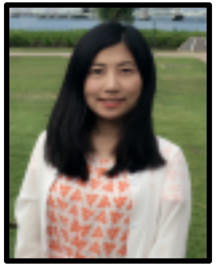
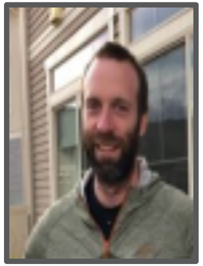
Knowledge-aware Text Generation: The Curious Case of Figurative Language and Argumentation

Smaranda Muresan (smara@columbia.edu)

Natural Language Processing
Columbia University

COLUMBIA UNIVERSITY
DATA SCIENCE INSTITUTE

Collaborators

Tuhin Chakrabarty	Aardit Trivedi	Nanyu (Violet)Peng	Debanjan Ghosh	Chris Hidey	Iryna Gurevych	Kevin Stowe
						

BELLARMINE COLLEGE PREPARATORY EST. 1851

Ucla NLP

ETS

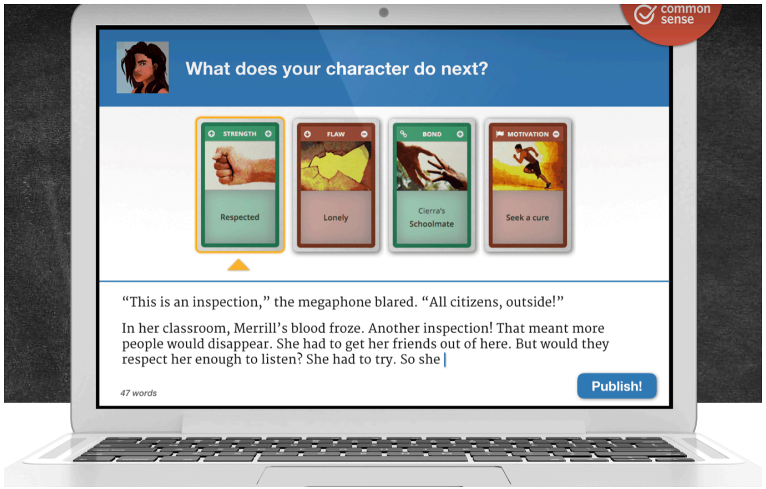
UBIQUITOUS KNOWLEDGE PROCESSING

TECHNISCHE UNIVERSITÄT DARMSTADT

Goals

- We want to generate figurative language (metaphors, similes, sarcasm) to promote more creative NLG output
  - Can make dialogue agents more engaging or humorous
  - Can be used as human-in-the loop tools, as writing assistants for creative (and argumentative/persuasive) writing process

STORIUMedu



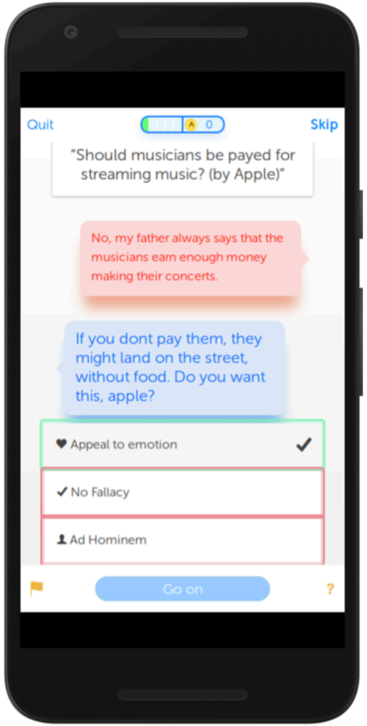
Metaphors are not to be trifled with. A single metaphor can give birth to love." (Kundera)

Goals

- We want to improve argument understanding by recovering implicit premises in an argument
- We want to improve the quality of civil discourse by reframing arguments in hyper-partisan or propagandistic discourse that contains logical and/or rhetorical fallacies (e.g., appeal to fear) to make them more trustworthy

=> Can be used as human-in-the-loop instructional assistants

Argotario (Habernal et al, 2017)



What we need!

- Addressing the lack of training data
- Getting insights from linguistic/argumentation theories
- Knowledge-aware models
- Evaluation methods and metrics

Our Recent Research Map

- Figurative Language Generation
  - **Metaphors** (NAACL 2021, ACL 2021)
  - Simile (EMNLP 2020)
  - Sarcasm (ACL 2020)

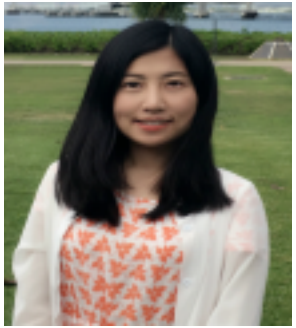
- Argument Generation
  - Argument Reframing (NAACL 2021)
  - **Generating Implicit Premises** (under submission EMNLP 2021)

MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding (NAACL 2021)

Collaborators:

Tuhin Chakrabarty

Nanyun Peng



Task Definition

- Given a literal input sentence generate a corresponding metaphoric sentence
- Simplifying assumption: focus on **verbs** as they are often the key component of metaphoric expressions (Steen et al., 2010; Martin, 2006).

Literal Input1	The wildfire **spread** through the forest at an amazing speed.
GenMetaphor1	The wildfire **danced** through the forest at an amazing speed.
Literal Input2	The window panes were **rattling** as the wind blew through them
GenMetaphor2	The window panes were **trembling** as the wind blew through them

"Metaphors are not to be trifled with. A single metaphor can give birth to love." (Kundera)

Key Challenges

- How to address lack of training data: (literal, metaphorical)
- How to ensure the generated metaphoric sentence has the same meaning as the literal one
- How to overcome the tendency of generative language models to produce literal text over metaphorical one

“Metaphors are not to be trifled with. A single metaphor can give birth to love.” (Kundera)

Insight

- Theoretically-grounded relation between *metaphors* and *symbols*

"A metaphor is not language, it is an idea expressed by language, an idea that in its turn functions as a symbol to express something" (Susanne Langer)

