ALMA MATER STUDIORUM UNIVERSITY OF BOLOGNA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING – DISI

Cogito Ergo Summ: Abstractive Summarization of Biomedical Papers via Semantic Parsing Graphs and Consistency Rewards

Giacomo Frisoni, Paolo Italiani, Stefano Salvatori, Gianluca Moro

DISI – University of Bologna, Cesena
Via dell'Università, 50 I-47522 Cesena (FC), Italy
{giacomo.frisoni, paolo.italiani, s.salvatori, gianluca.moro}@unibo.it

NSA workshop 2023 Presentation

Motivations – i

- **Biomedical Document Summarization**
  - Need for advanced tools to skim the literature efficiently and grasp salient contents
  - (i) Medical jargon truly hard to interpret
  - (ii) Precise domain information, narrow interpretation margin, no factual mistakes
  - (iii) Clauses' interdependence and complex interactions
  - (iv) Summarizers face issues in terms of succinctness, non-repetitiveness, fluency, informativeness, and faithfulness

*"…moderate-certainty evidence indicates that **actinomycin D** is **more likely to** lead to primary cure than **methotrexate** … there may be little or no difference in the risk of **severe adverse events (SAEs)** between the groups overall … subgroup analyses **suggests that actinomycin D may be** associated with a greater risk of **SAEs** than **methotrexate**…*

*Authors' conclusions: **Actinomycin D** is probably **more likely to** achieve a primary cure in women with low-risk **GTN** … however, **actinomycin D may be** associated with a greater risk of **severe adverse events**"*

●/●/● = biomedical entities; _ = modifiers

Motivations – ii

- **Beyond Lexical Superficiality**
  - Superficial text organization rather than underlying semantics
  - Uncaptured convoluted long-range dependencies between entities
  - Modern solutions are highly prone to hallucinating content or falling back on extraction

"Cogito, Ergo ~~Sum~~ Summ"
A neural network should think about the inner semantics of the text—via joint text-graph reasoning—before summarizing

- **Incorporating Explicit Semantic Structures**
  - Existing graph-augmented approaches have at least one of the following weaknesses
    - (i) Not designed for or evaluated in the biomedical domain
    - (ii) Graph-LSTMs architectures that struggle to compete with transformers
    - (iii) Built upon open-domain triplet-based extractions that are notoriously not adequate to represent the complete biological meaning of a document
    - (iv) Not ensuring document-summary consistency

