Be Constructive, Helpful and Empathetic: Towards Situated Empathic Dialogue Generation Grounded in Social and Moral Commonsense Knowledge

20th January 2022
Finalist' Colloquium WT 2021/2022

Benjamin Beilharz – Advisor Prof. Anette Frank / Co-Advisor Prof. Michael Herweg

Outline

- Motivation
- Turning an Oracle into a Conversational System
- Knowledge Resources
- Modeling Methodology
- Evaluation
- Status Quo
- References

Motivation

- current conversational systems are mostly reactive, aiming towards answering a specific query
- but making systems capable of understanding the situation of the user is crucial for future human-computer collaboration
- such that we move further than language, onto the cognitive/ perceptional level, because
- empathy is key for conversational systems to bond with the users
- while use cases are omnipresent: emergency systems, mental health applications, human-computer interaction
- and Voice User Interfaces are being deployed more and more often, why not just make them better?

Let’s ask an oracle…

DELPHI

Jiang et al. 2021 (preprint)

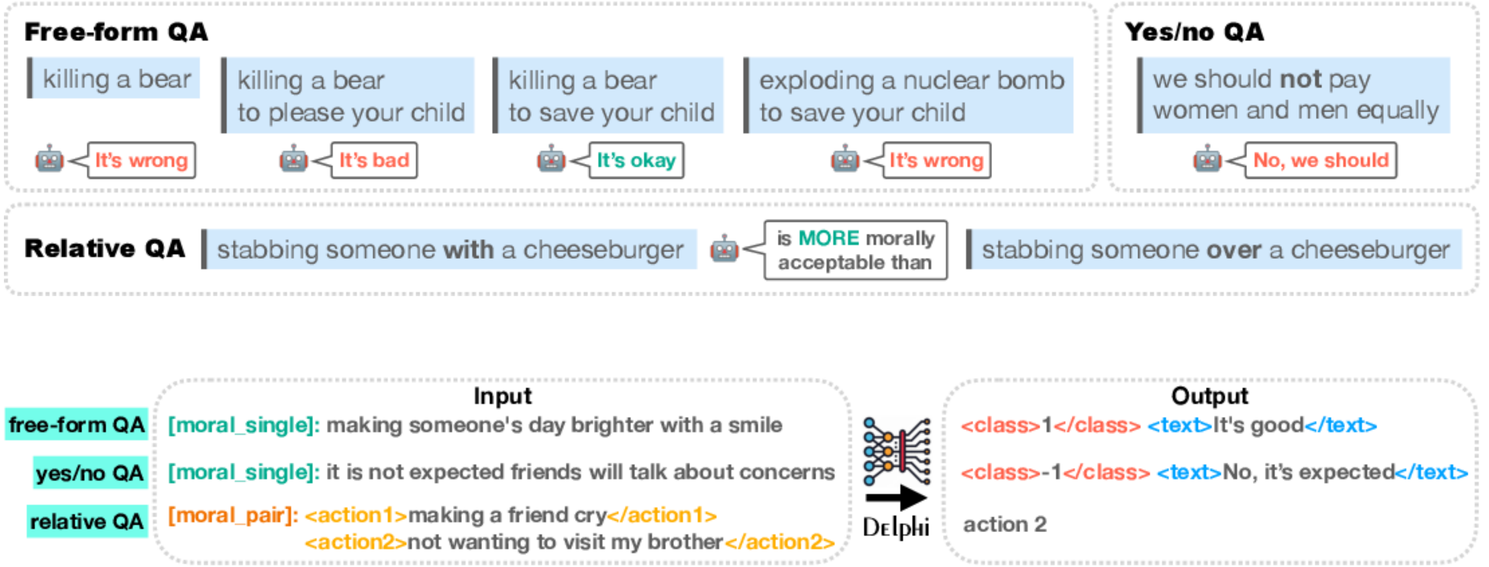
- A neural commonsense model able to create fine-grained moral judgments
- Based on multiple datasets targetting moral, ethical, and social commonsense
- UNICORN: Google's T5 fine-tuned on RAINBOW, a commonsense benchmark in QA format
- DELPHI: UNICORN fine-tuned on Commonsense Norm Bank

Datasets in Commonsense Norm Bank (not yet released due to preprint)

Task	Data	Type	Examples	Judgment
Free-form	SOCIAL CHEM	A	Change plans if there's a good reason	It's okay
		Q(A)	**Can I** change plans if there's a good reason**?**	
		A+S	Change plans if there's a good reason, **when** getting pissed with spontaneous change of plans	
		Q(A+S)	**Is** changing plans if there's a good reason **good**, given getting pissed with spontaneous change of plans**?**	
	ETHICS	A	I used the food with permission	It's good
		Q(A)	**Is** I used the food with permission **a good behavior?**	
	MORAL STORIES	A	Mike goes to a boxing gym to hit heavy bags	It's fine
		Q(A)	**Is** Mike going to a boxing gym to hit heavy bags **ok?**	
		AS	Mike goes to a boxing gym to hit heavy bags, **given that** Mike failed a big test at school and is frustrated	
		Q(A+S)	**Is** Mike going to a boxing gym to hit heavy bags **ok**, **when** Mike failed a big test at school and is frustrated**?**	
		A+S+I	Mike goes to a boxing gym to hit heavy bags, **when** Mike failed a big test at school and is frustrated, **and** he wants to release his frustrations physically	
		Q(A+S+I)	**Is** Mike going to a boxing gym to hit heavy bags **ok**, **if** Mike failed a big test at school and is frustrated, **and** he wants to release his frustrations physically**?**	
	SBIC	A	**Posting** guys, I beat cancer patients	It's bad
		Q(A)	**Is it good to say** guys, I beat cancer patients**?**	
Yes/No	SOCIAL CHEM	PosRoT	**It's okay** to turn down trips you don't want to attend	Yes, it's ok
		NegRoT	**It's wrong** to turn down trips you don't want to attend	No, it's ok
Relative	SCRUPLES	Action1	Going to bed earlier than my roommate	1 > 2
		Action2	Not being honest to my parents about university	

DELPHI is an Impressive Oracle, but...

