

Computational Argumentation — Part VI

Argument Generation

Henning Wachsmuth

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



Learning goals

▪ Concepts

- Selected basics of natural language generation
- Views of core building blocks of an argument
- Basics of transformers and neural language models



<https://commons.wikimedia.org>

▪ Methods

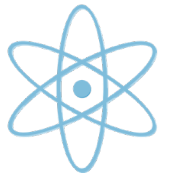
- Knowledge-based and neural composition and creation of arguments
- Neural language models for rewriting and countering arguments
- Extractive and abstractive summarization of argumentative texts



<https://pixabay.com>

▪ Associated research fields

- Natural language processing



<https://pixabay.com>

▪ Within this course

- How to reuse mined and assessed arguments in new arguments and how create fully new arguments



Outline: Introduction

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

- a) **Introduction**
- b) Argument composition and creation
- c) Argument rewriting and countering
- d) Argument and debate summarization
- e) Conclusion

Argument generation

▪ Argument generation

- The synthesis of new argumentative units, arguments, and argumentative texts

We use *synthesis* and *generation* largely interchangeably here.

▪ Argument generation tasks

- Encoding of knowledge in a new unit
- Reconstruction of implicit units
- Composition of units in an argument
- Creation of a new argumentative text
- Modification of existing units or arguments
- Writing of a summary of one or more texts

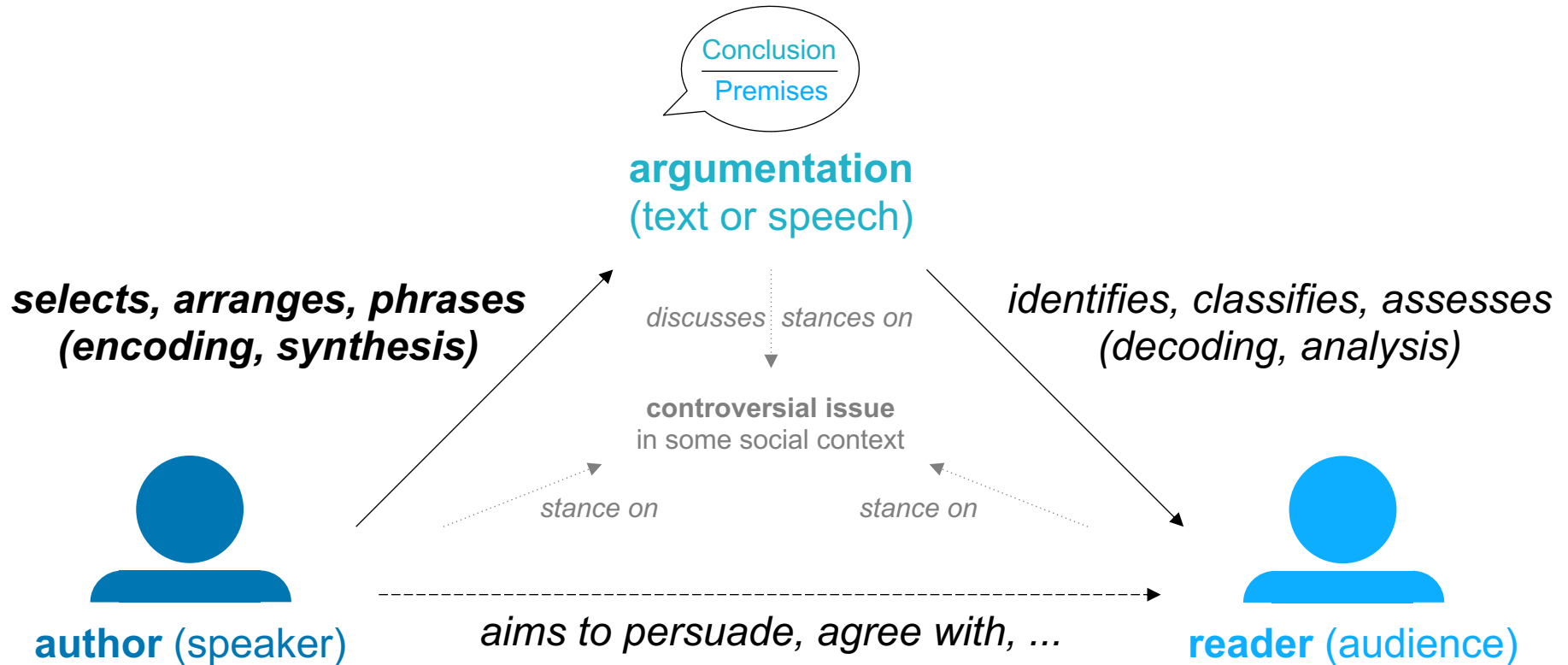
... along with variations of these

▪ Why argument generation?

- Technologies such as Project Debater should be able to form new arguments.
- Computers may be able to find new argumentative connections.



General argumentation setting (recap)



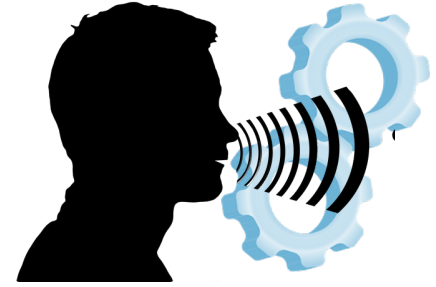
■ Generation vs. mining and assessment

- Argument generation refers to the encoding/synthesis side.
- Still, mining and assessment may be required to decide what to generate.
e.g., starting from what the opponent argued before

Natural language generation (NLG)

▪ Natural language generation (NLG)

- Algorithms for the synthesis of natural language (text)
- The goal is to encode structured or semi-structured information in an unstructured text



<https://de.wikipedia.org>
<https://pixabay.com>

▪ Two general types of NLG

- **Data-to-text.** Phrase a new text with data from a knowledge base.
- **Text-to-text.** Write an output text in response to a given input text.

▪ What makes NLG challenging?

- NLG requires to *choose* and *create* a specific text among many potential candidate texts.
- Its output needs to be relevant, coherent, truthful, and much more.

▪ Disclaimer

- Only a high-level introduction to selected NLG techniques is given below; more may be needed for working with NLG in general.

NLG process and techniques

- **A full NLG process** based on Reiter and Dale (1997)

- **Input.** A goal of what to generate, and knowledge represented in some way
- **Output.** A natural language text



- **Main steps**

- **Text planning.** Select content, arrange the discourse structure of sentences
- **Sentence planning.** Aggregate sentence content, make lexical choices, build referring expressions, ...
- **Linguistic realization.** Orthographic, morphological, and syntactic processing

- **NLG techniques detailed below**

- Composition, language modeling, summarization, and more
- Different techniques may also have to be combined adequately.

Evaluation of NLG

▪ How to evaluate NLG?

- **Goal.** Judge quality of generated texts.
- **Problem.** There is not *the* correct output.

Ground truth. “*Ban death penalty.*”

Generated text. “*We should ban the death penalty forever.*”

▪ Two types of NLG evaluation (details below)

- **Automatic.** Quantify similarity between ground-truth text and generated text based on their words or embeddings.
- **Manual.** Assess quality dimensions of generated texts with humans.

Another, recently trending approach is to use LLM judges for (additional) automatic evaluation.

▪ Main criticisms of automatic evaluation metrics

- **Uninterpretability.** Errors are not distinguishable, not all “errors” are wrong.
- **Unreliability.** Automatic and human assessment often do not correlate.

▪ Dilemma of evaluation

- Only manual evaluation is seen as reliable, but it costs time and money.
- Automatic evaluation is needed to observe progress during development.

Evaluation of NLG: Automatic metrics

Selected automatic metrics

- **BLEU**. Precision of n -gram overlap with brevity penalty
- **ROUGE**. Recall of n -gram overlap, either for a specific n or averaged
- **BERTScore**. F_1 -score derived from similarity matching of BERT embeddings

BiLingualEvaluation Understudy (BLEU) score (Papineni et al., 2002)

- Let $D_{gt}^{(i)}$ be the set of ground-truth n -grams for a test text i , $i \in \{1, \dots, m\}$
- Let $D_{gn}^{(i)}$ be the set of generated n -grams for a test text i .
- **Modified n -gram precision**. Fraction of each $D_{gn}^{(i)}$ that matches any n -gram in $D_{gt}^{(i)}$, counting each n -gram in $D_{gt}^{(i)}$ once only
- **Brevity penalty**. Prevents high scores for short texts to account for recall

Usually, $n \leq 4$ is used (or $n \leq 2$), case sensitivity is ignored. BLEU scores are in $[0,1]$, sometimes multiplied by 100.

$$BLEU := \frac{1}{m} \sum_{i=1}^m \left(\exp \left(\frac{1}{4} \sum_{n=1}^4 \log \frac{\#matches(D_{gn}^{(i)}, D_{gt}^{(i)})}{\#ngrams(D_{gn}^{(i)})} \right) \cdot \exp \left(\min \left\{ 1 - \frac{\#words(D_{gt}^{(i)})}{\#words(D_{gn}^{(i)})}, 0 \right\} \right) \right)$$

Diagram illustrating the BLEU score formula components:

- sum over test set
- Geometric mean, equal weight for each n
- modified precision of n -grams for ext i
- brevity penalty, if ground-truth text longer than generated text

Evaluation of NLG: Manual evaluation

- **Manual evaluation**

- Multiple human annotators assess the quality of a sample of generated texts.

- **Assessment**

- Absolute scores on a Likert scale (say, 1–5) or **relative ranking** of candidates

	Arg'ness	Style	Clarity	Rank
a. <i>“We should ban the death penalty forever.”</i>	5	4	4	#1
b. <i>“Death penalty is stupid.”</i>	3	2	5	#3
c. <i>“Anything good about the death penalty?”</i>	2	4	4	#2

- The mean or majority judgment of multiple annotators is used for evaluation.
- As for corpora, inter-annotator agreement can be used to assess reliability.

- **Quality dimensions** (van der Lee et al., 2019)

- What dimensions to be assessed, is to some extent task-specific.
- Some dimensions are very common, e.g., fluency or meaning preservation.

Outline: Argument creation and composition

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

- a) Introduction
- b) Argument composition and creation**
- c) Argument rewriting and countering
- d) Argument and debate summarization
- e) Conclusion

Argument composition and creation

▪ Argument composition

- The synthesis of argumentative units, arguments, and full argumentative texts from existing building blocks
- **Input.** An issue, possibly a stance, and possibly knowledge given in some way
- **Output.** A text arguing on the given issue

Pro
rescue boats



“If you wanna hear my view, I think that the EU should allow rescue boats in the Mediterranean Sea. Many innocent refugees will die if there are no such boats. While having such boats may make even more people die trying, nothing justifies to endanger the life of innocent people. Got it?”

▪ Argument creation

- The generation of new argument units, arguments, or full argumentative texts
- **Input/Output.** As above

Con
rescue boats



“Having rescue boats makes even more people die trying.”

Argument composition and creation: Examples

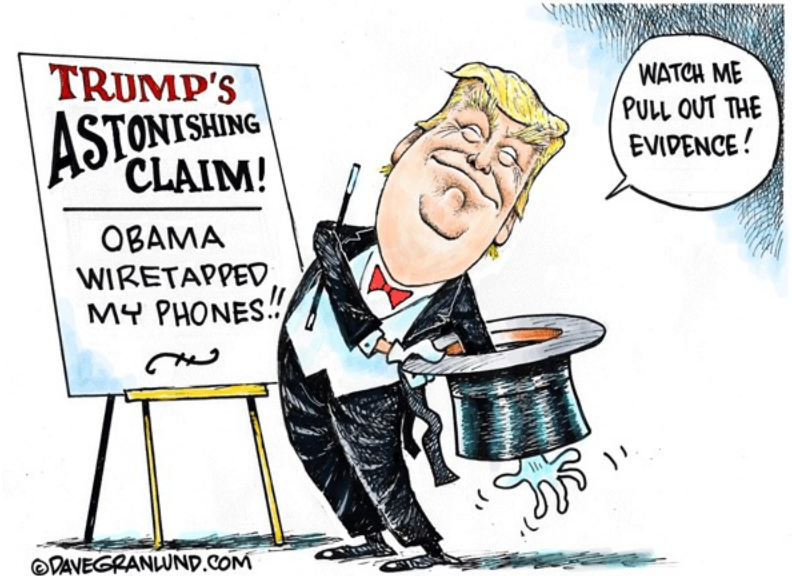
▪ Example: Reason creation

- Given the following claim, phrase a meaningful reason for it.

