

Computational Argumentation — Part IV

Argument Mining

Henning Wachsmuth

<https://ai.uni-hannover.de>



Learning goals

▪ Concepts

- Definitions, goals, and tasks in argument mining



<https://commons.wikimedia.org>

▪ Methods

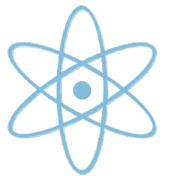
- Segmentation and classification of argumentative discourse units
- Detection and classification of relations between units and arguments
- Methods that tackle multiple mining tasks jointly



<https://pixabay.com>

▪ Associated research fields

- Natural language processing



<https://pixabay.com>

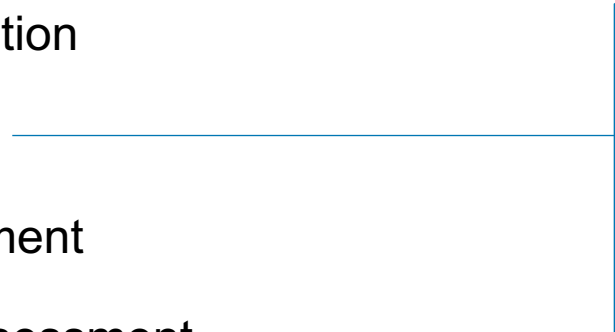
▪ Within this course

- The first of three main stages in computational argumentation



Outline: Introduction

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

- 
- a) Introduction**
 - b) Unit identification
 - c) Relation identification
 - d) Conclusion

Argument mining: Process

■ General process signature

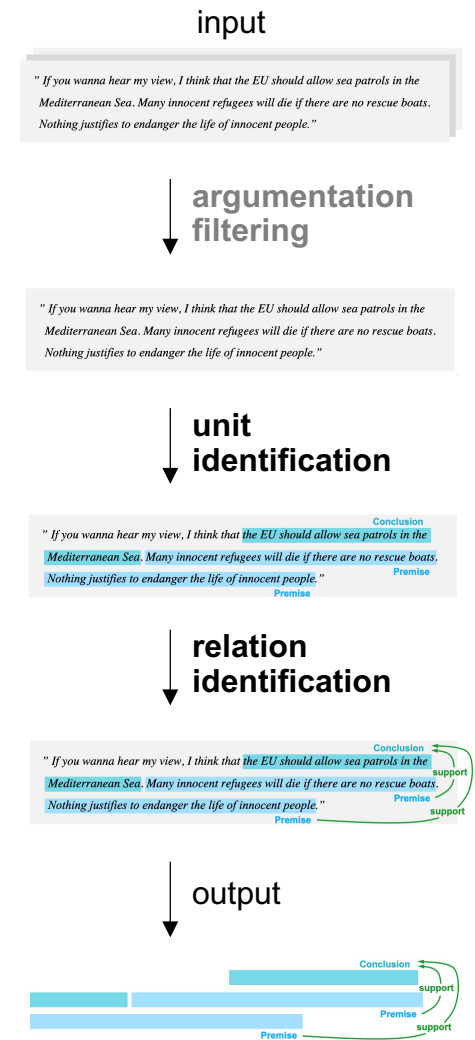
- **Input.** A set of (plain) texts
- **Output.** The argumentative structure of each text
What structure is mined exactly depends on the argument model employed.

■ Main high-level tasks

- Argumentation filtering. Finding argumentative texts
- **Unit identification.** Segmenting and classifying argumentative discourse units (ADUs)
- **Relation identification.** Detecting and classifying relations between argumentative units

■ Notice

- Different task decompositions and orderings exist.
- Some tasks may be tackled jointly, as we see below.
- Not all tasks always need to be tackled.



Outline: Unit identification

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

- a) Introduction
- b) Unit identification**
- c) Relation identification
- d) Conclusion

Unit identification

▪ Argument unit identification

- The segmentation and classification of all ADUs in a text
- **Segmentation.** Given a text, split it into argumentative units and other parts
- **Classification.** Given the units, assign each one type from a set of types

Conclusion

“*Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet. One who is living overseas will of course struggle with loneliness, living away from family and friends but those difficulties will turn into valuable experiences in the following steps of life. Moreover, the one will learn living without depending on anyone else.*”

Premise

Premise

Premise

example from Stab and Gurevych (2014a)

▪ Challenges

- What is argumentative may depend on the issue being discussed.
- Even humans may disagree on the correct segmentation.

Unit identification: Segmentation

▪ What is an argumentative unit?

- No clear general definition exists of what makes up the boundaries of units.
- In many genres, a unit is a clause or sentence (w/o discourse markers).
- However, some genres also includes multiple-sentence units.

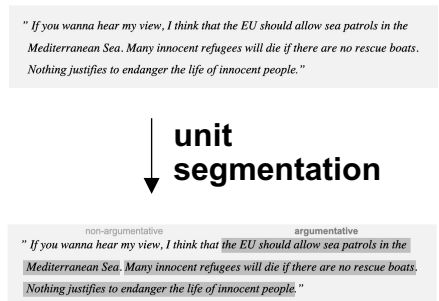
▪ Units across genres

- **Persuasive essays.** Nearly everything is argumentative. (Stab and Gurevych, 2014a)
- **Forum discussions.** Argumentativeness strongly varies. (Habernal and Gurevych, 2017)
- **Wikipedia articles.** Argumentativeness is issue-dependent. (Rinott et al., 2015)
- **News editorials.** Many ADUs rather have a *rhetorical* role. (Al-Khatib et al., 2016)

▪ Modeling unit segmentation

- **Individual classification** of candidate unit boundaries
- **Sequence labeling** of each token in a text

Some approaches also jointly tackle segmentation and classification.



Unit identification: Overview of approaches

▪ Selected segmentation approaches

- Rule-based boundary detection using parse trees (Persing and Ng, 2016)
- Conditional random field using diverse features (Stab, 2017)
- **Bi-LSTM using embeddings and diverse features** (Ajjour et al., 2017)

▪ Selected classification approaches

- Supervised classification with rich linguistic features (Stab and Gurevych, 2014a; Habernal and Gurevych, 2015; Rinott et al., 2015; Persing and Ng, 2016; Al-Khatib et al., 2017)
- Unit-level sequence labeling with rich linguistic features (Habernal and Gurevych, 2017)
- Zero-shot and few-shot prompting of large language models (Chen et al., 2024)

▪ Selected joint approaches

- LSTM using entity-relation information (Eger et al., 2017)
- Structure learning for graph prediction with SVMs and RNNs (Niculae et al., 2017)
- **Biaffine attention for unit-level sequence labeling** (Morio et al., 2020)

Bi-LSTMs for unit segmentation

- **Unit segmentation as token-level sequence labeling**

- Given a text, classify each token as belonging to an argumentative unit or not.
- Each token is beginning (B), inside (I), or outside (O) of a unit.

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

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ B I I I I I ○

- **Research questions**

1. What model is best to capture relevant context of a token?
2. What features are most effective in unit segmentation?
3. To what extent do models and features generalize across genres?

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

- **Presented approach** (Ajjour et al., 2017)

- A neural architecture where Bi-LSTMs capture the entire text as context
- Use of embeddings along with different types of features

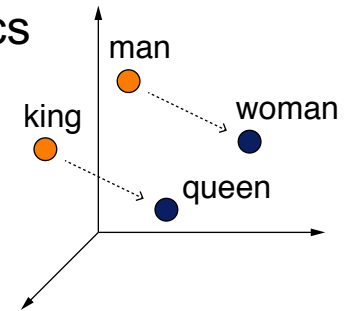
Background: Word embeddings

▪ **Word embedding** (aka word vector)

- A real-valued vector that represents the distributional semantics of a particular word in a high-dimensional space

$$\textit{queen} \rightarrow \mathbf{v}_{\textit{queen}} = (0.13, 0.02, 0.1, 0.4, \dots, 0.22)$$

- Words that occur in similar contexts have similar embeddings.
In other words, similarity can be observed even when different words are used.