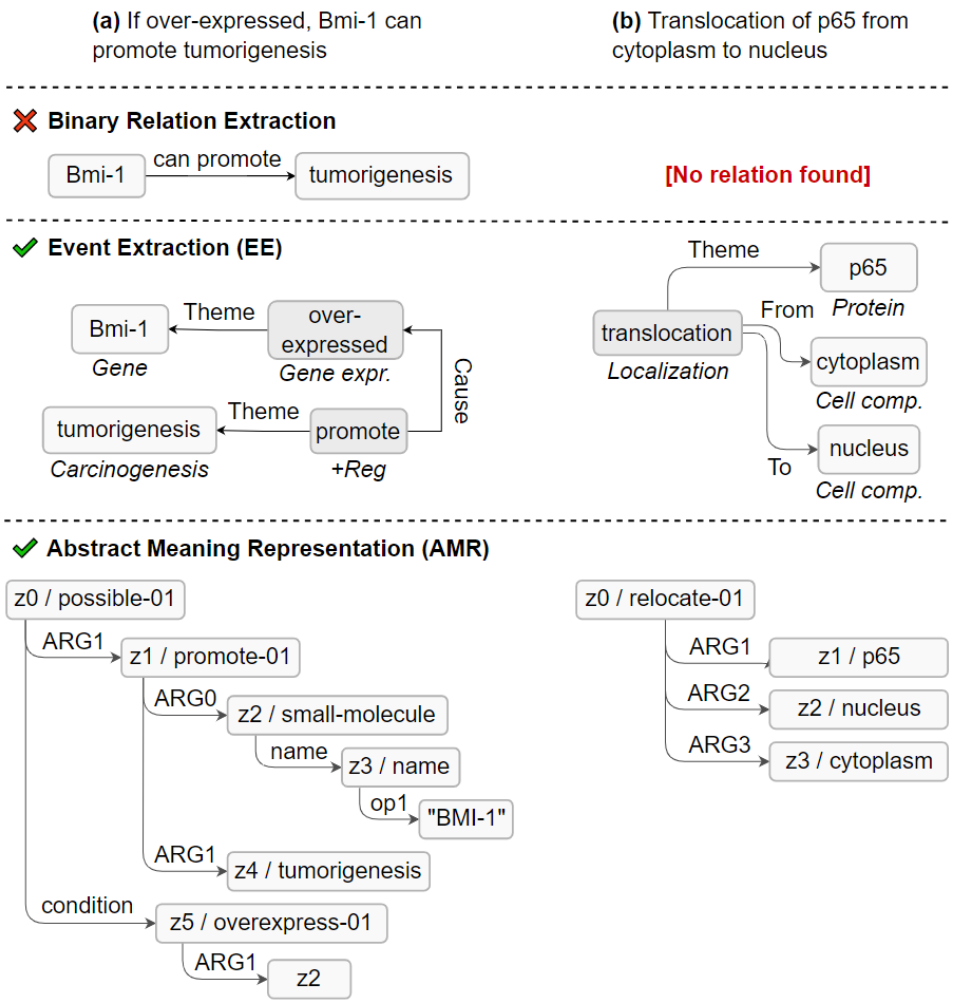
Contributions

- **CogitoErgoSumm**, the first semantics-aware transformer-based model for single document abstractive summarization in the biomedical domain
  - 💡 Combining PLMs and semantic parsing graphs providing formal meaning representations
    - Injecting semantic parsing graphs into an encoder-decoder PLM may help the latter to decouple concept units (*what to say*) from language competencies (*how to say it*)
  - Two different semantic parsing techniques with complementary strengths: Event Extraction (EE) and Abstract Meaning Representation (AMR)
  - Reinforcement Learning to ensure factuality and consistency
    - Reward function based on the average Smatch score between the original document and the generated summary

- **Neuro-Symbolic Problem Formulation**
  - Given a dataset $C = (d_1, d_2, \dots, d_k)$, where each document d_i consists of a sequence of n tokens $d = (x_1, x_2, \dots, x_n)$
  - The semantics of d_i is condensed in document-level event and AMR graphs (G_e and G_a)
  - The goal is to generate the target summary $y = (y_1, y_2, \dots, y_m)$, $m \leq n$ of each instance
  - By modeling the conditional distribution $p(y_1, y_2, \dots, y_m | x_1, x_2, \dots, x_n, G_e, G_a)$

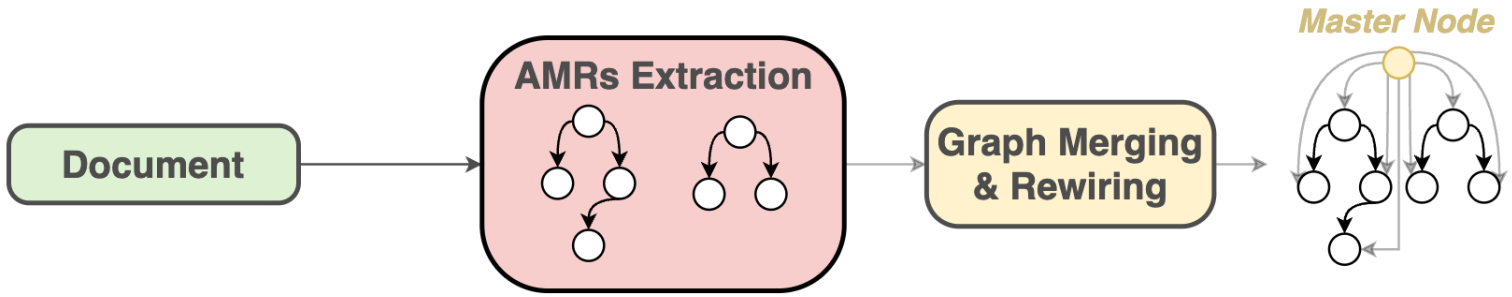
Graph Construction

- OpenIE is the most popular tool, however it is based on **binary relation extraction**



