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

Giacomo Frisoni, [1] **Paolo Italiani,** [1] **Stefano Salvatori** [1]**, Gianluca Moro** [1]**.**

[1]Department of Computer Science and Engineering, University of Bologna, Cesena Campus
Via dell'Università 50, I-47522 Cesena, Italy
{giacomo.frisoni, paolo.italiani, s.salvatori, gianluca.moro}@unibo.it

Abstract

The automatic synthesis of biomedical publications catalyzes a profound research interest elicited by literature congestion. Current sequence-to-sequence models mainly rely on the lexical surface and seldom consider the deep semantic interconnections between the entities mentioned in the source document. Such superficiality translates into fabricated, poorly informative, redundant, and near-extractive summaries that severely restrict their real-world application in biomedicine, where the specialized jargon and the convoluted facts further emphasize task complexity. To fill this gap, we argue that the summarizer should acquire semantic interpretation over input, exploiting structured and unambiguous representations to capture and conserve the most relevant parts of the text content. This paper presents COGITOERGOSUMM, the first framework for biomedical abstractive summarization equipping large pre-trained language models with rich semantic graphs. Precisely, we infuse graphs from two complementary semantic parsing techniques with different goals and granularities—Event Extraction and Abstract Meaning Representation, also designing a reward signal to maximize information content preservation through reinforcement learning. Extensive quantitative and qualitative evaluations on the CDSR dataset show that our solution achieves competitive performance according to multiple metrics, despite using 2.5× fewer parameters. Results and ablation studies indicate that our joint text-graph model generates more enlightening, readable, and consistent summaries.[1]

Introduction

Given the sheer number of biomedical publications, clinicians, patients, and researchers need advanced tools to skim the literature efficiently and grasp salient contents. Hence, automatically organizing everyday scientific discoveries or insights into natural, concise, and informative syntheses is essential to promote knowledge acquisition (Moradi and Ghadiri 2019). To this end, abstractive document summarization demands rephrasing and condensing long and often labyrinthine portions of text in a creative way, discarding redundant and unnecessary attributes. Compared with the open domain, performing this task in biomedicine raises substantial challenges and constraints (Karamanis 2007). Indeed, (i) medical jargon and professional language are truly hard to interpret; (ii) scientific documents convey precise domain information allowing for a narrow interpretation margin and non-tolerating factual mistakes; (iii) clauses are often interdependent and express complex interactions; (iv) knowledge rapidly evolves over time.

Despite unprecedented progress made possible by pre-trained transformer-based language models (Lin and Ng 2019), current summarizers still face issues in terms of succinctness, non-repetitiveness, fluency, informativeness, and faithfulness (Maynez et al. 2020). Most prior studies only depend on superficial text organization and ignore the deeper underlying semantic content (Bender et al. 2021), lacking structured representations to encapsulate the convoluted long-range associations between the mentioned entities (e.g., proteins, diseases, drugs). By contra, we argue that these connections are vital to document understanding and beneficial to knowledge selection.

Biomedical documents are usually composed of a series of events and factual evidence; understanding how to leverage such information in generative models is crucial. Notably, semantic parsing graphs normalize many lexical and syntactic variations by providing formal meaning representations capable of decoupling concept units (*what to say*) from language competencies (*how to say it*). Since these representations can be defined with a panoply of formalisms having different objectives and properties, we underline the importance of bridging the complementary strengths of two influential semantic parsing tasks: closed-domain Event Extraction (EE) and Abstract Meaning Representation (AMR). EE is task-driven and aims to derive n-ary and potentially nested interactions between participants having a specific semantic role, where event schemas (i.e., target event, entity, and role types) are pre-established conforming to a reference ontology; its history is very intertwined with health informatics (Frisoni, Moro, and Carbonaro 2021). AMR is linguistically-grounded and is conceived to graphically capture the general meaning of any sentence as high-level semantic relations between abstract concepts (Banarescu et al. 2013). Fig. 1 depicts and compares their expressive power.