Approach

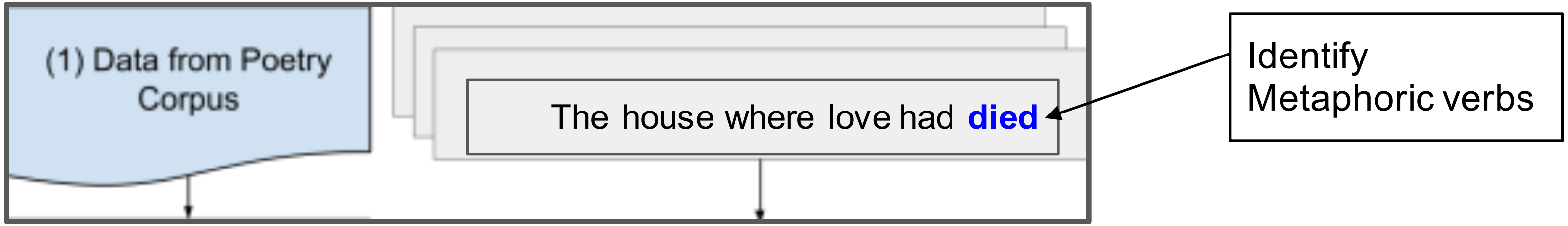
- 1) Automatically create a parallel dataset of sentence pairs (literal, metaphoric)
  - Identify metaphoric sentences (metaphoric verbs)
  - Generate literal equivalents that are *semantically consistent*

- 2) Fine-tune a seq2seq model (BART (Lewis et al 2019)) on our parallel data and use a discriminator to guide the decoding process

- Asses quality of generated metaphors through intrinsic and task-based evaluations

Automatic Creation of Parallel Data

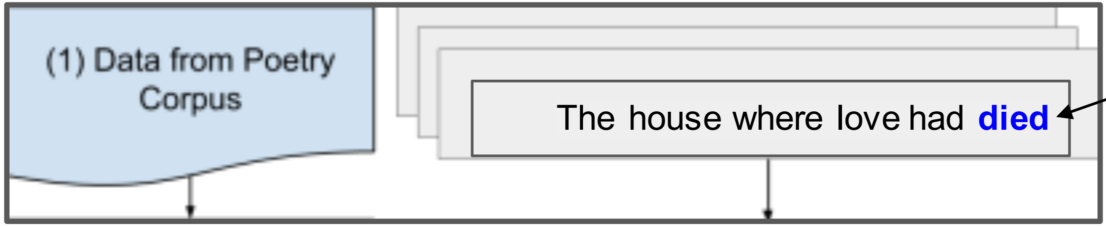
[Gutenberg Poetry Corpus]



- Use BERT model fine-tuned on VUA dataset (Steen et al 2010) to identify metaphoric verbs.
- Chose sentences where BERT model predicts verb(s) as metaphoric with confidence score of 95%(i.e., prediction probability 0.95).

Automatic Creation of Parallel Data

Gutenberg Poetry Corpus



Identify Metaphoric verbs

Convert to *literal* meaning that is *semantically consistent*

Generate Literal Meaning

- Use Masked Language Model infilling (e.g., BERT) to generate verbs that have a literal sense

Re-ranked based on inverse metaphoricity score

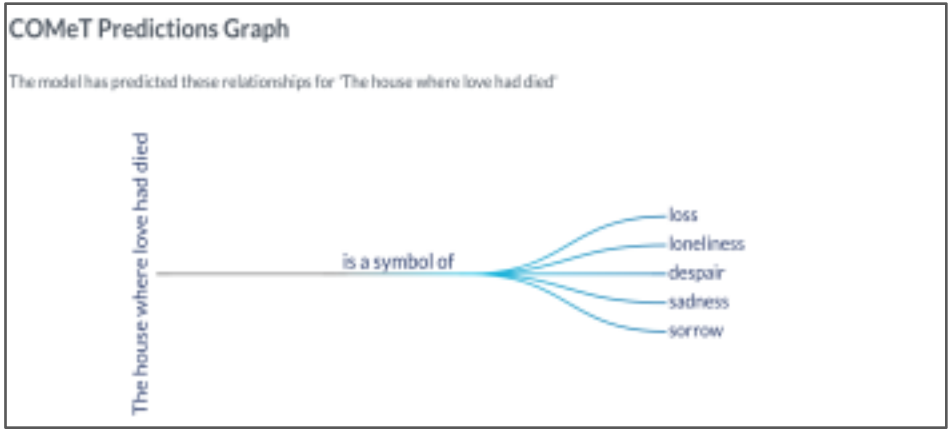
The house where love had **MASK**

started (0.003)
originated (0.004)
been (0.004)
…
ended (0.01)

Semantic Consistency

- We want semantic consistency with the metaphorical verb
- Use an *adapted knowledge model, COMeT (Bosselut et al., 2019)* (GPT-2 model fine-tuned on ConceptNet) with the *SymbolOf* relation

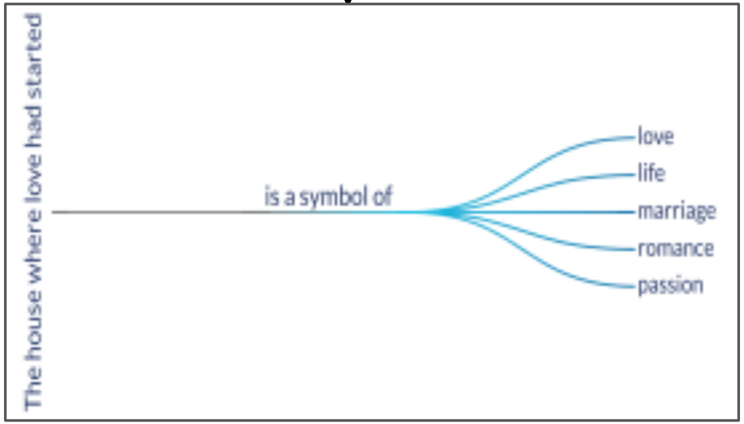
The house where love had **died**



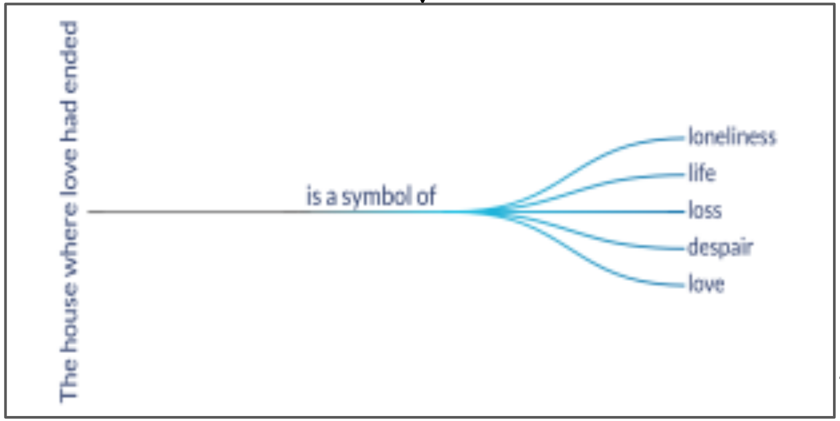
Generate literal meaning

- We want semantic consistency with the metaphorical verb
- Use an *adapted knowledge model, COMeT (Bosselut et al. 2019)* (GPT-2 model fine-tuned on ConceptNet) with the *SymbolOf* relation