A university degree is important for your career.



Employers look at what degree you have first.



▪ Challenges

- Knowledge bases might not contain suitable concepts for everything.
- Connections between different concepts build on world knowledge.
- Content, reasoning, and stance all need to be encoded properly.
- Linguistic adaptations of grammar may be necessary.

Argument composition and creation: Approaches

▪ Variations of argument composition and creation

- **Template filling.** Fill slots with concepts from knowledge base or other texts.
- **Language modeling.** Generate free text based on trigger concepts.
- **Controlled generation.** Generate free text that fulfills specified constraints.

▪ Selected approaches to argument composition

- Discourse planning for argumentative texts (Zukerman et al., 2000)
- **Predicate recycling for composing new claims** (Bilu and Slonim, 2016; 2019)
- Language modeling for rhetorical argument composition (El Baff et al., 2019)

▪ Selected approaches to argument creation

- **Neural target inference in conclusion generation** (Alshomary et al., 2020a)
- Neural knowledge encoding in argument generation (Al-Khatib et al., 2021)
- **Transformer-based generation of conclusions for assessment** (Gurcke et al., 2021)
- Conditioned neural generation of claims with beliefs (Alshomary et al., 2021a)
- Essay generation via prompting-based discourse planning (He et al., 2024)

Background: NLG via composition

▪ What is meant by composition?

- The generation of a text by selecting and arranging existing text fragments
- Phrasing is adjusted only to account for grammaticality and coherence.

Examples: Change from singular to plural, addition of discourse markers, capitalization

▪ How to compose?

- Simple rule-based techniques start from sentence and discourse templates whose slots are filled with information.

“I am <stance> <issue>, because <reason>.”

Issue. Death penalty

Stance. Con

Reason. *“The death penalty kills people”*

*“I am con death penalty,
because the death
penalty kills people.”*

- Composition can also be learned and encoded in statistical models, e.g., in language models.

Predicate recycling for claim synthesis (Bilu and Slonim, 2016)

▪ Task

- Given a target, generate a claim on the target.

great anarchy *a global language* *nuclear weapons*
democratization

▪ Data

- For 67 iDebate topics, 28 claims derived from Wikipedia.
- Each claim labeled as good or bad 5 times (majority label used)
“Good” means coherent and relevant to the topic.

▪ Approach

- **Identification.** Parse existing claims to extract stance-bearing predicates.
- **Composition.** Sort predicates by similarity to target, then construct candidate target-predicate pairs.
Linguistic adaptations via an off-the-shelf library
- **Selection.** Score each candidate using regression.
Features: Length, n -gram matches with Wikipedia, ...

cause lung cancer
contribute to stability *is a source of conflict*
lead to great exhaustion

Democratization
contributes to stability.

Nuclear weapons
cause lung cancer.

▪ Results

- Mean precision 0.93@1, 0.75@10

Target inference in conclusion generation (Alshomary et al., 2020a)

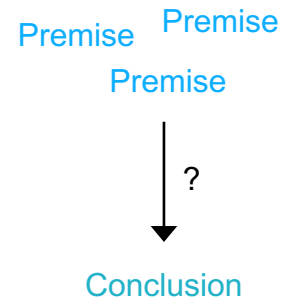
▪ Task

- Given the premises of an argument, infer (the target of) its conclusion.

Only the *target* inference tackled in this work

- **Motivation.** Humans often leave parts of arguments implicit.

Particularly, conclusions often left out (Habernal and Gurevych, 2015)



▪ Hypothesis

- The conclusion target is related to the targets of the premises.

▪ Data

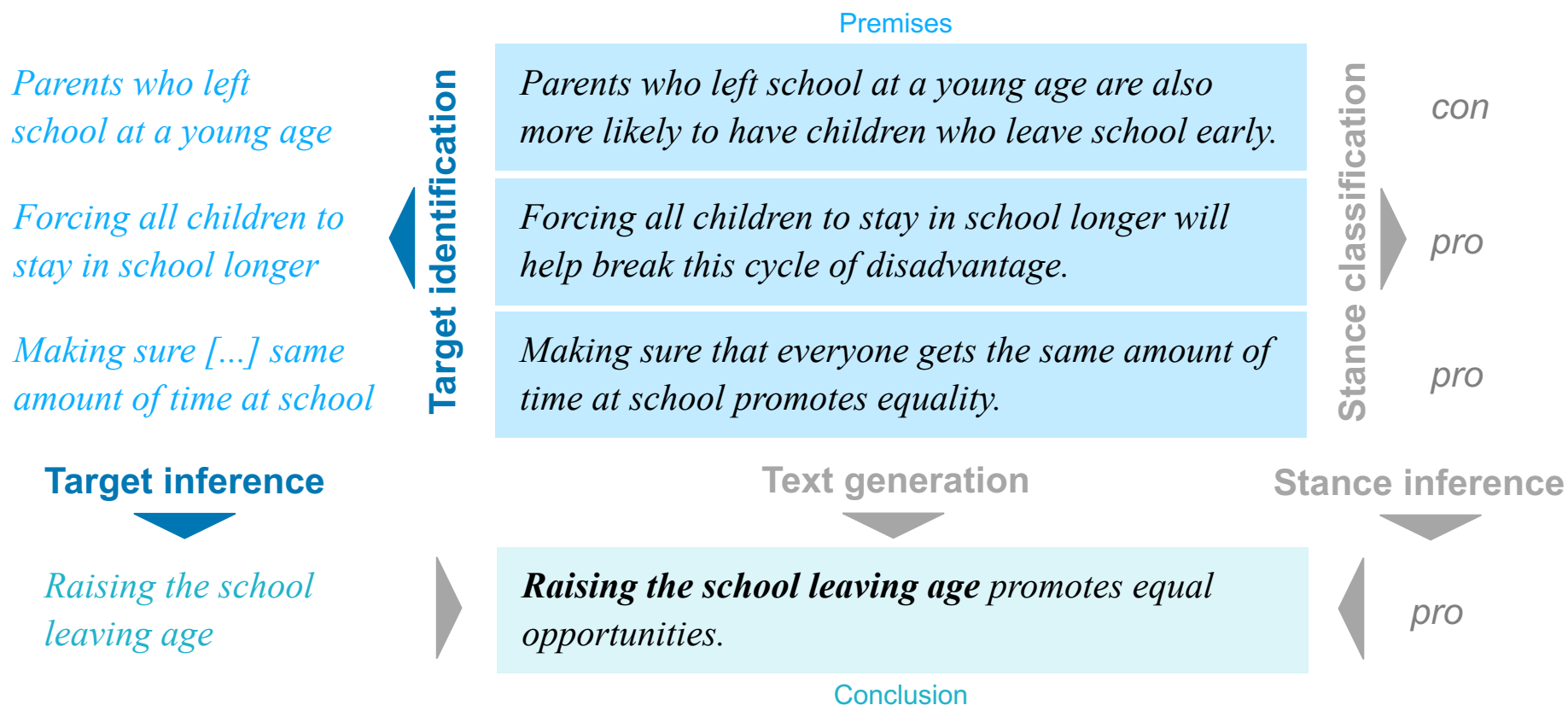
- **iDebate.** 2259 premise-conclusion arguments (Wang and Ling, 2016)
- **Essays-c.** 2020 premise-conclusion arguments (Stab, 2017)
- **Essays-t.** 402 conclusion-thesis arguments (Stab, 2017)

Each split into training, validation, and test set

▪ Two complementary approaches

- **Either,** rank identified premise targets by their representativeness.
- **Or,** match generated target embedding with target knowledge base.

Target inference in conclusion generation: Overall idea



▪ Target identification

- State-of-the-art tagger trained on existing data (Bar-Haim et al., 2017; Akbik et al., 2018)

Forcing all children to stay in school longer will help break this cycle of disadvantage .

B | | | | | | | | | | | | | | | | | | | |

Target inference in conclusion generation: Ranking

- **Inference hypothesis H_1**

- One of the premise targets represents an adequate conclusion target

- **Approach a_1 : Premise target ranking**

- **Model.** Prediction of a representativeness score for each candidate target
Trained on Jaccard similarity of ground-truth premise and conclusion targets, following Wang and Ling (2016)
- **Features.** Length, position, sentiment, and similarity to other candidates
- **Inference.** Pick most representative premise target

*Parents who left
school at a young age* 0.3

*Forcing all children to
stay in school longer* 0.7 ~

*Raising the school
leaving age*

*Making sure [...] same
amount of time at school.* 0.2

- **Implication**

- A target that is not given in the premises can never be predicted.

Target inference in conclusion generation: Learning

- **Inference hypothesis H_2**
 - The premise targets are semantically related to adequate conclusion targets

- **Approach a_2 : Embedding learning**

- **Learn** to map premise target embeddings to conclusion target embedding.

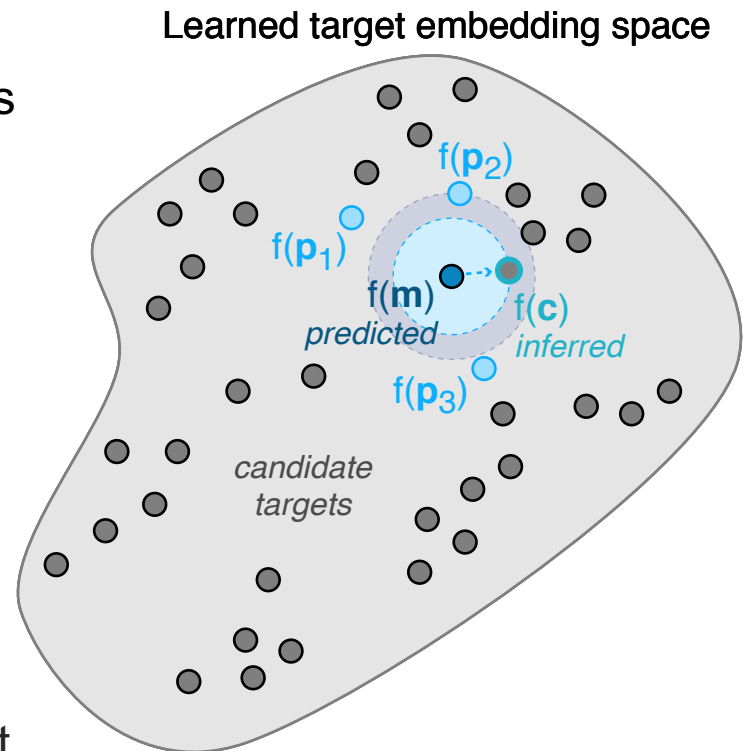
Details on next slides

- **Embed** candidate targets from some knowledge base.
- **Pick** candidate target whose embedding is closest to premise targets.