▪ **Word embedding model**

- A function that maps each known word to its embedding.
- Derived from a language model, trained on a (usually huge) corpus

The monarchy is ruled by the _____.

- Several pretrained embedding models can be found on the web.
Examples: GloVe, word2vec, BERT, DeBERTa, ...
- Embedding models can also be fine-tuned on a given task.

Background: Neural networks

▪ Neural network

- A layered machine learning model that takes a set of input values and computes one or more output values.
- **Layer.** Composes units that can learn complex functions
- **Unit.** Computes non-linear weighted sums of input values
Applies an activation function (e.g., *tanh*) to the sum, weights learned in training

▪ Input in NLP

- Tokens are represented in the form of embeddings.
- Other, human-defined features can be encoded as “one-hot” vectors.

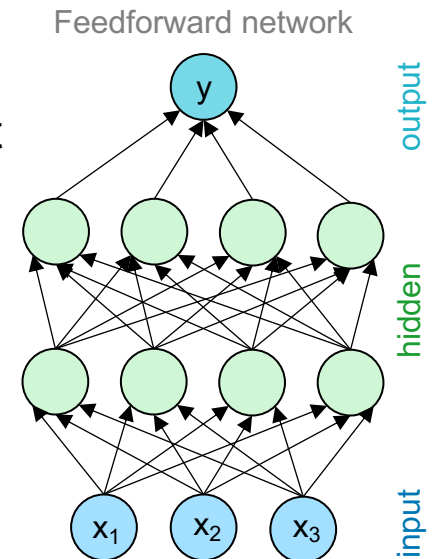
▪ Basic types of neural networks

- **Feedforward networks.** Used for classification and regression
- **Recurrent networks.** Used for sequence labeling and generation

Further neural network architectures follow, such as the transformer.

▪ Notice

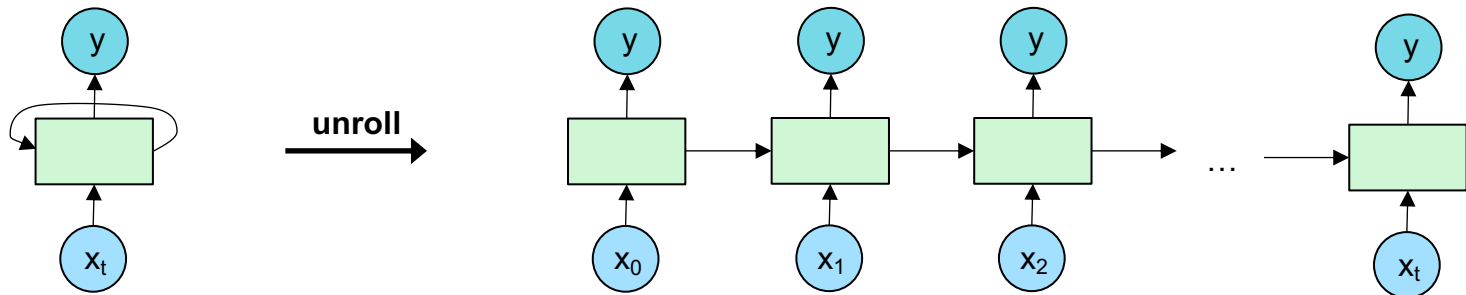
- In this course, neural network concepts are detailed only as far as needed.
For a more technical background on neural networks, see the slides of the course “Stastical NLP”.



Background: Recurrent neural networks (Jurafsky and Martin, 2026)

▪ Recurrent neural network (RNN)

- A network with cycles in its connections, that is, the value of a unit depends on earlier outputs as an input.



- A text is processed by presenting one token at a time to the network.
- The layer from step i serves as memory (or context) for decisions in step $j > i$.

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

▪ Limitations of simple RNNs

- **Unidirectionality.** Only past input is considered, not future input.
- **Limited memory.** Long-term dependencies are hard to learn.

Background: Bi-LSTM neural networks

▪ Bidirectional RNN

- Two RNNs, one processing a text from left to right, one from right to left.

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

- The outputs of both RNNs are combined into a single representation.
- By this, an entire input text can be considered as the context of a token.

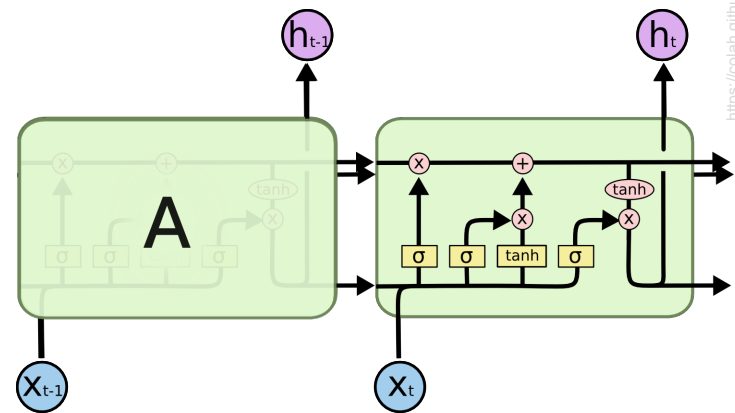
▪ Long short-term memory (LSTM) unit

- A context layer added to a hidden layer that explicitly manages context
- Gates that learn to decide what context to add, to forget and to use for output.

▪ Bi-LSTM

- A bidirectional RNN with LSTM units

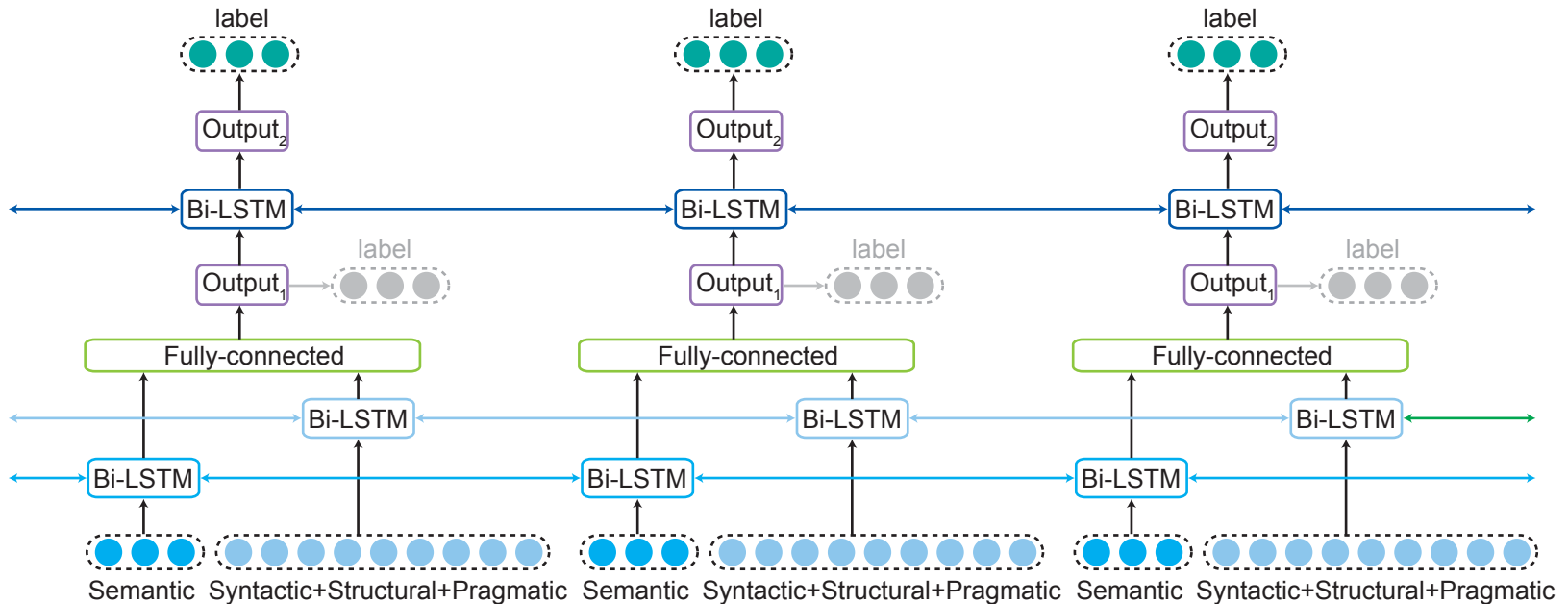
Multiple Bi-LSTMs (as well as other neural networks) can easily be stacked.