The house where love had **started**



The house where love had **ended**



Automatic Creation of Parallel Data

[Gutenberg Poetry Corpus]

Identify Metaphoric verbs

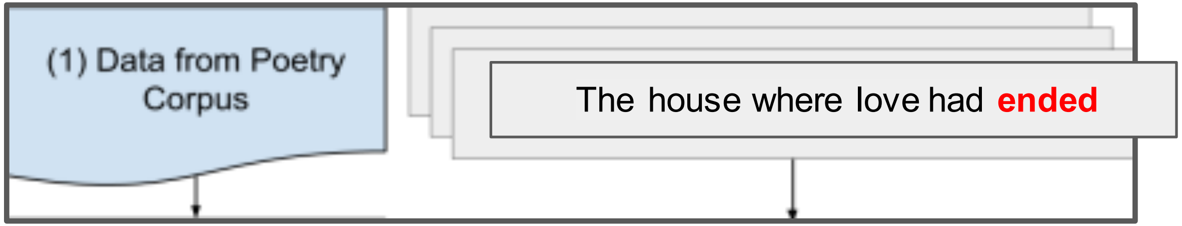
+

COMET

Symbol Of

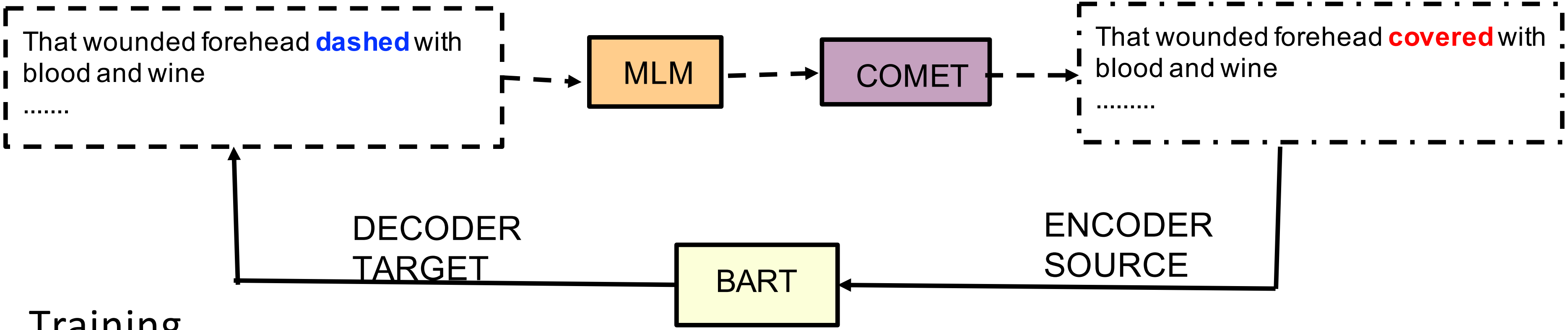
Loss, Loneliness, Despair,..

