Computational Argumentation — Part IV

Argument Acquisition

Henning Wachsmuth

henningw@upb.de

PADERBORN UNIVERSITY

Learning goals

- **Concepts**
  - Corpus design principles
  - Main text corpora for computational argumentation
  - Other argumentation-related resources
- **Methods**
  - How to create a corpus step by step
  - How to compute agreement between annotators
- **Associated research fields**
  - Corpus linguistics
  - Natural language processing
- **Within this course**
  - Learn about acquiring resources for computational argumentation, and understand their concepts

https://commons.wikimedia.org

https://pixabay.com

https://pixabay.com

Outline

What to acquire?

> *" It's not the one who has the best algorithm that wins. It's who has the most data."*

(Ng, 2018)

- **Data and language resources**
  - In data-driven research, the most important resources are corpora.
  - Corpora form the basis of development and evaluation.
  - We focus on *annotated text corpora* related to argumentation.
  - Other language resources. Lexicons, embedding models, and similar.

- **Web and software resources**
  - Online debate portals with tons of arguments "for free"
  - Community platforms where people collect argument resources
  - Code libraries for applying computational argumentation
  - Tools for creating, analyzing, and interacting with arguments

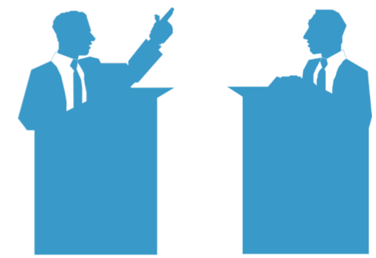
Argumentative genres (recap)

- **Written monologue**
  - Persuasive essays
  - Opinionated articles / Editorials
  - Argumentative blog posts
  - Customer and scientific reviews
  - Scientific articles
  - Law texts

  ... among others

- **Written dialogue**
  - Comments to news articles
  - Social media posts
  - Online forum discussions
  - eMail threads
  - Online debates

  ... among others



- **Spoken monologue** (possibly transcribed)
  - Political speeches
  - Law pleadings

  ... among others

- **Spoken dialogue** (possibly transcribed)
  - Classical debates
  - Everyday discussions

  ... among others

- **Notice**
  - The focus in this course is on *written* argumentation, i.e., argumentative texts.

Annotated text corpora (recap)


https://pixabay.com

- **Text corpus**
  - A collection of real-world texts with known properties, compiled to study a language problem
  - The texts are often *annotated* with respective meta-information.
  - Corpora are usually split into datasets for developing (training) and/or evaluating (testing) an algorithm.

- **Annotations**
  - Marks a text or text span as representing meta-information of a specific type

    Annotations may also be called *tags*, *labels*, or similar.
  - Types are specified by an *annotation scheme*.
  - Also used to specify relations between annotations

Time entity (2014) · Organization entity (Google)

" 2014 ad revenues of Google are going to reach

Reference (search company) · Time entity ('98)

$20B. The search company was founded in '98.

Founded relation

Reference (Its) · Time entity (in 2004)

Its IPO followed in 2004. [...] "

Topic: "Google revenues" **Genre:** "News article"

- **Corpora in NLP**
  - NLP approaches are developed and evaluated on text corpora.
  - Without, it's hard to develop a good approach, let alone to reliably evaluate it.

Manual, ground-truth, and automatic annotation

- **Manual annotation**
  - The annotations of a text corpus are usually created manually.
  - Annotation may be done by domain or language experts, but also by lay persons, e.g., using *crowdsourcing.*
  - To assess the quality of manual annotations, *inter-annotator agreement* is computed based on texts annotated multiple times.

- **Ground-truth annotations**
  - Manual annotations assumed to be correct are called the ground truth.
  - Sometimes, ground-truth annotations can also be derived from given data using *distant supervision.*
  - NLP algorithms are developed based on analyzing ground-truth annotations.

- **Automatic annotation**
  - Technically, NLP algorithms add annotations of certain types to input texts.
  - The automatic process usually aims to mimic the manual process.

    In this lecture part, automatic annotation is not in the focus.

Next section: Corpus creation

I. Introduction to computational argumentation

II. Basics of natural language processing

III. Basics of argumentation

IV. Argument acquisition

a) Introduction

b) Corpus creation

c) Existing argumentation-related resources

d) Conclusion

V. Argument mining

VI. Argument assessment

VII. Argument generation

VIII. Applications of computational argumentation

IX. Conclusion

Overview of corpus creation

- **Input**
  1. Text compilation. Choose the texts to be included.
  2. Annotation scheme. Define for what variables to annotate the texts.
  3. Text preprocessing. Prepare texts for annotation.
- **Annotation process**
  4. Annotation sources. Decide who provides annotations.
  5. Annotation guidelines. Define how to annotate.
  6. Pilot annotation. Test the annotation process.
  7. Inter-annotator agreement. Compute how reliable the annotations are.
- **Output**
  8. Postprocessing. Fix errors and filter annotations.
  9. File representation. Store the annotated texts adequately.
  10. Dataset splitting. Create subsets for training and testing.

1. Text compilation

- **Text compilation**
  - The first step in corpus creation is to collect the texts to be included.
  - The compilation should represent the application scenario of the studied task.
  - Several types of potential data bias need to be accounted for.
  - Also, copyrights may have to be considered.

- **Main compilation design decisions**
  - Size. Usually, the more the better, but annotation needs to remain doable
  - Domains. Topics, genres, languages, etc. (or combinations) to consider
  - Confounders. Variables to control for (via balancing, range restrictions, ...)
    Examples: Publication time, length, author, as well as many task-specific variables.