AMR Graph Construction

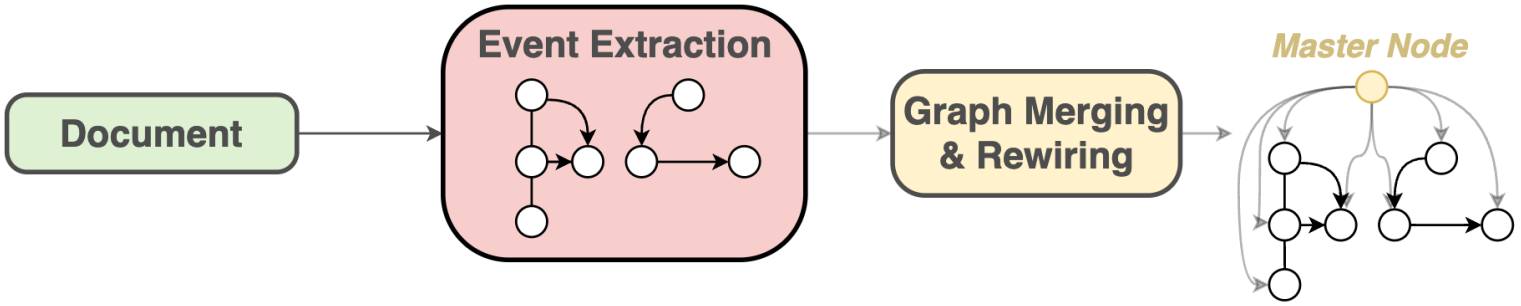
- AMR aims to graphically **capture the general meaning** of any sentence as high-level semantic relations
- We use the SOTA text-to-text AMR parser **SPRING** [Bevilacqua et al., 2021]
  - **AMRs supplement** not always exploitable event graphs
  - **Abstraction** from words to concepts (objects, attributes, etc.)
  - **Domain-general** with one amr graph for each sentence
- We operate **graph rewiring** by adding a master node connecting all nodes, to reflect the document structure and enhance the information flow, obtaining G_a



Bevilacqua, et al. "One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline." AAAI 2021.

Event Graph Construction

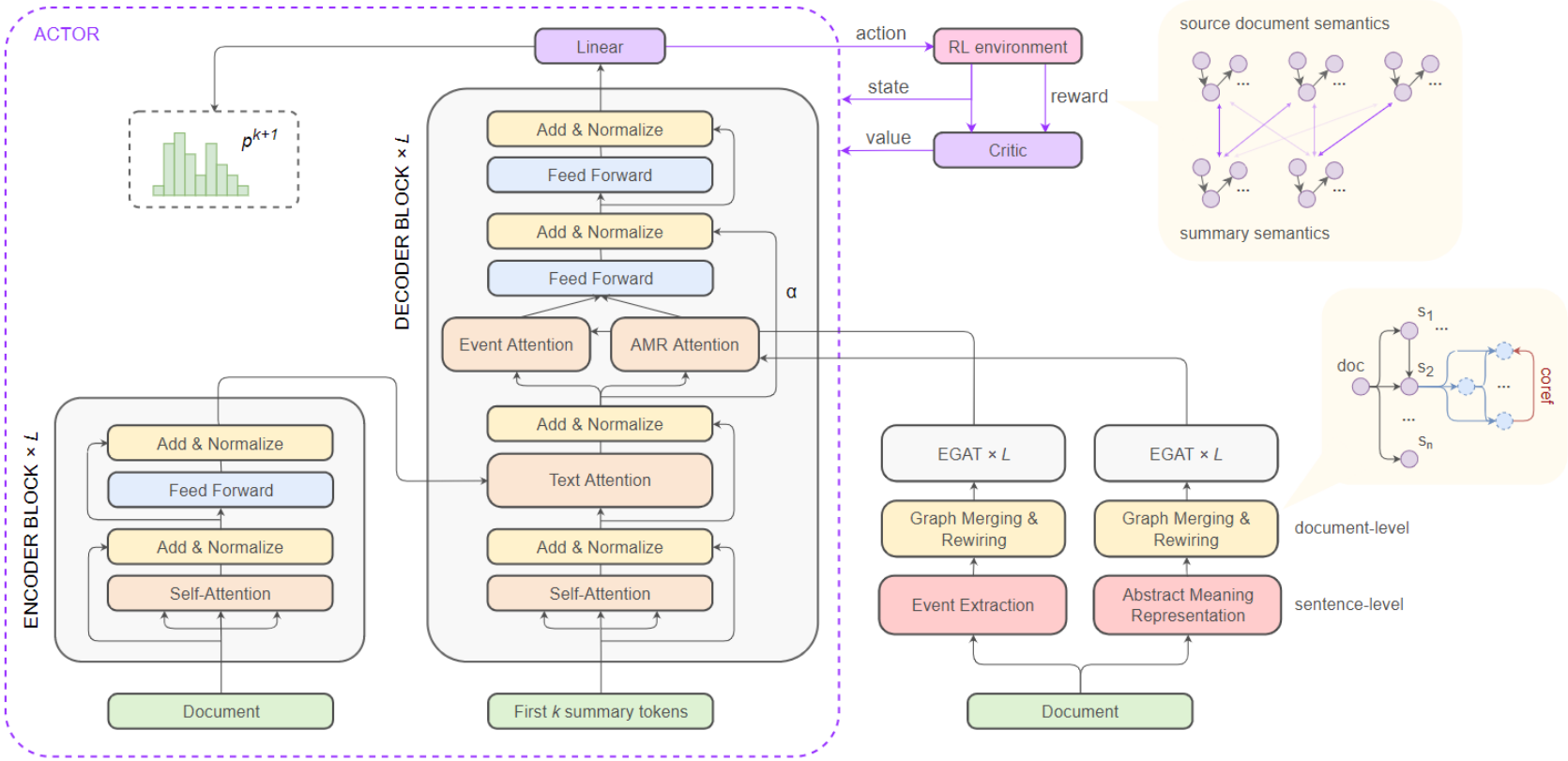
- Event graphs aim to **capture biomedical-specific** interactions
- We first obtain sentence-level event extraction using **DeepEventMine** [Trieu et al., 2020]
  - Derives **n-ary and potentially nested** interactions between participants
  - Events consist of a **trigger,** a **type** a **set of arguments** with a **semantic role**
  - A node represents a *trigger* or an *entity*
  - An edge models a *trigger* → *entity* or *trigger* → *trigger* (for nested events) relation
- We operate **graph rewiring** by adding a master node connecting all event nodes obtaining G_e