- **Implications**

- Guarantees to obtain a meaningful target
- Depends on quality of target knowledge base

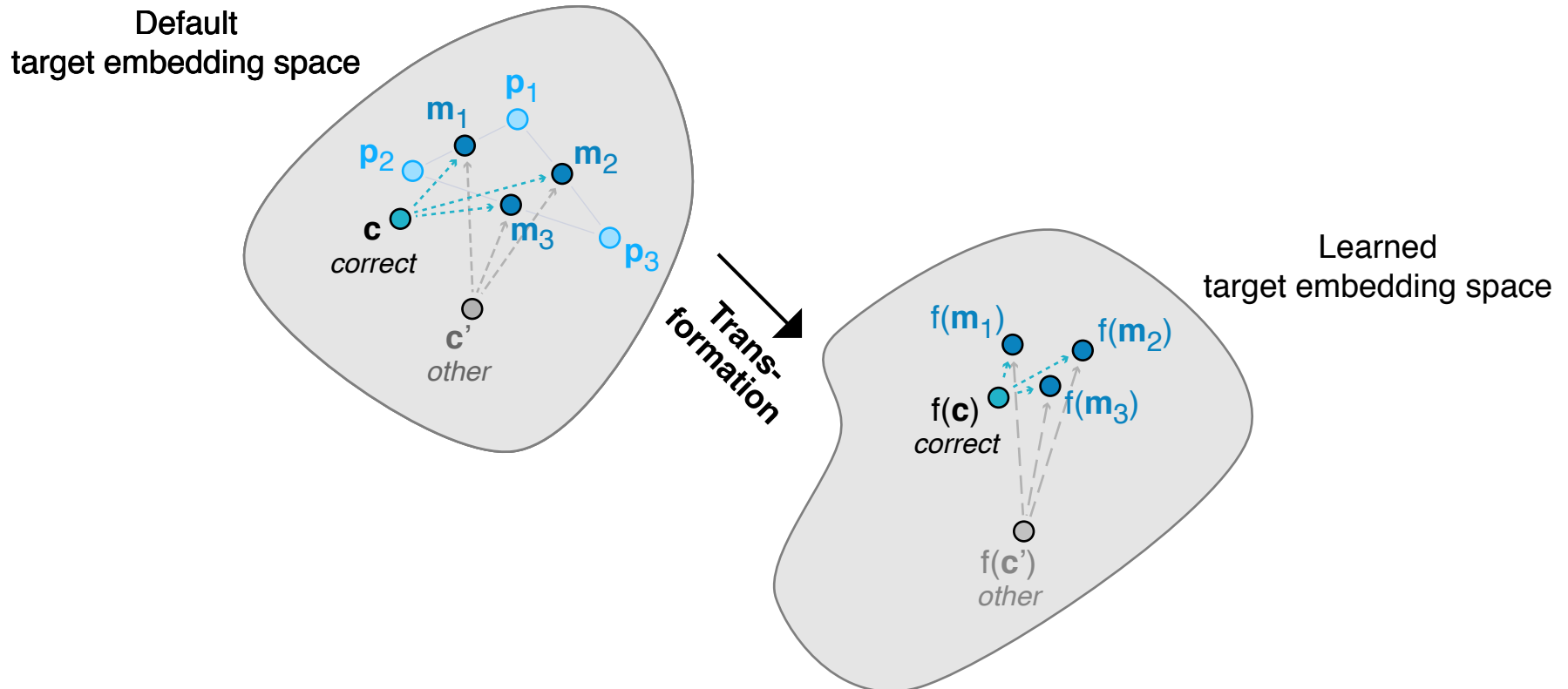
In the experiments, targets identified in training arguments used



Target inference in conclusion generation: Embeddings

▪ Contrastive learning of target embeddings

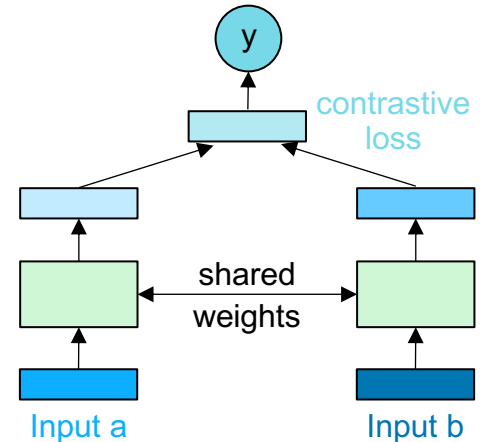
- **Compute** means $\mathbf{m}_1, \dots, \mathbf{m}_l$ of premise target embeddings $\mathbf{p}_1, \dots, \mathbf{p}_k$.
- **Learn** model f that makes \mathbf{m}_i more similar to correct conclusion target $f(\mathbf{c})$ and less similar to other targets $f(\mathbf{c}')$.



Background: Siamese and triplet neural networks

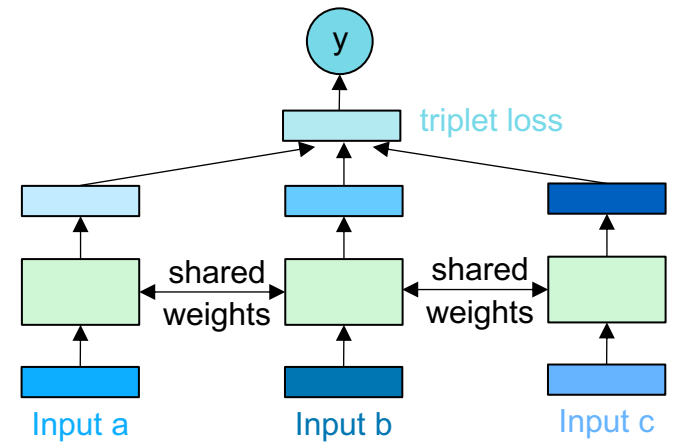
▪ Siamese neural network (SNN)

- Two networks sharing the same weights to transform two inputs a and b into two outputs
- The difference of outputs is quantified as a distance d .
- **Contrastive loss function.** The learning objective is to minimize d between similar inputs and to maximize it for dissimilar inputs.



▪ Triplet neural network (TNN)

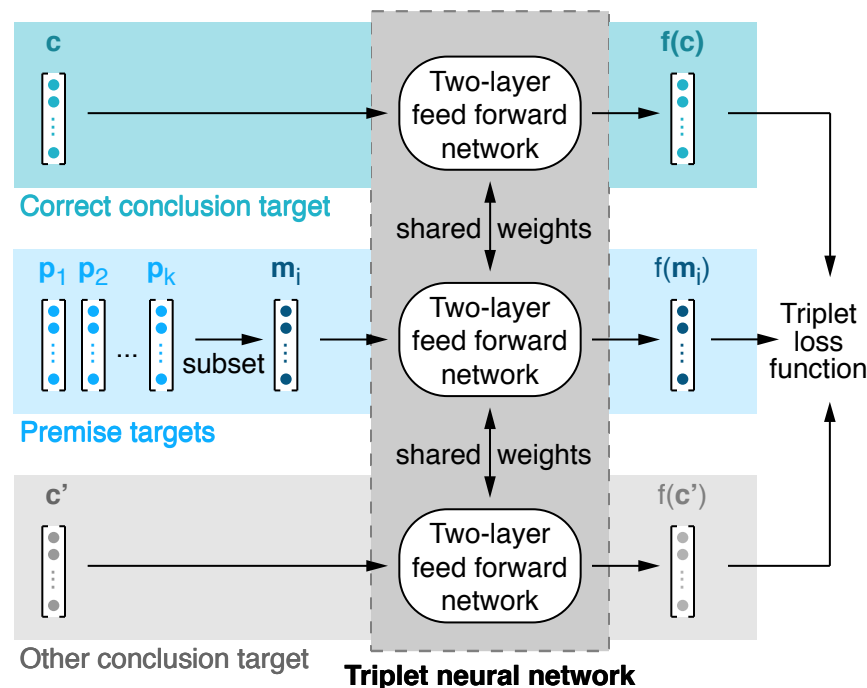
- A TNN follows a similar idea to an SNN for three inputs a , b , c .
- One input (say, b) is used to define distance.
- **Triplet loss function.** The learning objective is to minimize the distance d between a and b and to maximize it for b and c .



Target inference in conclusion generation: Optimization

- How to learn the mapping

- Train TNN on targets from complete arguments



- Optimize loss function based on distance to correct and to other target:

$$\mathcal{L} = \max \left\{ \underbrace{d(f(m_i), f(c))}_{\text{Distance to correct target}} - \underbrace{d(f(m_i), f(c'))}_{\text{to wrong target}} + \underbrace{d_{max}}_{\text{targeted maximum (hyperparameter)}}, 0 \right\}$$

Target inference in conclusion generation: Results

▪ Baselines and hybrid approach

- Seq2Seq*. Summarize premises, tuned to their targets. (Wang and Ling, 2016)
- Premise target (random). Pick one premise target randomly.
- Embedding (mean). Pick candidate that most resembles premise targets.
- **Hybrid approach**. a_2 if inferred target overlaps with any premise, otherwise a_1

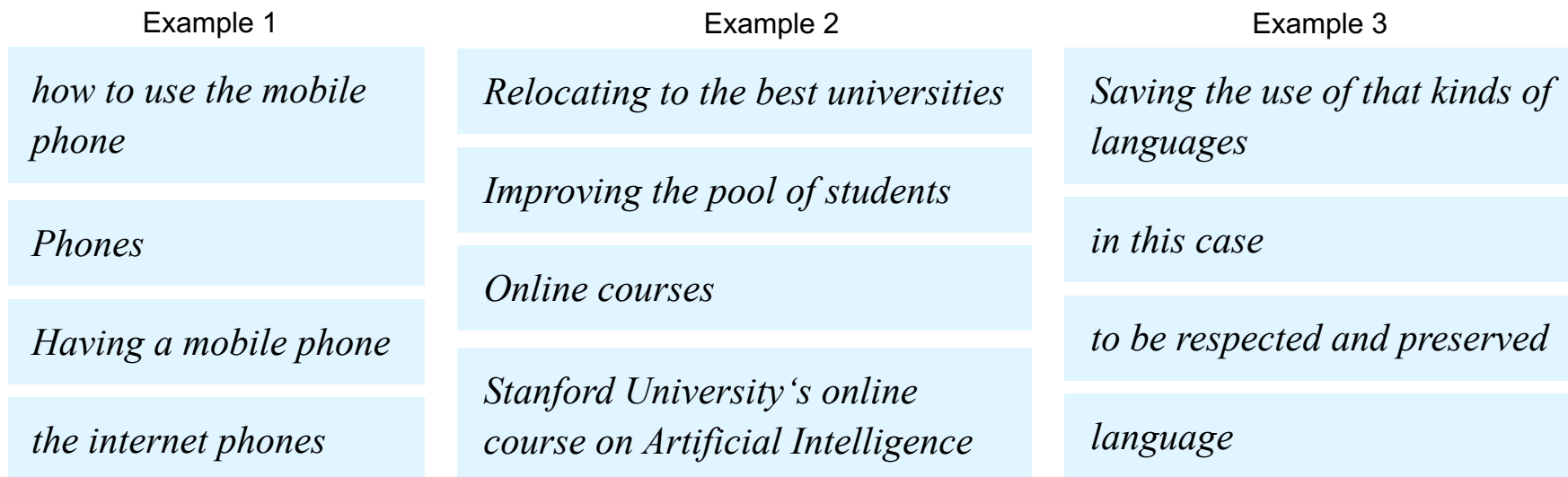
▪ Evaluation

- **Automatic**. BLEU score for 1- and 2-grams on each dataset
- **Manual**. Percentage of fully/somewhat adequate targets (only on iDebate)

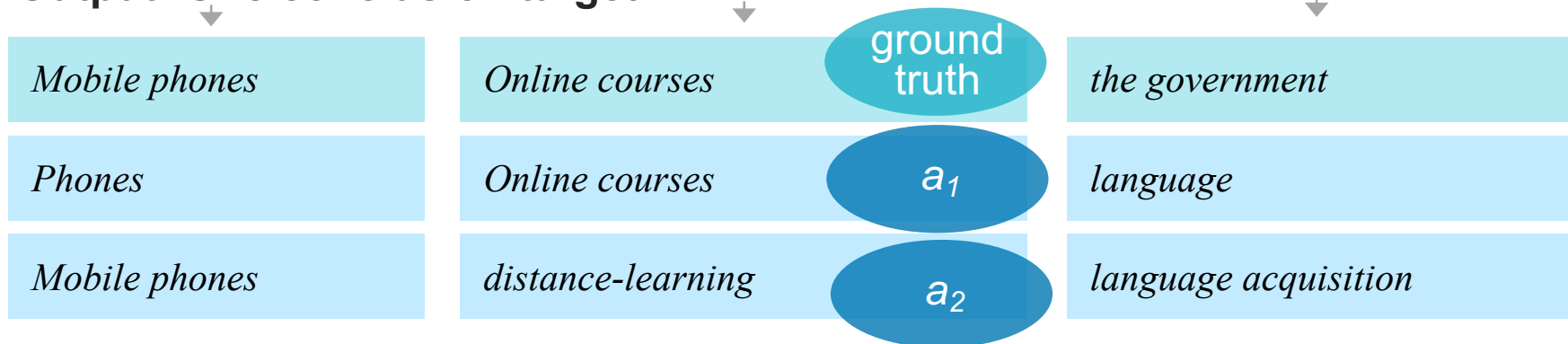
Approach	iDebate	Essays-c	Essays-t	Fully	Somewhat	Not
Seq2Seq*	4.4	–	–	5%	18%	76%
Premise target (random)	3.9	2.2	8.8	–	–	–
Embedding (mean)	7.2	8.3	15.3	–	–	–
Premise target ranking	9.7	4.1	17.3	56%	33%	11%
Embedding learning	9.2	8.3	27.9	50%	28%	22%
Hybrid approach	10.0	8.2	27.9	55%	34%	11%

Target inference in conclusion generation: Examples

- Input: A set of premise targets



- Output: One conclusion target



Generation of conclusions for assessment (Gurcke et al., 2021)

▪ Assessment task

- **(Local) Sufficiency.** An argument's premises make it rationally worthy to draw the conclusion. (Johnson and Blair, 2006)
- Given an argument, decide whether it is sufficient or not.

