Argumentation Technology for Artificial Intelligence
Part 2: Argument Mining and Assessment

Philipp Cimiano, Benno Stein, and Henning Wachsmuth

September 24th, 2019
Argumentation in natural language

- **Real-world arguments**
  - Mostly not logically valid
  - Leave much implicit
  - May be hidden in longer utterances
  - May be split over multiple utterances
  - May depend on the context

> "If you wanna hear my view I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."

- **Can we actually use formal approaches?**
  - Long story short: To a wide extent, yes
  - But we need to *mine* arguments and *assess* their properties before

*Alice.* I think a university degree is important. Employers always look at what degree you have first.

*Bob.* LOL ... everyone knows that practical experience is what does the trick.

*Alice.* Good point! Anyway, in doubt I would always prefer to have one!
Argument mining and assessment: Outline

1. Argument mining
   - How to find argument units and relations in text?

2. Stance classification
   - How to determine whether an argument is pro or con?

3. Overall argumentation analysis
   - How to analyze longer argumentative structures?

4. Argumentation quality assessment
   - How to judge whether an argument is good or bad?

5. Fallacy detection
   - How to identify argumentative flaws in discussions?
1. Argument Mining
Argument mining: Introduction

- **Argument mining** (aka argumentation mining)
  - Automatic identification of arguments in natural language text
  - Core task for natural language argumentation

- **Three main argument mining steps**
  - Segmenting a text into argument units and other parts
  - Classifying the type or role of each unit
  - Identifying and classifying relations between units

"If you wanna hear my view I think that **the death penalty should be abolished**. It legitimizes an irreversible act of violence. As long as human justice remains **fallible**, the risk of executing the innocent can never be eliminated."
Unit segmentation

• **Argument units** (aka argumentative discourse units)
  – Text segments with an argumentative function
    Usually, the premises and conclusions of arguments

• **Unit segmentation**
  – **Task.** Given a text, segment it into argument units and other parts
  – **Method.** Usually, token-level sequence labeling (more on this below)

  ”If you wanna hear my view I think that the death penalty should be abolished.”

  O O O O O O O O O O B I I I I I I O

• **Challenges**
  – Unit granularity differs: Anything between clauses and paragraphs
  – Usually the first step: Unclear what are the arguments

• **State of the art** (Ajjour et al., 2017)
  – Rather reliable on narrow genres ($F_1$ 0.72–0.82), unsolved across them
Background: Evaluation measures

• **Data-driven research**
  - Models and methods developed on training texts
  - Most methods not fully "correct"
  - Effectiveness evaluated on test texts

• **Effectiveness measures**
  - **Accuracy.** Used if both positives and negatives important
    \[
    \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
    \]
  - **Precision, recall, and }F_1\text{-score.** Used if positives in the focus
    \[
    \text{Precision (P)} = \frac{TP}{TP + FP} \quad \text{Recall (R)} = \frac{TP}{TP + FN} \quad \text{F}_1\text{-score} = \frac{2 \cdot P \cdot R}{P + R}
    \]
Unit classification

- **Unit classes**
  - Claim and evidence types (Rinott et al., 2015; Al-Khatib et al., 2017)
  - Roles within argumentation (Stab and Gurevych, 2014; Habernal and Gurevych 2015)
  - Often corpus-specific

- **Unit classification**
  - **Task.** Given an argument unit, assign one class from a set of classes
  - **Method.** Usually, supervised text classification (more on this below)

- **State of the art**
  - Reliable on "explicit" argumentation, such as essays ($F_1 0.87$) (Stab, 2017)
  - Still rather reliable on news editorials ($F_1 0.77$) (Al-Khatib et al., 2017)
  - Minority classes may be problematic, though
Relation identification and classification

• **Argumentative relations**
  - Within arguments (premises to conclusion) or between them (arg to arg)
  - **Types.** Support or attack, partly more fine-grained

• **Relation identification and classification**
  - **Task.** Given two units/arguments, what relation holds between them, if any
  - **Method.** Various, e.g., computing the minimum spanning tree (Peldszus and Stede, 2015)

• **State of the art**
  - Semi-reliable on narrow genres, such as essays ($F_1$ 0.73) (Stab, 2017)
  - Identification works better than classification
  - Relations hard to agree on for "hidden" arguments, such as in editorials

• **Related tasks**
  - Given an argument, classify its argumentation scheme (Feng and Hirst, 2011)
  - Given an argument, find the best counterargument (Wachsmuth et al., 2018)
2. Stance Classification
Stance classification: Introduction

- **Stance**
  - Overall position of a person towards some target, such as an issue or statement
  - To have/take a stance on a target means to be *pro* or *con* towards it

(Somasundaran and Wiebe, 2010)

**Pro** towards claim above

"If you wanna hear my view I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."