A growing body of research in natural language generation (NLG) calls attention to incorporating explicit semantic structures into the summarization process (Yu et al. 2020),

[1]The data and the code to fully reproduce our results are available at https://github.com/disi-unibo-nlp/cogito-ergo-summ

thereby unlocking a higher level of abstraction than bags of sentences and more accurate emulation of human interpretation, rewriting, and paraphrasing. However, existing graph-augmented approaches have at least one of the following weaknesses: (i) they have not been designed for or evaluated in the biomedical domain; (ii) they employ graph-LSTMs architectures that struggle to compete with transformers; (iii) they are based on open-domain triplet-based extractions that are notoriously not adequate to represent the complete biological meaning of a document; (iv) they do not include a module to ensure document-summary consistency.

We present COGITOERGOSUMM[2], the first semantics-aware transformer-based model for single-document abstractive summarization in the biomedical domain. Concretely, we condense source documents into sets of unambiguous EE and AMR graphs, using multi-relational graph neural networks (GNNs) to yield their dense embeddings without imposing linearization. From this foundation, we explore a fine-tuning recipe for merging predicted symbolic representations into pre-trained encoder-decoder language models. Specifically, we integrate semantic parsing graphs via EE- and AMR-specific attention mechanisms in the decoder, thus aiding key content selection and semantic understanding. We optimize the network via reinforcement learning (RL), devising a consistency-guided reward signal based on a soft alignment between the sets of meaning representations extractable from documents and generated summaries.

Automatic and human evaluations are carried out on CDSR (Guo et al. 2021), a popular dataset for generating lay language summaries of biomedical reviews. Our model brings substantial improvements in compliance with multiple quality criteria, achieving near state-of-the-art (SOTA) performance while using 2.5x fewer parameters. Empirical results corroborate the value of semantic graphs in helping the model to preserve the essential global context and keep the factual connections between the most relevant entities.

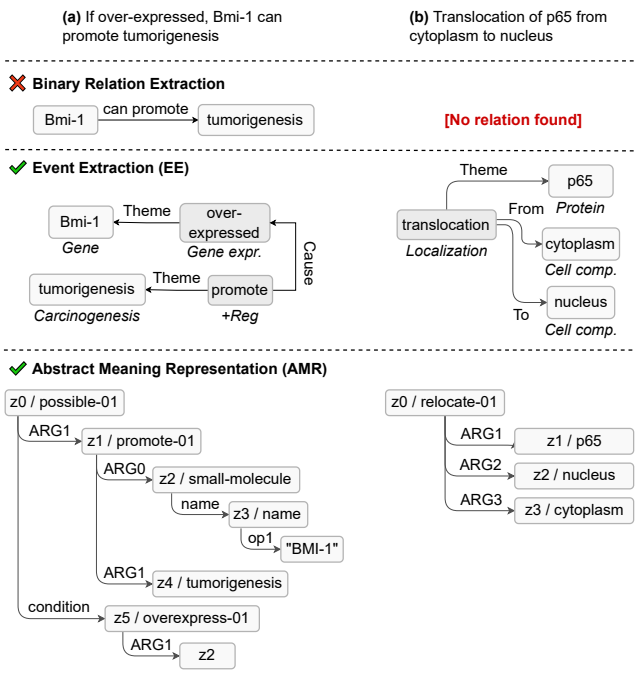

Figure 1: Expressiveness comparison between binary relation extraction (OpenIE 5.1), EE, and AMR on two example biomedical sentences. *Italic* denotes node types.

Related Work

Abstractive Summarization Sequence-to-sequence architectures founded on self-supervised pre-trained transformers have been responsible for a profound impetus in implicitly learning abstractive summarization procedures (Rothe, Narayan, and Severyn 2020; Zhang et al. 2020a; Lewis et al. 2020), comprising multi-document (Moro et al. 2022) and low-resource settings (Moro and Ragazzi 2022). Nevertheless, modern solutions are highly prone to hallucinating content (Cao et al. 2018; Maynez et al. 2020) or falling back on extraction (See, Liu, and Manning 2017). Gaining an understanding of semantics and context is becoming a prerogative, but a model trained purely on form cannot learn meaning (Bender and Koller 2020)—even with more data and huge architectural dimensions.

