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

Philipp Cimiano, Khalid Al-Khatib / Benno Stein, Henning Wachsmuth

September 21, 2020
Argumentation in natural language

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

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

- **Can we use formal approaches?**
  - Yes, but we need to *mine* arguments and *assess* their properties before
  - Natural language processing needed
Argument mining and assessment: Outline

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

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

3. Argument mining for writing support
   - How to leverage the output of argument mining?

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

5. Fallacy detection in online discussions
   - How to identify argumentative flaws and their triggers?

"Again, how old are you?"
1. 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)

• **Stance classification**
  – Determination of the stance encoded in a text span towards a target
  – Usually, the target is the (given) issue at discussion

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

  **Con** towards “banning rescue boats”
  **Pro** towards claim above

• **Notice**
  – Stance classification usually comes *after* argument mining
    Here discussed first, because of its conceptual simplicity
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

Feature vector for each text

natural language processing

Text corpus

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

Feature computation

average sentiment

position in text

here, the "traditional" approach feature-based classification

machine learning

Stance classifier

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

Argumentation Technology for Artificial Intelligence – KI 2020 Tutorial
Background: Evaluation measures

• NLP is data-driven
  – Methods developed on training texts, output not always correct
  – Effectiveness of methods 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-score.** Used if positives in the focus
    
    \[
    \text{Precision (P)} = \frac{TP}{TP + FP} \quad \text{Recall (R)} = \frac{TP}{TP + FN} \quad F_1\text{-score} = \frac{2 \cdot P \cdot R}{P + R}
    \]

  – **Mean absolute/squared error.** Often used for numerical scores
Modeling stance

- **Candidate features of the text** (Somassundaran 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
  
  → accuracy up to 0.70

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

  → 0.84 for most confident 10%, 0.65 overall (3 classes)
2. Argument Mining
Argument mining: Introduction

• **Argument mining** (aka argumentation mining)
  – Automatic identification of arguments in natural language text

• **Three main steps** (variations found in literature)
  – Segmenting a text into argument units and other parts
  – Classifying the type/role of each unit
  – Identifying and classifying relations between units

• **Why argument mining?**
  – Real-world arguments often "hidden" in longer text, possibly fragmented
  – Mining is the basis for any argument analysis and any application

  Exception: Arguments, and their structure, already given in the source data
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

---

<table>
<thead>
<tr>
<th>non-argumentative</th>
<th>argumentative</th>
</tr>
</thead>
<tbody>
<tr>
<td>“If you wanna hear my view, I think that the EU should allow sea patrols in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats. Nothing justifies to endanger the life of innocent people.”</td>
<td></td>
</tr>
</tbody>
</table>

- **State of the art** (Ajjour et al., 2017)
  - Rather reliable on narrow genres (F₁ 0.72–0.82), unsolved across them
  - Unit granularity differs: Anything between clauses and paragraphs
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)

- **Unit classification**
  - **Task.** Given an argument unit, assign one class from a set of classes
  - **Method.** Usually, supervised text classification

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

- **State of the art**
  - Reliable on "explicit" argumentation, such as essays (F₁ 0.87) (Stab, 2017)
  - Still rather reliable on news editorials (F₁ 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 do they have, if any?
  – Method. Various, e.g., with minimum spanning trees (Peldszus and Stede, 2015)

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

• 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
3. Argument mining for writing support
(Wachsmuth et al., 2016)
Argument mining for writing support: Introduction

Argumentation quality assessment

Assessment

argumentative structure

Minning

Essay scoring each quality dimension with a score from [1, 4].

Synthesis

Suggestion (output)

organization 2.0
clarity 3.0
adherence 4.0
strength 2.5

Henning Wachsmuth, Khalid Al-Khatib, Benno Stein

Modeling the Argument Annotated Essays corpus

Learning of mining analysis of common ADU change flows in all ICLE paragraphs.

Application of mining conclusion based on the output of mining argumentation

Statistical insights into argumentation essay

For poets and literate people of yore it was a common idea to transcend reality or to go beyond it by using their imagination. Man is a way transcends himself through his imagination. It is possible to achieve a higher level of understanding by using our imagination. Therefore, it is important to use our imagination in our daily lives.