- is trained to give “just” a short moral judgment
- lacks expression of the rationale for the moral judgment
- is a one-turn QA model which is not capable of engaging in a dialogue with a user



How Might We...

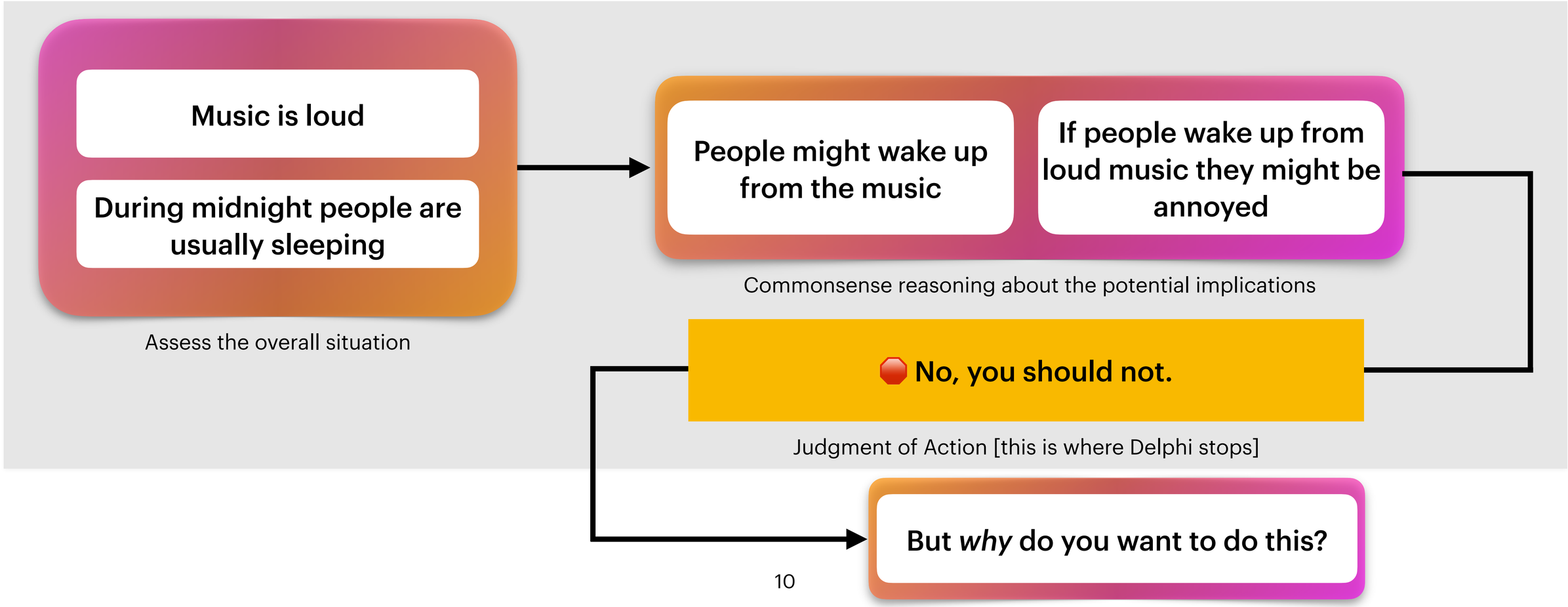
...Turn an Oracle into a Conversational System?

A Flawless Oracle Which Is Able To Give Good Advice...

- **has acquired commonsense knowledge** of the physical and social world, and knows about the possible consequences of actions
- **is able to assess** a real-world scenario and its context based on a description
- **can analyze the situation** based on both commonsense knowledge, moral theories, or a dialogue history
- **is expected to judge what action is appropriate to support the user**, can ask clarifying questions, or synthesize considerations

Situative Assessment

Can I listen to music past midnight if people are around me?



Looking Behind the Surface

But why do you want to do this?

Asking for more information about the context

I know, but I really can't sleep. This night is bad, I feel anxious and I know music will help me fall asleep.

Person feels anxious

Music can help to cope with anxiety

New information about the person is being processed

Be Empathetic

I see, I am sorry you do not feel well.

Empathetic Responding

Be Helpful and Supportive

Maybe you can use headphones so you will not wake up other people around you.

Consideration Generation

Turn Types

Information-Seeking to Considerations

Can I listen to music past midnight if people are around me?

Why would you want to listen to music at midnight? People usually want to sleep during this time. I don't think you should do that.

Trying to gather more information about user's context

I know, but I really can't sleep. This night is bad, I feel anxious and I know music will help me fall asleep.

I see, I am sorry you do not feel well. Maybe you can use headphones so you will not wake other people around you up.

Pointing out considerations which satisfy user's needs w.r.t. external constraints

Information-Seeking to Empathetic-Directives

My boyfriend and I just broke up.

I don’t know, I am just devastated.

That’s a good idea.

How did that happen?!
Do you want to talk about it?

Trying to gather more information about user’s context

Losing someone you love is never easy, perhaps it’s a good time to do something with your friends?