Graph-enhanced Summarization Graph structures have long been studied for implementing summarization subtasks (i.e., information extraction, content selection, surface realization), registering distinct benefits depending on their composition. Early techniques for extractive summarization build intra- (Mihalcea and Tarau 2004) and inter-document (Wan 2008) cosine similarity connectivity networks to identify salient sentences. Late hybrid neural systems mainly stand on the shoulders of run-of-the-mill GNNs (Wu et al. 2021a), exploiting graph-based attention (Tan, Wan, and Xiao 2017) and heterogeneous word-/sentence-level nodes (Wang et al. 2020). As for abstractive summarization, the community has attempted a medley of graphical document representations, from entailment (Mehdad, Carenini, and Ng 2014), dependency (Wu et al. 2021b), sentiment (Moro et al. 2018), and coreference links (Balachandran et al. 2021) to discourse relations (Li et al. 2020; Chen and Yang 2021) and citation networks (An et al. 2021). To better contemplate entity interactions, compositions of $<subject, predicate, object>$ triplets from the OpenIE framework (Angeli, Premkumar, and Manning 2015) have turned into a cornerstone (Fan et al. 2019; Huang, Wu, and Wang 2020; Ji and Zhao 2021; Zhu et al. 2021). On the flip side, open-domain binary relations are inadequate for biomedicine, risking deriving incorrect or incomplete facts that are difficult to compare and merge with post-processing (Frisoni, Moro, and Carbonaro 2021). Most pertinent to our work are summarizers enhanced by AMR semantic analysis (Dohare, Gupta, and Karnick 2018; Hardy and Vlachos 2018; Lee et al. 2021), underexplored in bioinformatics despite showcasing superior generation controllability in general domain. Frisoni. et al. (2022) investigate EE-

[2]Inspired by the first principle of René Descartes's philosophy, we coin the name *Cogito Ergo Summ* to emphasize the researched neural network capacity of "thinking" about the inner semantics of the text—via joint text-graph reasoning—before summarizing.

augmented summarization observing that performances are held back by the reduced number of event graphs. Furthermore, almost all solutions nowadays reckon on graph-LSTM architectures—that hardly compete with transformers—or graph-to-text verbalizers—that ignore source text contribution. To our knowledge, we are the first to combine text, EE, and AMR for transformer-based abstractive summarization, solving their mutual limitations. Another trend is context augmentation through external knowledge graphs, such as UMLS for medicine (Gigioli et al. 2018). However, unlike flexible meaning representations, these resources are known to have limited and static coverage of real-world entities.

Reinforcement Learning for Abstractive Summarization In a traditional encoder-decoder architecture, the network is trained to minimize the maximum-likelihood loss for next-token prediction but is evaluated on the not necessarily equivalent optimization of desired automatic metrics. Moreover, the decoder knows the ground-truth sequence during training but does not have such supervision when testing, leading to an exposure bias (Ranzato et al. 2016). RL methods have recently been prompted to mitigate these discrepancies by directly solving metric optimization problems that are not differentiable, not requiring gold summaries, and allowing global decisions rather than local (token-level) ones at each timestep. Kryscinski et al. (2018); Paulus, Xiong, and Socher (2018); Sharma et al. (2019) utilize ROUGE scores to encourage novelty and relevance, ignoring entity interactions. Scialom et al. (2019); Huang, Wu, and Wang (2020) propose fill-in-the-blank and triplet-reconstruction question answering rewards, needing ad-hoc models for generating artificial questions. In contrast, we reward predicted summaries according to their key content consistency with respect to the original document, measuring the alignment between their semantic graphs.

COGITOERGOSUMM Framework

Problem Statement

Given a dataset $C{=}(d_1, d_2, \ldots, d_k)$, each document d_i consists of a sequence of n tokens $d{=}(x_1.x_2, \ldots, x_n)$. The semantics of d_i is condensed in document-level event and AMR graphs (G_e and G_a, respectively). Formally, the goal is to generate the target summary $y{=}(y_1, y_2, \ldots, y_m)$, $m{\leq}n$, of each instance by modeling the conditional distribution $p(y_1, y_2, \ldots, y_m|x_1, x_2, \ldots, x_n, G_e, G_a)$, conceptualizable as a neuro-symbolic task. We keep G_e and G_a separate to let the model realize their unique characteristics. Fig. 2 illustrates a concrete example.