<https://colah.github.io>

Bi-LSTMs for unit segmentation: Approach

■ Bi-LSTM-based unit segmentation (Ajjour et al., 2017)



Architecture illustration for three consecutive tokens

- The first Bi-LSTM layers encode semantic features as word embeddings, others as one-hot vectors.
- Another Bi-LSTM layer models interdependencies of consecutive predictions.
- The output layers predict confidence values for the possible labels (B, I, O).

Bi-LSTMs for unit segmentation: Experiments

▪ Baselines

- **SVM.** Linear support vector machine that classifies each token independently
- **CRF.** Linear-chain conditional random field that classifies each token in the context of its $k = 5$ surrounding tokens

▪ Features

- **Semantic.** The token's embedding (for Bi-LSTM) or its text (for SVM, CRF)
- **Structural.** If token is at start, inside, or end of a sentence, clause, or phrase
- **Syntactic.** Part-of-speech tag of the token
- **Pragmatic.** If token is before or after a discourse marker, or in-between two

▪ Data

- **Essays.** 402 student essays (Stab, 2017)
- **News.** 300 news editorials
(Al-Khatib et al., 2016)
- **Web.** 340 forum posts, comments, ...
(Habernal and Gurevych, 2015)

Corpus	B	I	O
Essays	6 089	94 411	44 022
News	14 234	251 381	21 849
Web	1 129	40 042	44 814

Bi-LSTMs for unit segmentation: Experiments

▪ Cross-domain evaluation

- Train model on training set (and optimize on validation set) of one genre
- Apply model to test sets of all three genres

▪ Overall results (token-level macro F_1)

Approach	Test on essays			Test on news editorials			Test on web discourse		
	Essays	News	Web	Essays	News	Web	Essays	News	Web
SVM	61.4	50.9	31.3	58.8	79.9	22.6	39.1	37.4	42.8
CRF	79.2	52.5	21.7	69.8	82.0	8.0	37.1	37.6	37.7
Bi-LSTM	88.5	57.1	37.0	60.7	84.1	20.9	20.9	36.6	54.5

- 88.5 significantly better at $p < .001$ than best result before (86.7) (Stab, 2017)
- In general, cross-genre effectiveness limited

▪ Feature analysis

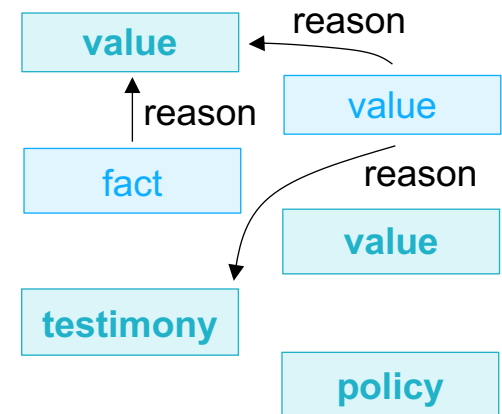
- Semantic features best in-genre (e.g., 87.9 on essays)
- Structural features most genre-robust (e.g., 35.5–39.5 on web discourse)

Biaffine attention for unit type classification

▪ Unit type classification in real-world argumentation

- Often, writers mix different claims and reasons with partial structure only.
- Classifying unit types may require knowledge about the units' relations.

I'm with Massachusetts on this one. Repetitive and robo-calls are annoying and not productive. Another fact about robo-calls is that their messages often start in the middle, or maybe this is done on purpose. When it has happened to me, I just hang up. Policies regulating the number of contacts made within a specific time period should include all modes of technology.



▪ Biaffine attention approach (Morio et al., 2020)

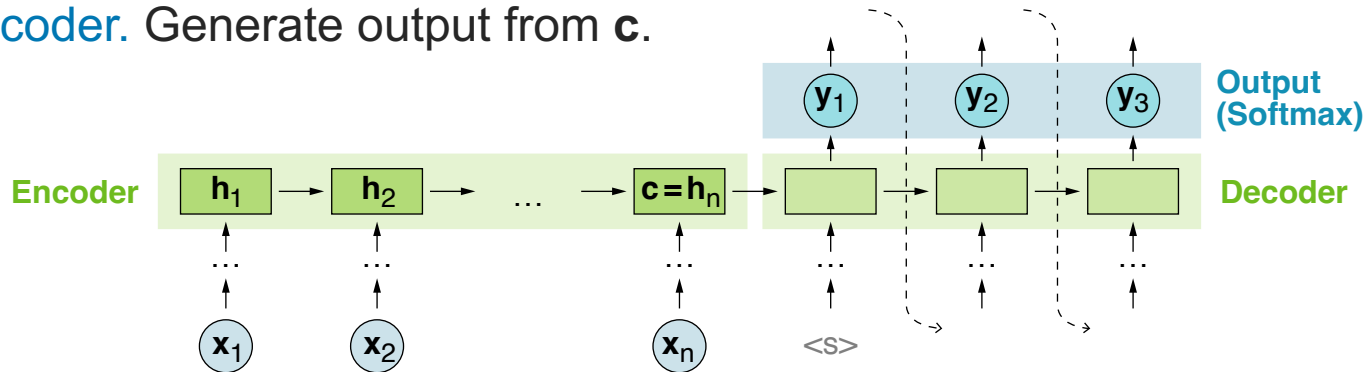
- Jointly classify unit types and identify relations to model interdependencies.
- Bi-LSTMs learn to put attention on related pairs of units.

While both unit types and relations are modeled, the approach could be used for either only.

Background: Attention in RNNs (Jurafsky and Martin, 2026)

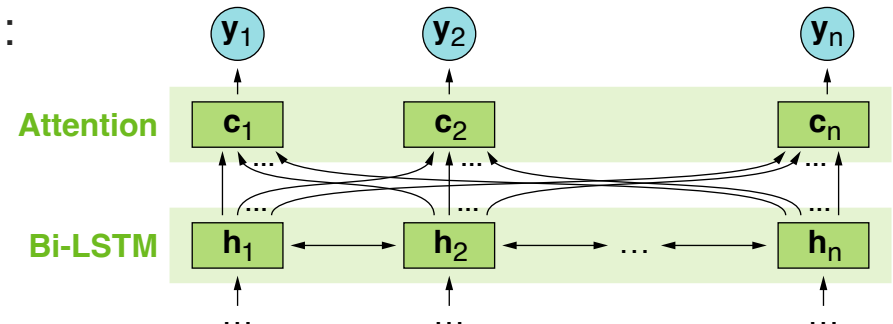
Encoder-decoder RNN

- An RNN that separates input encoding from output decoding
- **Encoder.** Process whole input to create a context representation $\mathbf{c} = \mathbf{h}_n$
- **Decoder.** Generate output from \mathbf{c} .



