum aggregation

- **Experiment on graph with 18k arguments**
  - 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.”

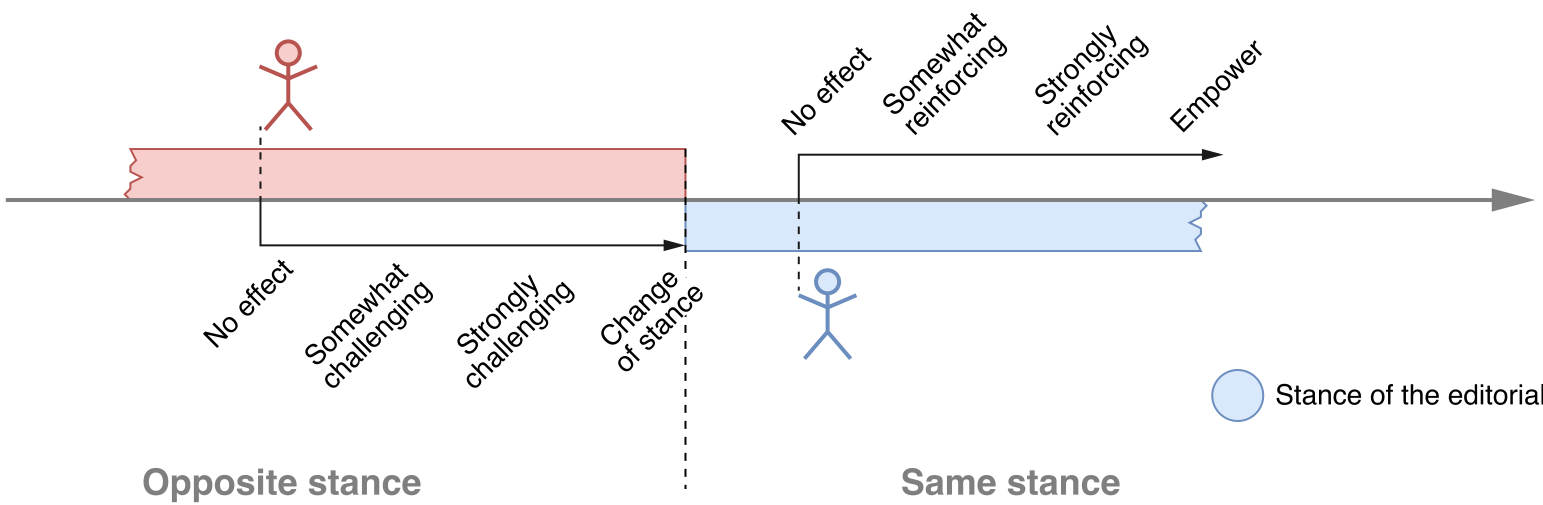
Inclusion of Subjectivity: Overview

- **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.
- **Studies**
  - Personalities. Effectiveness of emotional vs. rational arguments (Lukin et al., 2017)
  - Ideologies/Personalities. Challenging and reinforcing arguments (El Baff et al., 2018)
- **Selected approaches**
  - Debate winner prediction using logistic regression on debater and voter characteristics (Durmus and Cardie, 2018; Al-Khatib et al., 2020)
  - Audience-specific effectiveness prediction using style features (El Baff et al., 2020)