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - 2100 English hotel reviews to be annotated (+ 196,865 additional)
    All reviews were filtered from a previously published corpus (Wang et al., 2010).
  - 300 reviews each out of 7 locations, 420 each with user overall rating 1–5
  - At least 10 hotels per location, but as few as possible

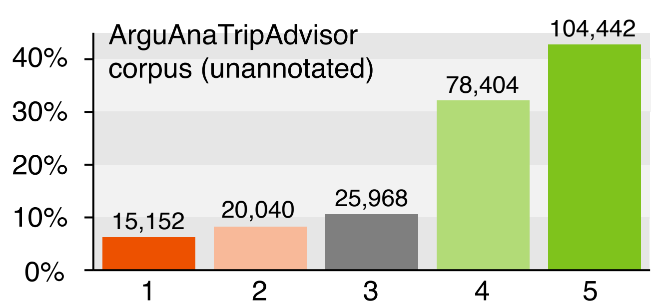
https://pixabay.com

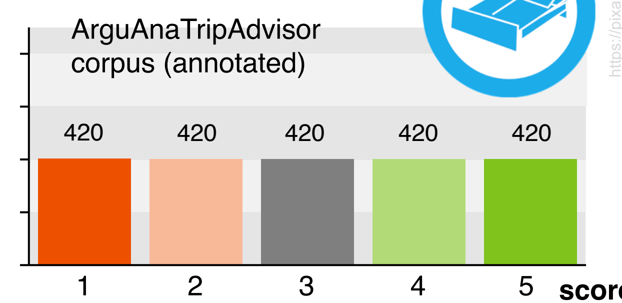
1. Text compilation: Representativeness and balance

- **Representativeness**
  - A text compilation is representative for some variable, if it includes the full range of variability of texts with respect to the variable.
  - Representativeness is important for generalization, since the corpus governs what can be learned about a given domain.

- **Representative vs. balanced distributions**
  - Evaluation. The distribution of texts over different values of a type should be representative for the real distribution.
  - Development. A balanced distribution where all values are represented evenly can be favorable (for machine learning and for analysis).

- **Example: ArguAna TripAdvisor corpus**
  (Wachsmuth et al., 2014)





2. Annotation scheme

- **Annotation scheme**
  - The definition of the annotation types to be considered within a task
  - Clarifies syntax, semantics, and possibly pragmatics behind each type
  - Represents the model of the given task and implies what can be studied on a corpus (in a supervised way)

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - Sentiment. Each statement classified as positive, negative, or neutral
    A statement was defined to be ≥ 1 clause and ≤ 1 sentence as well as to be meaningful on its own.
  - Aspects. Each aspect of a hotel marked
  - Ratings. Each review scored for several quality dimensions

title: *great* **location**, *bad* **service**

sentiment score: 2 of 5

body: *stayed at the darling harbour holiday inn. The* **location** *was great, right there at China town, restaurants everywhere, the monorail station is also nearby. Paddy's market is like 2 mins walk.* **Rooms** *were however very small. We were given the 1st floor* **rooms**, *and we were right under the monorail track, however* **noise** *was not a problem.* **Service** *is terrible.* **Staffs** *at the* **front desk** *were impatient, I made an enquiry about* **internet access** *from the* **room** *and the person on the phone was rude and unhelpful. Very shocking and unpleasant encounter.*

3. Text preprocessing

- **Text preprocessing**
  - The preparation of corpus texts for their manual annotation

- **Usual preprocessing steps**
  - The input files are converted into a common, usually simple format.
  - Metadata is stored, in case it is considered relevant.
  - The texts are analyzed, usually automatically, in order to create the instances to be annotated.

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)

https://pixabay.com

  - Originally, the input reviews were crawled HTML pages.
    Due to the resort to an existing corpus, the reviews had an intermediate format already.
  - The review contents were converted to plain text.
  - The review ratings and other metadata were stored in annotations.
  - Each text was automatically segmented into statements using a rule-based algorithm provided with the corpus.

4. Annotation sources

- **Expert annotation**
  - Experts for a task (or for linguistics, ...) manually annotate each corpus text.
  - Usually achieves the best results, but is often time and cost intensive

- **Crowd-based annotation**
  - Instead of experts, *crowdsourcing* is used to create manual annotation.
  - Access to many lay annotators (cheap) or semi-experts (not that cheap)
  - Distant coordination overhead; results for complex tasks unreliable

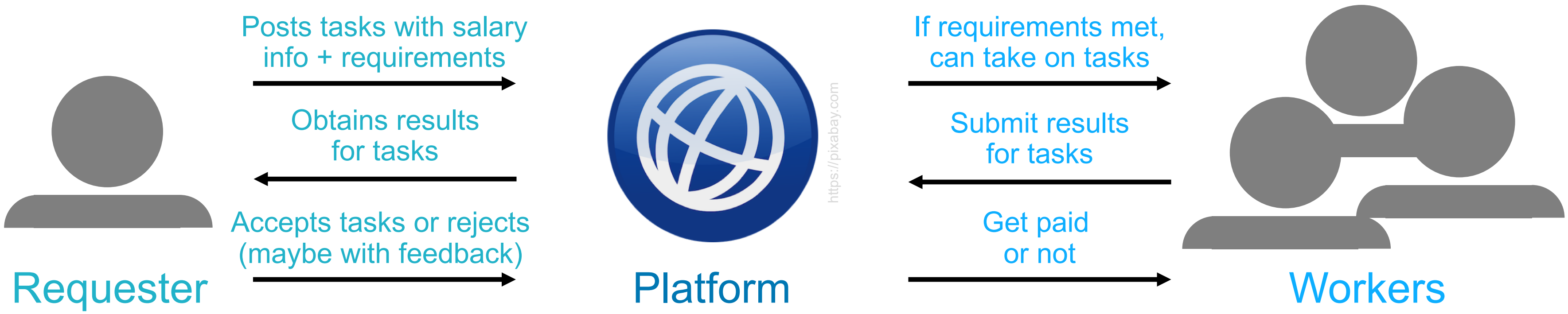
- **Distant supervision**
  - Annotations are (semi-) automatically derived from existing metadata.
  - Enables large corpora, but annotations may be noisy