**Con** towards "death penalty"

- **Stance classification**
  - Determination of the stance towards a target encoded in a text span
  - In argumentation. Conceptual overlap with relation classification, but usually stance refers to the issue at discussion
Background: Supervised text classification

- **Text classification**
  - *Task*. Given a text, assign one class from a set of classes
    Stance classification is a text classification problem
  - Usually done with supervised machine learning

- **Feature-based classification**
  - Map text to feature vector, map feature vector to class label
    Features engineered manually or semi-automatically
  - *Models*. Support vector machines, random forest, ... 

- **Neural classification** (usually works better, given enough data)
  - Features (weights in neural networks) learned automatically
  - *Models*. Convolutional neural networks, bi-directional LSTMs, ... 

- **Sequence labeling** (applicable, when a sequence of texts is classified)
  - Like other techniques, but considering previous classifications
  - *Models*. Conditional random fields, recurrent neural networks, ...
How to develop a stance classification algorithm

Texts with annotations

natural language processing

Text corpus

"If you wanna hear my view I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."

Machine learning

Feature vector for each text

average sentiment

Feature vector for each text position in text

Stance classifier

"If you wanna hear my view I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."

Here, the "classical" approach feature-based classification

Stance classification

Average sentiment here, the "classical" approach feature-based classification

Argumentation Technology for Artificial Intelligence – KI 2019 Tutorial
Modeling stance

• **Candidate features of the text** (Somasundaran and Wiebe, 2010, Hasan and Ng, 2013)
  - **Bag-of-words.** Distribution of words or word n-grams
  - **Core vocab.** Terms from subjectivity lexica
  - **Discourse.** Connectives + relations between units
  - **Sentiment.** Aspect or topic-directed polarity
    ... and many more...

  $\rightarrow$ accuracy up to 0.70

• **Candidate features of the context**
  - Exploit author knowledge in dialog $\rightarrow$ up to 0.74
    (Ranade et al., 2013)
  - Exploit opposing views in dialog $\rightarrow$ up to 0.75
    (Hasan and Ng, 2013)
  - Connections between topics of claim and target
    (Bar-Haim et al., 2017)

  $\rightarrow$ 0.84 for most confident 10%, 0.65 overall (3 classes)
3. Overall Argumentation Analysis
(Wachsmuth et al., 2017c)

Giovanni Da San Martino
Dora Kiesel
Benno Stein
Overall argumentation analysis: Introduction

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)

• Hypothesis
  - Overall structure is decisive for downstream analysis tasks
Analysis tasks and corpora

- **Myside bias on AAE corpus** (Stab and Gurevych, 2016)
  - 402 persuasive student essays
  - 15.1 units/text, proprietary argument model
  - 251 one-sided, 151 two-sided

- **Stance on Arg-Microtexts** (Peldszus and Stede, 2016)
  - 112 short argumentative texts
  - 5.1 units/text, model of Freeman (2011)
  - 46 pro stance, 42 con stance, 24 unlabeled

- **Genre on Web Discourse** (Habernal and Gurevych, 2015)
  - 340 argumentative web texts
  - 3.4 units/text, modified model of Toulmin (1958)
  - 216 comments, 46 blog posts, 73 forum posts, 5 articles
Study: The impact of overall argumentation

• Research questions
  1. How to jointly model sequential and hierarchical overall structure?
  2. How important is overall structure in analysis tasks?

• Background: Kernel methods in machine learning
  - Representation of instances in implicit feature space
  - Similarity function used by classifier (e.g., SVM)
  - Strong when good features unknown and/or data limited

• Kernels for structured data
  - Subsequence kernel for sequential structure
    (Mooney and Bunescu, 2006)
  - Tree kernel for hierarchical structure
    (Collins and Duffy, 2001)
  - Route kernel: Tree kernel with positions
    (Aiolli et al., 2009)
Experiments for each analysis task

- Overall argumentation approaches

- Baseline approaches

- Experiments on ground-truth argument corpora
  - SVM for each kernel in repeated 10-fold cross-validation
  - Hyperparameter tuning, fairness in training
Accuracy results

• Myside bias on AAE-v2

77.0 (Stab and Gurevych, 2016)

- 70.5 tokens
- 83.4 frequencies
- 87.9 sequences
- 97.1 hierarchies
- 95.8 routes

- best blue + best red

• Stance on Arg-Microtexts

- 52.3 majority
- 58.8 pos
- 65.2 tokens
- 49.7 frequencies
- 52.2 sequences
- 59.8 hierarchies
- 66.7 routes

- best blue + best red

• Genre on Web Discourse

- 64.5 majority
- 74.0 pos
- 75.6 tokens
- 62.6 frequencies
- 64.5 sequences
- 58.1 hierarchies
- 53.4 routes

- best blue + best red
4. Argumentation quality assessment
Argumentation quality assessment: Introduction

- **Argumentation quality**
  - Natural language arguments rarely logically valid
  - Need to quantify how *strong* an argument or argumentation is.