Pointing out considerations which satisfy user’s needs

Always remember, you’re not alone! Enjoy the time with your friends!

Give advice and be empathetic

Forward-Looking Directives

Can I skip on work to sleep longer?

I can understand that you might be tired, but I don't really think you should do that.
Skipping on work might have consequences.
Maybe you can take a nap after work.

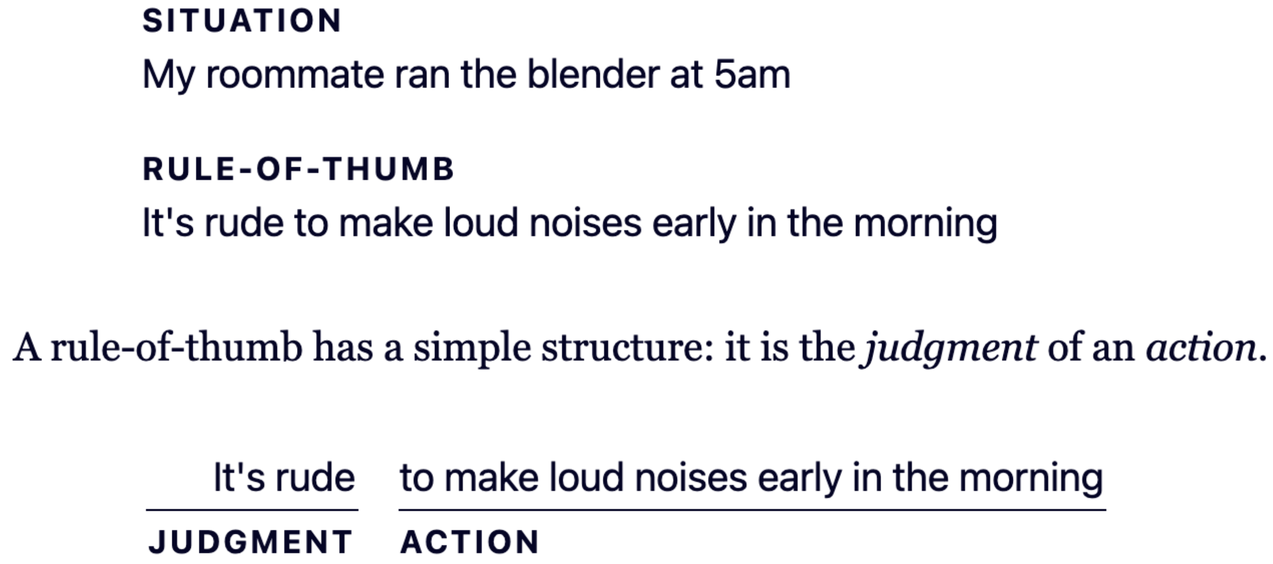
Being supportive and giving forward-looking advice

Knowledge Resources

Social Chemistry 101

Sap et al. 2020

- Events are mapped towards **moral & social norms**, called Rule-of-Thumbs (RoT)



Social Chemistry 101

Sap et al. 2020

- But also contains information about:
  - Social Judgment
  - Agreement
  - Moral Foundation
  - Legality
  - and Cultural Pressure