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - Sentiment. Crowd-based annotation, with three annotators each
  - Aspects. Expert annotations, one expert per review (two for a sample)
  - Ratings. Distant supervision; ratings directly obtained from review metadata

https://pixabay.com

4. Annotation sources: Crowdsourcing

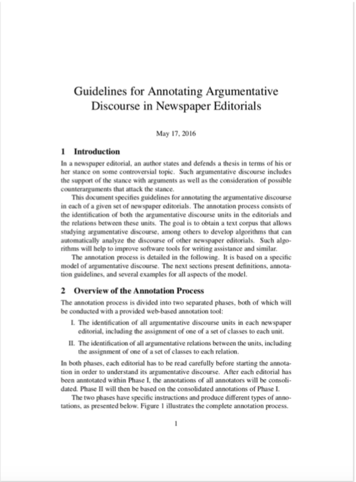
- **Crowdsourcing**
  - Outsourcing of (usually micro) jobs to people around the world
  - Tasks and results are submitted to a crowdworking platform.



- **Selected platforms**
  - mturk.com (Amazon Mechanical Turk, AMT). Biggest platform, lay workers
  - upwork.com. Semi-professional freelancers for several areas

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - AMT, $0.05 per 12 sentiment classifications, 328 workers involved

5. Annotation guidelines



Guidelines for Annotating Argumentative Discourse in Newspaper Editorials

May 17, 2016

1 Introduction

In a newspaper editorial, an author states and defends a thesis in terms of his or her stance on some controversial topic. Such argumentative discourse includes the support of the stance with arguments as well as the consideration of possible counterarguments that attack the stance.

This document specifies guidelines for annotating the argumentative discourse in each of a given set of newspaper editorials. The annotation process consists of the identification of both the argumentative discourse units in the editorials and the relations between these units. The goal is to obtain a text corpus that allows studying argumentative discourse, among others to develop algorithms that can automatically analyze the discourse of other newspaper editorials. Such algorithms will help to improve software tools for writing assistance and similar.

The annotation process is detailed in the following. It is based on a specific model of argumentative discourse. The next sections present definitions, annotation guidelines, and several examples for all aspects of the model.

2 Overview of the Annotation Process

The annotation process is divided into two separated phases, both of which will be conducted with a provided web-based annotation tool:

I. The identification of all argumentative discourse units in each newspaper editorial, including the assignment of one of a set of classes to each unit.

II. The identification of all argumentative relations between the units, including the assignment of one of a set of classes to each relation.

In both phases, each editorial has to be read carefully before starting the annotation in order to understand its argumentative discourse. After each editorial has been anntotated within Phase I, the annotations of all annotators will be consolidated. Phase II will then be based on the consolidated annotations of Phase I.

The two phases have specific instructions and produce different types of annotations, as presented below. Figure 1 illustrates the complete annotation process.

1

- **Annotation guidelines**
  - To obtain reliable annotations, annotators get guidelines that clarify what and how to annotate.
  - Guidelines define concepts, explain the annotation scheme, prescribe the annotation process, and often give examples.
    For experts, they may span dozens of pages; for lay persons they are often kept short.

- **Length as a design decision**
  - The more complete, the more guidelines will represent the authors‘ view.
  - The more concise, the more they will represent the annotators‘ view.

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - For crowd-based sentiment, we had the following simple guidelines:
    (along with a set of carefully chosen examples)

”When visiting a hotel, are the following statements positive, negative, or neither?“

Notes. *(1)* Pick “neither” only for facts, not for unclear cases. *(2) Pay attention to subtle statements where sentiment is expressed implicitly or ironically. (3) Pick the most appropriate answer in controversial cases.*

6. Pilot annotation

- **Pilot annotation**
  - Before a complete corpus is annotated, annotation guidelines are usually tested on a small sample.
  - The goal is to identify unclear parts, overseen and hard cases, as well as general annotation problems.
  - Guidelines are often written incrementally based on multiple pilot studies.
    The cases identified in pilot studies often serve as examples in the guidelines.

https://pixabay.com

- **Annotators in pilot study**
  - Rule of thumb. If authors don‘t achieve *agreement*, annotators won‘t either.
    In (Al-Khatib et al., 2016b), the annotation of argumentative relations were dropped for this reason.
  - Experts may discuss and align their annotation based on pilot results.
  - Sometimes, the set of annotators is chosen based on pilot results.

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - Sentiment. The guideline above was best among multiple variations.
  - Aspects. The decision to use experts was based on pilot crowdsourcing tests.

https://pixabay.com

7. Inter-annotator agreement

- **Inter-annotator agreement** (aka *inter-rater reliability* or *inter-coder agreement*)
  - A quantification of the similarity of annotations of the same instances by two or more annotators

    Common numbers of annotators are 2, 3, or 5. Sometimes, also way more are used (especially in crowdsourcing).
  - Between 1.0 (perfect agreement) and –1.0 (systematic disagreement)

    0.0 then means random agreement (i.e., no agreement).