Convert to Literal: MLM + re-ranked by semantic meaning overlap



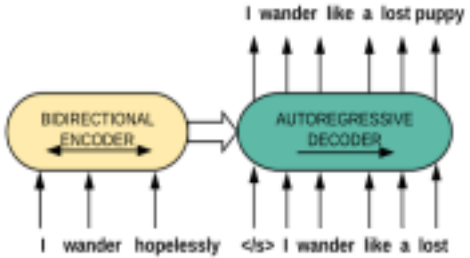
Metaphor Generation

- Created parallel data: 90K training, ~3k validation



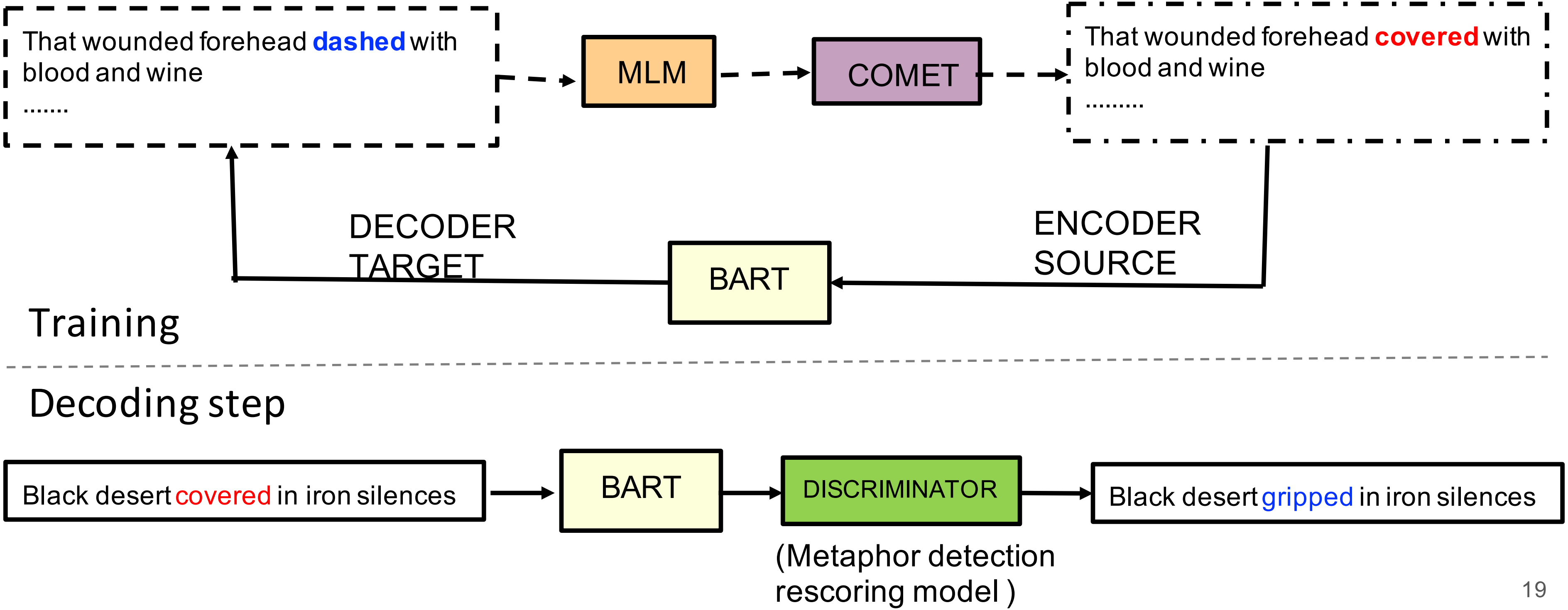
Training

Fine-tune BART (Lewis et al 2019): pre-trained seq2seq model



Metaphor Generation

- Created parallel data: 90K training, ~3k validation



Intrinsic Evaluation

- Test set
  - Source1: literal examples from Mohammad et al (2016)
  - Source2: literal examples from r/WRITINGPROMPT and r/OCPOETRY
  - Randomly select 150 examples
  - Ask 2 literary experts to generate metaphors
- Baselines
  - Lexical Replacement (LexRep): MLM+COMET
  - Metaphor Masking (META_M) (Stowe et al, 2020)
  - Fine-tuned BART (our model without the discriminator)
- Evaluation Criteria:
  - Fluency, Meaning Preservation, Creativity, Metaphoricity
  - Scale 1 (worst) – 5 (best)
- Mturk: 5 crowdsource workers per HIT

Intrinsic Evaluation

System	Flu	Mea	Crea	Meta
HUMAN1	**3.83**	**3.77**	**4.02**	**3.52**
HUMAN2	3.29	3.43	3.58	3.16
LEXREP	2.21	2.59	2.16	1.98
META_M	2.10	1.91	2.00	1.89
BART	3.33	3.08	3.16	2.85
MERMAID	3.46	3.35	3.50	3.07

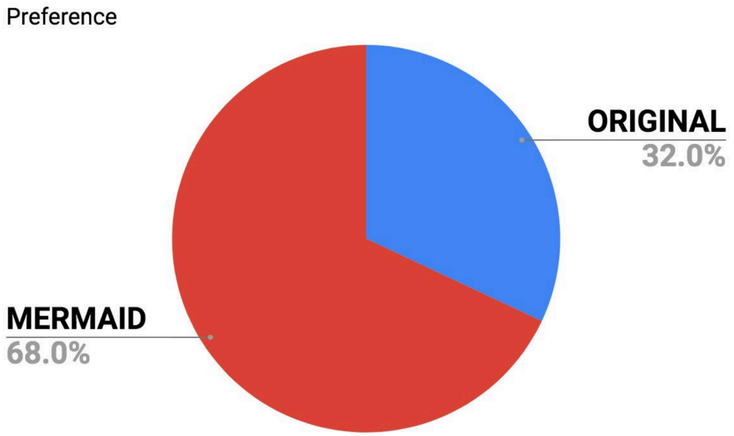
Intrinsic Evaluation

Input	System	Output	FL	MP.	C.	MF
My heart *beats* when he walks in the room	HUMAN1	My heart *skips* when he walks in the room	4.7	**5.0**	4.0	**4.3**
	HUMAN2	My heart *sings* when he walks in the room	**5.0**	4.3	3.7	3.3
	LEXREP	My heart *made* when he walks in the room	1.0	1.0	1.0	1.0
	META_M	My heart *came* when he walks in the room	1.7	1.0	1.3	1.3
	BART	My heart *sings* when he walks in the room	**5.0**	4.3	3.7	3.7
	MERMAID	My heart *jumps* when he walks in the room	4.7	4.7	**4.3**	4.0
After a glass of wine, he *relaxed* up a bit	HUMAN1	After a glass of wine, he *loosened* up a bit	**4.7**	**5.0**	**5.0**	**4.0**
	HUMAN2	After a glass of wine, he *unfurled* up a bit	2.0	**5.0**	2.0	3.7
	LEXREP	After a glass of wine, he *followed* up a bit	3.7	1.0	2.7	1.7
	META_M	After a glass of he *touched* up a bit	1.3	1.0	1.7	2.0
	BART	After a glass of wine, he *dried* up a bit	2.7	1.0	2.3	2.0
	MERMAID	After a glass of wine, he *loosened* up a bit	4.3	**5.0**	**5.0**	3.7

Task-based Evaluation

- Replace literal verbs in poems with the metaphorical verbs
- Collect poems from r/OCPoetry (limed to 4 sentence stanza)

- Ask Turkers whether the original version or the re-written version is better



What are we still missing?

- *More theoretical insights*

- **Conceptual Metaphor Theory (CMT)** (Lakoff and Johnson, 1980) holds that we use conceptual mappings between domains (conceptual structures that group related concepts) to generate linguistic metaphors.

- Metaphoric mappings consist of a **Source** and a **Target** conceptual domain. The source domain is the conceptual domain from which we draw the metaphorical expressions, while the target domain is the conceptual domain that we try to understand.

What are we still missing?