Trieu, et al. "DeepEventMine: end-to-end neural nested event extraction from biomedical texts." Bioinformatics 2020.

Method – Overview

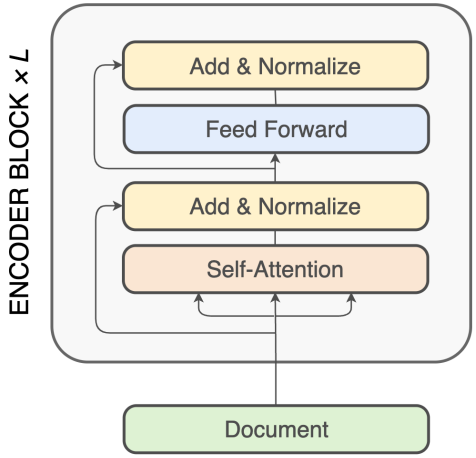
- We extend a pre-trained BART-base architecture with the nimble ability to **attend semantic parsing graphs** during decoding and preserve the most relevant information via **reinforcement learning** (RL)

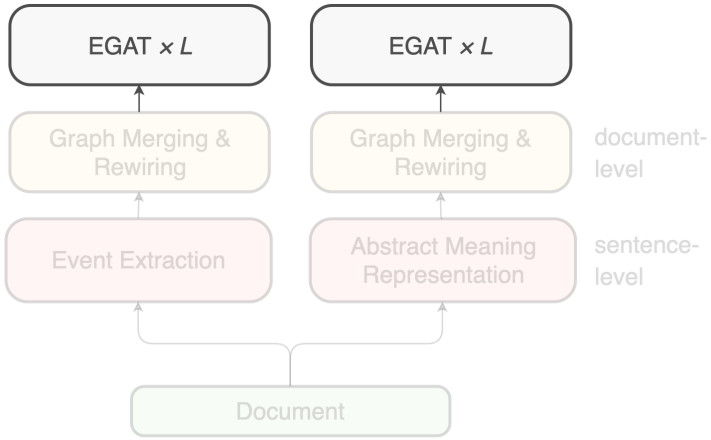


Method – Encoder

- **Text Encoder**
  - We feed the input document to **a text bidirectional encoder**, obtaining a contextual hidden representation for each token

- **Graph Encoders**
  - We initialize node features with embeddings outputted by the text encoder
  - Through two EGATs $\boldsymbol{E}_{\boldsymbol{G}_e}$ and $\boldsymbol{E}_{\boldsymbol{G}_a}$ we take $\boldsymbol{G}_e$ and $\boldsymbol{G}_a$ to learn supervised **node embeddings** and tap **implicit relations**
  - EGATs extend the graph attention groundwork (GAT) by **considering the edge type** connecting two nodes