- **Why inter-annotator agreement?**
  - Captures the reliability (or homogeneity) of the annotations of a corpus
  - Gives a rough idea of how effective an algorithm may become

    It is unlikely in general that an algorithm will more agree with humans than humans agree with each other.
  - Dilemma. Low agreement may indicate bad guidelines or insufficient training — but also just a subjective task.

- **Basis for computing agreement**
  - Either, each corpus instance is annotated by multiple annotators.
  - Or, a sample is annotated multiple times, and the rest once each.

    The former is statistically more reliable and allows annotation filtering, majority agreement, etc.; the latter is cheaper.

7. Inter-annotator agreement: Overview of measures

- **Joint probability measures**
  - Simply represent percentage agreement on nominal annotations
  - Percentage. Proportion of instances where pairs of annotators agreed
  - Full. Proportion of instances where $k \geq 3$ annotators all agreed
  - Majority. Proportion of instances where >50% of the annotators agreed
- **Chance-corrected measures**
  - More robust, taking into account that agreement may be due to chance
  - Cohen's κ. Difference between observed and chance agreement (see below)
  - Fleiss' κ. "Generalization" of Cohen's κ to $k \geq 3$ annotators
  - Krippendorff's α. Focus on *dis*agreement cases, any *k*, any type of scale
- **Correlation measures**
  - Quantify the (mean) pairwise correlation among annotators for ordinal scale
  - Kendall's τ. Concordance of ranks of two orderings of instances (see below)
  - Spearman's ρ. Monotonicity of the relation between two orderings
  - Pearson's *r*. Linear correlation between two sets of *continuous* values

7. Inter-annotator agreement: Cohen‘s κ

- **Cohen‘s κ**
  - Given *n* instances annotated by annotators, *A* and *B*, for a set of nominal categories *C*:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad \text{where} \quad p_e = \frac{1}{n^2} \sum_{c \in C} a_c \cdot b_c \,,$$

  - p_o is the observed percentage agreement on instances
  - p_e is the expected agreement by chance
  - a_c and b_c are the numbers of times *A* and *B* chose class *c*, respectively

κ range	Agreement
[-1.0, 0.0]	No
(0.0, 0.2]	Slight
(0.2, 0.4]	Fair
(0.4, 0.6]	Moderate
(0.6, 0.8]	Substantial
(0.8, 1.0]	„Perfect”

- **Example**
  - $n = 100$, two categories *c* and *c*‘, $a_c = b_c = 80$, $a_{c‘} = b_{c‘} = 20$, $p_o = 0.75$

$$p_e = \frac{1}{10000} \cdot (6400 + 400) = 0.68 \quad \text{and thus} \quad \kappa = \frac{0.75 - 0.68}{1 - 0.68} \approx 0.22$$

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - Sentiment. Fleiss‘ κ = 0.67, 73.6% full, 98.3% majority
  - Hotel aspects. Cohen‘s κ = 0.73 (based on 546 cases)

https://pixabay.com

7. Inter-annotator agreement: Kendall‘s τ

- **Kendall‘s τ rank correlation coefficient**
  - Given n instances to be ranked, let $(a_1, b_1), \ldots, (a_n, b_n)$ be the joint ranks of all instances assigned by two annotators, A and B:

$$\tau = \frac{\#\, concordant\ pairs \;-\; \#\, discordant\ pairs}{n \cdot (n-1)/2}$$

  - Concordant. Any $(a_i, b_i), (a_j, b_j), i < j$: $a_i < a_j$ and $b_i < b_j$, or $a_i > a_j$ and $b_i > b_j$
  - Discordant. Any $(a_i, b_i), (a_j, b_j), i < j$: $a_i < a_j$ and $b_i > b_j$, or $a_i > a_j$ and $b_i < b_j$

- **Adjustment for ties**
  - The default τ ignores the number of *ties,* t_A (for $a_i = a_j$) and t_B (for $b_i = b_j$).
  - A common adjustment, τ‘, replaces the denominator of τ by:

$$\sqrt{(\#concordant + \#discordant + t_A) \cdot (\#concordant + \#discordant + t_B)}$$

- **Example**
  - $n = 3$, rank pairs: $(1, 2), (2, 3), (3, 3)$ — $\tau = (2 - 0) / 3 \approx 0.67$
  - $\#\, concordant = 2, \#\, discordant = 0, t_A = 0, t_B = 1$ — $\tau‘ = (2 - 0) / \sqrt{6} \approx 0.82$

8. Postprocessing

- **Postprocessing**
  - The consolidation of the annotated texts for the final corpus
  - Includes *cleansing* of potentially wrong or inconsistent cases
  - May be manual and/or automatic

- **Common postprocessing steps**
  - Resolution (or discarding) of cases where annotators disagreed
  - Removal of noise in the data observed during annotation
  - Merging of labels that have been assigned only rarely with others
  - Conversion of the instance format into the final corpus *file representation*

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - Each statement was assigned its majority sentiment where available.
  - The 1.7% sentiment disagreement cases were resolved manually in the context of their associated reviews.
  - Wrong hotel aspect annotation boundary errors were automatically fixed.

https://pixabay.com

9. File representation

- **File representation**
  - Usually, each text in a corpus is stored in a separated file.
    Often, each dataset (or other subsets of the corpus) in a separated folder
  - Large corpora may be stored in databases or indexes.
  - Various file formats and instance representations exist

https://commons.wikimedia.org

- **Common file formats**
  - Plain text file only. One line per token, one tab per token-level annotation
  - Plain text + annotation file. Only text in file, extra file specifies annotations
  - XMI/XML file. One file for each text, one tag per annotation
  - Spreadsheet. One row per text, one additional column per annotation

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - XMI files preformatted for the Apache UIMA framework
  - Each annotation is stored as a tag with attributes and character indices.
  - The annotation scheme is specified in a global type system descriptor file.

https://pixabay.com

10. Dataset splitting

- **Dataset splitting**
  - Deciding how to split a corpus into training, validation, and test set (or similar) is not trivial, but depends on the task.
  - The goal is to mimic the real-world situation to be studied.
  - A good split minimizes bias that can be exploited in learning.
  - The annotations within a text should usually not be put in different datasets, as they overlap in terms of content (explicitly or implicitly).