  "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)

- **Argumentation quality assessment**
  - Absolute or relative judgment of specific quality dimensions
  - Identification of flaws and fallacies

- **Critical for any application**
  - **Argument search.** What argument to rank highest?
  - **Writing support.** What argumentative flaws does a text have?
  - **Decision making.** Which arguments outweigh others?

https://args.me
What to assess and how to assess it

• **What to assess**
  - Several, partly very subjective quality dimensions
  - Different granularity levels

  "If you wanna hear my view I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."  

  acceptability: 4 out of 5

  more acceptable than

  "Human beings never act freely and thus should not be punished for even the most horrific crimes."

• **How to assess**
  - Absolute or relative?
  - How *should* we argue vs. how *do* we argue?
  - Based on manual assessments or "objective" properties?
  - Include model of audience?
Three aspects of argumentation quality

Logic

"An argument is cogenent if its premises are relevant to its conclusion, individually acceptable, and together sufficient to draw the conclusion."

Blair (2012)

Dialectic

"A dialectical discussion derives its reasonableness from a dual criterion: problem validity and intersubjective validity."

van Eemeren (2015)

Rhetoric

"In making a speech, one must study three points: the means of producing persuasion, the style or language to be used, and the proper arrangement of the various parts."

Aristotle (2007)
A taxonomy of argumentation quality (Wachsmuth et al., 2017b)

- Rhetoric
- Argumentation Technology
- Artificial Intelligence – KI 2019 Tutorial

- local relevance
- global relevance

- local acceptability
- global acceptability

- cogency
- reason-ability

- local sufficiency
- global sufficiency

- effectiveness
- appropriateness

- credibility
- emotional appeal

- clarity

- organization
- argument strength
- persuasiveness
- winning side
- convincingness

- level of support
- argument acceptability

- local relevance
- Cabri and Villata (2012)

- global relevance
- argument prominence
- Bolužić and Šnajder (2015)

- argument relevance
- Wachsmuth et al. (2017a)

- argument acceptability
- Cabri and Villata (2012)

- global sufficiency
- argument prominence
- Bolužić and Šnajder (2015)

- global sufficiency
- argument relevance
- Wachsmuth et al. (2017a)

- organization
- Persing et al. (2010)

- global coherence
- Rahimi et al. (2014)

- global coherence
- Boltužić and Šnajder (2015)

- amount of evidence
- Feng et al. (2014)

- amount of evidence
- Stab and Gurevych (2017)

- sufficiency
- Stab and Gurevych (2017)

- evaluability
- Braunstain et al. (2016)

- thesis clarity
- Persing and Ng (2013)

- prompt adherence
- Persing and Ng (2014)

- global coherence
- Feng et al. (2014)

- global coherence
- Braunstain et al. (2016)

- evaluability
- Persing et al. (2010)

- Rahimi et al. (2014)

- Rahimi et al. (2015)

- Rahimi et al. (2015)

- Rahimi et al. (2015)

- Persing and Ng (2015)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persing et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)

- Persig et al. (2010)
The role of participants in argumentation

Author (or speaker)
• Argumentation is connected to the person who argues
• The same argument is perceived differently depending on the author

Reader (or audience)
• Argumentation often targets a particular audience
• Different arguments and ways of arguing work for different persons

"University education must be free. That is the only way to achieve equal opportunities for everyone."

"According to the study of XYZ found online, avoiding tuition fees is beneficial in the long run, both socially and economically."
Selected assessment approaches

- **Absolute assessment** (Wachsmuth et al., 2016)
  - Regression of 4 essay quality dimensions
  - Features based on argument mining

- **Relative assessment** (Zhang et al., 2016)
  - Classification of the winner of a debate
  - Modeling own and attack of opponent’s points

- **"Objective" assessment** (Cabrio and Villata, 2012)
  - Graph analysis to determine acceptability (Dung, 1995)
  - Textual entailment to obtain attacking arguments

- **Audience-specific assessment** (El Baff et al., 2018)
  - Analysis of editorial effectiveneseness for audience
  - Model of audience’s ideology and personality
5. Fallacy detection
(Habernal et al., 2018)
Fallacy detection: Introduction