Attention

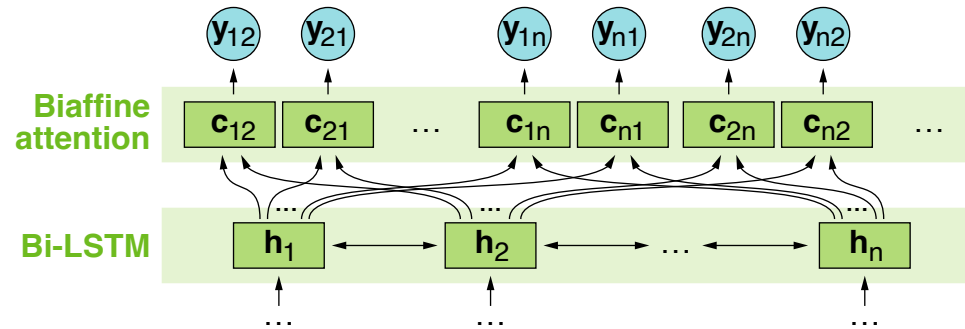
- Retain hidden states to learn which inputs are relevant to which outputs.
- Condition \mathbf{c} on *all* outputs of layer \mathbf{h} :
 $\mathbf{c} := f(\mathbf{h}_1, \dots, \mathbf{h}_n)$.
Often, \mathbf{h} is modeled as a Bi-LSTM layer.
- Use separate context \mathbf{c}_t in each decoding step t .



Background: Biaffine attention (Dozat and Manning, 2017)

▪ Biaffine attention

- Represent all possible pairs of inputs (instead of single inputs).
- Learn the relation of input pairs to outputs.



▪ Notice

- Attention and biaffine attention model input-output connections.
- This resembles how transformers work, but is not the same.

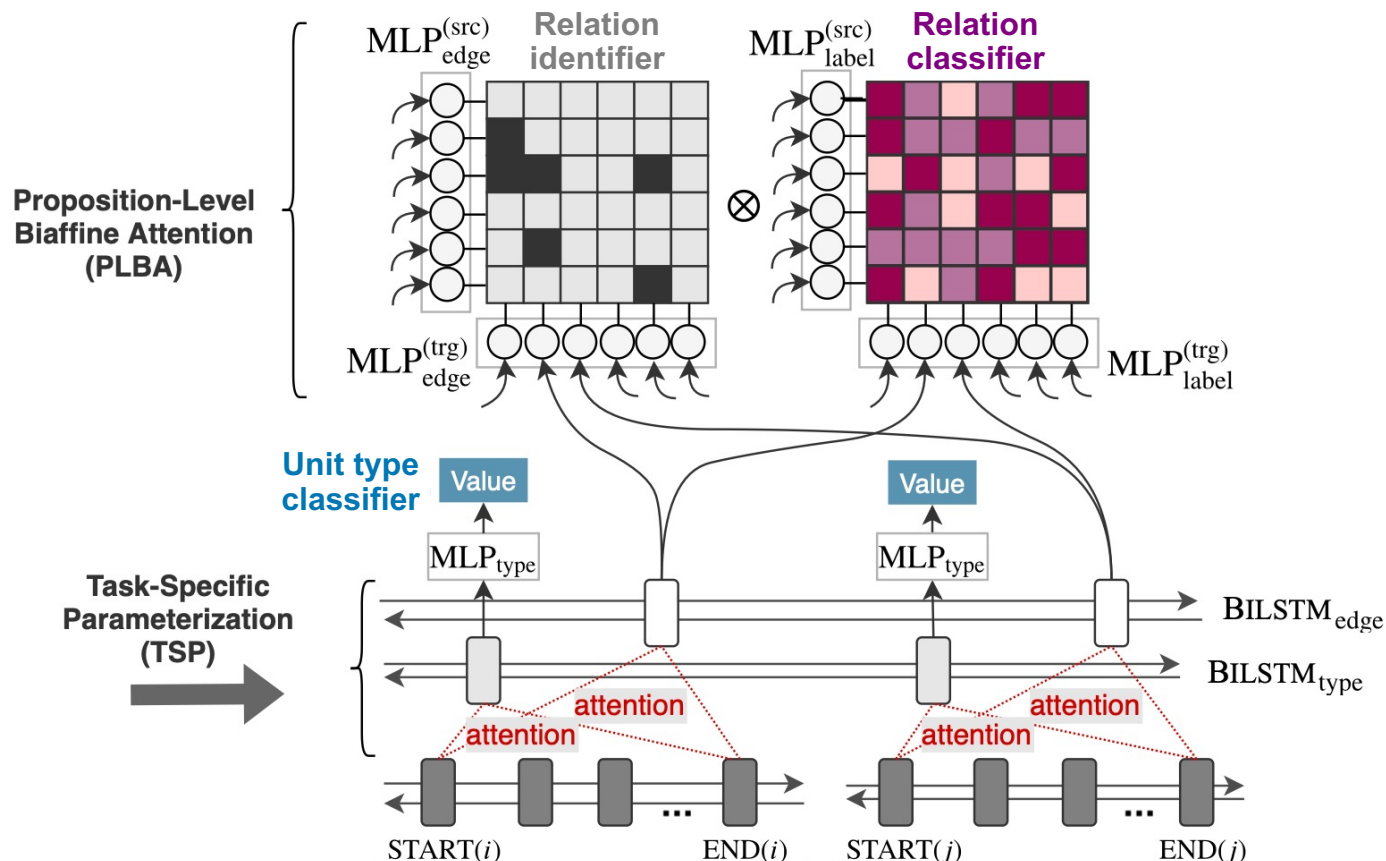
Details later in the course

▪ Unit-level biaffine attention

- Each input is one argumentative unit of the text.
- Relations may exist between any pair of units that also affect the units' types.

Biaffine attention for unit type classification: Approach

- **Biaffine attention for unit type classification** (Morio et al., 2020)
 - **Unit-level biaffine attention.** Model relations and their types
 - **Task-specific parameterization.** Use separate attention layer for unit types



Biaffine attention for unit type classification: Experiments

- **Data** (Park and Cardie, 2018)

- **eRulemaking**. 731 forum comments, labeled for evaluability model

Policy
Value
Fact
Testimony
Reference

- **Baselines** (Niculae et al., 2017)

- **SVM-based graph prediction**. Previous state of the art, jointly predicting entire argument graph structures using SVMs
- **RNN-based graph prediction**. Same idea, prediction with RNN

- **Results** (F_1 -score)

Approach	Unit types	Relations
SVM-based graph prediction	73.2	26.7
RNN-based graph prediction	72.7	14.4
Biaffine attention	78.9	34.0

- Strong improvements for unit types
- Effectiveness on relation identification task unsatisfying

Unit identification: Discussion

▪ **State-of-the-art unit identification**

- The context of units may be critical to assess their argumentativeness.
- High effectiveness is possible, at least in rather explicit argumentative genres.
- Still, minority unit types may be hard to identify accurately.

▪ **Definitions of concepts**

- Units may simply not be the same across genres.
- Units and their types may depend on what is seen as the discussed issue.
- Conceptually, classifying the argumentative *role* of a unit is questionable, as a unit may have different roles in different arguments.

▪ **Ordering of tasks**

- Segmenting first means that no knowledge about what is being argued about.
- Similarly, without relations, deciding about unit types may be pointless.
- Joint approaches may often be preferable in practice.