Training

Validation

Test

- **Common splitting criteria**
  - Random. Split done (pseudo-) randomly
  - Topic. Datasets (more or less) disjunct in terms of topic
  - Time. Oldest texts for training, newest for testing

    Other metadata relevant in a given task may equally serve as splitting criteria.

- **Example: ArguAna TripAdvisor corpus** (Wachsmuth et al., 2014)
  - Location. 3 locations for training, 2 for validation, 2 for test

    This way, location-specific information that may influence sentiment cannot be exploited.

https://pixabay.com

Amsterdam, Seattle, Sidney

Berlin, San Francisco

Barcelona, Paris

Next section: Existing argumentation-related resources

Overview of argumentation-related corpora 1

- **Argumentation-related corpora**
  - Corpora with annotations of argument structure
  - Corpora with assessments of argumentation quality
  - Corpora with classifications of stance and similar

  ... and possibly others

- **Selected corpora on argument structure**
  - AAE-v2. Persuasive essays, proprietary argument model (Stab, 2017)
  - Arg-microtexts. Short texts, Freeman model (Peldszus and Stede, 2015)
  - Araucaria. Mixed argumentative texts, Walton‘s schemes (Reed and Rowe, 2004)
  - AZ. Scientific articles, argumentative zones (Teufel, 1999)
  - IBM Debater. Wikipedia articles, claims and evidence (Rinott et al., 2015)
  - Web discourse. Mixed web arguments, Toulmin model (Habernal and Gurevych, 2015)
  - Webis-Debate-16. Debate portal texts, argumentativeness (Al-Khatib et al., 2016a)
  - Webis-Editorials-16. News editorials, six unit types (Al-Khatib et al., 2016b)

  ... and some others

Overview of argumentation-related corpora 2

- **Selected corpora on argumentation quality**
  - ArgQuality. Debate portal arguments, 15 quality scores (Wachsmuth et al., 2017b)
  - Cornell ChangeMyView. Discussion posts, effectiveness labels (Tan et al., 2016)
  - IBM Arg-Q. Crowdsourced arguments, preference pairs (Toledo et al., 2019)
  - UKP-ConvArg. Debate portal arg's, convincingness pairs (Habernal et al., 2016)
  - Webis-ArgRank-17. Mixed arguments, relevance rankings (Wachsmuth et al., 2017a)
  - Webis-Editorials-18. News editorials, effectiveness ratings (El Baff et al., 2018)

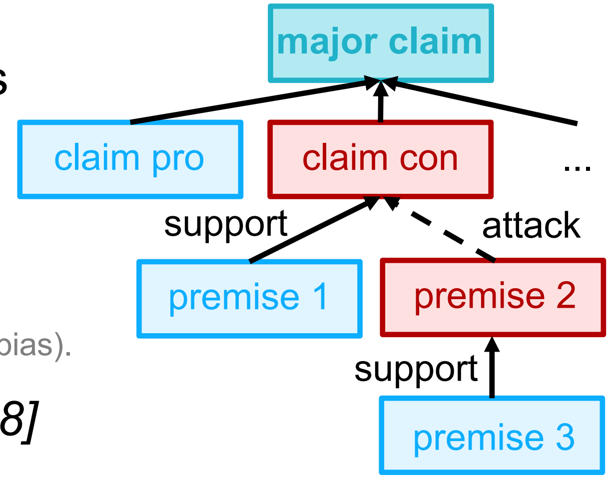
  ... and some others

- **Selected corpora on stance and similar**
  - ArguAna Counterargs. Debate portal counterargument pairs (Wachsmuth et al., 2018a)
  - ArguAna TripAdvisor. Hotel reviews with sentiment flows (Wachsmuth et al., 2014)
  - IBM Debater. Wikipedia articles, claim-related stance (Bar-Haim et al., 2017)
  - Ideological debates. Online discussions with stance (Hasan and Ng, 2013)
  - Internet arguments. Web discussions with topic and stance (Walker et al., 2012)

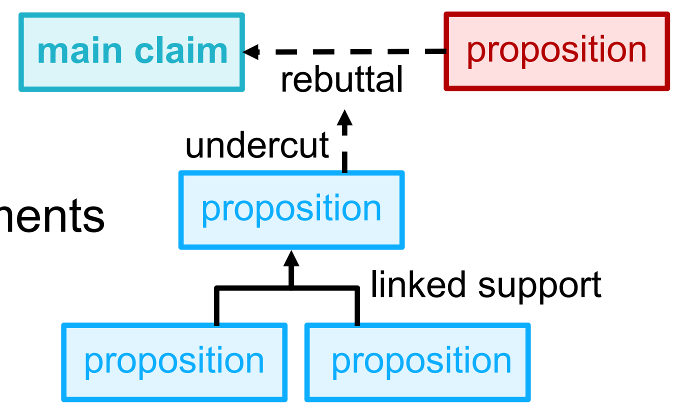
  ... and many others

Examples: AAE-v2 and Arg-microtexts

- **AAE-v2** (Stab, 2017)
  - Texts. 402 mixed-topic persuasive student essays from a web portal
  - Annotations. 6089 argumentative units of three types and 5687 relations of two types
    Extensions also cover quality-related annotations (sufficiency and myside bias).
  - Creation. 3 experts, Krippendorff‘s α in *[0.63, 0.88]*