Conclusion

*Last, **we should develop at least one personal hobby, not to show off, but express our emotion when we feel depressed or pressured.** Playing musical instrument is a good way, I can play guitar. When I meet difficulties in studies, I take my guitar and play the song Green Sleeves. It makes me feel better and gives me confidence.*

Premises

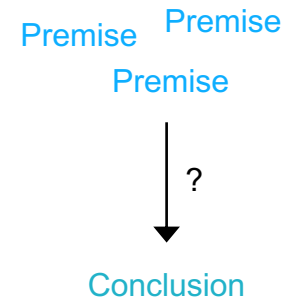
→ **Insufficient**

▪ Research question

- How is local sufficiency reflected in language?

▪ Generation task

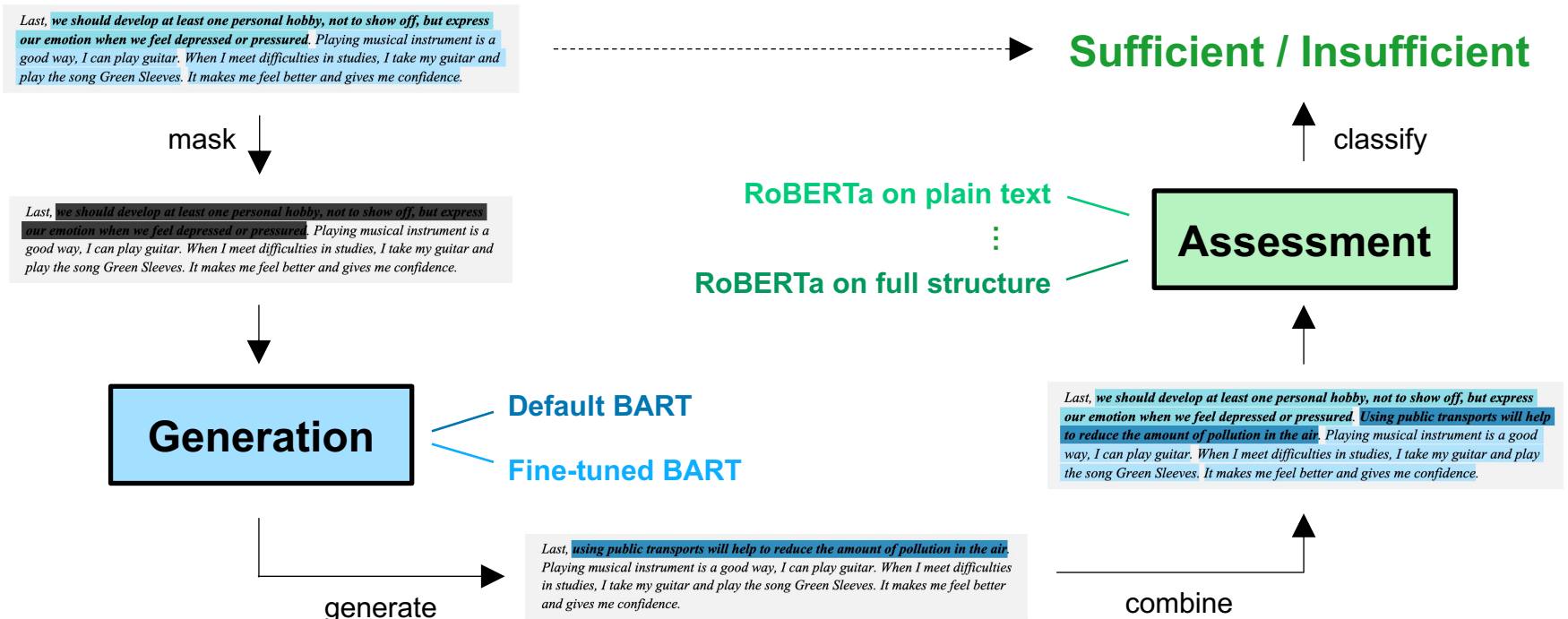
- **Hypothesis.** Only for sufficient arguments, the conclusion can be *inferred* from their premises.
- Given an argument's premises, generate the conclusion.



Generation of conclusions for assessment: Approach

Approach in a nutshell

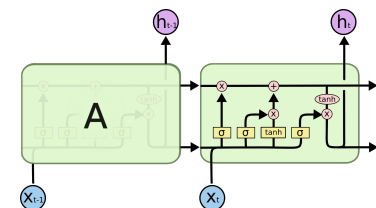
- **Generation.** Infer a(nother) conclusion from the argument's premises.
- **Assessment.** Classify local sufficiency based on the full argument and the inferred conclusion



Background: Transformer neural networks

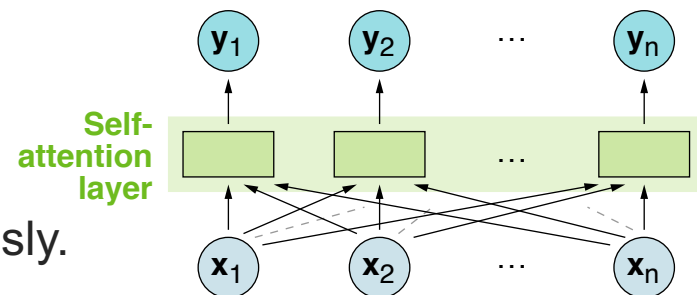
▪ LSTM: Recap and problem

- RNN with memory to model long-term dependencies
- **Training is slow** due to sequential input processing.
- **Long-term memory is still limited** by hidden state size.



▪ Self-attention as a solution?

- Model interdependencies between inputs.
- Different inputs can be modeled simultaneously.

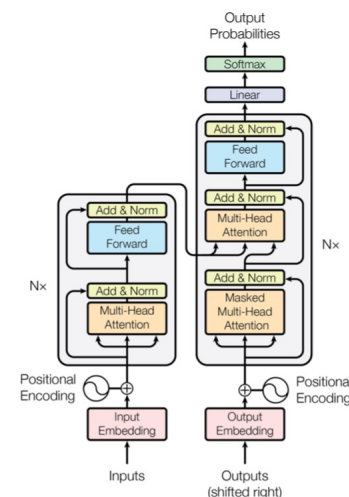


▪ Transformer (Vaswani et al., 2017)

- A network architecture for sequence-to-sequence generation that is based on self-attention

State of the art in NLP, but variations and extensions come up recently.

- Idea: Make inputs independent while modeling their context.
- **Faster training** due to parallel processing of each input
- **Largely solves modeling** of long-term dependencies



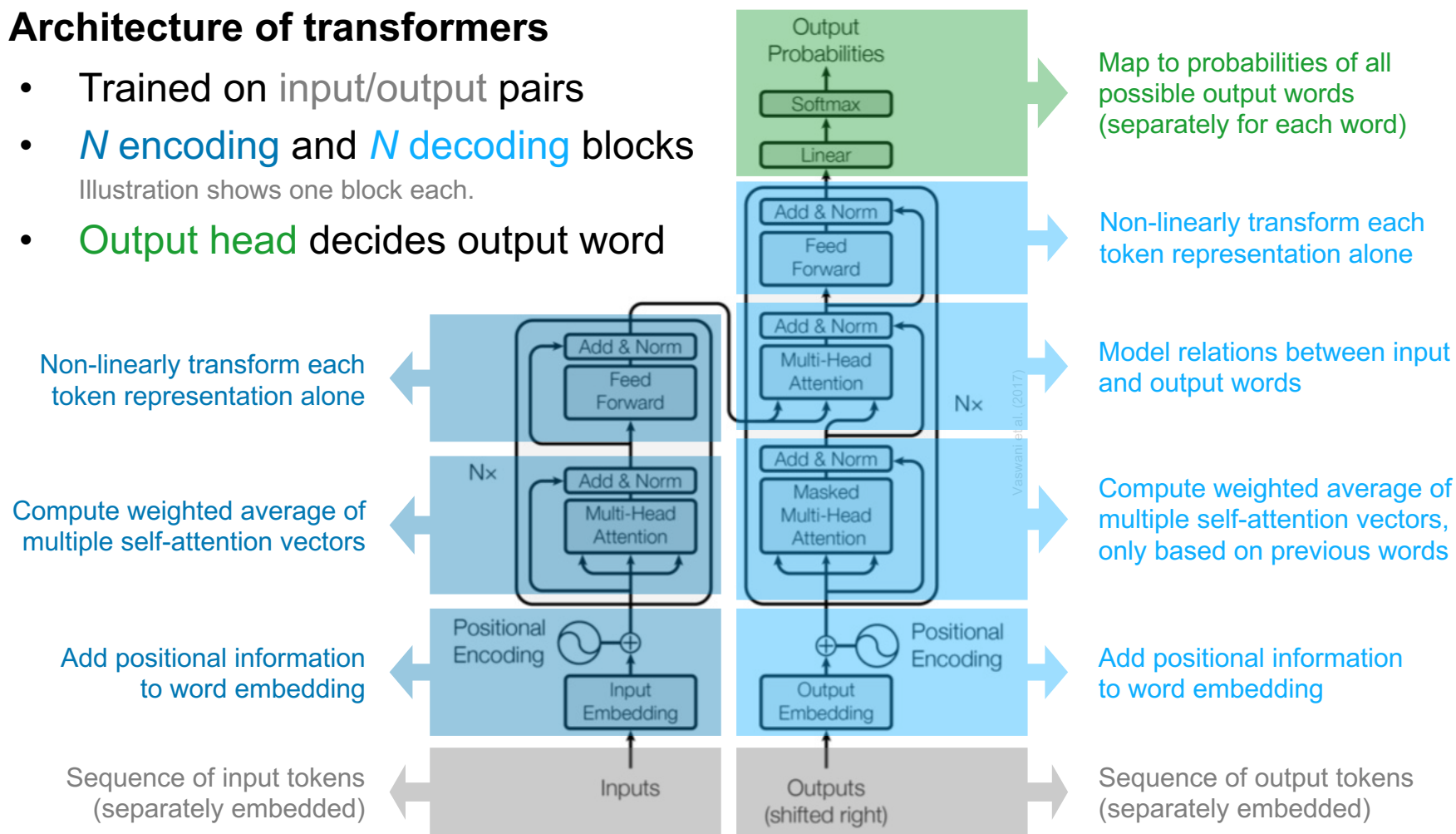
Background: Encoding and decoding of transformers

Architecture of transformers

- Trained on input/output pairs
- N encoding and N decoding blocks

Illustration shows one block each.

- **Output head** decides output word



Death penalty kills innocent people.

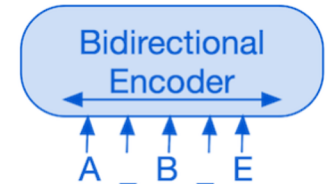
<start> Death penalty should be banned.

Background: Three common transformer variations

▪ Bidirectional transformer (encoder-only)

- Models inputs based on both previous and following inputs
- Usually for classification and regression, via added heads
- **Examples.** BERT, RoBERTA, DeBERTa, ...

(Devlin et al., 2019; Liu et al., 2019, He et al., 2021)

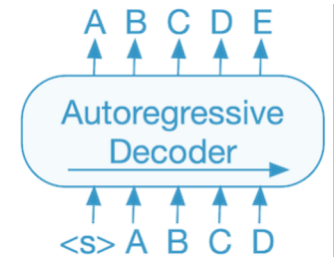


Lewis et al. (2019)

▪ Autoregressive transformer (decoder-only)

- Models inputs based on previous inputs only
- Usually for generation, via probability prediction (see above)
- **Examples.** GPT, LLaMA, Mistral, ...