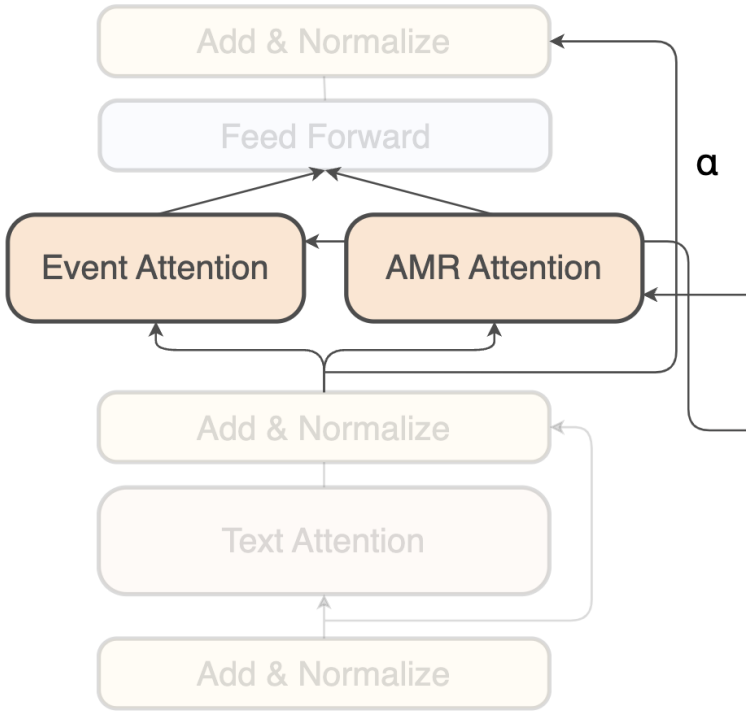




Method – Decoder

- We aggregate **different levels** of encoded representations via a multigranularity decoder

- **We supplement the BART transformer decoder** with two extra cross-attentions conducted over the node representations learned by $\boldsymbol{E}_{G_e}$ and $\boldsymbol{E}_{G_a}$
- The token-, event-, and -AMR attended vectors are **combined into a semantics-aware representation**
- A learnable parameter **α modules updates** from cross-attention over semantic graphs



Method – Reinforcement Learning

- Summaries should **preserve as much pivotal information as possible from the original document**
  - More Consistency and Factuality, Less Hallucinations
  - *How?* Maximize the degree of meaning overlap between the AMRs of the input document and the AMRs of the generated summary using a document-level Smatch [Cai et al., 2013] (Fscore)
  - **Smatch is not differentiable → We see it as a reward function and use RL to maximize it**

Reward: $\psi(doc, summ) = AvgSmatch(doc, summ) - \beta \log \frac{\pi_\theta(a_t|s_t)}{\pi_{base}(a_t|s_t)}$

KL-divergence *avoid deviating too much from the pretrained model*

- **Proximal Policy Optimization (PPO)** [Schulman et al., 2017]

Ratio

$$\frac{new\ policy}{old\ policy} = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

Generalized Advantage Estimation

$$L_{ppo} = \mathbb{E}[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$$

Classic Policy Gradient

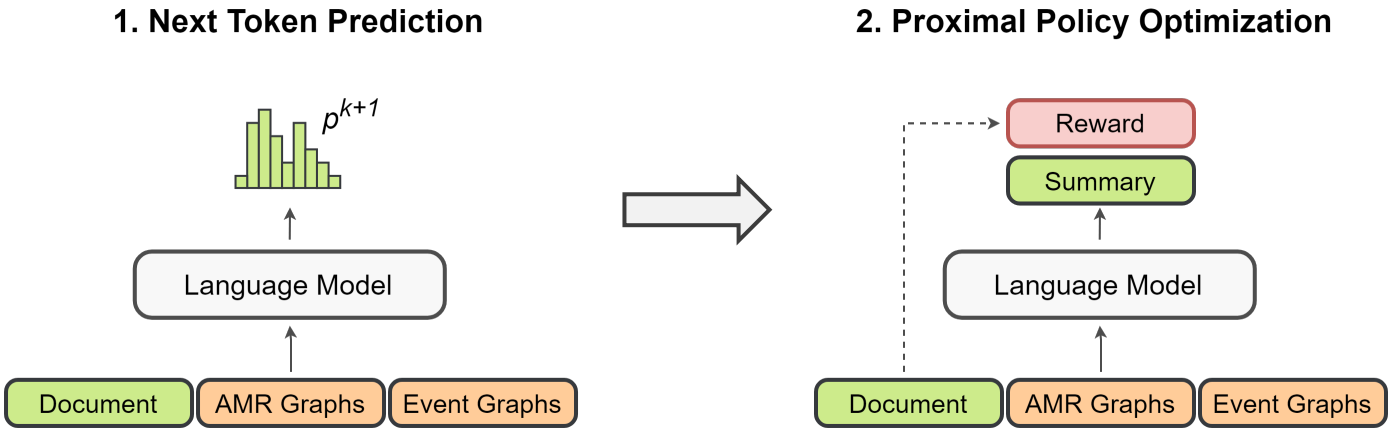
Clipped Policy Gradient
improves stability and convergence

Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347, 2017
Cai and Knight "Smatch: an Evaluation Metric for Semantic Feature Structures." ACL 2013

Method – Training Phases

- **1st Phase: Summarization with Semantic Parsing Graphs**
  - Integration of AMR Graphs and Event graphs to improve reasoning
  - Train with *Cross entropy Loss (=Next Token Prediction)*

- **2nd Phase: Reinforcement Learning with Consistency Reward**
  - Improve consistency and factuality by comparing AMRs of the input document and the generated summary
  - Train with *Proximal Policy Optimization Loss*



Dataset

- **We evaluate on CDSR [Guo et al., 2020]**
  - Designed for assessing the automated generation of **lay language summaries from biomedical scientific reviews**
  - Besides creating accurate and factual summaries, this task also requires a **joint style transformation** from the original professional language to that of the general public
  - 5,178 (training), 500 (eval), 999 (test)

Source	Target
. . . We considered all randomised controlled trials (RCTs) comparing EVLA, endovenous RFA or UGFS with conventional surgery in the treatment of SSV varices for inclusion. . . . (Paravastu, Horne, and Dodd 2016)	. . . We found three randomised controlled trials (clinical studies where people are randomly put into one of two or more treatment groups) that compared endovenous lasers (EVLA) with surgery. . . .
. . . Abnormal blood flow patterns in fetal circulation detected by Doppler ultrasound may indicate poor fetal prognosis. . . . (Alfirevic, Stampalija, and Dowswell 2017)	. . . Doppler ultrasound detects changes in the pattern of blood flow through the baby's circulation. These changes may identify babies who have problems. . . .

Guo, et al. "Automated Lay Language Summarization of Biomedical Scientific Reviews." AAAI 2021.

Experimental Setup

- **Ablations**
  - **w/o RL** = reinforcement learning exclusion
  - **w/o event** = event cross-attention exclusion
  - **w/o AMR** = AMR cross-attention exclusion
- **Extractive baseline**
  - **Oracle**: oracle summary created by selecting the set of sentences in the document that generates the highest ROUGE-2 score with respect to the gold standard summary
  - **BERT**: inter-sentence transformer layers and sigmoid classifier on top of BERT outputs, with oracle extractive used as supervision for training
- **Abstractive baseline**
  - **Pointer generator**: standard seq2seq model with a pointer network that allows copying words from the source and generating words from a fixed vocabulary
  - **BART**: full-transformer pretrained on large corpora by reconstructing text after a corruption phase with an arbitrary noising function
  - **EASumm:** an event-augmented graph-LSTM architecture for abstractive summarization

Metrics

Quantitative Metrics

- **ROUGE-n**: measures n-grams overlaps
- **ROUGE-L**: measures common subsequence overlaps
- **BERTScore**: computes a similarity score for each token in the candidate sentence with each token in the reference sentence
- **FactCC** [Kryscinski et al., 2018]: verifies the factual consistency between generated summary and input document
- **Novel n-grams**: proportion of novel n-grams with $n \in [1-4]$
- **Flesch-Kincaid and Coleman-Liau**: estimate the years of education generally required to understand the summary (*Included in the paper, but not discussed in this presentation*).

Qualitative Metrics

- **Informativeness**: conveying salient content
- **Factualness**: being faithful to the article
- **Fluency**: being fluent, grammatical, and coherent
- **Succinctness**: non containing redundant and unnecessary information

Kryscinski, et al. "Evaluating the Factual Consistency of Abstractive Text Summarization." ACL 2020.