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

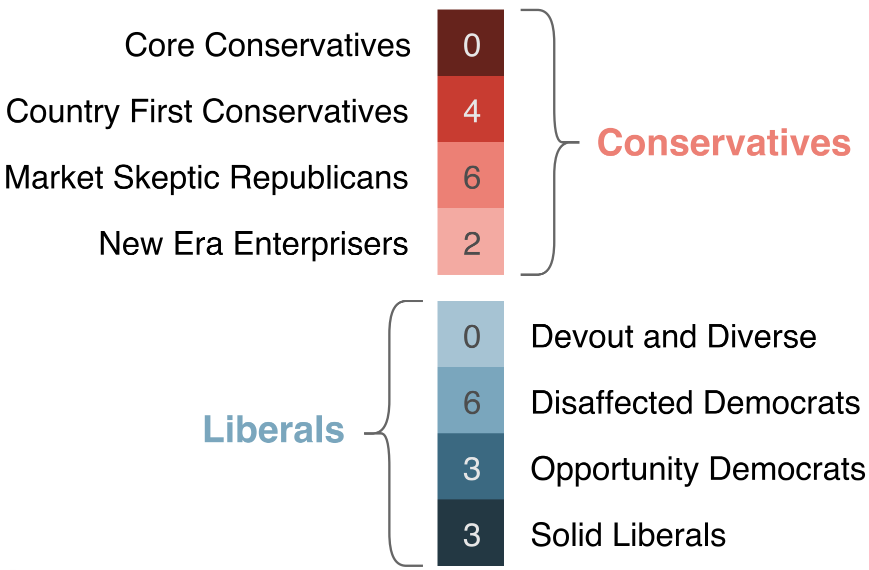
- **Effects of news editorials**
  - News editorials are said to shape public opinion. However:
  - They rarely *change* readers‘ prior stance; rather, they challenge or reinforce it

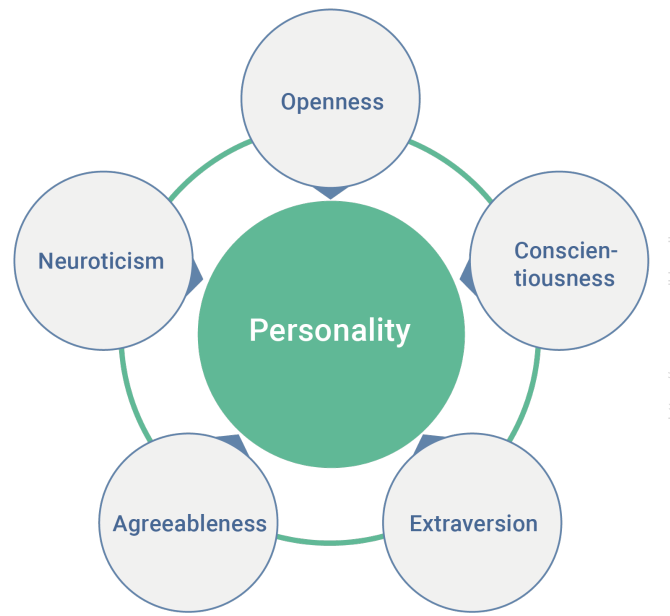


- **Dialectical notion of argumentation quality**
  - A good editorial reinforces one side and challenges the other.
  - Or it challenges both sides.
  - Hypothesis. Different effect depending on political ideology and personality

Effectiveness based on target audience: Study

- **Study**
  - Impact of ideology and personality on the effectiveness of news editorials
  - Ideology. Conservative vs. liberal (as measured by Political Typology Quiz)
  - Personality. Five dimensions (as measured by Big Five Test)