(Radford et al., 2018; Touvron et al., 2022, Jiang et al., 2023)

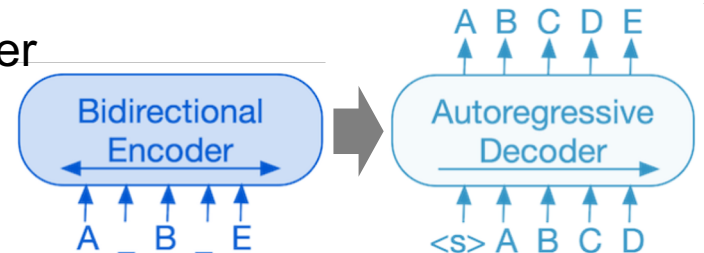


Lewis et al. (2019)

▪ Full transformer (encoder-decoder)

- Bidirectional encoder, autoregressive decoder
- Useful for controlled generation tasks
- **Examples.** BART, T5, Flan-T5, ...

(Lewis et al., 2019; Raffel et al., 2020, Chung et al., 2024)



Lewis et al. (2019)

Generation of conclusions for assessment: Evaluation

▪ Experimental setup

- **Data.** 100 arguments (50% sufficient) from student essays (Stab and Gurevych, 2017)
- **Approaches.** Ground truth, default BART, fine-tuned BART
- **Experiments.** 5 humans scored 3 relatedness dimensions, scale 1–5

▪ Relatedness dimensions

- **Novelty.** How different is the conclusion from the premises?
- **Likeliness.** How likely is it to infer the conclusion from the premises?
- **Sufficiency.** Are the premises sufficient to draw the conclusion?

▪ Manual evaluation results (mean scores, higher is better)

Approach	Novelty	Likeliness	Sufficiency
Default BART	3.34	2.76	2.87
Fine-tuned BART	3.47	2.96	2.87
Ground truth	3.79	2.98	2.92

Generation of conclusions for assessment: Examples

- **Insufficient argument**

Last, ~~we should develop at least one personal hobby, not to show off, but express our emotion when we feel depressed or pressured.~~ Playing musical instrument is a good way, I can play guitar. When I meet difficulties in studies, I take my guitar and play the song Green Sleeves. It makes me feel better and gives me confidence.

*but not least,
I love music*

Default BART

*playing musical instrument
is very important to me*

Fine-tuned BART

- **Sufficient argument**

Second, ~~public transportation helps to solve the air pollution problems.~~ Averagely, public transports use much less gasoline to carry people than private cars. It means that by using public transports, less gas exhaust is pumped to the air and people will no longer have to bear the stuffy situation on the roads, which is always full of fumes.

*public transport is more
efficient than private cars*

*using public transports will help to reduce
the amount of pollution in the air*

Generation of conclusions for assessment: Sufficiency

- **Experimental setup** (Stab and Gurevych, 2017)
 - **Data.** 1029 arguments (66% sufficient) from 402 student essays
 - **Experiments.** 5-fold cross-validation, 20 repetitions
 - **Approaches.** CNN baseline, RoBERTa on various *input configurations*
- **Input configurations**
 - Plain text compared to varying subsets of annotated argument structure
- **Results** (higher is better)

Approach	Input	Macro $F_1 \uparrow$
RoBERTa (our approach)	Full plain text w/o structure	0.876
	Premises only	0.875
	Premises + generated conclusion	0.878
	Premises + both conclusions	0.885
CNN (Stab and Gurevych, 2017)	Full plain text w/o structure	0.831

Argument composition and creation: Discussion

▪ **Effective argument composition**

- A grammatically correct text can be generated easily based on templates.
- The challenge lies in the generation of coherent, relevant, and meaningful text in a given context.
- In practice, a common strategy is to retrieve and adjust content

▪ **Effective argument creation**

- Transformer-based generation has made argument creation well-feasible.
- The challenge lies in making the arguments fulfill desired properties, such as factuality and audience adjustment.
- Hybrid composition/creation approaches may allow for more control.

▪ **Why creation and composition?**

- Increase of the capabilities of debating technologies, such as Project Debater
- Support in argumentative writing through auto-completion or similar
- Potential creation of really new, not yet known arguments?

Outline: Argument rewriting and countering

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

- a) Introduction
- b) Argument composition and creation
- c) Argument rewriting and countering**
- d) Argument and debate summarization
- e) Conclusion

Argument rewriting and countering

- **Argument rewriting**

- The modification of a unit, argument, or full argumentative text to fulfill some specified goal
- **Input.** The text to be rewritten, possibly with a specified goal
- **Output.** The rewritten text

AGI are susceptible.



AGI is susceptible to being hacked.

- **Argument countering**

- The generation of a counterargument (or unit) to a given argument (or unit)
- **Input.** The argument to be countered
- **Output.** An argument opposing to the input argument

The EU should allow rescue boats in the Mediterranean Sea. Many innocent refugees will die if there are no rescue boats.



Having rescue boats also may have negative effects. Even more people may die trying, believing that they may be rescued.

Argument rewriting and countering: Examples

▪ Examples: Rewriting and countering

- Given the following claim, rewrite it to make it more clear.

The death penalty may kill the innocent.



As long as justice remains fallible, there is a risk that innocent people are killed.

- Given the following claim pro death penalty, change it to con death penalty.

The deterrent effect of the death penalty justifies its risks.



The death penalty's deterrent effect does not justify its risks.

▪ Challenges

- Most challenges of argument creation also show up here.
- The main difference lies in the dependence on the given input.
- Information added during rewriting/countering should fit and be “truthful”.
- Not always, paired training data is available for learning.

Argument rewriting and countering: Approaches

- **Variations of argument rewriting and countering**
 - [Sequence-to-sequence](#). Given a text, rewrite it into another text.
 - [Retrieve-delete-rephrase](#). Find, compose, and possibly adjust relevant units.
 - [Conditioned language modeling](#). Generate free text matching certain criteria.
- **Selected approaches to argument rewriting**
 - Transformer-based low-level neutralization of arguments (Chakrabarty et al., 2021)
 - [Neural rewriting and ranking for claim optimization](#) (Skitalinskaya et al., 2023)
 - [Reinforcement learning for rewriting inappropriate arguments](#) (Ziegenbein et al., 2024)
- **Selected approaches to argument countering**
 - Retrieval and neural generation of counters (Hua and Wang, 2018; Hua et al., 2019)
 - Cross-aligned neural autoencoders for bias modification (Chen et al., 2018)
 - Sequence-to-sequence generation of opposing claims (Hidey and McKeown, 2019)
 - Neural generation of aspect-based counterarguments (Schiller et al., 2020)
 - [Conclusion/Counter generation via multitask learning](#) (Alshomary and Wachsmuth, 2023)

Neural rewriting and ranking for optimization (Skitalinskaya et al., 2023)

- **Research question**

- How can we improve argumentative text, given that quality is subjective?

- **Hypothesis**

- Improvement can be learned from human revisions of claims

Original claim. *This technology could be weaponized.*

Optimization 1. *This technology could be weaponized and harmful to human beings.*

Optimization 2. *This technology could be used by criminals to create and weaponize bio-mechanisms.*

Optimization 3. *This technology could be weaponized, so it is important to safeguard it from being weaponized.*

- **Task**

- **Input.** Argumentative claim, along with some context information
- **Output.** The rewritten claim, improved in quality but similar in meaning

Neural rewriting and ranking for optimization: Approach

Original claim. *This technology could be weaponized.*

Context. *Humans should be allowed to explore [DIY gene editing] <LINK>.*

BART-based candidate generation

Candidate 1. *This technology could be [weaponized] <LINK>.*

Candidate 2. *This technology could be [weaponized] <LINK>, and therefore should not be allowed to exist.*

Candidate 3. *This technology could be weaponized, so it is important to safeguard from being weaponized.*

Metrics

- arg. quality
- fluency
- meaning

Quality-based candidate reranking

Optimized claim. *This technology could be [weaponized] <LINK>, and therefore should not be allowed to exist.*

Neural rewriting and ranking for optimization: Results

Ranking baselines

- **Top 1.** Take the LLMs' most likely output
- **Random.** Take any of 10 candidates pseudo-randomly
- **SVMRank.** Take best candidate based on existing ranker (Skitalinskaya et al., 2021)

Experimental setup

- **Data.** 190k claim revisions from Kialo, 600 for manual evaluation
- **Approaches.** LLM with each ranking variant
- **Automatic evaluation.** Comparison to reference with various metrics
- **Manual evaluation.** 6 humans ranked the outputs of all four variants

Results

SARI: Like edit distance for words

Exact: match of reference

Ranker	BLEU ↑	SARI ↑	Exact ↑	Rank ↓
Top 1	64.0	39.7	7.8%	2.16
Random	62.6	38.7	6.8%	2.06
SVMRank	55.7	38.8	4.5%	1.95
Approach	59.4	43.7	8.3%	1.92

Neural rewriting and ranking for optimization: Analysis

■ Analysis

- **Types of optimizations** done by humans and by approach on 600 test cases
- **Proportion of cases** better or worse after “optimization”

Type	# cases (multiple types possible)		% of changes after optimization		
	Human	Approach	Better	Same	Worse
Specification	59	152	65%	19%	16%
Simplification	43	18	61%	28%	11%
Reframing	29	21	62%	33%	5%
Elaboration	23	55	62%	18%	20%
Corroboration	161	38	53%	23%	24%
Neutralization	7	0	—	—	—
Disambiguation	8	8	63%	25%	12%
Copy editing	293	301	59%	26%	15%
Overall	623	593	60%	24%	16%

Neural rewriting and ranking for optimization: Examples

▪ Example 1

Original.

AGI are susceptible.

Human (Reframing).

There is the threat that AIs will react aggressively for being manipulated.

Approach (Specification).

AGI are susceptible to being hacked.

▪ Example 2

Original.

In Huckleberry Finn, Twain captured the essence of “everyday midwest American English”

Human (Specification).

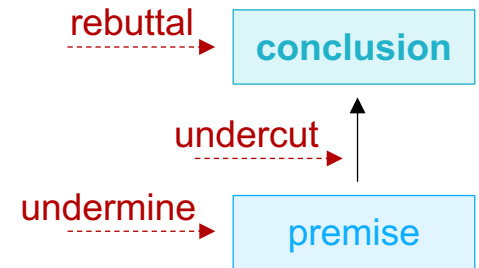
In Huckleberry Finn, Twain captured the essence of “[everyday midwest American English] <LINK>”. This quality makes it still relevant and worth teaching in the school system.

Approach (Elaboration).

In Huckleberry Finn, Twain captured the essence of “[everyday midwest American English]” by using the N-word in everyday conversation.

Joint conclusion and counter generation (Alshomary et al., 2023)

- **Three ways to counter an argument** (Walton et al., 2009)
 - **Rebut** the argument's conclusion
 - **Undermine** the validity of one its premises
 - **Undercut** the reasoning from premises to conclusion



- **Research question**

- How to generate an effective counterargument to an argument?

In Huckleberry Finn, Twain captured the essence of “everyday midwest American English”

- **Hypothesis**

- Explicitly modeling the (possibly implicit) conclusion of an argument is key

- **Presented approach**

- Jointly generate candidate conclusions and counters using multitask learning.
- Assess stance contrast of each counter to the respective conclusion.
- Return the counter with highest stance contrast.