- **What is a fallacy?** (Tindale, 2007)
  - 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 his or her arguments
  - **Red herring.** Reasoning based on an unrelated issue
  - **Appeal to ignorance.** Taking lack of evidence as proof for the opposite

  *My girlfriend won’t give me a gift for my birthday. I have received no indication to the contrary from her.*

  *My flight tomorrow won’t be delayed. I have received no indication to the contrary from the airline.*

  (credit to Mario Treiber for this example)

- **Fallacies are hard to detect**
  - Structure identical to other arguments
  - Understanding and context knowledge needed
A study of ad-hominem arguments on the web

• **Ad-hominem arguments**
  - Attacking the opponent instead of his or her arguments
  - 20% of all online news comments uncivil (Coe et al., 2014)

• **Research questions**
  - Can we identify ad-hominem automatically?
  - What are triggers of ad-hominem?

• **Data**
  - 2M posts from Reddit ChangeMyView
  - 3866 posts (0.2%) contain an ad-hominem argument
    Ad-hominem is deleted by moderators, but we obtained all comments from them

• **Reddit ChangeMyView (CMV)**
  - An opinion poster (OP) states a view
  - Others argue for the opposite
  - OP gives Δ to convincing posts
Ad-hominem on CMV

"Reading comprehension is your friend"
"Ever have discussions with narcissistic idiots on the internet? They are so tiring"
"You still refuse to acknowledge that you used a strawman argument against me"
"little buddy"

"Thank you so much for all your pretentious explanations"
"To say that people intrinsically understand portion size is idiotic."
"You started with a fallacy and then deflected."
"Please don’t waste peoples time pretending to know what you’re talking about"

"boy"
"Did you even read this?"
"Read what I posted before acting like a pompous ass"
"Do you even know what you’re saying?"

"Again, how old are you?"
"You’re making the claims, it’s your job to prove it. Don’t you know how debating works?"
"You’re obviously just Nobody with enough brains to operate a computer could possibly believe something this stupid"
"Your second paragraph is fairly idiotic"

"You have no capability to understand why"
"Wow. Someone sounds like a bit of an anti-semite"
"You’re obviously just Nobody with enough brains to operate a computer could possibly believe something this stupid"
"Possible lie any harder?"

"How can you explain that? You can’t because it will hurt your feelings to face reality"
"You’re just a dishonest troll"
"you dumb fuck"
"Can you also use Google?"

"Willful ignorance is not something I can combat"
"You’re trash at debating."
"Your just an asshole"
"You’re using troll tactics"

"You’re too dishonest to actually quote the verse because you know it’s bullshit"
"sir"
"If you can’t grasp the concept, I can’t help you"
Identification of ad-hominem

- **Distribution of ad-hominem**
  - 75% threads with \( \leq 2 \) ad-hominem
    - (but some with >50)
  - 49% threads stop after ad-hominem
  - 66% ad-hominem "out of the blue"
    - (but one after 57 posts)
  - 23% ad-hominem against OP
  - 12% ad-hominem from OP

- **Types of ad-hominem**
  - Ad-hominem annotated in 400 arguments by 7 crowdworkers
  - 15 types derived manually from their annotations

- **Identification of ad-hominem**
  - **Manual.** 100 balanced arguments (50 ad-hominem) classified by 6 workers
  - **Automatic.** 7242 balanced arguments classified by 2 neural classifiers (Bi-LSTM & CNN)

Accuracy:
- Manual: 0.88
- Bi-LSTM: 0.78
- CNN: 0.81
Analysis of triggers of ad-hominem

- Prediction of ad-hominem
  - Attentive LSTM trained on 2852 argument 3-tuples
  - Accuracy 0.72
  - Manual attention analysis

(00V_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 .

(00V_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, such as ”racist“
  - Partly meta about argumentation

vulgar intensifiers "... the fuck..."
direct imperatives "You should..."
missing evidence "unsupported claims!"
bad argumentation "You’re grasping at straws"
Conclusion
Conclusion

- **Argument mining and assessment**
  - Finding arguments in natural language text
  - Classifying stance, schemes, and similar
  - Assessing quality dimensions and flaws

- **State of the art**
  - Most tasks now tackled with neural approaches
  - In narrow domains, reasonable effectiveness achievable
  - Robust "off-the-shelf" algorithms rare so far

- **Role within argumentation technology**
  - Builds on argumentation theory and computational linguistics
  - Needed to process natural language arguments
  - Converts arguments to (semi-) structured information
References


References


References


References


References


References