Graphs Construction

We first obtain sentence-level EE and AMR outputs, i.e., multi-relational directed acyclic graph structures symbolizing the core concepts of each sentence from two viewpoints. Then, we address graph fusion to build G_e and G_a, ensuring to end up with single wisely-interconnected document graphs rather than two sets of small disjoint networks.

Event Graphs In consonance with BioNLP-ST competitions, events are forged from a trigger (a span that testifies their occurrence, e.g., "interacts", "regulates"), a type (e.g., "binding", "regularization"), and a set of arguments—classed entities or events themselves—playing a certain role (e.g., "theme", "cause"). The specificity of the interactions sought expectedly makes EE domain-specific. We adopt DeepEventMine (Trieu et al. 2020), an end-to-end framework holding the SOTA on seven biomedical benchmarks. We convert output standoff .a* annotation files to heterogeneous event graphs following Frisoni et al. (2022). A node indicates a trigger or an entity, while an edge acts for an entity-trigger or trigger-trigger relation, with the second applying for nested events. Unlike AMRs, event graphs are not available for all sentences but only for the ones expressing desired interactions for which the model has been trained.

AMR Graphs AMR aims to produce a language-neutral representation of meaning, abstracting away from English and providing a layer of abstraction from words to concepts (objects, attributes, etc.) in a rooted graph. It covers ≈100 widespread PropBank semantic roles; as EE, annotations include entity/role identification and typing. We use a SOTA text-to-text AMR parser (Bevilacqua, Blloshmi, and Navigli 2021) with an 83.0 Smatch score on the AMR 3.0 (LDC2020T02) sembank. Since many EE outputs may be empty and damage model robustness (Frisoni. et al. 2022), AMRs supplement not always exploitable event graphs.

Graphs Merging and Rewiring Inside EE and AMR formalisms, an entity, trigger, or concept is canonicalized and represented by a single graph fragment, regardless of how many times it recurs in the sentence (semantic integrity). If a node fulfills multiple roles, AMRs cover within-sentence coreference edge types. On top of event and edited AMR representations, we separately operate graph rewiring to reflect the document structure and enhance the information flow. Mechanically, we introduce artificial *sentence nodes*, each connected to all the event/AMR vertices that originate from that mention. Sentence nodes are linked to each other following their positional order and collected by a *master node*. The resulting G_e and G_a graphs formulate intra- and cross-sentence information, allowing for document traversal by narrative order, concept association, or proximity.

Model

Motivated by the limitations of graph-empowered LSTMs (Frisoni. et al. 2022), COGITOERGOSUMM extends a pre-trained BART-base architecture (Lewis et al. 2020) with the nimble ability to attend to semantic parsing graphs during decoding and preserve the most relevant information via RL. Fig. 3 sketches the overall architecture, built upon four modules, namely text encoding, graph encoding, semantics-driven multi-view decoding, and consistency rewarding.

Text Encoder We feed the input document to a text bidirectional encoder $E_t(\cdot)$ through the learnable BART encoding channel. The l tokens of a record d_i are converted in their contextual hidden representations:

$$\{h_{t_{i,0}}, \ldots, h_{t_{i,l}}\} = E_t\left(\{x_{i,0}, \ldots, x_{i,l}\}\right). \quad (1)$$

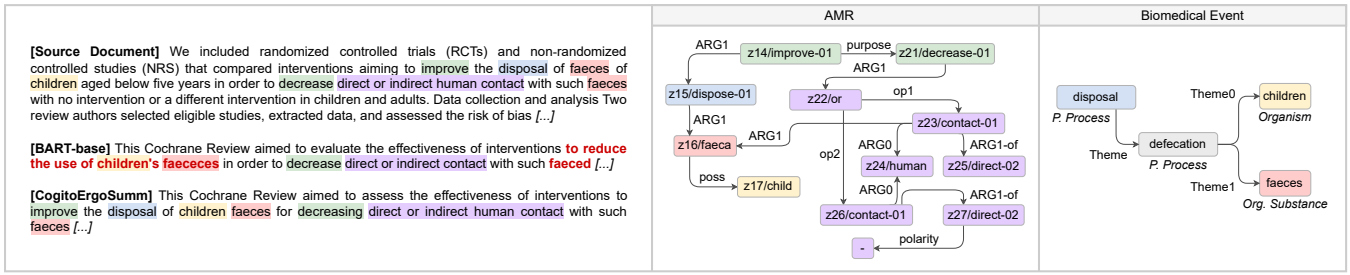

Figure 2: Qualitative example of induced semantic parsing graphs and their assistance to high-quality summarization. Red text indicates hallucinated facts. Token background highlights node alignments.