Outline: Relation identification

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

- a) Introduction
- b) Unit identification
- c) Relation identification**
- d) Conclusion

Relation identification

▪ Relation identification

- The detection and classification of all argumentative relations between the argumentative units in a text
- **Detection.** Given the units, find all pairs that form an argument
- **Classification.** Given the relations, assign each a type (e.g., *support* or *attack*)

“*Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet. One who is living overseas will of course struggle with loneliness, living away from family and friends but those difficulties will turn into valuable experiences in the following steps of life. Moreover, the one will learn living without depending on anyone else.*”

attack

support

example from Stab and Gurevych (2014a)

▪ Challenges

- In some genres, related units may be far away from each other.
- Subtle argumentation leaves relations implicit on purpose.

Relation identification: Detection

▪ What is an argumentative relation?

- Conceptually, an $n:1$ relation with n premise units and 1 conclusion unit
- Many works focus on 1:1 relations.
- Some works also consider relations between complete arguments.

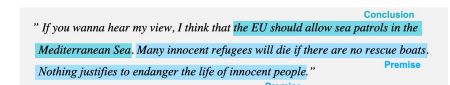
▪ Relations across genres

- **Persuasive essays.** Complete hierarchical tree structure (Stab and Gurevych, 2014a)
- **Forum discussions.** Disconnected argumentative threads (Park and Cardie, 2018)
- **Wikipedia articles.** Relations are issue-dependent as well. (Rinott et al., 2015)
- **News editorials.** Few and often implicit relations (Al-Khatib et al., 2016)

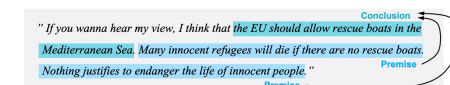
▪ Modeling relation detection

- **Supervised classification** of candidate unit pairs
- **Optimization of the graph** induced by units and relations

Tackling relation detection and classification is rather the default.



↓
**relation
detection**



Relation identification: Classification

▪ What types of relations exist?

- The distinction of support and attack exists across all argumentative genres.
- Some argument models consider different support and attack sub-types.
- Other relation types include reasoning schemes or unit reuse.

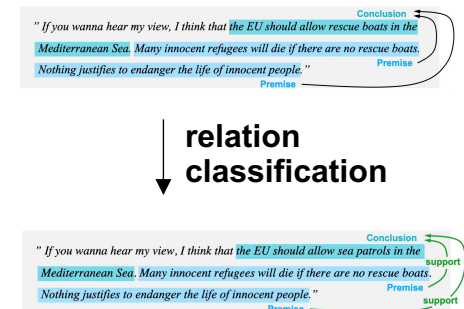
▪ Relation types across models

- **Freeman.** Multiple types of support and attack (Peldszus and Stede, 2013)
- **Walton.** Inference relations of argument schemes (Lawrence and Reed, 2017)
- **Essay-specific.** Simple support and attack (Stab and Gurevych, 2014a)
- **Evaluability-oriented.** Reason and evidence (Park and Cardie, 2018)

▪ Modeling relation classification

- Techniques fairly similar to those used for detection
- Besides, conceptual overlap with stance classification

As indicated above, sometimes also tackled jointly with unit identification



Relation identification: Overview of approaches

▪ Selected joint approaches

- [Maximum spanning tree on classified roles and functions](#) (Peldszus and Stede, 2015)
- Structure learning for graph prediction with SVMs and RNN (Niculae et al., 2017)
- Transition-based parsing using BERT and LSTMs (Bao et al., 2021)

▪ Selected detection approaches

- Topic modeling based on inferential topic pairs (Lawrence and Reed, 2017)
- Transformer-based classification of unit pairs (Poudyal et al., 2020)
- Embedding-based Bi-LSTM with maximum spanning trees (Putra et al., 2021)

▪ Selected classification approaches

- Transformer-based classification with active learning (Hua and Wang, 2022)
- Transformer-based classification across languages/domains (Ruiz-Dolz et al., 2024)
- [Few-shot prompting of large language models](#) (Gorur et al., 2025)

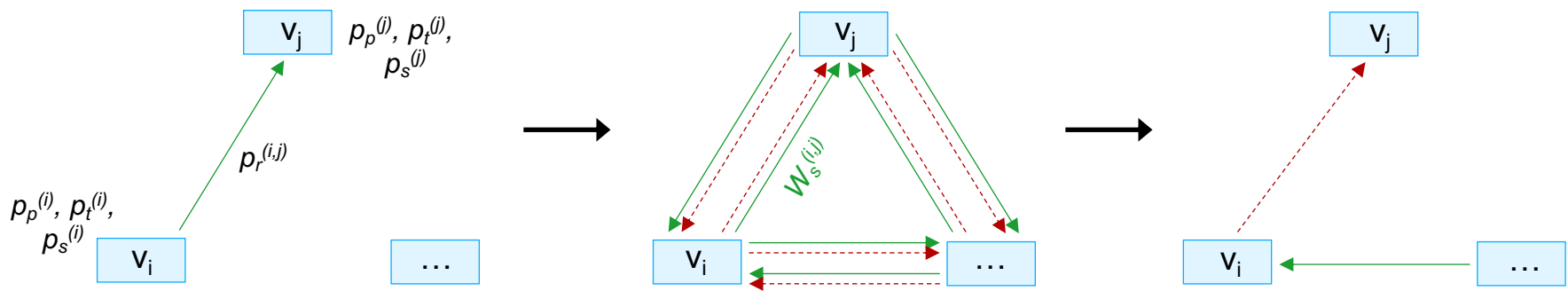
MST-based relation identification

- **Joint unit classification and relation identification**

- **Input.** An argumentative text, segmented into units
- **Output.** Type of each unit (proponent or opponent), relations between units, and type of each relation

- **Presented approach** (Peldszus and Stede, 2015)

- **Classification** of units obtain unit and relation type probabilities
- **Aggregation** of probabilities to obtain evidence graph
- **Determination** of maximum spanning tree (MST) to obtain relations



MST-based relation identification: Data

- **Data** (Peldszus and Stede, 2015)
 - **Arg-microtexts**. 112 texts with 576 units, in both English and German
 - Annotated for Freeman's model, but simplified to (single) support and attack
- **Example**

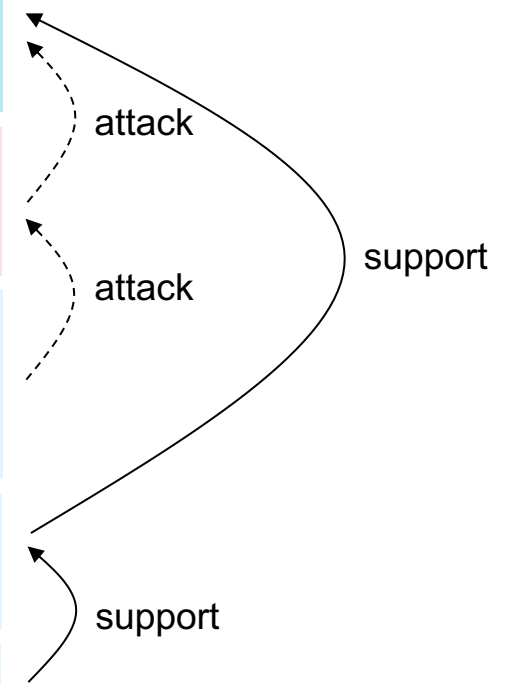
Health insurance companies should naturally cover alternative medical treatments.

Not all practices and approaches that are lumped together under this term may have been proven in clinical trials.

Yet it's precisely their positive effect when accompanying conventional 'western' medical therapies that's been demonstrated as beneficial.