- **Arg-microtexts** (Peldszus and Stede, 2015)
  - Texts. 112 ”pure“ arguments, explicitly written for 18 different controversial issues
  - Annotations. 576 units composed in 443 arguments according to Freeman‘s model
    Extensions also cover RST discourse structure.
  - Creation. 3 experts, Fleiss κ = *0.83*



Examples: IBM Debater and Webis-16-Editorials

- **IBM Debater** (Rinott et al., 2015; Bar-Haim et al., 2017)
  - Texts. 2394 claims and 3057 evidence statements for 58 controversial issues from Wikipedia articles
  - Annotations. Stance of claims on issue, target in each claim, claim-evidence support relations
  - Creation. 5 annotators, Cohen's κ = *0.4* for claims, *92.5%* majority agreement for target, rest not reported

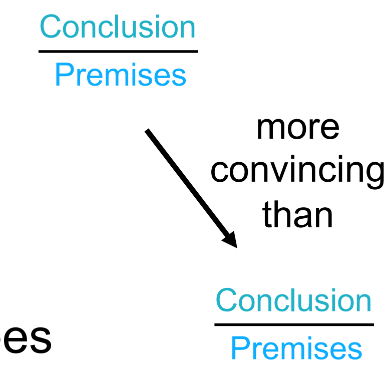
pro claim	55%
con claim	45%
anecdotal	12%
expert	58%
study	31%

- **Webis-16-Editorials** (Al-Khatib et al., 2016)
  - Texts. 300 mixed-topic news editorials, 100 each from three very different online news portals
  - Annotations. 14,313 argumentative units of six types
  - Creation. 3 semi-professional crowdworkers each, Fleiss' κ = *0.56*, ranging from *0.11* to *0.68*

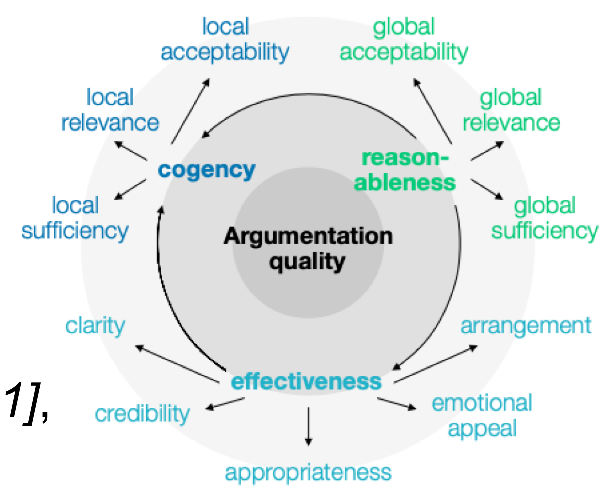
assumption	68%
common ground	2%
anecdote	18%
testimony	8%
statistics	3%
other	1%

Examples: UKP-ConvArg and ArgQuality

- **UKP-ConvArg** (Habernal et al., 2016)
  - Texts. 16,927 argument pairs (based on 1052 arguments) for 32 issue-stance pairs from a debate portal
  - Annotations. Each pair annotated as to which argument is more convincing (+ free text reasons)
  - Creation. Five lay crowdworkers each, best annotator agrees in 93.5% of the cases with "majority" (Hovy et al., 2013)



- **ArgQuality** (Wachsmuth et al., 2017b)
  - Texts. 320 arguments from UKP-ConvArg, 10 each per issue-stance pairs
  - Annotations. Scores in {1, 2, 3} for 15 different quality dimensions
  - Creation. 3 experts, Krippendorff's α in *[0.26, 0.51]*, majority agreement in *[0.87, 0.98]*



Other language resources

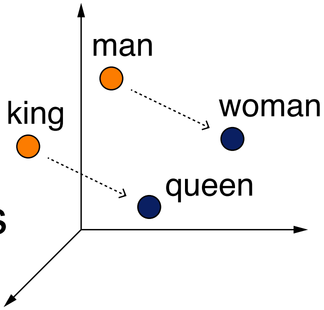
- **Argumentation-related lexicons**
  - Term repositories capturing specific aspects of argumentative language
  - Often come with much useful meta-information
  - Often can be created from annotated corpora

    Notice, though, that lexicon generation is a research area itself.

- **Lexicon types**
  - Argument-specific. Still rare and often published only as part of a code library

    Example: www.hlt.utdallas.edu/~persingq/ICLE/ (lexicons related to argumentation in persuasive essays)
  - Subjective language. Some powerful lexicons exist that include sub-lexica related to argumentation

    Examples: https://liwc.wpengine.com, http://www.wjh.harvard.edu/~inquirer/



- **Argument-specific embedding models**
  - Mappings from arguments, or their parts, to real-valued vectors

    Examples: aclanthology.org/2021.konvens-1.3.pdf, arxiv.org/pdf/2106.10832.pdf#page=60

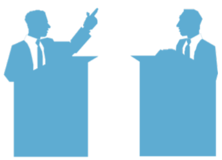
Online debate portals

- **Online debate portals**
  - Platforms where arguments are directly given for debates on several issues
  - Constitute a rich source of ”ground-truth“ argumentation

    Our argument search engine https://args.me and several corpora are based on debate portal arguments.

- **Two types of portals**
  - Debating forums. In each debate, users argue against each other.

    Examples: debate.org, reddit.com/r/changemyview/, createdebate.com, theworlddebating.com


https://de.wikipedia.org

  - Argument ”wikis“. Each ”debate“ collects arguments on an issue.