A classical mapping is ARGUMENT IS WAR

They ~~argued~~ fought against the contract.
They ~~supported~~ defended their new proposal

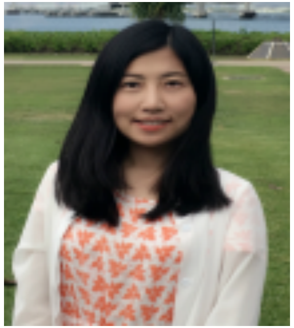
Metaphor Generation with Conceptual Mappings (ACL 2021)

Collaborators:

Kevin Stowe

Tuhin Chakrabarty

Nanyun Peng



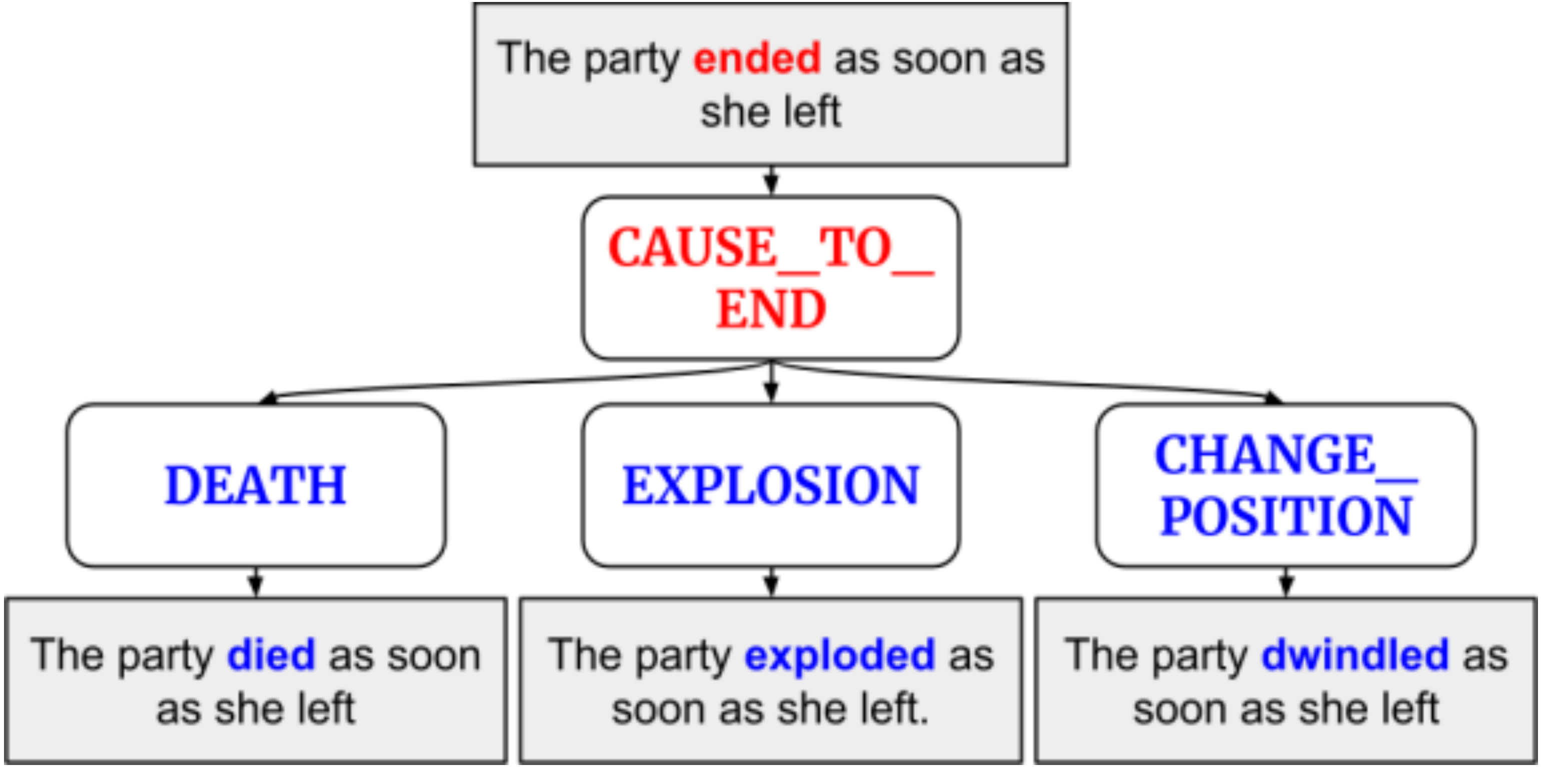
Iryna Gurevych

Task Definition

- Given a literal input sentence that evokes a target domain we generate metaphoric sentences that evoke desired output corresponding to the source domain.

- We propose a novel framework for metaphor generation informed by conceptual metaphor theory (CMT).

- We focus only on **verbs** as they are often the key component of metaphoric expressions (Steen et al., 2010; Martin, 2006).

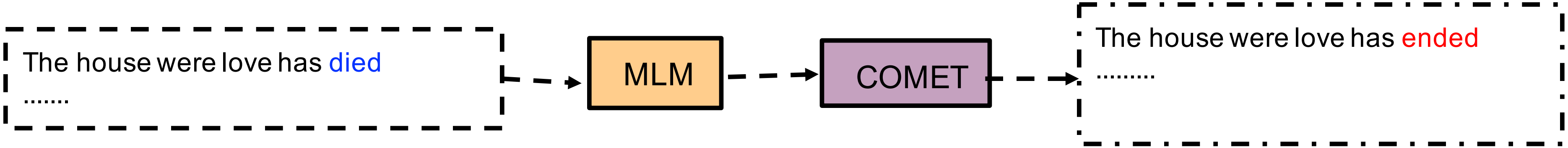

The party ended as soon as she left
CAUSE_TO_END
DEATH
EXPLOSION
CHANGE_POSITION
The party died as soon as she left
The party exploded as soon as she left.
The party dwindled as soon as she left

Approach

- 1) Automatically create a parallel dataset of sentence pairs (literal, metaphoric)
  - Identify metaphoric sentences (metaphoric verbs)
  - Generate literal equivalents that are semantically consistent
  - ***Label the metaphoric and literal verb with SOURCE and TAGERGET domains***
- 2) Fine-tune a seq2seq model (BART) on our parallel data where input is augment with SOURCE and TARGET info ***as controlled codes (CM-BART)***