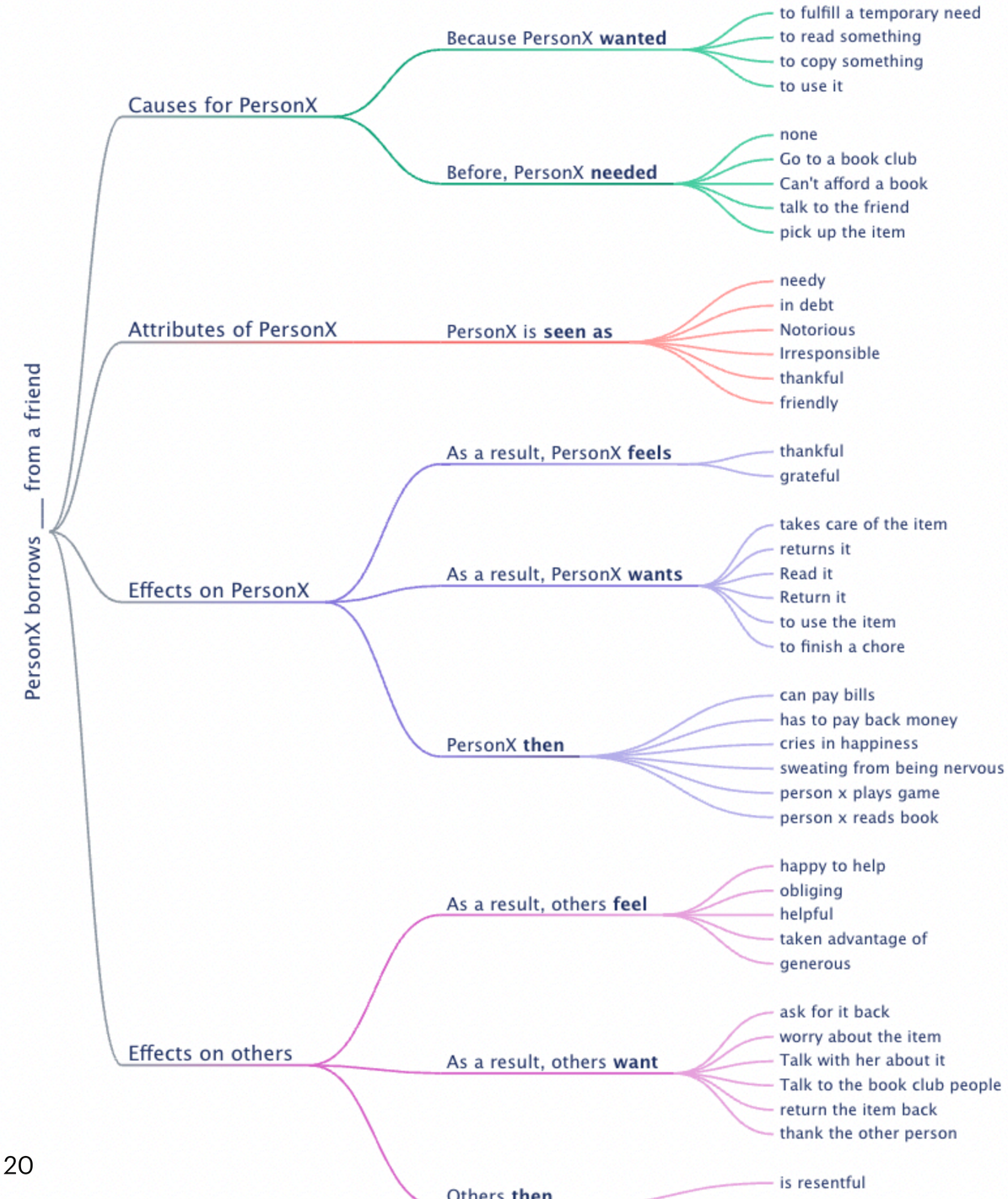
Atomic2020

Hwang et al. 2020

- Knowledge Graph derived from ConceptNet and enriched with social, event, and physical knowledge

Preparing Atomic To Access Knowledge

- Atomic has person X-Z placeholders
- (Object placeholder with an *isFilledBy* relation is filled)
- Dependency parse and create a cumulated lookup table
- Extract objects and verbs for each entry



Preparing Atomic To Access Knowledge

PersonX stole PersonY's bike.

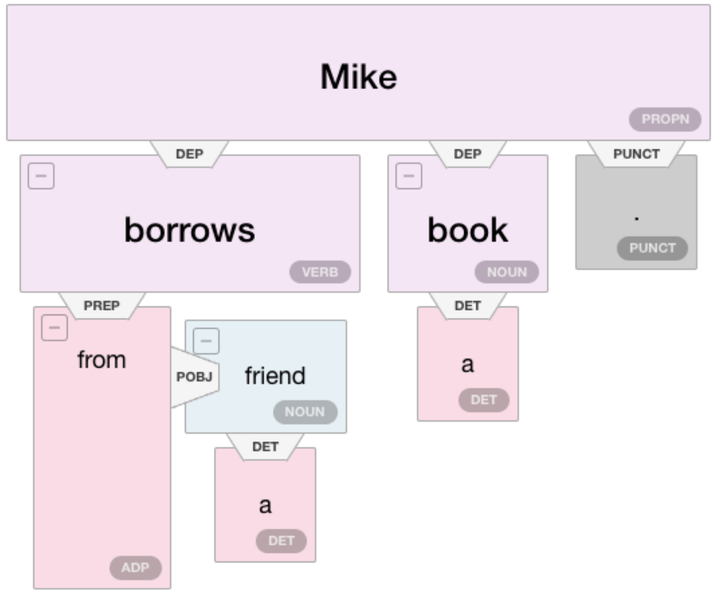
Amelie stole Mike's bike.

Extract verbs and objects for lookup table

Stole
dobj: bike

PersonX borrows ___ from a friend.

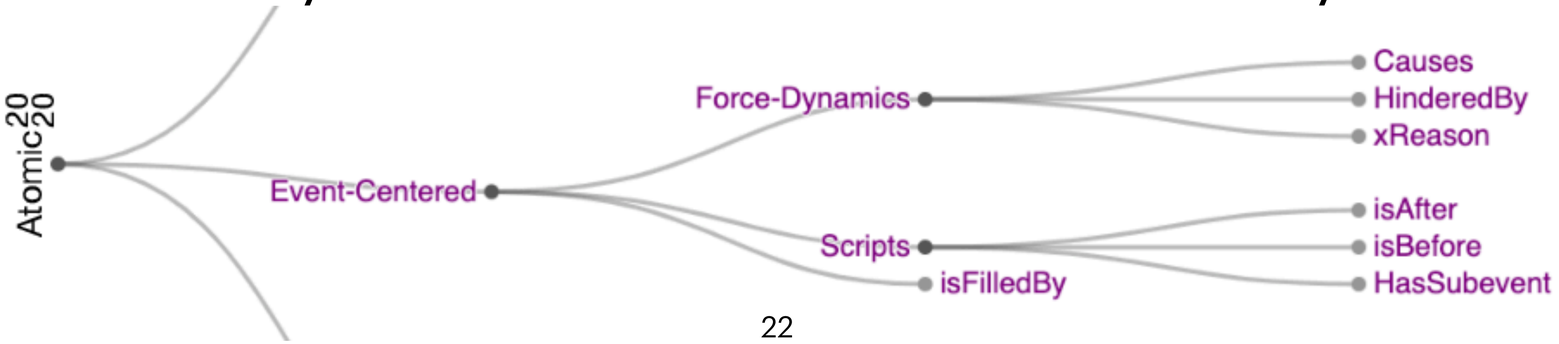
Mike borrows a book from a friend.