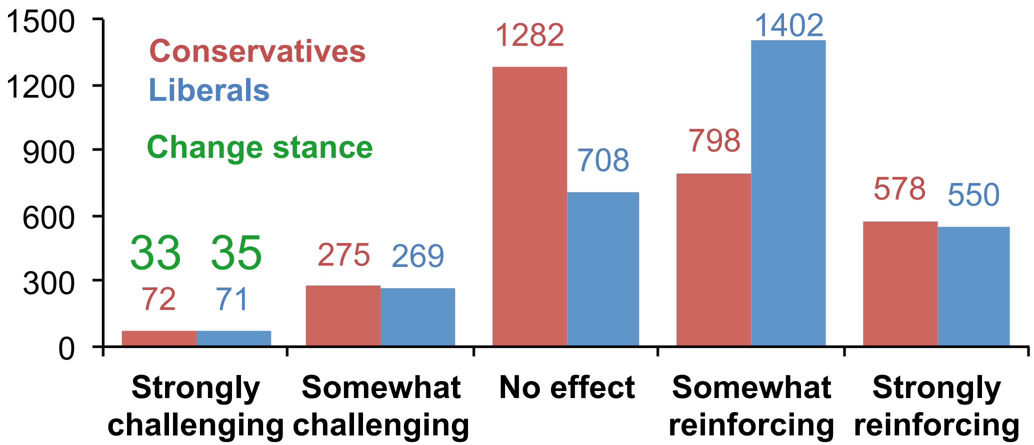


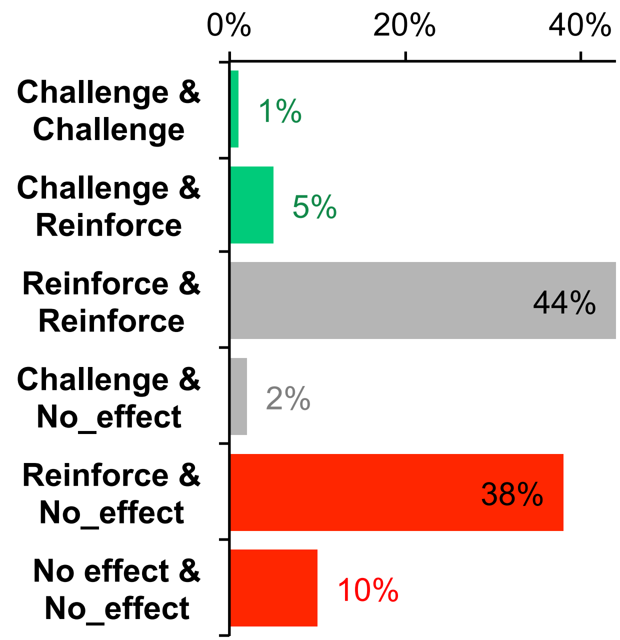


- **Data**
  - 1000 editorials from the New York Times
  - Persuasive effect, annotated by 3 conservatives and 3 liberals (24 in total)

Effectiveness based on target audience: Results

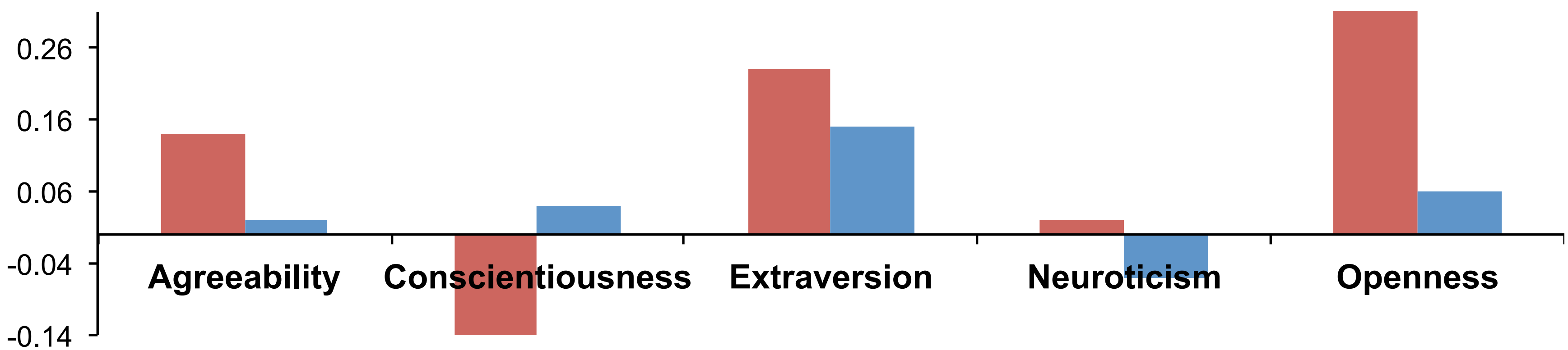
- **Majority effect distribution in the corpus**





- **Effect depending on ideology and personality**

Kendall's τ correlation with challenge/reinforce



Outlook: LLMs for quality assessment? (May 3, 2024)

You

Look at this argument on rescue boats in the Mediterranian Sea:

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 such boats, but nothing justifies to endanger the life of innocent people.

Do you think it is rhetorically effective in persuading a conservative person?