As a logical conclusion to my essay I would like to put only one thing. "Wouldn't it be better if imagination makes the world go where we want?" Man thinks that at last he has succeeded in making travelling without boundaries very easy. He can make a trip to any place he wants to. The world is a small village now. The ability to travel without boundaries is a dream for many people. Moreover, the ability to travel without boundaries is a dream for many people. The world is a small village now. The ability to travel without boundaries is a dream for many people. Moreover, the ability to travel without boundaries is a dream for many people.

Additionally, many people believe that the matrix in particular the scientific type of Man thinks that at last he has succeeded in making travelling without boundaries very easy. He can make a trip to any place he wants to. The world is a small village now. The ability to travel without boundaries is a dream for many people.

For poets and literate people of yore it was a common idea to transcend reality or to go beyond it by using their imagination. Man is a way transcends himself through his imagination. It is possible to achieve a higher level of understanding by using our imagination. Therefore, it is important to use our imagination in our daily lives.

As a logical conclusion to my essay I would like to put only one thing. "Wouldn't it be better if imagination makes the world go where we want?" Man thinks that at last he has succeeded in making travelling without boundaries very easy. He can make a trip to any place he wants to. The world is a small village now. The ability to travel without boundaries is a dream for many people. Moreover, the ability to travel without boundaries is a dream for many people.

Additionally, many people believe that the matrix in particular the scientific type of Man thinks that at last he has succeeded in making travelling without boundaries very easy. He can make a trip to any place he wants to. The world is a small village now. The ability to travel without boundaries is a dream for many people. Moreover, the ability to travel without boundaries is a dream for many people. The world is a small village now. The ability to travel without boundaries is a dream for many people.
Essay scoring based on argument mining

• Research question
  – Does argument mining help to score the quality of persuasive essays?

• Quality dimensions (Persing et al., 2010; Persing and Ng, 2013–2015)
  – Organization. How well is the essay arranged?
  – Thesis clarity. How easy to understand is the thesis?
  – Prompt adherence. How close does the essay stay to the issue?
  – Argument strength. How strong is the argument made for the thesis?

• Data
  – 800–1003 essays with scores in [1,4] for each dimension

• Approach
  1. Mine argument structure
  2. Analyze patterns in the structure
  3. Assess quality based on patterns
Essay scoring approach: Mining and analysis

• **Mining**
  – **Task.** Classify each sentence as thesis, conclusion, premise, or none
  – **Data.** AAE corpus (Stab and Gurevych, 2014a)
  – **Approach.** Support vector machine (SVM), several standard features

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority baseline</td>
<td>52.5</td>
<td>36.1</td>
</tr>
<tr>
<td>Stab and Gurevych (2014b)</td>
<td><strong>77.3</strong></td>
<td>72.6</td>
</tr>
<tr>
<td><strong>Our SVM classifier</strong></td>
<td>74.5</td>
<td><strong>74.5</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Unit role flows</th>
<th>Average</th>
<th>First</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusion, Premises</td>
<td>25.1%</td>
<td>–</td>
<td>13.1%</td>
</tr>
<tr>
<td>Conclusion, Premises, Conclusion</td>
<td>17.0%</td>
<td>–</td>
<td><strong>27.2%</strong></td>
</tr>
<tr>
<td>None, thesis</td>
<td>3.4%</td>
<td><strong>25.9%</strong></td>
<td>–</td>
</tr>
<tr>
<td>Premises, Conclusion</td>
<td>2.9%</td>
<td>–</td>
<td>2.7%</td>
</tr>
</tbody>
</table>
Argument mining on example essay

Prompt

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

Essay

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

Conclusion

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

Premise

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

Body

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

• **Assessment**
  – **Approach.** SVM regression, argument-specific and standard features

  ![Diagram]

  - **Unit role flows**
    - Intro: 0.25
    - Body: 0.50
    - Conc: 0.25
  - **Unit role composition**
    - 3: 0.25, 0: 0.75
    - 2: 0.25, 1: 0.50, 0: 0.25
    - 3: 0.25, 2: 0.50, 0: 0.25
  - **Function flows**
    - (Intro, Body, Body, Conc)
    - (Intro's, Bodies, Conc's)
  - **Content features**
    - Token {1, 2, 3}-grams
    - Prompt similarity