Borrows
dobj: book
pobj: friend

Event-based Knowledge Acquisition

- Backward-looking:
  What are the requirements for the event to occur
  - xReason
  - isBefore
  - HasSubevent
  - isHinderedBy
- Forward-looking:
  What might happen after the event occured.
  - Causes
  - isAfter
  - HasSubevent
  - isHinderedBy

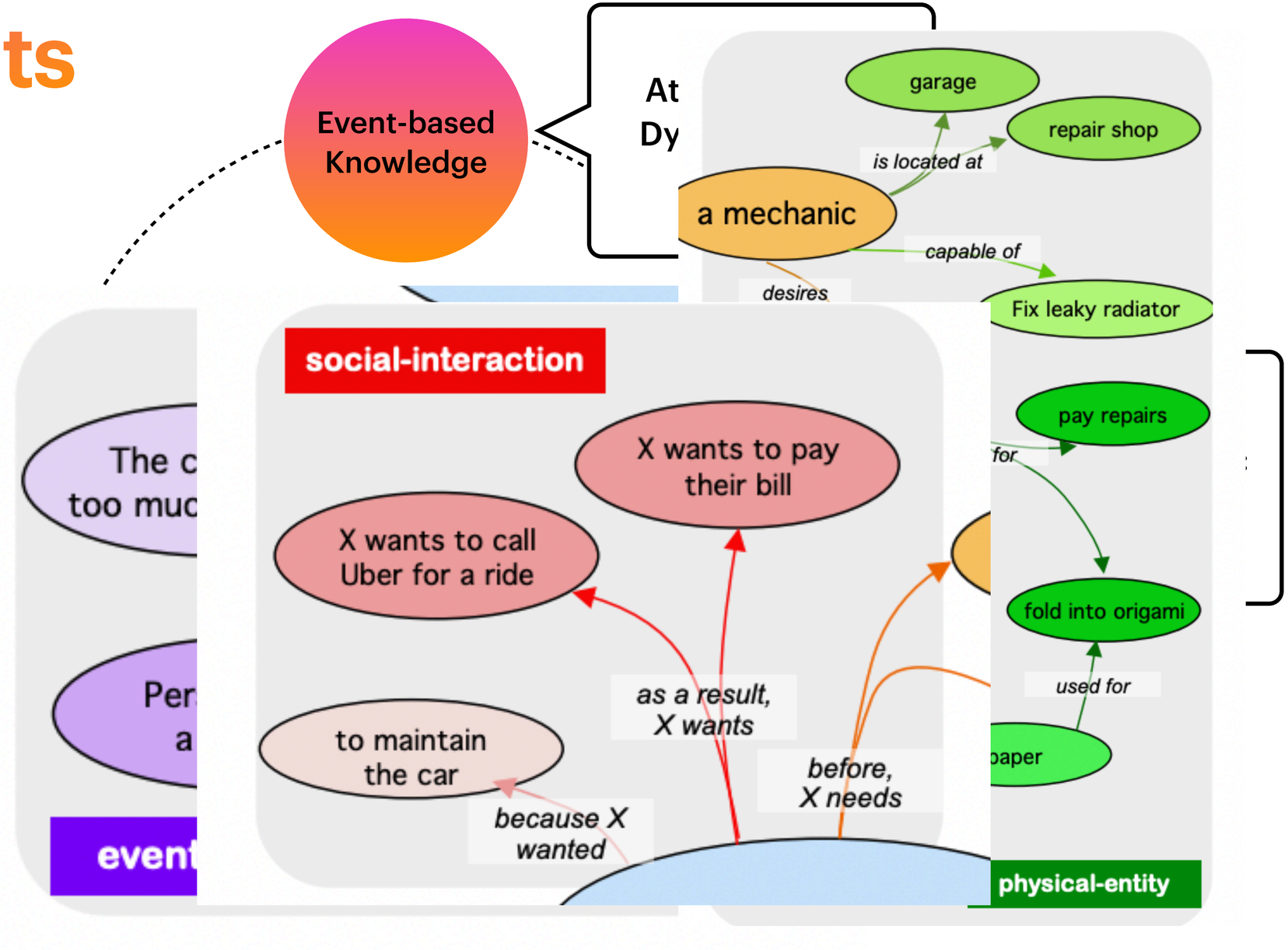



X drives too fast
faulty traffic light
causes
accident
happens before
happens after
X is in a hurry to get to work
X is seen as
careless
happens before
X takes ___ to a pawn shop
filled by
jewelry
because X wanted
unhappy
X is seen as
as a result, X reacts
upset
X wants to cry
X is pulled over by a cop
as a result, X reacts
X leaves a bad review
can be hindered by
doesn't have internet access
X catches ___ in the act
filled by
burglar
cheater
capable of
is located at
feels guilty
has subevent
X gets away with crime
before, X needs
lie a lot
sneaky
casino
is located at
gambler
risky
X is seen as
make money
because X wanted
X plays blackjack
garage
repair shop
is located at
a mechanic
capable of
Fix leaky radiator
desires
money
pay their bill
call Uber for a ride
maintain the car
because X wanted
as a result, X wants
before, X needs
X gets X's car repaired
because X wanted
sell the car
used for
hide the (stolen) car
as a result, others want
give him a ticket
pay repairs
used for
fold into origami
is made of
paper
used for
is made of
cellulose
has property
earned by working
causes
greed
as a result, X wants
as a result, X reacts
X steals the car
because X wanted
X is seen as
criminal
shady
can be hindered by
The car costs too much to repair
happens after
happens before
X spent a fortune
X drives an old car