Results – i

- **Automated evaluation** on the full test set of CDSR with ROUGE (R in short)

Model	#params	R-1	R-2	R-L
ORACLE†	—	53.56	25.54	49.56
BERT-base†	110M	26.60	11.11	24.59
POINTER GENERATOR†	22M	38.33	14.11	35.81
BART-base (PubMed)	139M	51.20	19.77	48.47
BART-large (PubMed)†	406M	**52.66**	**21.73**	**49.97**
EASUMM‡	8M	46.30	18.73	43.78
COGITOERGOSUMM	181M	52.23	<u>20.63</u>	49.44
- w/o RL	180M	<u>52.30</u>	20.47	<u>49.46</u>
- w/o event and RL	155M	52.13	20.42	49.30
- w/o AMR and RL	157M	52.02	20.54	49.25

Top: extractive models. **Middle:** abstractive models.
Bottom: our semantics-augmented abstractive model.

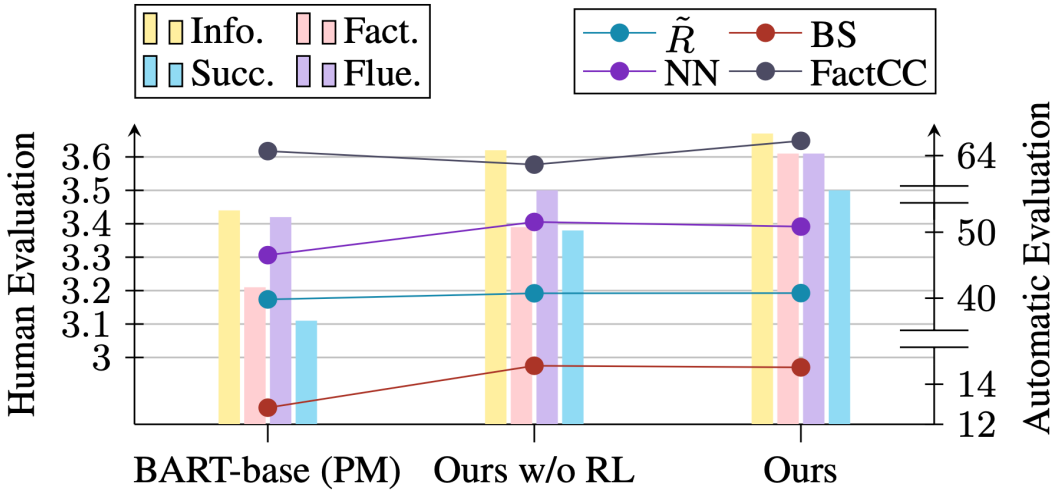
Results – ii

Model	#params	R-1	R-2	R-L
ORACLE†	—	53.56	25.54	49.56
BERT-base†	110M	26.60	11.11	24.59
POINTER GENERATOR†	22M	38.33	14.11	35.81
BART-base (PubMed)	139M	51.20	19.77	48.47
BART-large (PubMed)†	406M	**52.66**	**21.73**	**49.97**
EASUMM‡	8M	46.30	18.73	43.78
COGITOERGOSUMM	181M	52.23	20.63	49.44
- w/o RL	180M	52.30	20.47	49.46
			20.42	49.30
			20.54	49.25

ROUGE scores significantly higher, except for BART-large, for which our model is still competitive with 2x fewer parameters