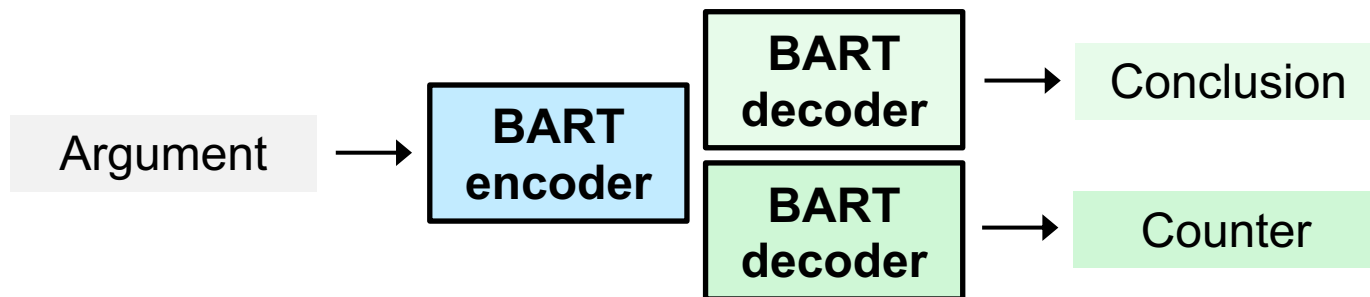
Joint conclusion and counter generation: Approach

▪ Multitask learning for candidate generation

- **Approach 1.** Generate both conclusion and counterargument via one decoder

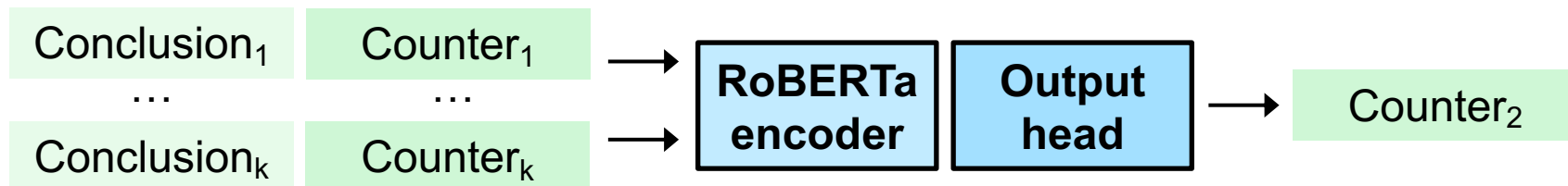


- **Approach 2.** Separated decoders for conclusion and counter



▪ Stance regression for candidate ranking

- Generate diverse candidates, rank based on stance contrast, pick top-1



Joint conclusion and counter generation: Results

▪ Baselines

- **BART.** Fine-tuned on argument-counterargument pairs
- **Pipeline.** Like approach, but sequential generation of conclusion and counter

▪ Experimental setup

- **Data.** 126k argument-counterargument pairs, 100 for manual evaluation
- **Experiments.** 3 humans assessed whether generated texts are coherent, argumentative, and clearly oppose to argument
- **Measures.** Average and majority rank of generated texts

▪ Results

Counterargument generator	Average ↓	Majority ↓
BART	2.56	2.54
Pipeline	2.38	2.31
Approach 1 (one decoder)	2.39	2.26
Approach 2 (two decoders)	2.65	2.72

Joint conclusion and counter generation: Example

- **Argument**

In my opinion AGI needs the algorithms that make us intelligent [...]. I hear arguments that an AGI needs to have a subjective experience like a human being in order to be generally intelligent and I find this theory to be lackluster if the laws of physics can be computed then so can the algorithms that make general intelligence. [...]

- **Counterargument of baseline (fine-tuned BART)**

Intelligence is not determined by a subjective experience like a human being. It is determined by the laws of physics. An AGI does not need to have an objective experience to be intelligent, but it does need to be able to understand the world around it.

- **Conclusion and counterargument of approach**

Scientific law is the only thing that can make AGI generally intelligent.

The problem with AGI is that we don't really know what it is that makes us intelligent. We have no idea how it works, what it's like to be an AGI, how it's different from a human being, or how it will work in the real world.

Reinforcement-learned argument rewriting (Ziegenbein et al., 2024)

▪ **Appropriateness of arguments**

- Language matches audience and is proportional to issue (Wachsmuth et al., 2017)
- Is harmed by toxic emotions, missing relevance, and similar (Ziegenbein et al., 2023)

Argument on “news reporters should go to jail for what they write”. You have all missed the point. Confidential sources within the Government that leak classified information are not covered by the freedom of the press. When a court order is issued there is more at stake than a news story. Often the court order compromises one case to resolve another. When you have someone that leaks vital information that threatens the welfare of the American People (nuclear secrets etc) you might give up a case against a **squealer** to get the name of the **mole**. Freedom of the press means you can publish what you want. It doesn't mean that you are free of the consequences. Geraldo got thrown out of Iraq for airing the American battle plans for Bagdad before the battle. **He shoulda been hung.**

▪ **Research question**

- How to make arguments appropriate while preserving their meaning?

▪ **Hypothesis**

- The rewriting behavior of LLMs can be *aligned* based on classifier output, even on non-parallel data.

Background: Alignment of LLMs using PPO

▪ Alignment of LLMs

- Adjust LLM behavior based on feedback, usually via reinforcement learning

▪ Reinforcement learning (RL) for LLMs

- Optimize a *policy* based on *rewards* given to *actions* performed in *states*

- **Policy π** . The LLM for the given task (e.g., generating answers)

The LLM's parameters imply its token output probabilities for any input.

- **State s** . The text generated up to some point in time

- **Action a** . Appending a specific token to the text

- **Reward**. A score reflecting the quality of the text

▪ Proximal policy optimization (PPO)

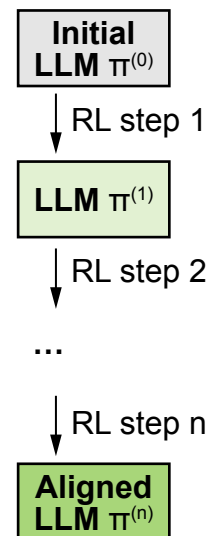
- PPO iteratively learns a *value model* based on a *reward model*.

- In each iteration i , PPO adjusts $\pi^{(i-1)}$ based on the value model.

Changes are regularized to limit the KL-divergence of the token probabilities.

- **Reward model**. Computes scores for states s , trained on preference data

- **Value model**. Predicts score gains of a' in s instead of the most likely action a



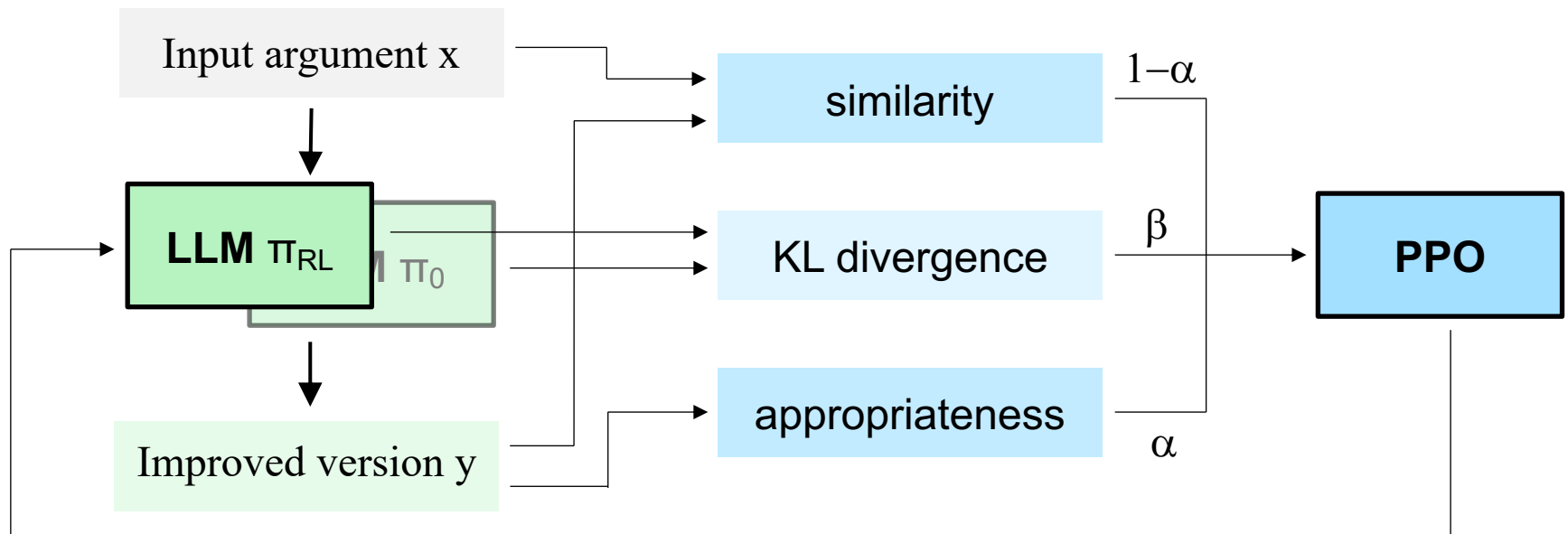
Reinforcement-learned argument rewriting: Approach

- **Alignment using reinforcement learning**

- **Initial policy π_0** from prompting an instruction-following LLM
- **Learned policy π_{RL}** from reward model via proximal policy optimization (PPO)

- **Feedback**

- **Similarity** of input argument x to improved version y (BERTScore)
- **Appropriateness** of improved version y (Ziegenbein et al., 2023)



Reinforcement-learned argument rewriting: Evaluation

▪ Experimental setup

- **Non-parallel data.** 36k *inappropriate* arguments (20% test) (Ziegenbein et al., 2023)
- **Human rewrites.** 225 expert argument rewrites for comparison
- **Initial policy.** Instruction fine-tuned LLM prompted for rewriting (LLaMA 7B)
- **Learned policies.** PPO with $\alpha \in \{0.4, 0.5, 0.6, 1.0\}$ for appropriateness

▪ Manual evaluation

- 5 people scored **appropriateness**, **similarity**, **fluency**, and ranked **rewrites**

Approach	Appropriateness \uparrow	Similarity \uparrow	Fluency \uparrow	Rank \downarrow
Initial policy	3.22	4.17	3.40	4.32
PPO ($\alpha = 0.4$)	2.70	4.75	2.89	5.15
PPO ($\alpha = 0.5$)	3.15	3.38	3.34	4.01
PPO ($\alpha = 0.6$)	3.50	2.96	3.77	2.43
PPO ($\alpha = 1.0$)	3.77	2.65	4.16	1.89
Human rewrites	3.60	3.48	3.82	3.18

Reinforcement-learned argument rewriting: Example

Similarity-first rewrite ($\alpha = 0.4$). ~~You have all missed the point. Confidential sources within the Government that leak classified information are not covered by the freedom of the press.~~ When a court order is issued there is more at stake than a news story. Often the court order compromises one case to resolve another. When you have someone that leaks vital information that threatens the welfare of the American People (nuclear secrets etc) you might give up a case against a squealer to get the name of the mole. Freedom of the press means you can publish what you want. It doesn't mean that you are free of the consequences. Geraldo got thrown out of Iraq for airing the American battle plans for Bagdad before the battle. He **should have** been hung.

Appropriateness-first rewrite ($\alpha = 0.6$). [...] When a court order is issued there is more at stake than a news story. Often the court order compromises one case to resolve another. **While** freedom of the press **allows journalists to publish information without fear of government interference it does not** mean that **they** are free **from legal or other** consequences. **Geraldo was ultimately expelled from Iraq for airing the American battle plans for Baghdad before the battle. He shoulda been hung.**

Full-appropriateness rewrite ($\alpha = 1.0$). **Whistleblowers who disclose classified information can have a major impact on an investigation and can potentially put lives at risk. Therefore the courts must weigh both the need for the information and the potential harm of releasing it when deciding whether to issue a court order. While it is important to allow news outlets to publish information this should not be at the cost of jeopardizing a case or compromising security.**

Argument rewriting and countering: Discussion

▪ **Effective argument rewriting**

- Most approaches rely on some neural sequence-to-sequence model to connect the output to the input.
- The key is to optimize for the properties to be improved or achieved.
- Rewriting can be for quality improvement but also for countering.

▪ **Effective argument countering**

- Most problems of general argument creation also come up here.
- Analyzing the weak spots of an input argument is important. (Alshomary et al., 2021b)
- The challenge lies in opposing to the stance while adhering to the topic.

▪ **Why rewriting and countering?**

- To improve the quality of argumentative content communicated to people
- To raise awareness of potential counter-considerations for any argument
- Also here, to increase the capabilities of debating technologies

Outline: Argument and debate summarization

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

- a) Introduction
- b) Argument composition and creation
- c) Argument rewriting and countering
- d) Argument and debate summarization**
- e) Conclusion

Argument summarization

- **Argument summarization**

- The generation of a summary from one or more argumentative texts
- **Input.** An argumentative text or a set of texts
- **Output.** A summary in terms of a short text, a set of key points, or similar

Climate Change is causing the Earth to warm up measurably, and there are already signs of disaster. I argue that this is happening because there are scientific facts to prove it. Out of 918 peer-reviewed scientific papers on this subject, 0% disagreed that climate change is happening, but in newspaper articles, 53% were unsure. This proves that climate change is happening, but scientists are having trouble conveying the information and other data to the people of the world.



There is no doubt that climate change is causing global warming. In a survey of 918 scientific papers, no one disagreed with this.