ChatGPT

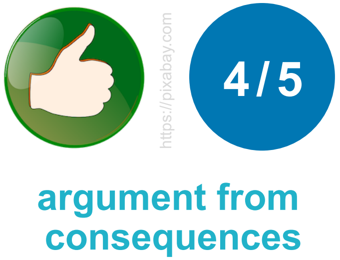
[…] However, whether this argument would be rhetorically effective in persuading a conservative person depends on various factors, including the specific beliefs and priorities of the individual in question. Some conservative individuals may indeed be swayed by appeals to humanitarian values and the sanctity of life, especially if they have a strong moral compass or religious convictions that emphasize compassion and charity.

On the other hand, some conservative individuals might prioritize concerns such as national security, border control, or the rule of law over humanitarian considerations. They might view the presence of rescue boats as facilitating illegal migration or incentivizing dangerous journeys, potentially leading to more deaths in the long run. In such cases, appeals to humanitarianism alone may not be sufficient to sway their opinion. […]

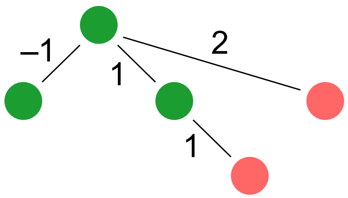
Next section: Conclusion

Conclusion

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



- **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 may be done.
  - Subjectiveness may be included or somehow circumvented.

References

- **Aiolli et al. (2009).** Fabio Aiolli, 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.
- **Al Khatib et al. (2020).** Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7067–7072, 2020.
- **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.
- **Barrow et al. (2021).** Joe Barrow, Rajiv Jain, Nedim Lipka, Franck Dernoncourt, Vlad Morariu, Varun Manjunatha, Douglas Oard, Philip Resnik, and Henning Wachsmuth. Syntopical Graphs for Computational Argumentation Tasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1583–1595, 2021.
- **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.
- **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

- **Collins and Duffy (2001).** Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. In Advances in Neural Information Processing Systems 14, pages 625–632.
- **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.
- **Durmus and Cardie (2018).** Esin Durmus and Claire Cardie. Exploring the Role of Prior Beliefs for Argument Persuasion. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1035–1045, 2018
- **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, pages 3154–3160, 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.
- **Freeman (2011).** Argument Structure: Representation and Theory. Springer, 2011.
- **Goffredo et al. (2022).** Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious Argument Classification in Political Debates. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, pages 4143-4149, 2022.
- **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.

References

- **Habernal and Gurevych (2015)*.*** Exploiting Debate Portals for Semi-supervised Argumentation Mining in User-generated Web Discourse. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2127– 2137, 2015.
- **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.
- **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.
- **Jin et al. (2022).** Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical Fallacy Detection. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7180–7198, 2022.

References

- **Lauscher et al. (2020).** Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, Logic, and Dialectic: Advancing Theory-based Argument Quality Assessment in Natural Language Processing. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- **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.
- **Mirzakhmedova et al. (2024).** Nailia Mirzakhmedova and Marcel Gohsen and Chia Hao Chang and Benno Stein. Are Large Language Models Reliable Argument Quality Annotators?. In arXiv 2404.09696, 2024.
- **Mooney and Bunescu (2006).** Raymond J. Mooney and Razvan C. Bunescu. 2006. Subsequence kernels for relation extraction. In *Advances in Neural Information Processing Systems 18*, pages 171–178. MIT Press.
- **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.
- **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

- **Skitalinskaya et al. (2021).** Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1718–1729, 2021.
- **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.
- **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.
- **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.
- **Toledo-Ronen et al. (2019).** Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Automatic Argument Quality Assessment - New Datasets and Methods. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5625–5635, 2019.

References

- **Toulmin (1958).** Stephen E. Toulmin. The Uses of Argument. Cambridge University Press, 1958.
- **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. (2017c).** 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. (2017d).** 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 et al. (2008).** Douglas Walton, Christopher Reed, and Fabrizio Macagno. Argumentation Schemes. Cambridge University Press, 2008.

References

- **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.