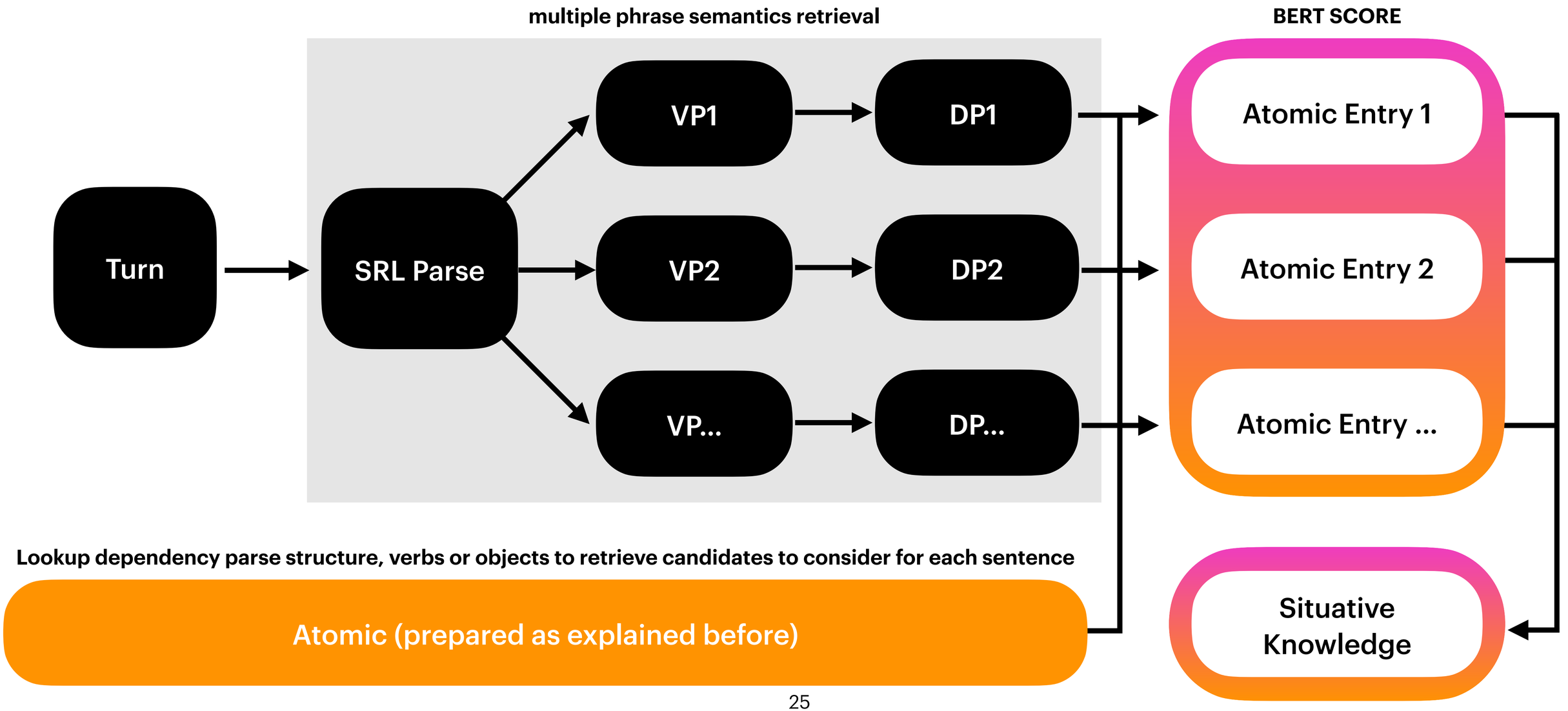
Requirements

Event-based Knowledge

Atomic: Social-Interaction based relations

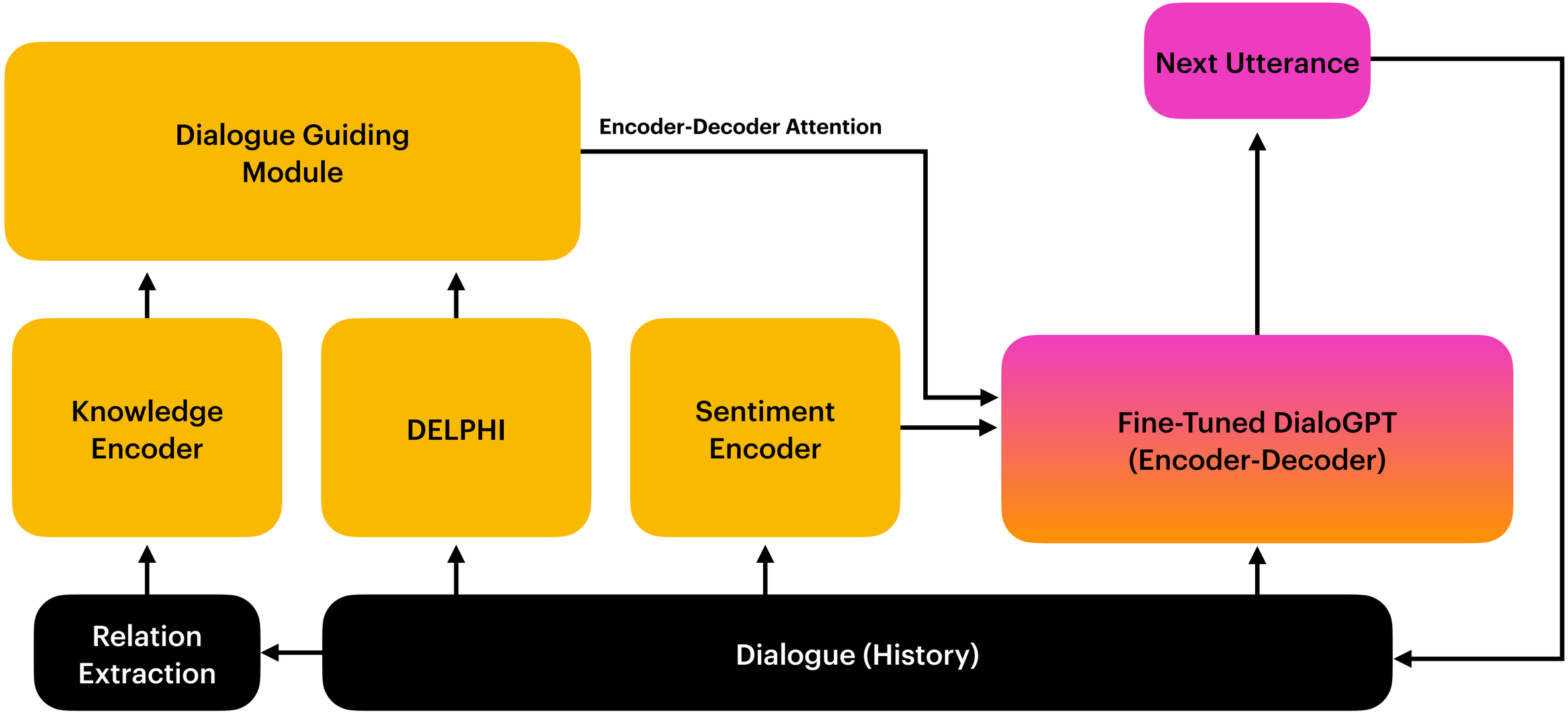


How to Extract Knowledge from Atomic



Modeling Methodology

Conditioned Dialogue Generation

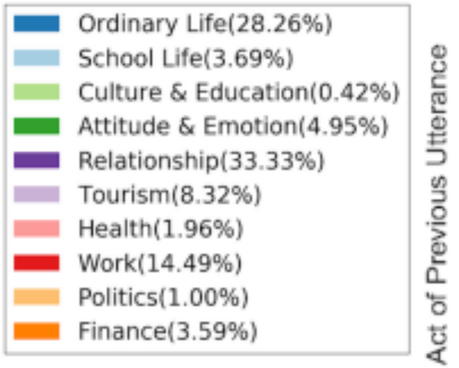


Fine-tuning Dialogue Datasets

Daily Dialog Li et al. 2017

- The purpose is to exchange and share ideas
- Enhance social bonding
- Labels explicitly turn types as:
  - Questions
  - Inform
  - Directives
  - Commissives

A: I'm worried about something.
B: What's that?
A: Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.
B: That's annoying, but nothing to worry about. *Just breathe deeply when you feel yourself getting upset.*
A: Ok, I'll try that.
B: Is there anything else bothering you?
A: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.
B: Do you have any other plans this weekend?
A: I'm supposed to work on a paper that'd due on Monday.
B: *Try not to take on more than you can handle.*
A: You're right. I probably should just work on my paper. Thanks!