- Asses quality of generated metaphors through intrinsic evaluations

Automatic Creation of Parallel Data



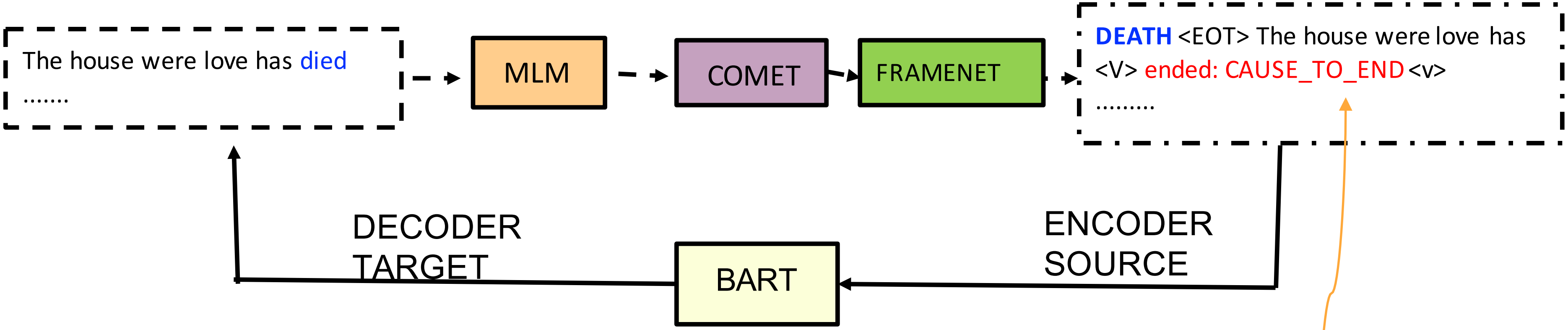
+

Tagging sentences with SOURCE and TARGET domain using FRAMENET (Open-SESAME parser (Swayamdipta et al., 2017))

*The house where love had died/**DEATH***

*The house where love had ended/**CAUSE_TO_END***

Metaphor Generation

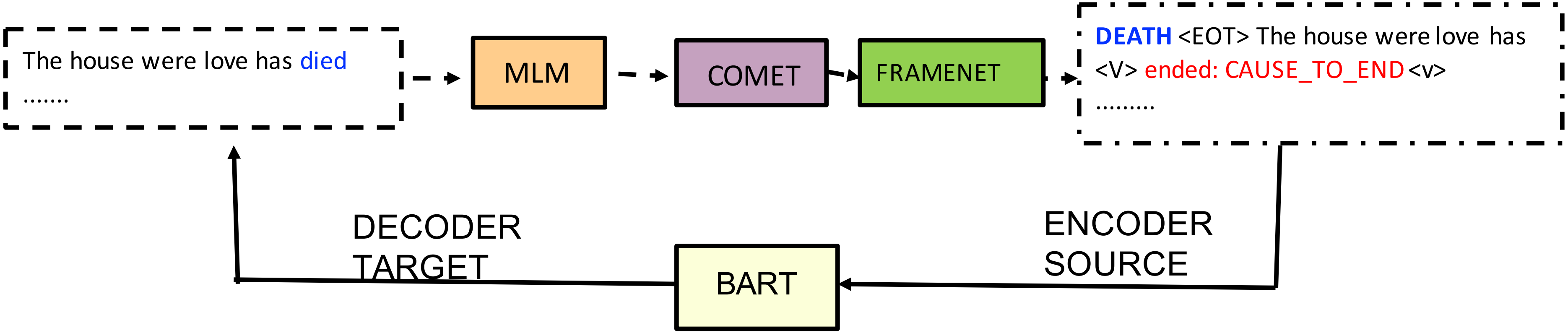


Training

Add the SOURCE and TARGET domains (FrameNet frames) as **controlled codes** in the input following idea by Shiller et la (2020)

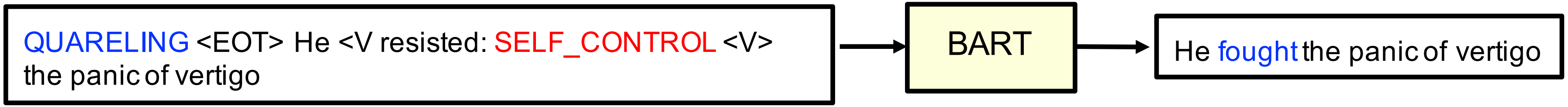
Fine-tune BART (Lewis et al, 2019) on this augmented parallel data

Metaphor Generation



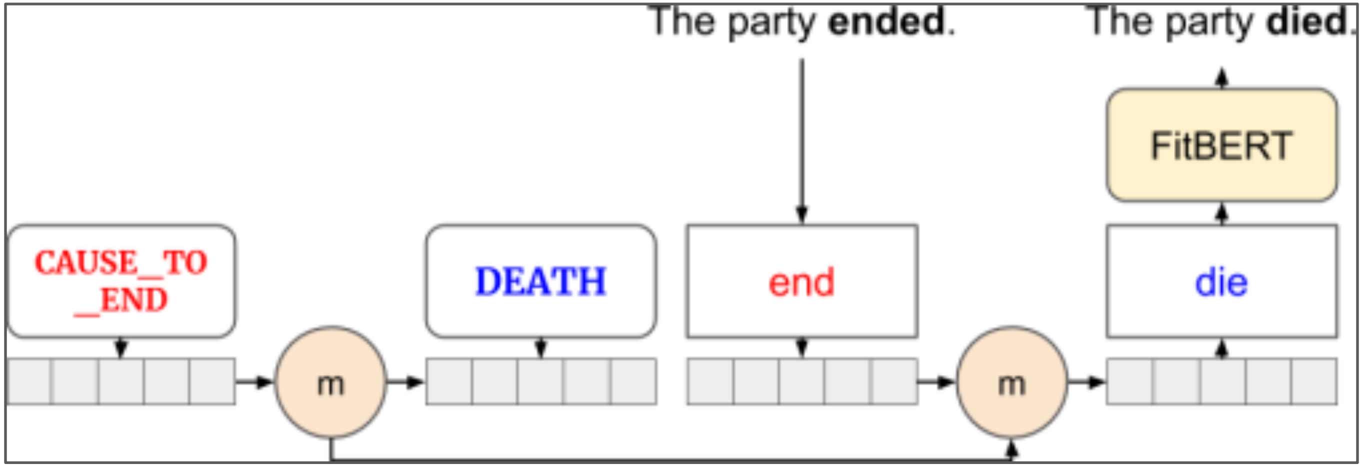
Training

Decoding step



Baselines

CMLEX: unsupervised lexical model relying on frame embeddings learned from corpora tagged with FrameNet frames



MERMAID: BART based model based on discriminative decoding

Intrinsic Evaluation

- Test set (Gold)
  - Several sources for metaphors: Gutenberg Poetry Corpus, Mohammad et 2016, Brown Corpus
  - Construct literal meaning and label with FrameNet frames.
- Test for unknown metaphors
  - Unseen and rare conceptual mappings