Besides many general practitioners offer such counselling and treatments in parallel anyway

and who would want to question their broad expertise?



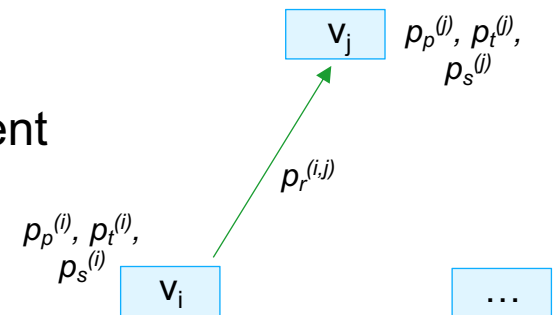
MST-based relation identification: Classification

▪ Unit and relation types

- **Role p .** Whether unit i is on the proponent or opponent side
- **Thesis t .** Whether unit i is a thesis or not
- **Function s .** Whether unit i has a supporting (or attacking) function
- **Relation r .** Whether unit i is in relation to unit j

▪ Supervised classification

- Linear log-loss model with stochastic gradient descent
- Model predicts probabilities $p_p^{(i)}$, $p_t^{(i)}$, $p_s^{(i)}$, and $p_r^{(i,j)}$ for the four types



▪ Employed features

- **Content.** Lemma unigrams
- **Style.** POS tags, dependency parse triples, and discourse connectives
- **Structure.** Length and position of unit, distance and order of unit pair

MST-based relation identification: Aggregation

▪ From node labels to edge labels

- For aggregation, all four probabilities are mapped to edge labels

$$p_p^{(i,j)} := \begin{cases} p_p^{(i)} \cdot p_p^{(j)} + (1 - p_p^{(i)}) \cdot (1 - p_p^{(j)}) & \text{if } (i, j) \text{ support edge} \\ p_p^{(i)} \cdot (1 - p_p^{(j)}) + (1 - p_p^{(i)}) \cdot p_p^{(j)} & \text{if } (i, j) \text{ attack edge} \end{cases}$$

$$p_t^{(i,j)} := 1 - p_t^i \qquad p_s^{(i,j)} := \begin{cases} p_s^{(i)} & \text{if } (i, j) \text{ support edge} \\ 1 - p_s^{(i)} & \text{if } (i, j) \text{ attack edge} \end{cases}$$

▪ Weighted probability aggregation

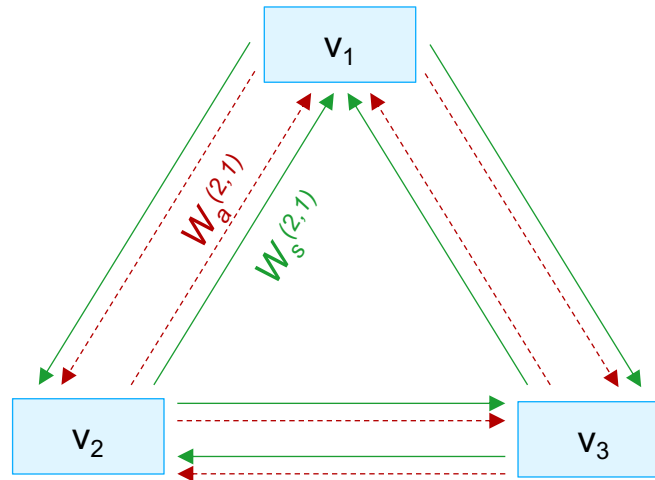
- Add weight to each probability of a given candidate unit pair.
- Learn weights for each probability on training set.

$$w^{(i,j)} = \frac{w_p \cdot p_p^{(i,j)} + w_t \cdot p_t^{(i,j)} + w_s \cdot p_s^{(i,j)} + w_r \cdot p_r^{(i,j)}}{\sum_k w_k}$$

MST-based relation identification: Evidence graph

▪ Evidence graph

- A complete double-connected weighted directed graph $G = (V, E)$
- **Nodes.** Each node v in V represents a unit.
- **Support edges.** Any pair of nodes v_i, v_j is connected with an edge e_s .
- **Attack edges.** Any pair of nodes v_i, v_j is connected with an edge e_a .
- **Weights.** Each e is labeled with a weighted pair score $w^{(i,j)}$ as defined above.

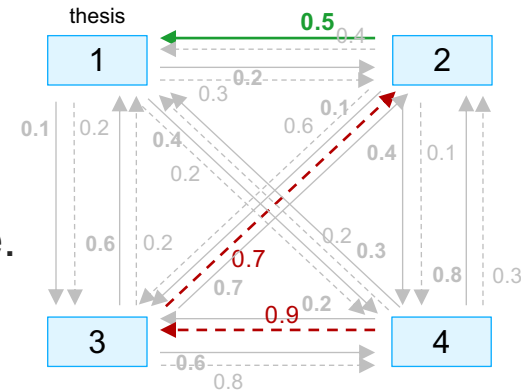


MST-based relation identification: Determination

Maximum spanning tree (MST)

- A subgraph G^* of a weighted graph $G = (V, E)$ whose edges E connect all nodes V with maximum weight
- MSTs have $|V|-1$ edges; the root has no outgoing edge.

For directed graphs, can be found in $O(|E| + |V| \log |V|)$



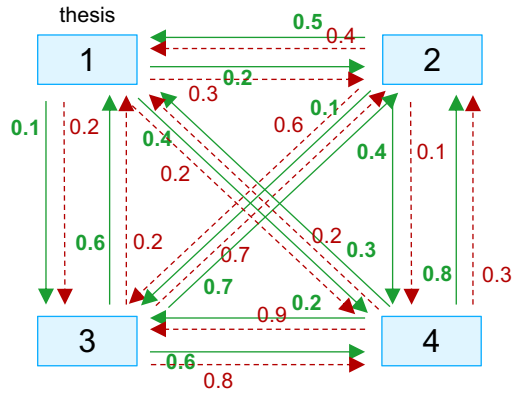
Chu-Liu-Edmonds MST determination

(Chu and Liu, 1965; Edmonds, 1967)

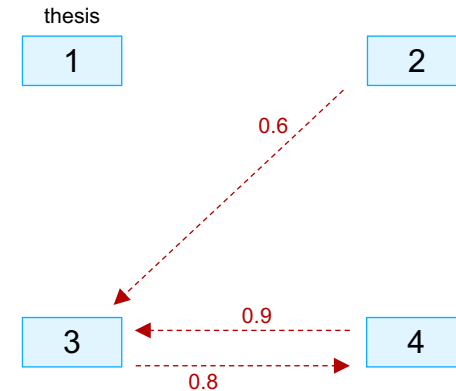
- **Input.** A directed weighted graph $G = (V, E)$ with root v_R in V
 - **Output.** An MST G^* of G
1. Initialize G^* : For each node $v \neq v_R$, add edge (v, v') of v with maximum score.
In case of ties, prefer support edges. Pick randomly from these.
 2. For each cycle: Replace edge (v, v') with minimum score by new edge (v, v'') with maximum score, $v' \neq v''$.
 3. Repeat Step 2 until no more circle exists.
 4. Return G^* .