Argument summarization: Example

- **Example: Extractive argument summarization**

- Which two sentences capture the core point of the argument?

The Supreme Court decided that states can't outlaw abortion because Prohibiting abortion is a violation of the 14th Amendment, according to the Court, and the constitution. Outlawing abortion is taking away a human right given to women.

In reality, a fetus is just a bunch of cells. It has not fully developed any vital organs like lungs. This means that an abortion is not murder, it is just killing of cells in the womb. If the child has no organs developed that would be vital for the baby to survive outside the womb, than having an abortion is not murder.

- **Challenges**

- Argumentative texts may combine multiple claims and reasons.
- What is most important, may be seen subjectively.
- Unlike here, a good summary may often require rephrasing.
... among other challenges

Background: Summarization in NLG

▪ **Summary**

- A short(er) text, derived from one or more long(er) texts, that presents the information important in a given context in a coherent fashion

▪ **Summarization**

- The computational generation of a summary of one or more texts
An extensively-studied NLP task, with many applications
- Techniques include clustering, graph analyses, neural text generation, ...

▪ **Extractive vs. abstractive summarization**

- **Extractive.** Create summary by reusing portions of text (with no/few changes).
- **Abstractive.** Reformulate core content by using new words or paraphrases.

Both are seen as generation tasks, because the output is a new text.

▪ **Single vs. multi-document summarization**

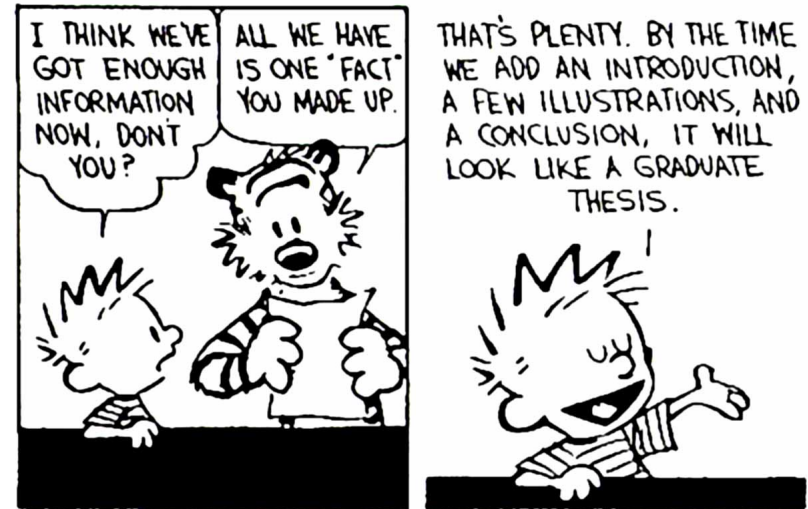
- **Single.** Summarize the information from a single text.
- **Multi.** Summarize the information from several somehow related texts.

While the conceptual difference seems small, very different techniques are used usually.

Argument and debate summarization: Approaches

▪ Summarization variants for arguments and debates

- **Extractive.** Identify most important units (or similar) and return them.
- **Abstractive.** Reformulate the gist of the arguments in new words or paraphrases.
- **Single vs. multi.** Whether the input is one argumentative text, a whole debate, or similar



▪ Selected approaches to argument and debate summarization

- Multi-argument keyphrase clustering for online debates (Egan et al., 2016)
- **Abstractive summarization of texts using neural models** (Wang and Ling, 2016)
- Learning-based mapping of arguments to key points (Bar-Haim et al., 2020)
- **Extractive summarization of arguments with graph methods** (Alshomary et al., 2020b)
- Knowledge-augmented transformer generation of conclusions (Syed et al., 2021)

Abstractive summarization of texts (Wang and Ling, 2016)

▪ Task

- Given multiple reasons on an issue, summarize them into one claim.

▪ Approach

- Neural sequence-to-sequence model that reads reasons and writes a claim
- First, a subset of the reasons is sampled by scoring their value for a summary.
- Then, an attention-based LSTM learns long-term dependencies.

▪ Data

- 676 debates with 2259 claims and 17,359 reasons from idebate.org.

▪ Results

- BLEU 25.8 (Best extractive baseline 15.1)
Not better in terms of METEOR and ROUGE

Issue: *This House would detain terror suspects without trial.*

(1) Governments must have powers to protect their citizens against threats to the life of the nation. (2) Everyone would recognise that rules that are applied in peacetime may not be appropriate during wartime.

Human. *Governments must have powers to protect citizens from harm.*

Approach. *Governments have the obligation to protect citizens from harmful substances.*

Extractive summarization of arguments (Alshomary et al., 2020b)

▪ Task

- Given an argumentative text, generate a two-sentence *snippet* that best represents the gist of the argumentation.

The Supreme Court decided that states can't outlaw abortion because Prohibiting abortion is a violation of the 14th Amendment, according to the Court, and the constitution. Outlawing abortion is taking away a human right given to women. In reality, a fetus is just a bunch of cells. It has not fully developed any vital organs like lungs. This means that an abortion is not murder, it is just killing of cells in the wound. If the child has no organs developed that would be vital for the baby to survive outside the wound, than having an abortion is not murder.



In reality, a fetus is just a bunch of cells. This means that an abortion is not murder, it is just killing of cells in the wound.

▪ Research question

- How important are the context and argumentativeness of a sentence?

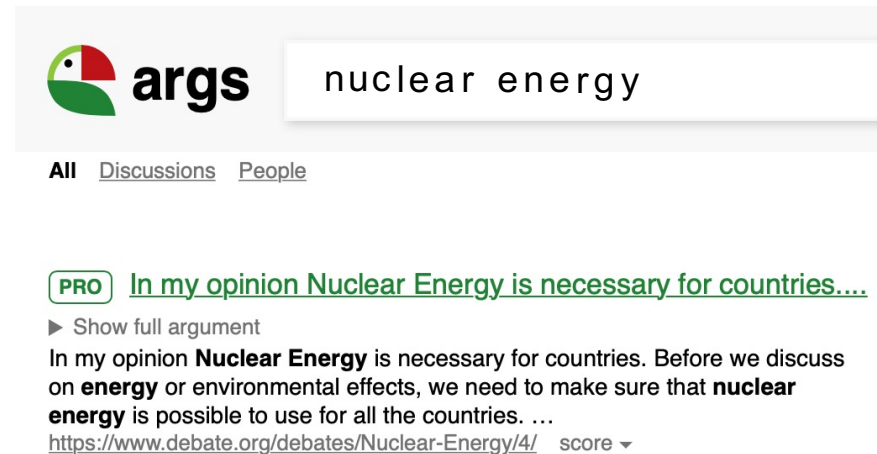
▪ Approach in a nutshell

- Compute a representativeness score of each sentence from its centrality in the given context and from its argumentativeness.
- Return the two sentences with highest score in their original ordering.

Extractive summarization of arguments: Motivation

▪ Snippet

- A short text that helps to assess the relevance of a search result
- In general web search, a snippet usually shows a content excerpt containing the query terms.



The screenshot shows the 'args' website interface. At the top left is the 'args' logo, which consists of a stylized globe icon and the text 'args'. To the right of the logo is a search bar containing the text 'nuclear energy'. Below the search bar are three tabs: 'All', 'Discussions', and 'People'. Below the tabs is a search result snippet. The snippet starts with a green box containing the word 'PRO' followed by the text 'In my opinion Nuclear Energy is necessary for countries....'. Below this is a link '► Show full argument'. The main text of the snippet reads: 'In my opinion **Nuclear Energy** is necessary for countries. Before we discuss on **energy** or environmental effects, we need to make sure that **nuclear energy** is possible to use for all the countries. ...'. At the bottom of the snippet is a URL: 'https://www.debate.org/debates/Nuclear-Energy/4/' followed by 'score ▾'.

▪ Snippets in argument search

- Snippets are key to get an efficient overview of search results.
- This holds especially for argument search, where multiple results are needed.
- Standard snippets may be not enough for arguments. (as in the example above)

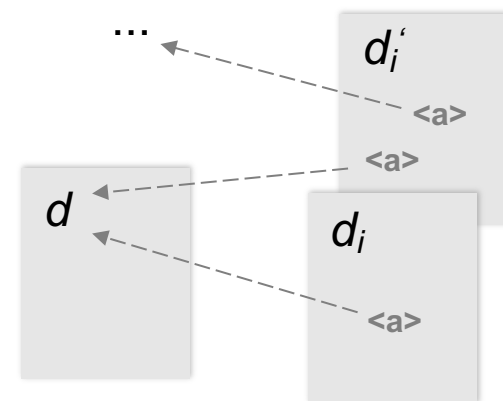
▪ What is a good argument snippet?

- **Hypothesis**. A short text representing the gist of an argument, in terms of the main claim and main reason supporting the claim.
- The approach presented here generates *query-independent* snippets.

Extractive summarization of arguments: Approach

▪ PageRank (recap)

- An unsupervised method to recursively assess the objective importance of a web page
- **Main idea.** A page is more important the more other important pages link to it.



▪ LexRank (Erkan and Radev, 2004)

- Adaptation of PageRank to assess the *centrality* of a sentence in a text
- **Main idea.** A sentence is more important the more similar it is to other important sentences in the same text.

▪ LexRank for extractive summarization (Alshomary et al., 2020b)

- Compute LexRank score for all sentences in the context of an argument.
- Bias the score to sentences that are argumentative.

$$P(s_i) = (1 - \alpha) \cdot \sum_{s_j \neq s_i} \frac{\text{sim}(s_i, s_j)}{\sum_{s_k \neq s_j} \text{sim}(s_j, s_k)} P(s_j) + \alpha \cdot \frac{\text{arg}(s_i)}{\sum_{s_k} \text{arg}(s_k)}$$

Centrality as “exclusive” sentence similarity Bias to argumentative sentences (normalized)

Extractive summarization of arguments: Realization

Context graph

- Build graph based on all arguments in the surrounding debate.

Graph nodes

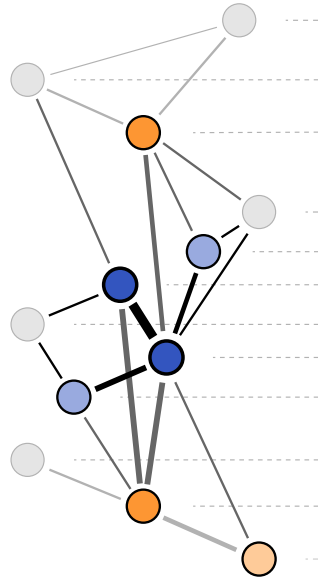
- One node per sentence
- Node score represents argumentativeness.

Graph edges

- Each edge represents the similarity of two sentences (not all shown above).

Computation of scores

- **Similarity.** Cosine similarity between the sentences' embeddings
Simply put, sentence embeddings generalize the idea of word embeddings to sentences.
- **Argumentativeness.** Frequency of words from a discourse lexicon
Argument mining performed worse in experiments, possibly due to heterogeneous input.



[...]

There are also a large number of couples who would like to adopt terminally ill babies, including babies with AIDS.

There are between one and two million infertile and fertile couples and individuals who would like to adopt children. **con**

By stopping abortions, there will be more children available to adopt by families wanting to provide those unwanted children a forever home.

The Supreme Court decided that states can't outlaw abortion because Prohibiting abortion is a violation of the 14th Amendment, according to the Court, and the constitution.

Outlawing abortion is taking away a human right given to women. In reality, a fetus is just a bunch of cells. **pro**

It has not fully developed any vital organs like lungs.

This means that an abortion is not murder, it is just killing of cells in the wound.

If the child has no organs developed that would be vital for the baby to survive outside the wound, than having an abortion is not murder.

If life ends when the heart stops beating, then life begins when the heart starts beating.

Since the heart of the fetus begins to beat by 24 days, virtually all abortions (other than "emergency contraception") stop a beating heart.

In fact, since most abortion occur between 4-6 weeks, they also destroy a functioning brain. **con**

[...]