Intrinsic Evaluation

- For gold test set
- Automatic Metrics
  - Distance from gold metaphor (GM) – use SBERT (Reimers and Gurevych, 2019)
  - Relation distance: minimize distance between cos(L,M) and cos(L, GM)
  - % of times when generated metaphor is exactly the same as gold metaphor
- Human Evaluation
  - 3 experts
  - 2 criteria: Metaphoricity and Relevance to Source Domain
  - Scale 1-4 (with 0 for unintelligeable output)

Intrinsic Evaluation

System	Distance from Gold	Relation distance	% same as GM
MERMAID	0.147	0.087	13.3
CM-LEX	0.151	0.086	10.7
CM-BART	**0.085**	**0.047**	**29.3**

System	Metaphoricity	Relevance Source
MERMAID	2.56	2.12
CM-LEX	2.34	2.43
CM-BART	**2.72**	**2.87**

INPUT / TARGET /SOURCE	SYSTEM	OUTPUT	MET	SRC
A dim aurora **rises** in my east **CHANGE_POSITION_ON_A_SCALE** **IS** **RESIDENCE**	Gold	A dim aurora **lives** in my east		
	CMLEX	A dim aurora **stands** in my east	3	3
	MERMAID	A dim aurora **hangs** in my east	3	2
	CM_BART	A dim aurora **lives** in my east	4	4
He **resisted** the panic of vertigo **SELF_CONTROL** **IS** **QUARALLING**	Gold	He **fought** the panic of vertigo		
	CMLEX	He **confrontations** the panic of vertigo	0	0
	MERMAID	He **felt** the panic of vertigo	1	2
	CM_BART	He **disputed** the panic of vertigo	3	4

Overview

- Figurative Language Generation
  - Metaphors (NAACL 2021, ACL 2021)
  - Simile (EMNLP 2020)
  - Sarcasm (ACL 2020)

- Argument Generation
  - Argument Reframing (NAACL 2021)
  - **Generating Implicit Premises** (under submission EMNLP 2021)

Implicit Premise Generation with Discourse-aware Commonsense Knowledge Models (under submission EMNLP 2021)

Collaborators:

Tuhin Chakrabarty

Aadit Trivedi

Enthymeme Reconstruction

- **Enthymeme**: an incomplete argument found in discourse, where some components are explicit, but other *propositions are left implicit* and need to be *filled in* as premises or conclusions to fully understand what the argument is

- *Sherlock Holmes' Silver Blade case*

  "A dog was kept in the stable, and yet, though someone had been in and fetched out a horse, he had not barked enough to rouse the two lads in the loft. Obviously, the midnight visitor was someone whom the dog knew well."

  Missing Premise: *Dogs generally bark when a person enters an area unless the dog knows the person well.*

Task Definition and Key Challenges

- **Task**: given an enthymeme consisting of a stated conclusion and a stated premise, generate the implicit/missing premise.
- **Key Challenges**:
  - The lack of large scale data of incomplete arguments together with annotated missing premises needed to train a sequence-to-sequence model
  - The inherent need to model commonsense or word knowledge.

Insight

- **Argumentation Theory:**
  - Incomplete arguments in naturally occurring discourse, more often than not, require abductive reasoning (plausible explanations) rather than the more strict form of reasoning based on deductive logic (Walton and Reed, 2005; Sabre, 1990)
  - Silver Blaze case is such an example

Approach

- Leverage abductive reasoning as an auxiliary task
    - Fine-tune BART (Lewis et al 2019) on the *Abductive Reasoning in Narrative Text (ART)* dataset (Bhagavatula et al. 2020)
- Encode discourse-aware common sense knowledge
    - Use PARA-COMET (Gabriel et al., 2021), a discourse-aware knowledge model that incorporates paragraph-level information to generate coherent commonsense inferences from narratives.

Example

Reason	Vaccinations save lives
Claim	Vaccination should be mandatory for all children
ZeroShot	Vaccines save lives, they save money
Fine-tuned on *ART*	Vaccinations are the best way to protect children.
Fine-tuned on *ART +PARA-C*	Vaccinations are the best way to prevent childhood diseases.

Fine-tune BART on ART

- *Abductive Reasoning in Narrative Text (ART) dataset* (Bhagavatula et al. 2020)
    - *Generated using crowdsourcing:* given 2 observations (O1 and O2) generate the most plausible and implausible hypotheses that explain the observations (O1 and O2 are taken from ROCStories dataset)
    - Adversarial filtering to keep most challenging plausible and implausible hypothesis
    - 50,481 training instances
- Fine-tune BART on ART dataset
    - Encoder input: *O1 [SEP] O2* Decoder Output: *O1. And since* ***H.*** *O2*

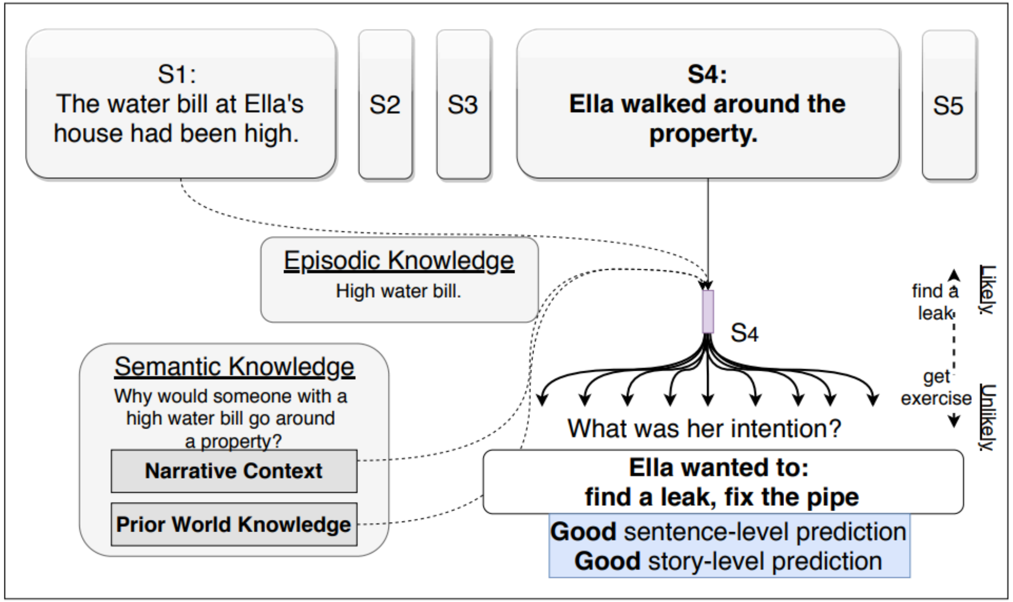
Amy was looking through her mother's old scrapbooks. [SEP] Amy realized her mother had dated her history professor.