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>Organization</th>
<th>Clarity</th>
<th>Adherence</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average baseline</td>
<td>0.349</td>
<td>0.469</td>
<td>0.291</td>
<td>0.266</td>
</tr>
<tr>
<td>Persing et al. (2010–2015)</td>
<td>0.175</td>
<td><strong>0.369</strong></td>
<td><strong>0.197</strong></td>
<td><strong>0.244</strong></td>
</tr>
<tr>
<td><strong>Our SVM regressor</strong></td>
<td><strong>0.164</strong></td>
<td>0.425</td>
<td>0.216</td>
<td><strong>0.226</strong></td>
</tr>
<tr>
<td>- Unit role flows</td>
<td>0.234</td>
<td>0.461</td>
<td>0.247</td>
<td>0.242</td>
</tr>
<tr>
<td>- Unit role composition</td>
<td><strong>0.194</strong></td>
<td>0.457</td>
<td>0.239</td>
<td>0.239</td>
</tr>
<tr>
<td>- Function flows</td>
<td>0.220</td>
<td>0.478</td>
<td>0.255</td>
<td>0.251</td>
</tr>
<tr>
<td>- Content features</td>
<td>0.336</td>
<td><strong>0.425</strong></td>
<td><strong>0.231</strong></td>
<td><strong>0.236</strong></td>
</tr>
</tbody>
</table>
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?
What to assess and how to assess it

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

• How to assess
  – Absolute or relative?
  – How *should* we vs. how *do* we argue?
  – Based on manual assessments or ”objective“ properties?
  – Include model of audience?
"An argument is cogent if its premises are relevant to its conclusion, individually acceptable, and together sufficient to draw the conclusion."

Blair (2012)

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

van Eemeren (2015)

"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)

- Local relevance
  - Local acceptability
  - Cogency
  - Clarity
  - Effectiveness
  - Credibility
  - Appropriateness

- Global relevance
  - Global acceptability
  - Reasonableness
  - Arrangement
  - Emotional appeal
  - Persuasiveness
  - Winning side
  - Convincingness
The role of participants in argumentation

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

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

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

"According to a recent UN study, the number of rescue boats had no effect on the number of refugees who try."
Selected quality assessment approaches

• **Absolute assessment** (Wachsmuth et al., 2016)
  - Regression of four 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., 2020)
  - Analysis of editorial effectiveneness for audience
  - Model of audience‘s ideology and personality
5. Fallacy detection in online discussions
(Habernal et al., 2018)

Ivan Habernal
Iryna Gurevych
Benno Stein
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

- **Fallacies are hard to detect**
  - Structure identical to other arguments
  - Understanding and context knowledge needed

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)
A study of ad-hominem arguments on the web

- **Ad-hominem arguments**
  - Attacking the opponent instead of his or her arguments
  - 20% of 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

That’s an ad hominem fallacy

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

Deltas(s) from OP

CMV: Trump has done nothing of substance since being elected to office.

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

He pointed out things like the economy, which was growing.
"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?"

"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"

"Again, how old are you?"

"You have no capability to understand why"

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

"You're just a dishonest troll"

"you dumb fuck"

"Your just an asshole"

"Possible lie any harder?"

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

"You're trash at debating."

"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"

"Willful ignorance is not something I can combat"

"Can you also use Google?"

"You’re using troll tactics"
Identification of ad-hominem

- **Distribution of ad-hominem**
  - 75% threads with \( \leq 2 \) ad-hominems (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 Graph]

- Vulgar insult: 31%
- Illiteracy insult: 13%
- Condescension: 7%
- Ridiculing and sarcasm: 7%
- "Idiot" insults: 7%
- Accusation of stupidity: 4%
- Denial of no arguing skills: 4%

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>0.88</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>0.78</td>
</tr>
<tr>
<td>CNN</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Analysis of triggers of ad-hominem

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

(VOV_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.

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

• Terms with much attention
  – Mostly topic-independent rhetorical devices
  – A few loaded keywords, such as ”racist“
  – Partly meta about argumentation

(OOV means out-of-vocabulary)
Conclusion
Conclusion

• **Argument mining and assessment**
  - Finding arguments in natural language text
  - Classifying stance and other properties
  - 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