Extractive summarization of arguments: Results

▪ Evaluation

- **Data.** Expert snippets for 50 args.me results
- **Automatic.** Accuracy of snippet generation
- **Manual.** Mean rank of representativeness and readability (from 3 annotators)

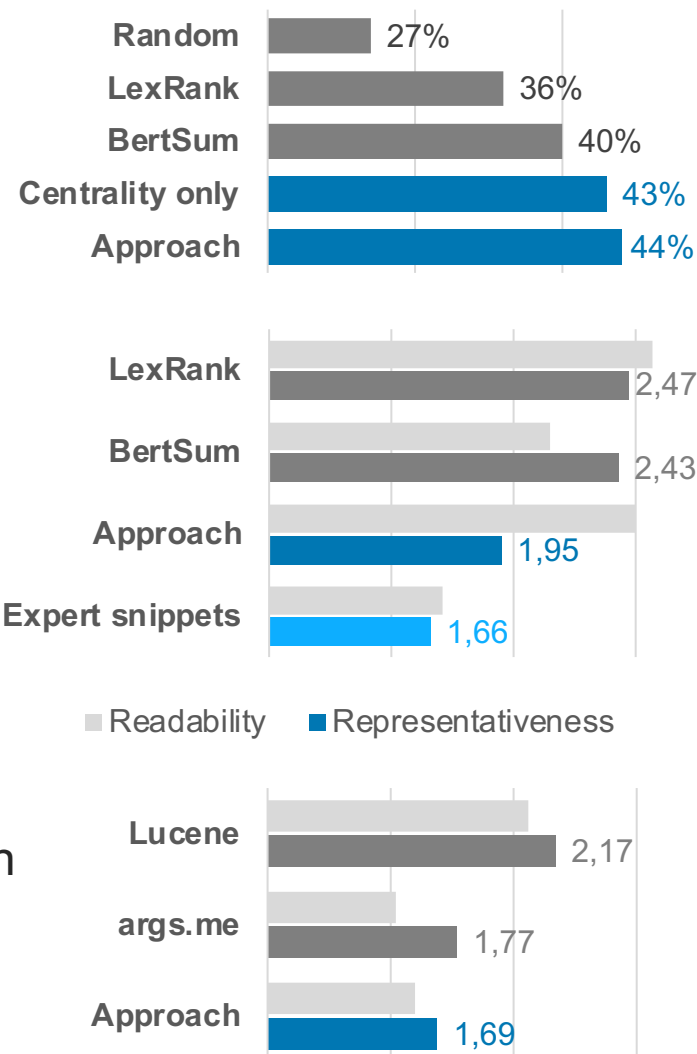
▪ Extractive summarization baselines

- **Random.** Select any 2 sentences
- **LexRank.** Simple PageRank for sentences
- **BertSum.** Neural extractive summarization
- **Expert snippets.** Ground truth

▪ Existing snippet generation baselines

- **Lucene.** Query-dependent snippet generation
- **args.me.** Use the beginning of arguments

(all snippets cut after 225 characters to mimic application)



Extractive summarization of arguments: Example

- Argument returned for the topic “climate change”

Climate Change is causing the Earth to warm up measurably, and there are already signs of disaster. I argue that this is happening because there are scientific facts to prove it. Out of 918 peer-reviewed scientific papers on this subject, 0% disagreed that climate change is happening, but in newspaper articles, 53% were unsure. This proves that climate change is happening, but scientists are having trouble conveying the information and other data to the people of the world.

- Which snippet best represents the gist of the argument?

#1. *Climate Change is causing the Earth to warm up measurably, and there are already signs of disaster... I argue that this is happening because there are scientific facts to prove it...*

args.me

#2. *Out of 918 peer-reviewed scientific papers on this subject, 0% disagreed that climate change is happening, but in newspaper articles, 53% were unsure... This proves that climate change is happening, ...*

approach

#3. *Climate Change is causing the Earth to warm up measurably, and there are already signs of disaster ... reviewed scientific papers on this subject, 0% disagreed that climate ...*

Lucene

Argument and debate summarization: Discussion

▪ **General text summarization**

- Text summarization used to be said to be hard, but transformers have turned out very good on this task.
- Abstractive summarization is more complex, but more human-like.
- Extractive summarization may be more accepted in settings such as search.

▪ **Argument and debate summarization**

- An argument summary should represent the main reasoning well.
- In debates, it may be important to cover all main viewpoints.
- How much subjectiveness should be kept, depends on the application.

▪ **Why argument and debate summarization?**

- Not only in argument search, short argument summaries are needed.
- Getting an overview of different or longer arguments is important in many applications of computational argumentation.

Rationale behind: We cannot always consume all information out there.

Outline: Conclusion


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

- a) Introduction
- b) Argument composition and creation
- c) Argument rewriting and countering
- d) Argument and debate summarization
- e) Conclusion**

Conclusion

Argument generation

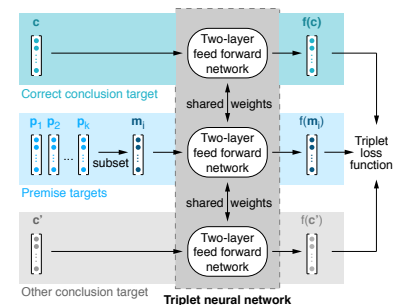
- Composition and creation of arguments
- Rewriting and countering of arguments
- Summarization of arguments and debates



The EU should allow rescue boats...

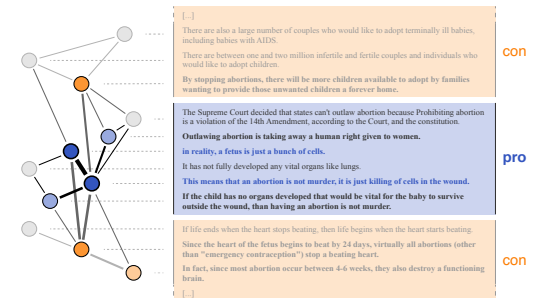
Composition and creation

- Classifier-based reuse of predicates in new units
- Contrastive learning for target reconstruction
- Controlled conclusion generation for assessment



Rewriting, countering, and summarization

- Joint conclusion and counterargument generation
- Reinforcement learning for appropriateness rewriting
- Extractive or abstractive summaries of texts



References

- **Akbik et al. (2018)**. Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.
- **Al-Khatib et al. (2021)**. Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. Employing Argumentation Knowledge Graphs for Neural Argument Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, to appear, 2021.
- **Alshomary et al. (2020a)**. Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. Target Inference in Argument Conclusion Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, 2020.
- **Alshomary et al. (2020b)**. Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth. Extractive Snippet Generation for Arguments. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1969–1972, 2020.
- **Alshomary et al. (2021a)**. Milad Alshomary, Wei-Fan Chen, Timon Gurcke, Henning Wachsmuth. Belief-based Generation of Argumentative Claims. *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, 2021.
- **Alshomary et al. (2021b)**. Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, Henning Wachsmuth. Counter-Argument Generation by Attacking Weak Premises. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, 2021.
- **Alshomary et al. (2023)**. Milad Alshomary and Henning Wachsmuth. Conclusion-based Counter-Argument Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, to appear, 2023.

References

- **Bar-Haim et al. (2017a)**. Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance Classification of Context-Dependent Claims. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 251–261, 2017.
- **Bar-Haim et al. (2020)**. Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. From Arguments to Key Points: Towards Automatic Argument Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, to appear 2020.
- **Bilu and Slonim (2016)**. Yonatan Bilu and Noam Slonim. Claim Synthesis via Predicate Recycling. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 525–530, 2016.
- **Bilu et al. (2019)**. Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkovich, Anael Malet, Assaf Gavron, Noam Slonim. Argument Invention from First Principles. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1013–1026, 2019.
- **Chakrabarty et al. (2021)**. Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. ENTRUST: Argument Reframing with Language Models and Entailment. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4958–4971, 2021.
- **Chung et al. (2024)**. H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL <http://jmlr.org/papers/v25/23-0870.html>.
- **Devlin et al. (2019)**. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages 4171–4186, 2019.

References

- **Egan et al. (2016)**. Charlie Egan, Advait Siddharthan, and Adam Wyner. Summarising the points made in online political debates. In Proceedings of the Third Workshop on Argument Mining (ArgMining2016), pages 134–143, 2016.
- **El Baff et al. (2019)**. Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, Manfred Stede, and Benno Stein. Computational Argumentation Synthesis as a Language Modeling Task. In Proceedings of the 12th International Conference on Natural Language Generation, pages 54–64, 2019.
- **Erkan and Radev (2004)**. Günes Erkan and Dragomir R Radev. LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization. Journal of Artificial Intelligence Research 22:457–479, 2004.
- **Gatys et al. (2015)**. Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style. CoRR abs/1508.06576, 2015.
- **Gero et al. (2019)**. Katy Gero, Chris Kedzie, Jonathan Reeve, Lydia Chilton. Low Level Linguistic Controls for Style Transfer and Content Preservation. In Proceedings of the 12th International Conference on Natural Language Generation, pages 208–218, 2019.
- **Gurcke et al. (2021)**. Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. Assessing the Sufficiency of Arguments through Conclusion Generation. In Proceedings of the 8th Workshop on Argument Mining, pages 67–77, 2021.
- **Habernal and Gurevych (2015)**. Ivan Habernal and Iryna Gurevych. 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.
- **He et al. (2021)**. P. He, J. Gao, and W. Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. CoRR, abs/2111.09543, 2021. URL <https://arxiv.org/abs/2111.09543>.

References

- **He et al. (2024).** Yuhang He, Jianzhu Bao, Yang Sun, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. Decomposing Argumentative Essay Generation via Dialectical Planning of Complex Reasoning. In Findings of the Association for Computational Linguistics: ACL 2024, pages 12305–12322, 2024.
- **Hidey and McKeown (2019).** Christopher Hidey and Kathy McKeown. Fixed That for You: Generating Contrastive Claims with Semantic Edits. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1756–1767, 2019.
- **Hua and Wang (2018).** Xinyu Hua and Lu Wang. Neural Argument Generation Augmented with Externally Retrieved Evidence. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 219–230, 2018.
- **Hua et al. (2019).** Xinyu Hua, Zhe Hu, and Lu Wang. Argument Generation with Retrieval, Planning, and Realization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2661–2672, 2019.
- **Jiang et al. (2023)** A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- **Johnson and Blair (2006).** Ralph H. Johnson and J. Anthony Blair. 2006. Logical Self-defense. International Debate Education Association.
- **Lewis et al. (2019).** Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, 2020. 45–55, 2015.

References

- **Liu et al. (2019).** Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, 2019.
- **Papineni et al. (2002).** Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, 2002.
- **Radford et al. (2018).** Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. OpenAI Blog, 2018.
- **Reiter and Dale (1997).** Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. Natural Language Engineering, 3(1):57–87.
- **Schiller et al. (2020).** Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Aspect-Controlled Neural Argument Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, to appear 2020.
- **Shen et al. (2017).** Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In Advances in Neural Information Processing Systems, pages 6833–6844, 2017.
- **Skitalinskaya et al. (2021).** Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1718–1729, 2021.
- **Skitalinskaya et al. (2023).** Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. Claim Optimization in Computational Argumentation. In Proceedings of the 16th International Natural Language Generation Conference, pages 134–152, 2023.

References

- **Stab (2017)**. Christian Stab. Argumentative Writing Support by means of Natural Language Processing, Chapter 5. PhD thesis, TU Darmstadt, 2017.
- **Stab and Gurevych (2017)**. Christian Stab and Iryna Gurevych. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 980–990, 2017.
- **Touvron et al. (2022)**. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL <https://arxiv.org/abs/2307.09288>.
- **van der Lee (2019)**. Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In Proceedings of the 12th International Conference on Natural Language Generation, pages 355–368, 2019.
- **Vaswani et al. (2017)**. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. Attention Is All You Need. In 31st Conference on Neural Information Processing Systems, 2017.
- **Walton (2009)**. Douglas Walton. 2009. Objections, rebuttals and refutations. pages 1–10.
- **Wang and Ling (2016)**. Lu Wang and Wang Ling. Neural Network-Based Abstract Generation for Opinions and Arguments. In: Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics, pages 47–57, 2016.

References

- **Ziegenbein et al. (2023).** Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. Modeling Appropriate Language in Argumentation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4344–4363, 2023.
- **Ziegenbein et al. (2024).** Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, Henning Wachsmuth. LLM-based Mitigation of Inappropriate Argumentation using Reinforcement Learning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, to appear, 2024.
- **Zukerman et al. (2000).** Ingrid Zukerman, Richard McConachy, Sarah George. Using Argumentation Strategies in Automated Argument Generation. In INLG'2000 Proceedings of the First International Conference on Natural Language Generation, pages 55–62, 2000.