MST-based relation identification: MST example

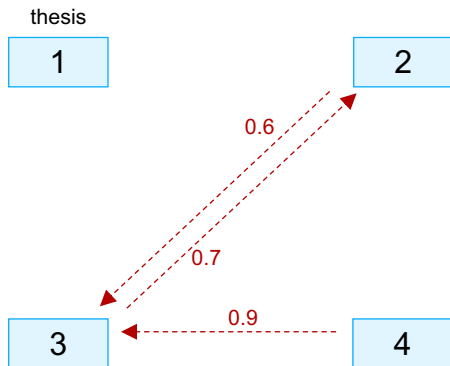
1. Evidence graph G



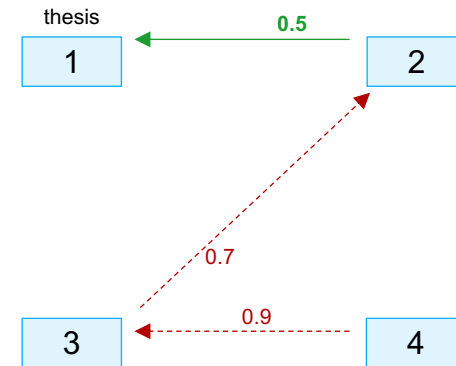
2. Initialize G*



3. Replace edge to remove cycle



4. Replace edge to remove cycle



MST-based relation identification: Experiments

▪ Baselines

- **Classifiers.** Determine whether one unit supports or attacks another, or not.
- **Discourse.** Fine-tuned discourse parser (instead of evidence-based MST)
- **Classifiers + Discourse.** Discourse parser, using classifier outputs as features

▪ Results (macro F_1 -score)

- 5-fold cross validation (10 times repeated), on **English** and **German** test set

Approach	Unit	Relation	Rel. type	Unit	Relation	Rel. type
Classifiers	0.82	0.66	0.67	0.85	0.68	0.70
Discourse	0.78	0.71	0.49	0.83	0.72	0.50
Classifiers + Discourse	0.83	0.72	0.68	0.86	0.73	0.72
Classifiers + MST	0.87	0.69	0.71	0.89	0.71	0.74

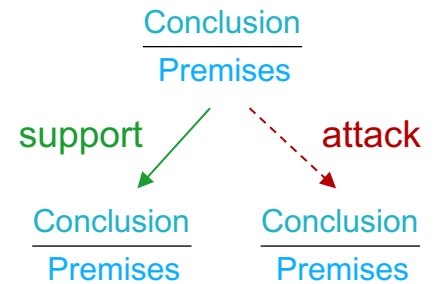
▪ Discussion

- Approach best for unit and relation types (not relations) in both languages
- The MST idea makes sense, if full argumentative structure can be expected.
- Otherwise, some kind of argument decomposition is needed before.

Prompting-based relation identification

▪ Task

- Given two arguments, does the second support or attack the first?



▪ Data

- 10 existing argument corpora, boiled down to binary attack vs. support
- Corpora include persuasive essays, forum arguments, Arg-Microtexts, Kialo debate portal arguments, and more

(Stab, 2017; Park and Cardie, 2018; Peldszus and Stede, 2015)

“ ChatGPT will reach AGI level by 2030. ”

← attack

“ To reach AGI, it should be able to generate its own goals and intentions: where would it draw these from? ”

▪ Presented approach (Gorur et al., 2025)

- Simple use of an instruction fine-tuned large language model (LLM)
- Few-shot prompting to prime the LLM on the task

Background: Large language models (LLMs)

▪ Language model (LM)

- Represents a probability distribution over word sequences, derived from data
- Probabilities can be used to generate most likely *next* words

Prompt: Can you tell me an argument in favor of having rescue boats?

LM: Without such boats, many innocent refugees will <?> →

P(“die” | dialogue) = .04
P(“drown” | dialogue) = .03
P(“suffer” | dialogue) = .01

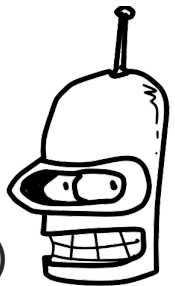
- Its input representation can also be used for classification and regression.

▪ Large language model (LLM)

- Large is not exactly defined, but most LLMs have billions of parameters.
- Mostly, a pretrained transformer LM is meant that follows instructions, that is, it answers to prompts.

▪ Types of LLMs

- **Base.** Transformer-based (GPT-3, BART, ...)
- **Instruct/Chat.** Instruction fine-tuned and aligned (GPT-3.5, LLaMA, ...)

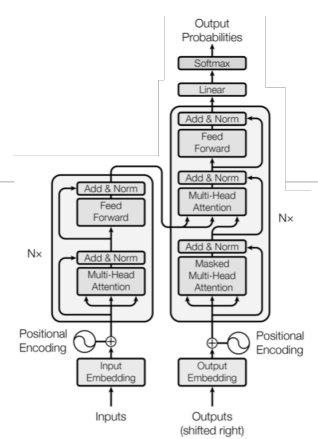


<https://letsdraw.it>

Background: Core concepts of LLMs

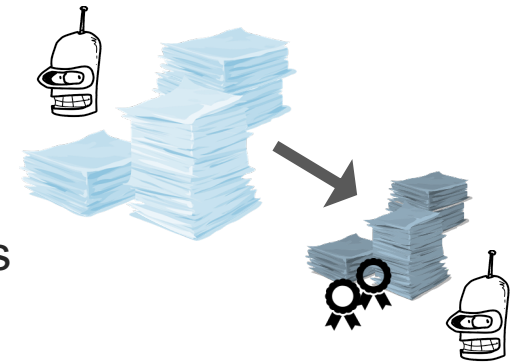
Transformer

- A neural network architecture for parallel input processing
The actual architecture is detailed later in this course.



Transfer learning

- Pretrain** network self-supervised on huge text data.
- Fine-tune** it supervised on task-specific training data.
- Enables LLMs to leverage knowledge across contexts



Instruction fine-tuning (IFT) and alignment

- IFT.** Train LLMs to create answer-like output to any instruction
- Alignment.** Optimize answers towards human-defined preferences



- Enables LLMs to give reasonable answers to nearly any prompt

Background: Prompts and few-shot prompting

■ Prompt

- Input given to an LLM, serving as context for output generation
- **Prompting.** The act of phrasing a prompt to tackle a given task
- **Prompt engineering.** The (manual) tuning of prompts to boost effectiveness

Persona

Imagine you are doing the customer relationship management of a hotel, analyzing what pasts guests think about your hotel.

Task description

You should classify the sentiment polarity of this opinion: <INPUT>

Definition

An opinion is a statement that evaluates a specific aspect of the hotel.

Directional stimulus

You should output one of two label as the polarity: “positive” or “negative”.

Reasoning steps

To do so, first identify the aspect being talked about in the statement. Then, identify what sentiment is expressed towards the aspect and decide whether this is positive or negative for the hotel. The resulting label is

■ Few-shot prompting

- The inclusion of $k \geq 1$ examples of the task (shots) in the prompt
- This affects the LLM’s behavior and how the output looks like.

Shot 1

Opinion: the room was clean and cozy. Polarity: positive

Shot 2

Opinion: this alone never justifies the price. Polarity: negative

Opinion: <INPUT>. Polarity: _____

Prompting-based relation identification: Approach

- **Approach** (Gorur et al., 2025)
 - **Training data.** Start from training set of argument relations.
 - **Few-shot prompting.** Combine four training examples with new instance.
 - **Classification.** Let LLM classify new instance as attack or support.
- **Used LLMs**
 - **Llama 2.** LLM developed by meta; models with 13B and 70B parameters (Touvron et al., 2023)
 - **Mistral.** LLM by French company; models with 7B and 8x7B parameters (Jiang et al., 2024)

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.

Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).

Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.

Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.

Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

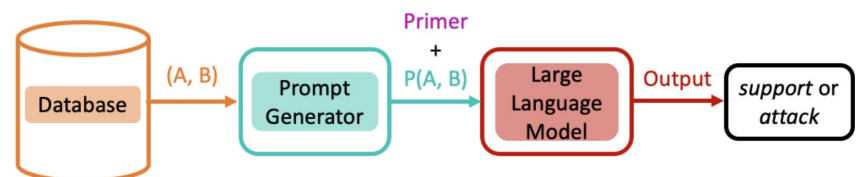
Primer

Prompt

Arg1: Parent Argument (B)

Arg2: Child Argument (A)

Relation:



Prompting-based relation identification: Experiments

- **Baseline**

- **RoBERTa**. Transformer-based encoder, fine-tuned on one dataset (Kialo)

- **Results** (macro F1-score)

Approach	Essays	CDCP	Arg-Microtexts	...	Kialo
RoBERTa	0.80	0.75	0.67	...	n/a
Llama 2 (13B)	0.82	0.87	0.67	...	0.65
Llama 2 (70B)	0.90	0.92	0.73	...	0.86
Mistral (7B)	0.85	0.75	0.67	...	0.83
Mistral (8x7B)	0.89	0.93	0.70	...	0.84

- **Discussion**

- The approach is as simple as it gets with LLM prompting.
- Prompt engineering could further evolve such an approach.
- The binary decision presupposes that the arguments are in relation.
- Still, results indicate general impact of LLMs.

Relation identification: Discussion

▪ **State-of-the-art relation identification**

- The best approaches detect and classify all units and relations jointly.
- LLMs are not straightforward to apply to argument mining tasks.
- Identification often works reasonably but not fully reliable in general.

▪ **Definition of concepts**

- In some genres, relations are explicitly marked by linguistic indicators.
- In others, relations may be hidden and hard to agree on, even for humans.
- Also relations may depend on what is seen as the discussed issue.

▪ **Difference to stance**

- Attack/support relations and pro/con stance conceptually overlap.
- Unlike relations, stance refers to the author's position on an issue.
- Still, some works model support/attack as pro/con premises.

Outline: Conclusion

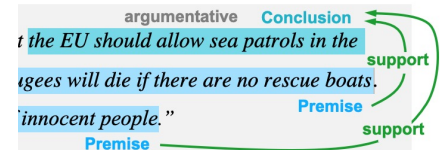
- I. Introduction to computational argumentation
- II. Basics of natural language processing
- III. Basics of argumentation
- IV. Argument mining**
- V. Perspective assessment
- VI. Argument quality assessment
- VII. Argument generation
- VIII. Applications of computational argumentation
- IX. Conclusion

- a) Introduction
- b) Unit identification
- c) Relation identification
- d) Conclusion**

Conclusion

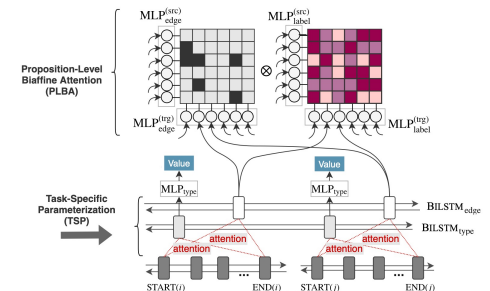
Argument mining

- Computational identification of argumentative structure
- Aims at different types of units and their relations
- May be based on different argument models



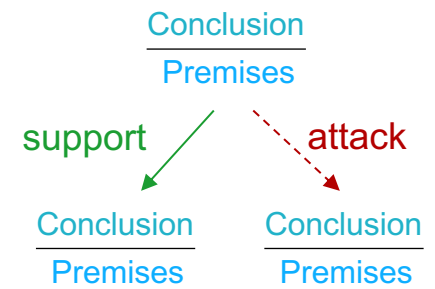
Selected mining approaches

- Unit segmentation using Bi-LSTMs
- Unit type classification using biaffine attention
- Relation identification using MSTs or LLM prompting



Effectiveness of argument mining

- May work pretty reliable within narrow, explicit genres
- Hard on subtle argumentation and across genres
- Simple argument models may enable more robustness



References

- **Ajjour et al. (2017)**. Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit Segmentation of Argumentative Texts. In Proceedings of the Fourth Workshop on Argument Mining, pages 118–128, 2017.
- **Al-Khatib et al. (2016)**. 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.
- **Al-Khatib et al. (2017)**. Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of Argumentation Strategies across Topics. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1362–1368, 2017.
- **Bao et al. (2021)**. Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. A Neural Transition-based Model for Argumentation Mining. 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 6354–6364, 2021.
- **Chen et al. (2024)**. Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. Exploring the Potential of Large Language Models in Computational Argumentation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2309–2330, 2024.
- **Chu and Liu (1965)**. Y. J. Chu and T. H. Liu. 1965. On the Shortest Arborescence of a Directed Graph. *Science Sinica*, 14:1396–1400.
- **Dozat and Manning (2017)**. Timothy Dozat and Christopher D. Manning. Deep Biaffine Attention for Neural Dependency Parsing. In Proceedings of the 5th International Conference on Learning Representations, 2017.
- **Eger et al. (2017)**. Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11–22, 2017.

References

- **Freeman (2011)**. *Argument Structure: Representation and Theory*. Springer, 2011.
- **Gorur et al. (2025)**. Deniz Gorur, Antonio Rago, and Francesca Toni. Can Large Language Models perform Relation-based Argument Mining? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534, 2025.
- **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 (2017)**. Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1), pages 125–179, 2017.
- **Hua and Wang (2022)**. Xinyu Hua and Lu Wang. Efficient argument structure extraction with transfer learning and active learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 423–437, 2022.
- **Jiang et al. (2024)**. Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. *Mixtral of experts*, 2024.
- **Jurafsky and Martin (2026)**. Daniel Jurafsky and James H. Martin (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 3rd edition draft. <https://web.stanford.edu/~jurafsky/slp3/>

References

- **Lawrence and Reed (2017).** John Lawrence and Chris Reed. Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models. In Proceedings of the 4th Workshop on Argument Mining, pages 39–48, 2017.
- **Morio et al. (2020).** Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. Towards Better Non-Tree Argument Mining: Proposition-Level Biaffine Parsing with Task-Specific Parameterization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3259–3266, 2020.
- **Niculescu et al. (2017).** Vlad Niculescu, Joonsuk Park, and Claire Cardie. Argument Mining with Structured SVMs and RNNs. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 985–995, 2017.
- **Park and Cardie (2018).** Joonsuk Park and Claire Cardie. A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018.
- **Poudyal et al. (2020).** Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. ECHR: Legal corpus for argument mining. In Elena Cabrio and Serena Villata, editors, Proceedings of the 7th Workshop on Argument Mining, pages 67–75, 2020.
- **Putra et al. (2021).** Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. Multi-task and multi-corpora training strategies to enhance argumentative sentence linking performance. In Khalid Al-Khatib, Yufang Hou, and Manfred Stede, editors, Proceedings of the 8th Workshop on Argument Mining, pages 12–23, 2021.
- **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.

References

- **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.
- **Persing and Ng (2016)**. Isaac Persing and Vincent Ng. End-to-End Argumentation Mining in Student Essays. In Proceedings of NAACL-HLT 2016, pages 1384–1394, 2016.
- **Ruiz-Dolz et al. (2024)**. Ramon Ruiz-Dolz, Chr-Jr Chiu, Chung-Chi Chen, Noriko Kando, and Hsin-Hsi Chen. Learning strategies for robust argument mining: An analysis of variations in language and domain. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pages 10286–10292, 2024.
- **Stab (2017)**. Christian Stab. Argumentative Writing Support by means of Natural Language Processing, Chapter 5. PhD thesis, TU Darmstadt, 2017.
- **Toulmin (1958)**. Stephen E. Toulmin. The Uses of Argument. Cambridge University Press, 1958.
- **Touvron et al., (2023)**. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.