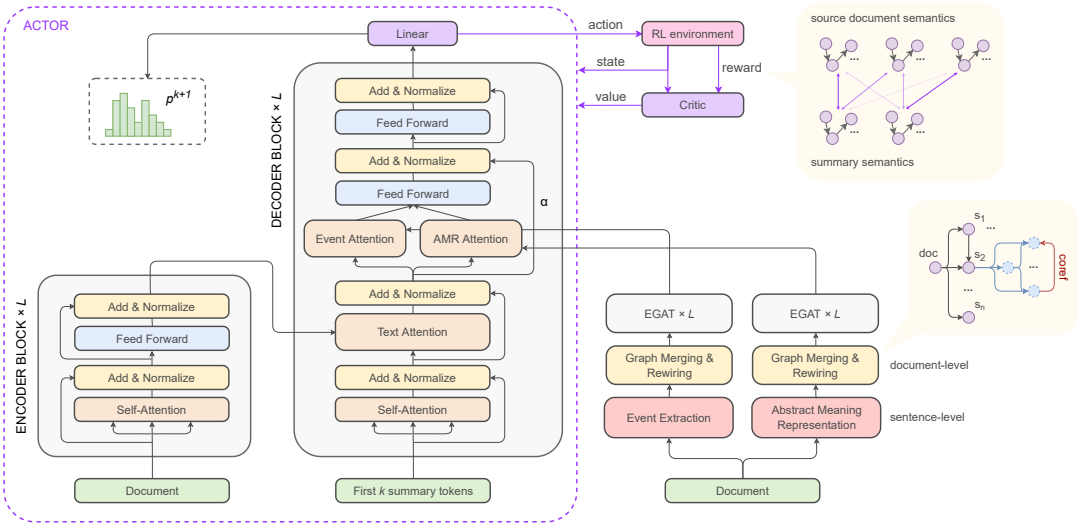

Figure 3: Illustration of the CogitoErgoSumm architecture. We dig out semantic parsing graphs from the input text by means of event extraction and AMR. Each source document is fed to a transformer encoder; event and AMR graph embeddings are learned by edge-aware graph attention networks. A multi-granularity decoder then generates summaries based on all information levels: text, events, and abstract meaning representations. Document-summary semantic graph alignments are used to optimize model consistency via reinforcement learning.

Graph Encoders *Node Initialization*. Since all nodes are accompanied by text, we initialize their features with embeddings from a frozen pre-trained BART language model $E_n(\cdot)$. Pointedly, given a trigger/entity/concept node i with a tokenized text attribute $\texttt{tokens}(i)$, we take the average of its per-token embeddings contextualized on the long input document, i.e., $E_n([x_1, x_2, \ldots, x_n \| \texttt{tokens}(i)])$, where $[.\|.]$ denotes the concatenation operator. To reduce structured prediction noise, we set the maximum node length to 5 tokens. For event graphs, we also prepend the entity and trigger types learned by DeepEventMine. Finally, for sentence and master nodes, we average the token embeddings of the sentence span and the entire document, respectively.

Edge-Aware Graph Attention Network. Through two graph encoders $E_{G_e}(\cdot)$ and $E_{G_a}(\cdot)$, we take the multi-view semantics modeled by G_e and G_a to learn supervised node embeddings and tap implicit relations. Our GNN modules lean on the graph attention groundwork (GAT) (Velickovic et al. 2018), which induces node representations by way of L layers of message passing (shared parameters) and multi-head attention neighborhood aggregation. Message passing in graphs made of cooperating nodes (Lodi, Moro, and Sartori 2010; Cerroni et al. 2013) is actually an established work mode borrowed from communication networks and distributed algorithms. As G_e and G_a are multi-relational graphs, we extend GAT to consider the edge type e connecting two nodes. In the ℓ-th layer, we update the representation $\vec{h}_i^{(\ell)} \in \mathbb{R}^D$ of each node i by:

$$t_{ij}^k = \sigma\left(\vec{\mathbf{a}}^{\mathbf{k}T}\left[\mathbf{W}^{\mathbf{k}}\vec{h}_i \parallel \mathbf{W}^{\mathbf{k}}\vec{h}_j \parallel \mathbf{W}_{\mathbf{r}}^{\mathbf{k}}\vec{e}_{ij}\right]\right), \quad (2)$$

$$\alpha_{ij}^k = \frac{\exp\left(t_{ij}^k\right)}{\sum_{z\in\mathcal{N}(i)}\exp\left(t_{iz}^k\right)}, \quad (3)$$

$$\vec{h}_i^{(\ell+1)} = \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^k \mathbf{W}^{\mathbf{k}}\vec{h}_j^{(\ell)}\right), \quad (4)$$

where $\|_{k=1}^K$ denotes the concatenation of K attention heads,

α_{ij}^k is the normalized attention weight computed by the k-th attention head, $\mathcal{N}(i)$ is the set containing the first-order neighbors of i (including i), $\vec{e}_{ij}$ is the one-hot embedding of the relation type[3] connecting nodes i and j, $\mathbf{W^k}$, $\mathbf{W_r^k}$, and $\vec{\mathbf{a^k}}$ are trainable parameters, σ is a LeakyReLU nonlinearity.

Decoder Inspired by Chen and Yang (2021), we aggregate different levels of encoded representations via a multi-granularity decoder $D(\cdot)$, which predicts the l-th token as:

$$\hat{y} = D(y_{<l}, E_t(d), E_{G_e}(G_e), E_{G_a}(G_a)), \quad (5)$$

$$P(y_l|y_{<l}, d, G_e, G_a) = softmax(\mathbf{W_p}\hat{y}), \quad (6)$$

where $\mathbf{W_p}$ stands for a trainable linear projection. We supplement the BART transformer decoder with two extra cross-attentions (Event Attention and AMR Attention), conducted over the node representations learned by E_{G_e} and E_{G_a} in parallel. We incorporate them into each layer after the original text cross-attention. The token-, event-, and -AMR attended vectors (a^t, a^e, a^a) are combined into a semantics-aware representation a^s through a feed-forward network. To accelerate the training of the new modules and alleviate the negative impact of randomly initialized graph encoders and cross-attentions at early stages, we apply ReZero (Bachlechner et al. 2021) to the residual connection after attending to semantic graphs in each decoder layer: $a^s = a^t + \alpha a^s$, where α is a learnable parameter modulating updates from cross-attention over semantic graphs.

Training Objectives and Consistency Reward During training, we seek to maximize the estimated probability of the actual summary. We adopt the common negative log-likelihood loss function using the teacher-forcing strategy:

$$\mathcal{L}_{nll} = -\sum logP_\theta(y_l|d, G_e, G_a), \quad (7)$$

with θ denoting the set of model parameters. In addition to being syntactically correct, we would like the generated summaries to factually preserve as much pivotal information as possible from the original document. Drawing inspiration from the success of PICO (DeYoung et al. 2021), we believe that structured semantic representations are suitable not only for improved text generation but also for biomedical consistency evaluation. Standing on this intuition, we design a lightweight reward function ψ to maximize the non-differentiable degree of document-summary meaning overlap, which is made possible with second-stage RL training. We refer to Smatch (F-score) (Cai and Knight 2013), metric computing the matching triples between two AMR graphs, benefitting from a higher correlation with human factuality judgments than summary-target ROUGE (Ribeiro et al. 2022). Document-level consistency rewards are obtained via an average pooling; each AMR graph of a predicted summary sentence is compared with all the AMR graphs of the source sentences (one-to-many soft alignments):

$$\phi = \frac{\sum_{y_i \in \text{AMRs}(y)} \left(\frac{\sum_{d_i \in AMRs(d)} smatch(y_i, d_i)}{|\text{AMRs}(d)|} \right)}{|\text{AMRs}(y)|}, \quad (8)$$

[3]Given the non-talking nature of the AMR labels, we make use of categorical edge features, agglomerating numbered arguments (e.g., :op*, :quant*), except for the core ARG* ones.

where $AMRs(\cdot)$ stands for the sentence-level AMR graphs.