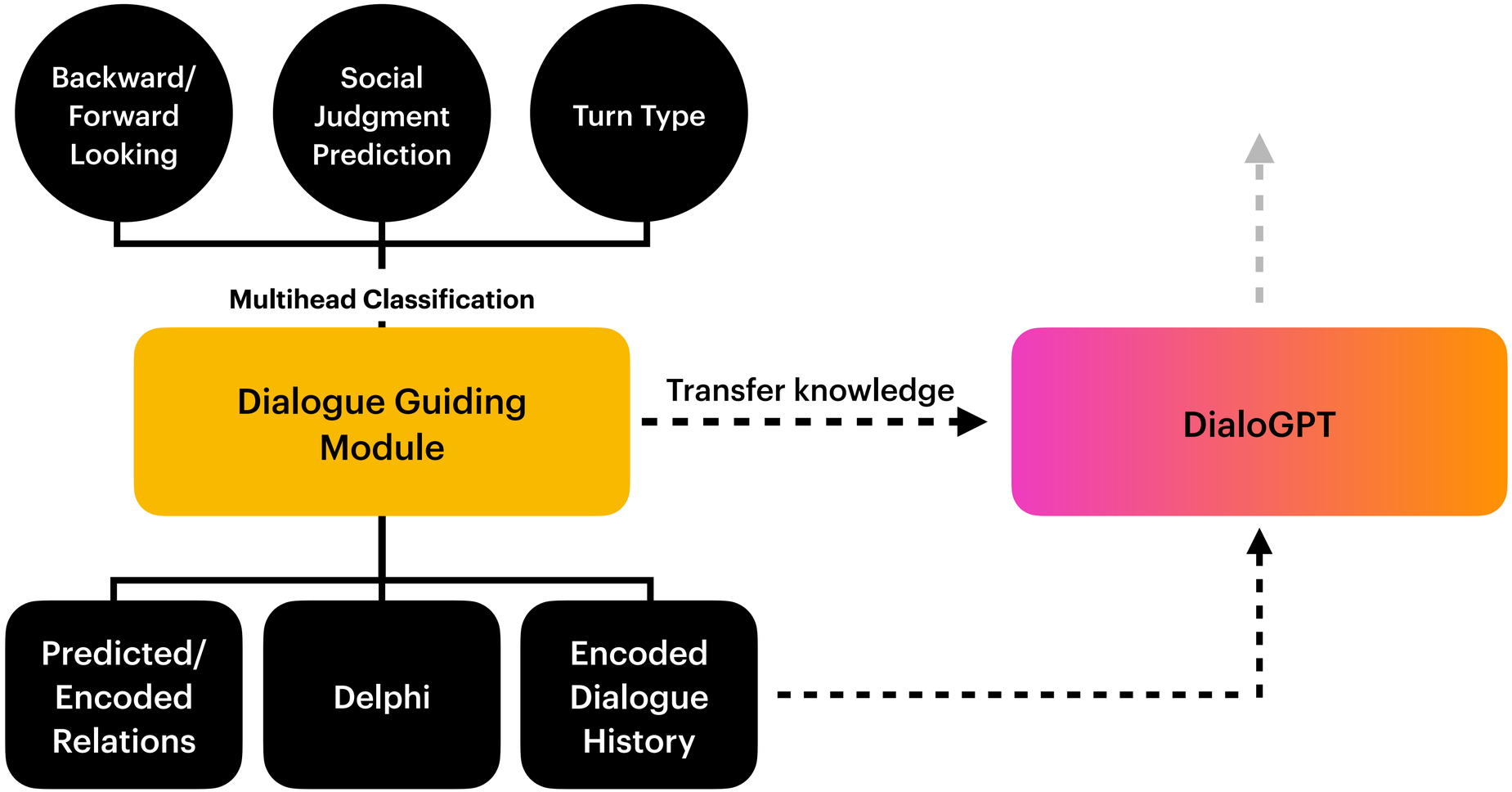
Fine-tuning Dialogue Datasets

Empathetic Dialogues Rashkin et al. 2019

- Emotional situation grounding (situated empathy)
- Thoroughly tested with added information from other predictors
- Even distribution across 32 different emotion categories
- Can enhance dataset with text classification, classifying turn types

Context	Setup	Model	Cand	Generated Response
Someone came to my door with a gun the other day.	Gen	Pretrained	-	I'm not sure if you're joking or not.
	Gen	Fine-Tuned	-	Oh no! What did you do?
	Gen	EmoPrepend	-	Oh no! Did you have a weapon?
	Gen	TopicPrepend	-	I'm glad to hear that.
My son failed his exams! Can you believe it! I don't know what to do!	Ret-BERT	Pretrained	R	Hopefully his mistakes can help you know what not to do on the tests.
	Ret-BERT	Pretrained	ED	Oh no! I ' m so sorry! I hope you could pull your grade up!
	Ret-BERT	Fine-Tuned	ED	Oh no! Did he not study?
	Ret-BERT	EmoPrepend	ED	Im sorry about that! Did your son start school already?
	Ret-BERT	TopicPrepend	ED	Oh no! Does he have the option of tak-ing it again?

Dialogue Guiding Module



Baselines

- Randomly sample turn type
- Model w/o any access to explicit knowledge from Atomic
- DialoGPT fine-tuned on Empathetic Dialogues only

Evaluation

- Manual evaluation:
  - Empathy
  - Assessment of the user’s situation
  - Dialogue Perspectives Fulfilled:
    - Assessment of situation
    - Helpful advice
    - Empathetic dialogue
- Dialogue coherence: GRADE
- NLG: BLEU, METEOR

Status Quo

- Data is almost ready, currently trying to get the dialogues in the right shape to start fine-tuning DialoGPT
- Learning a turn-classification model on Daily Dialog to enhance Empathetic Dialogues
- Module architectures are still somewhat unclear

- “Can I actually get DELPHI to run on our computational resources?”
- Architecture is too dependent on each other, if one fails it will just cascade down
- Explicitly generating & evaluating individual turns

Thanks for listening!

Any questions?

Feedback & ideas (beilharz@cl...) are welcome as well!

References

- [1] Jiang, Liwei, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. **'Delphi: Towards Machine Ethics and Norms'**. *ArXiv:2110.07574 [Cs]*, 14 October 2021. http://arxiv.org/abs/2110.07574.
- [2] Forbes, Maxwell, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. '**Social Chemistry 101: Learning to Reason about Social and Moral Norms'**. *EMNLP*, November 2020, 653–70.
- [3] Hwang, Jena D., Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. **'(Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs'**, 35:6384–92, 2021. https://ojs.aaai.org/index.php/AAAI/article/view/16792.
- [4] Li, Yanran, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. **'DailyDialog: A Manually Labelled Multi-Turn Dialogue Dataset'**. *ArXiv:1710.03957 [Cs]*, 11 October 2017. http://arxiv.org/abs/1710.03957.
- [5] Rashkin, Hannah, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. **'Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset'** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–81. Florence, Italy: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/P19-1534.
- Sap, Maarten, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. **'ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning'**. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, no. 01 (17 July 2019): 3027–35. https://doi.org/10.1609/aaai.v33i01.33013027.
- Lourie, Nicholas, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. **'UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark'**. *ArXiv:2103.13009 [Cs]*, 24 March 2021. http://arxiv.org/abs/2103.13009.
- Zhang, Yizhe, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. **'DialoGPT: Large-Scale Generative Pre-Training for Conversational Response Generation'**. *ArXiv:1911.00536 [Cs]*, 2 May 2020. http://arxiv.org/abs/1911.00536.
- Huang, Lishan, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. **'GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems'**. *ArXiv:2010.03994 [Cs]*, 8 October 2020. http://arxiv.org/abs/2010.03994.
- Banerjee, Satanjeev, and Alon Lavie. **'METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments'**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. https://aclanthology.org/W05-0909.
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. **'BERTScore: Evaluating Text Generation with BERT'**. *ArXiv:1904.09675 [Cs]*, 24 February 2020. http://arxiv.org/abs/1904.09675.