→ BART →

Amy was looking through her mother's old scrapbooks. And since ***Amy found pictures of her history professor and mother together****. Amy realized her mother had dated her history professor.*

Discourse-aware common sense knowledge

- Use PARA-COMET (Gabriel et al., 2021): an extension of COMET pre-trained on ATOMIC (Sap et al., 2019) able to generate discourse-aware common sense knowledge.
  - ATOMIC: inferential knowledge organized as typed if-then relations with variables (centered on events)



Example from (Gabriel et al., 2021)

Discourse-aware common sense knowledge

- Use PARA-COMET (Gabriel et al., 2021): an extension of COMET pre-trained on ATOMIC (Sap et al., 2019) able to generate discourse-aware common sense knowledge.
  - ATOMIC: inferential knowledge organized as typed if-then relations with variables (centered on events)
- Input: a ”discourse” formed from the two observation form ART [O1, O2]
- Output: 9 common sense relations for both O1 and O2; after experimentation we chose *xIntet for O1 (xIntent = PersonX wanted to e2)*
- Fine-tune BART on [O1, commonSense, O2]

Amy was looking through her mother's old scrapbooks. [SEP] **to find something** [SEP] *Amy realized her mother had dated her history professor.*

→ BART →

*Amy was looking through her mother's old scrapbooks. And since **Amy found pictures of her history professor and mother together**. Amy realized her mother had dated her history professor.*

Experimental Setup

- **Test sets**: 3 different datasets of enthymemes annotated with human generated implicit premises
  - D1: 1651 enthymemes from Argument Reasoning Comphrehension Task (Habernal et al 2018)
  - D2: 494 enthymemes from online forum + human generated implicit premises (Boltužic and Šnajder, 2016)
  - D3: 112 enthymes from MicroText Corpus + human generated implicit premises (Becker et al. (2020)
- **Models:**
  - BART (zero-shot);
  - BART finetuned on ART;
  - BART finetuned on ART+PARA-COMET

Experimental Setup

- Automatic Evaluation
  - BLEU metric
  - BERTScore: a metric for evaluating text generation using contextualized embeddings.
- Human Evaluation: crowdsourcing on AMTurk
  - 50 enthymemes from each test set (total of 150 enthymemes)
  - Models: fine-tune BART (with or without PARA-COMET)
  - Given an enthymemes Turkers were asked if the generated implicit premises were plausible or not (agreement: 0.56 Krippendorff's α)

Results

Data	System	BLEU1	BLEU2	BS
D1	ZeroShot	6.02	2.17	42.88
	ART	9.16	3.11	48.35
	+PARA-COMET	**10.56**	**3.90**	**50.22**
D2	ZeroShot	28.24	15.13	46.96
	ART	37.77	18.76	60.63
	+PARA-COMET	**44.12**	**24.14**	**67.75**
D3	ZeroShot	12.58	6.25	44.64
	ART	14.89	6.34	51.78
	+PARA-COMET	**15.56**	**7.50**	**53.38**

Data	System	Plausibility
D1	ART	50%
	+PARA-COMET	**58%**
D2	ART	48%
	+PARA-COMET	**66%**
D3	ART	38%
	+PARA-COMET	**56%**

Generated Output

D1	St Premise	Deaf students need more specialized education
	St Claim	States need special schools for the deaf
	Gold	Their parents can't always enroll them in a deaf private school
	Zero-shot	We can't afford it, we shouldn't
	ART	The deaf students are not getting enough education.
	+PARA-COMET	Deaf students are not **being served well in the schools**

D2	St Premise	Bush new spending in 8 years? $5.07 TRILLION Obama total New Spending (projected out for the next 8 years)? $1.44 TRILLION. And of that total, only $430 billion is non-recession related.
	St Claim	Fixed the economy
	Gold	Obama spends less money than Bush.
	Zero-shot	We are talking about the economy
	ART	The Obama administration has spent $1 trillion.
	+PARA-COMET	The Obama's **spending is much less** than Bush's.

D3	St Premise	The morning-after pill has a number of side effects.
	St Claim	The morning-after pill should only be prescribed after counselling by a physician or pharmacist.,
	Gold	Physicians and pharmacists inform about side effects.
	Zero-shot	Morning-after pills are not FDA approved, they should be avoided .
	ART	The morning- after pill can cause depression.
	+PARA-COMET	The side effects **can be very serious**.

Take-Aways

- Blindly using large LM will not lead to good generation of figurative language and/or arguments

- Use *theoretical insights* about the phenomena

- *Use common sense knowledge/ connotative knowledge*, to plan content and add control to NLG systems.
  - open question: other type of knowledge (e.g., *social and cultural norms*)

- *Evaluation metrics and methods* are important (human-based evaluation is needed; task-based)
  - open question: what about appropriate automatic metrics?

Last Words

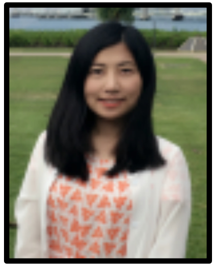
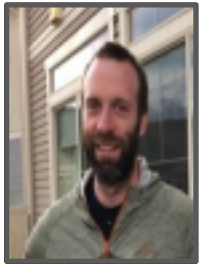
"Metaphors are not to be trifled with." (Kundera)

↓

Theoretical insights/Knowledge-aware models/Evaluation Metrics are not to be trifled with (Smara)

They can give birth to love of NLG for figurative language and argumentation!!!!

THANK YOU!!

Tuhin Chakrabarty	Aardit Trivedi	Nanyu (Violet)Peng	Debanjan Ghosh	Chris Hidey	Iryna Gurevych	Kevin Stowe
						

Data and Code: https://github.com/tuhinjubcse/