Given a starting policy π corresponding to the model trained following Eq. 7, we see autoregressively predicted tokens as actions and employ Proximal Policy Optimization (Schulman et al. 2017) to maximize:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, \quad (9)$$

$$\mathcal{L}_{ppo} = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t], \quad (10)$$

where $r_t(\theta)$ denotes the probability ratio between the action under the policy at iteration t (i.e., the token a_t generated conditionally to the previous ones s_t) and the action under the previous policy. $\hat{\mathbb{E}}[\dots]$ indicates the empirical average over a finite batch of samples; the $clip$ function, combined with the ϵ hyperparameter, ensures that the policy does not change too much among iterations; $\hat{A}$ is the advantage function, an estimation of the relative improvement obtained from the selected action in the current state. To compute advantages, we use the Generalized Advantage Estimation formula (Schulman et al. 2016):

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \dots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \quad (11)$$

where $\delta_t = \psi_t + \gamma V(s_{t+1}) - V(s_t)$; λ and γ are hyperparameters and T is the length of the generated text. The value function V is modeled using a second neural network trained in parallel with the main model (Actor-Critic paradigm). As proposed in (Ziegler et al. 2019), we add a penalty term to the reward function that prevents the new policy from deviating too much from the already pre-trained language model by reducing in expectation their $KL-$divergence:

$$KL = \log \frac{\pi_\theta(a_t|s_t)}{\pi_{base}(a_t|s_t)}, \quad (12)$$

$$\hat{\psi}(a_t|s_t) = \psi(a_t|s_t) - \beta KL, \quad (13)$$

where π_{base} is the starting policy, and β is an adaptive coefficient (initially set via a hyperparameter) that dynamically changes during training to target a specific value of the $KL-$divergence (termed KL_t, see (Ziegler et al. 2019)). The final loss function is:

$$\mathcal{L}_{rl}(\theta) = \hat{\mathbb{E}}[\mathcal{L}_{ppo}(\theta) - c_1\mathcal{L}_{vf}(\theta) + c_2 S[\pi_\theta](s_t)], \quad (14)$$

where $\mathcal{L}_{vf}$ is the mean square error of the value function in charge of updating the critic network; $S[\pi_\theta](s_t)$ is an entropy term used to ensure enough exploration during training; $c1$ and $c2$ are hyperparameters used to tune the importance of each component in the loss.

Experimental Setup

Dataset

We train and evaluate our model on the CDSR dataset (Guo et al. 2021), a publicly available corpus acquired by the widely-used Cochrane Database of Systematic Reviews[4]. CDSR is intended for health literacy, assessing the automatic generation of lay language summaries from

[4]https://www.cochranelibrary.com/cdsr/reviews

biomedical scientific reviews. As far as we know, it is the only biomedical summarization benchmark with manageable sizes and already-known graph-augmented baselines. CDSR comprises 6,677 high-quality pairs, where the source is a long abstract in professional language, and the target is a plain general-public version written by review authors or Cochrane staff. Besides creating accurate and factual summaries, this task also requires a joint style transition, imposing obstacles like terminology explanation and sentence structure simplification that outline a perfect testbed for COGITOERGOSUMM. Details are presented in Table 1.

	Train		Val		Test	
	Source	Target	Source	Target	Source	Target
# pairs	5,178		500		999	
# tokens	878	437	873	434	894	444
# sentences	26	16	26	16	26	16
# words	643	348	643	347	653	352
# extracted events per document	2.63	–	2.54	–	2.70	–
# nodes per G_e/G_a (w/o coref)	6/547	–	5/550	–	6/556	–

Table 1: CDSR dataset statistics. All values are mean except for "# pairs". Sentence segmentation is made with spaCy (Honnibal et al. 2020).