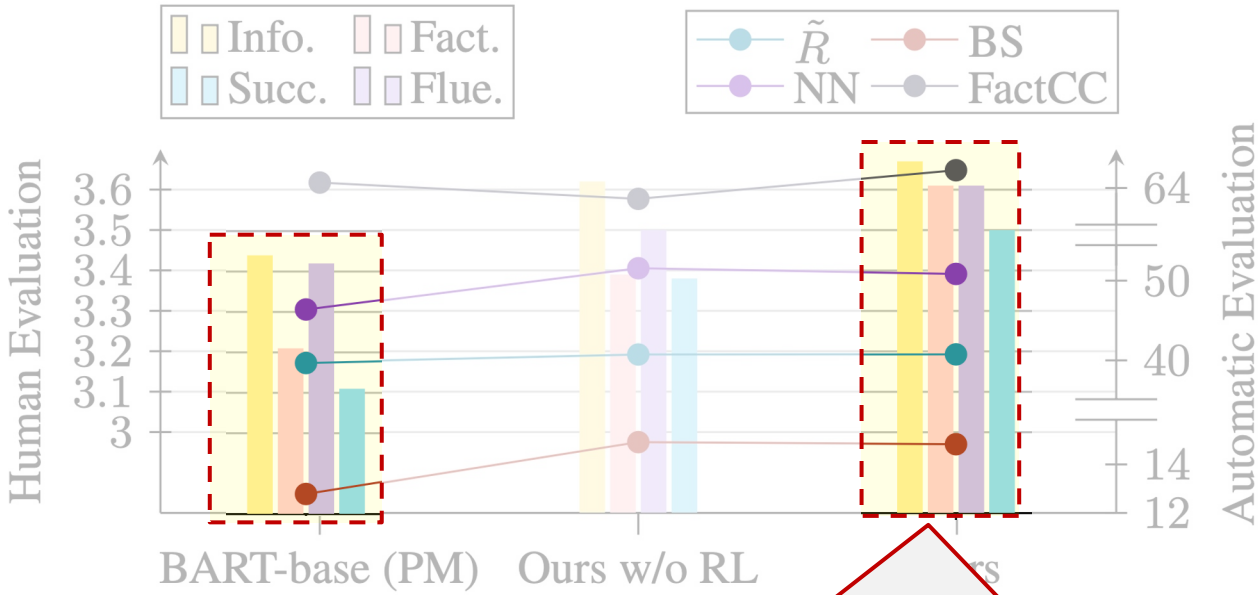
Results – iii

- **Human evaluation** scores
- Average Kendall coefficient among all evaluators' inter-rater agreement = 0.16



Informativeness, Factualness, Fluency, and Succinctness
compared to ROUGE-1/2/L average (R̃), % of novel n-grams (NN), BERTScore (BS) and FactCC

Results – iv


Info.
Succ.
Fact.
Flue.
$\tilde{R}$
NN
BS
FactCC
Human Evaluation
3.6
3.5
3.4
3.3
3.2
3.1
3
Automatic Evaluation
64
50
40
14
12
BART-base (PM)
Ours w/o RL
+12.46% factualness, +6.69% informativeness

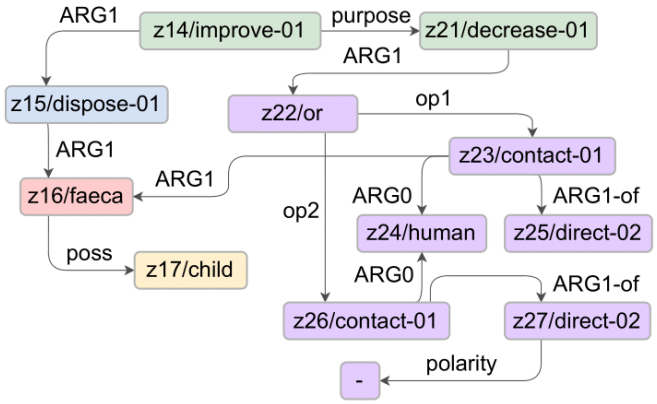
Results – v

[Source Document] We included randomized controlled trials (RCTs) and non-randomized controlled studies (NRS) that compared interventions aiming to improve the disposal of faeces of children aged below five years in order to decrease direct or indirect human contact with such faeces with no intervention or a different intervention in children and adults. Data collection and analysis Two review authors selected eligible studies, extracted data, and assessed the risk of bias *[...]*

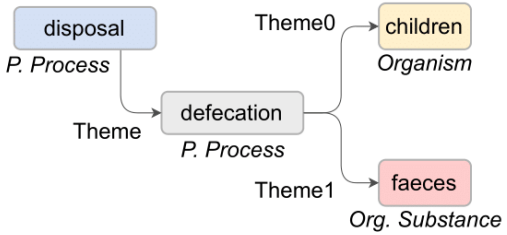
[BART-base] This Cochrane Review aimed to evaluate the effectiveness of interventions **to reduce the use of children's faececes** in order to decrease direct or indirect contact with such **faeced** *[...]*

[CogitoErgoSumm] This Cochrane Review aimed to assess the effectiveness of interventions to improve the disposal of children faeces for decreasing direct or indirect human contact with such faeces *[...]*

AMR



Biomedical Event



Conclusions

- In this paper, we introduce a framework for **infusing** domain-specific and -general **semantic parsing graphs**
- We propose new decoder **cross-attention modules** and **reward signals**
- Our framework **sets new marks** in informativeness, factuality, and readability, better selecting and preserving summary-worth content
- Qualitative evaluation unveils that our **models surpass current baselines** on all metrics associated with **human judgment** while still being **competitive** on **recall-based scores**
- Our results substantiate the hypothesis that **semantic awareness** through graph injection draws a **complementary path to architectural scaling**
- For future work, we plan to model extracted semantics through **logic representations** so as to **enable reasoning**

Thanks for the attention

(is all you need)