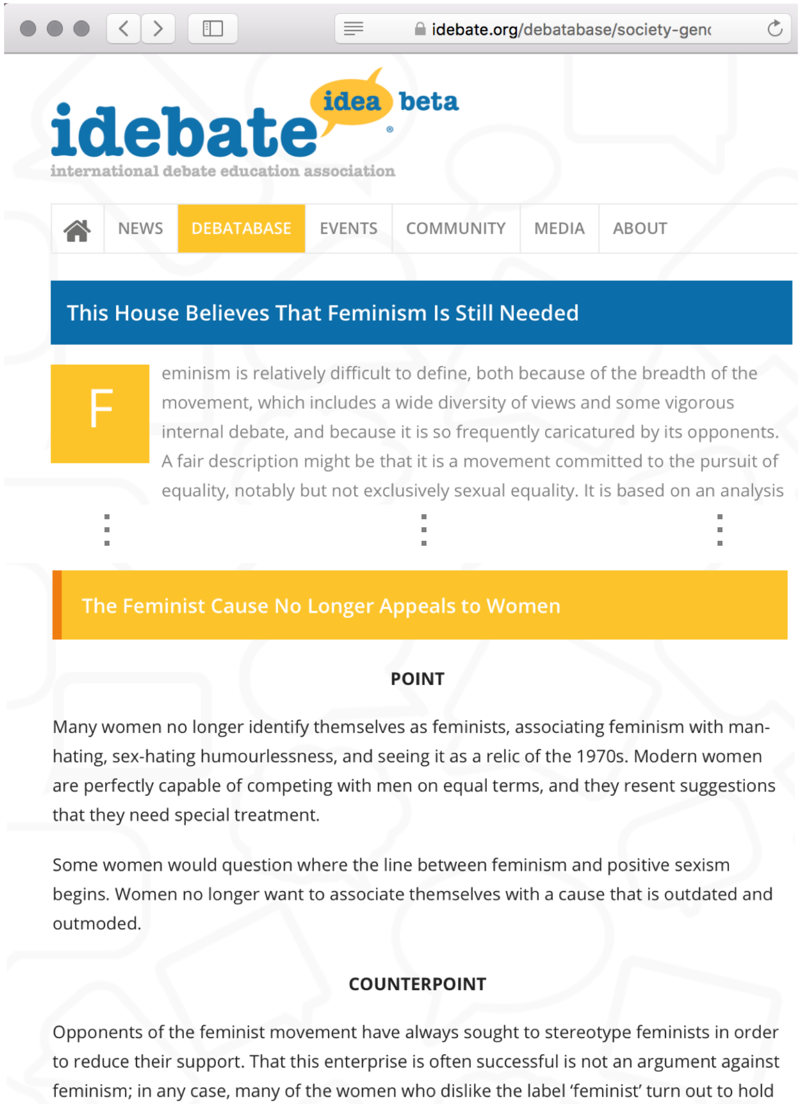
    Examples: idebate.org, debatepedia.org, debatewise.org, kialo.com


https://commons.wikimedia.org

- **Information found on debate portals**
  - On nearly all. Pro and con stance of arguments
  - On most. An introductory text on each issue
  - On several. Literature or web source of the arguments
  - On some. Meta-information on the authors of arguments
  - On some. User votings on arguments or stances

Example debate portal: iDebate

- **Web portal iDebate.org**
  - ”Debates“ on controversial issues
    e.g., Feminism is still needed
  - Categorized into 15 themes
    economy, religion, society, ...
- **Arguments on the portal**
  - Up to six pro and con points on each issue
    Each with conclusion and premise.
  - Collected by a community and revised multiple times
  - A counterpoint to every point is given
- **Size of iDebate (in January 2018)**
  - 1069 debates
  - 6753 point-counterpoint pairs



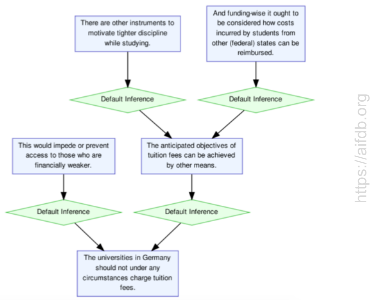
screenshots from https://idebate.org

Argumentation-related projects

- **ArguAna** www.arguana.com
  - Corpora, Java code, and tools for argumentation research
- **Argument Web** www.argumentinterchange.org
  - Tools to create, analyze, and interact with arguments
- **RATIO** www.spp-ratio.de
  - Priority program of the German research foundation with several projects
- **Project Debater Service API** https://early-access-program.debater.res.ibm.com/
  - API for main algorithms from Project Debater (early access)
- **VisArgue** visargue.inf.uni-konstanz.de
  - Tools to visualize dialogical argumentation, with built-in text analyses
- **And many more...**
  - rbutr.com, www.rationaleonline.com, cohere.open.ac.uk, www.archelogos.com, debategraph.org, www.argunet.org, evidence-hub.net, argumentz.com, www.truthmapping.com, https://diggingintodata.org, ...

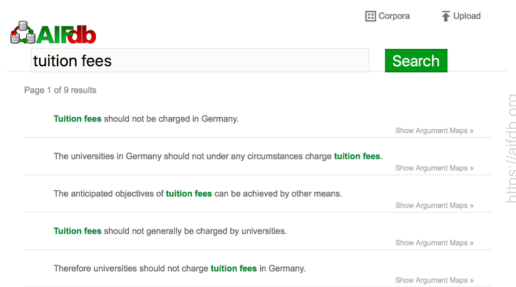
Example project: Argument Web

AIFdb Corpora



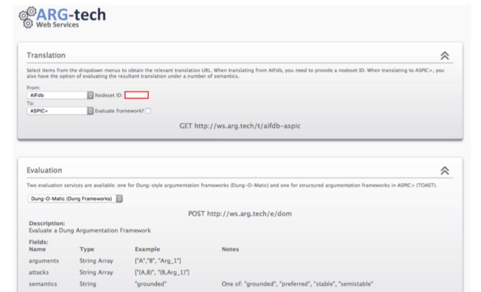
Structured argument data in uniform format

AIFdb Browser



Search interface for argument resources

ARG-tech API

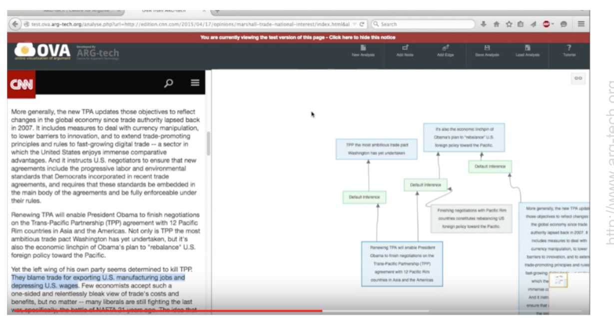


Several argument web services

Argublogging

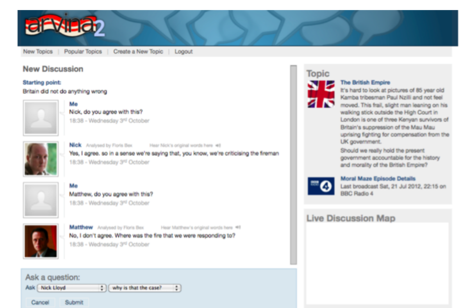
Widget for argument annotation in blogs

OVA



Online visualization and analysis of arguments

Arvina



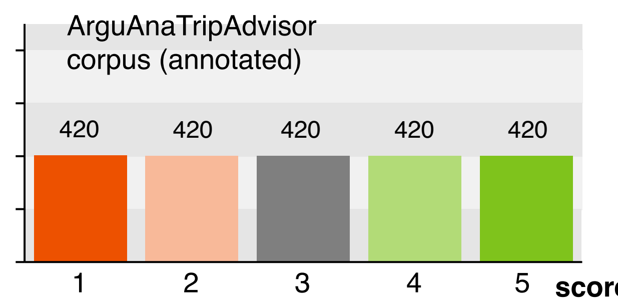
Dialog platform based on AIFdb

Next section: Conclusion

Conclusion

- **Resources for computational argumentation**
  - Text corpora annotated for arguments, stance, quality, ...
  - Argument-specific lexicons, embedding models, and similar
  - Web resources, code libraries, and tools

- **Corpus creation**
  - Compilation of texts suitable to study a task
  - Preprocessing and annotation of the input texts
  - Analysis and postprocessing of annotated texts



- **Important resources**
  - Particularly corpora with argument structure often used
  - Debate portals are a rich source of argumentation
  - No standard software, but some libraries and tools exist

ArguAna

References

- **Al-Khatib et al. (2016a).** Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. Cross-domain Mining of Argumentative Text through Distant Supervision. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1395–1404, 2016.
- **Al-Khatib et al. (2016b).** Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A News Editorial Corpus for Mining Argumentation Strategies. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3433–3443, 2016.
- **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.
- **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.
- **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. 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, 2016.
- **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. 2004.

References

- **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.
- **Ng (2018).** Andrew Ng. Machine Learning. Lecture slides from the Stanford Coursera course. 2018.https://www.coursera.org/learn/machine-learning.
- **Peldszus and Stede (2015).** Andreas Peldszus and Manfred Stede. Joint Prediction in MST-style Discourse Parsing for Argumentation Mining. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 938–948, 2015.
- **Reed and Rowe (2004).** Chris Reed and Glenn Rowe. Araucaria: Software for Argument Analysis, Diagramming and Representation. International Journal of AI Tools, 14:961– 980, **Rinott et al. (2015).** Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. Show Me Your Evidence — An Automatic Method for Context Dependent Evidence Detection. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 440–450, 2015.
- **Stab (2017).** Christian Stab. Argumentative Writing Support by means of Natural Language Processing, Chapter 5. PhD thesis, TU Darmstadt, 2017. http://tuprints.ulb.tu-darmstadt.de/6006/1/PhD-Thesis-ChristianStab.pdf
- **Tan et al. (2016).** Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In Proceedings of the 25th International Conference on World Wide Web, pages 613–624, 2016.
- **Teufel (1999).** Simone Teufel. Argumentative Zoning: Information Extraction from Scientific Text. Ph.D. thesis, University of Edinburgh, 1999.

References

- **Toledo et al. (2019).** Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, 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, pages 5624–5634, 2019.
- **Wachsmuth et al. (2014).** Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palarkarska. A Review Corpus for Argumentation Analysis. In Proceedings of the of the 15th International Conference on Intelligent Text Processing and Computational Linguistics, pages 115–127, 2014.
- **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. (2018a).** Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the Best Counterargument without Prior Topic Knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 241–251, 2018.
- **Walker et al. (2012).** Marilyn A. Walker, Pranav Anand, Jean E. Fox Tree, Rob Abbott, Joseph King. A Corpus for Research on Deliberation and Debate. In Proceedings of the 8th International Conference on Language Resources and Evaluation, pages 812–817, 2012.
- **Wang et al. (2010).** Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In: Proceedings of the 16th SIGKDD. pages 783–792, 2010.