Implementation Details

We extend the HuggingFace implementation of BART-base[5] and initialize weights from a model pre-trained on PubMed[6], leaving 42 as the default training seed. We truncate input documents and set the maximum output length to 1024. We utilize the DeepEventMine model pre-trained on MLEE[7]—the biomedical benchmark most aligned to CDSR (Frisoni. et al. 2022). On average, extracting all the sentence-level AMR and event graphs from a source document takes 3 and 1.2 seconds, respectively. We implement GNNs with PyTorch Geometric (Fey and Lenssen 2019). During RL, we freeze the encoder parameters and train the model decoder only; the critic network is a multilayer perceptron with one hidden layer (hidden size 256). Hyperparameters are listed in Appendix. Each experiment is performed on a workstation having one Nvidia GeForce RTX3090 GPU with 24GB of dedicated memory, 64GB of RAM, and an Intel® Core™ i9-10900X1080 CPU @ 3.70GHz. Training our best model requires 20GB VRAM and 32 hours (19 for RL); 3.18 kg CO2e carbon footprint, 9.83 kWh energy needed, 25.77 Υ_m average carburacy (Moro, Ragazzi, and Valgimigli 2023).

Baselines

We head-to-head compare COGITOERGOSUMM to representative extractive and abstractive summarization models.

- *Oracle*. It creates an extractive summary by selecting the sentences in the document having the highest ROUGE-2 score with the target (syntactic match upper bound).
- *BERT* (Liu and Lapata 2019). Inter-sentence encoder with classification head, supervised by Oracle extractive.
- *Pointer generator* (See, Liu, and Manning 2017). Seq2seq model trained both to copy words from the source and generate new ones from a fixed vocabulary.
- *BART* (Lewis et al. 2020). We take into account models pre-trained on PubMed.
- *EASumm* (Frisoni. et al. 2022). An event-augmented graph-LSTM architecture for abstractive summarization.

Quantitative and Qualitative Evaluation

On the trails of common practice, we automatically evaluate model performance in terms of ROUGE-1, ROUGE-2, and ROUGE-L F1 scores. These recall-based metrics evaluate informativeness by measuring unigrams, bigrams, and longest common subsequence overlaps. A standard ROUGE metric does not shed meaningful light on other important dimensions like semantic coherence, abstractness, and intelligibility—which are focal for our work. To better gauge summary quality, we apply BERTScore (Zhang et al. 2020b), FactCC (Kryscinski et al. 2020), and report n-grams novelty[8] in tandem with readability metrics, namely Flesch-Kincaid grade-level (Kincaid et al. 1975) and Coleman-Liau (Coleman and Liau 1975) indices[9]. We utilize default hyperparameters for all the metrics except for BERTScore, where we use the IDF weighting and specify `rescale_with_baseline=True` to increase interpretability and avoid small range variations. In the quest to analyze our generated summaries qualitatively, we operate an in-depth human evaluation study. We randomly select 30 CDSR test set instances and invite 3 annotators—proficient English speakers with biomedical competencies—to access our models' outputs, along with those of BART-base (presented in random order), i.e., the skeleton model oblivious to semantics. After reading the articles, each judge scores summaries on a Likert scale from 1 (worst) to 5 (best) in conformity with four independent perspectives: (i) *informativeness*, i.e., conveying salient content; (ii) *factualness*, i.e., being faithful to the article; (iii) *fluency*, i.e., being fluent, grammatical, and coherent; (iv) *succinctness*, i.e., non containing redundant and unnecessary information.

Results and Ablation Studies

Overall quantitative results are delighted in Table 2. Zooming out, we find that structured semantic information can greatly help a pre-trained language model recognize salient parts in source documents. Although our focus is on improving semantic consistency, the ROUGE scores attained by COGITOERGOSUMM are significantly higher than previous extractive and abstractive methods, except for BART-large, for which our models are still competitive despite having more than 2× fewer trainable parameters. Remarkably, we beat graph-LSTMs by 6 R-1/-L points. The per-

[5]https://huggingface.co/docs/transformers/model_doc/bart

[6]https://huggingface.co/gayanin/bart-mlm-pubmed

[7]http://nactem.ac.uk/MLEE (from molecules to organisms)

[8]Percentage of new word-level unigrams in the predicted summaries compared to the source (See, Liu, and Manning 2017).

[9]They estimate the years of education generally required to understand the summary. We compute them using https://pypi.org/project/textstat/. Lower scores indicate that the text is easier to read (U.S. college-level readability belongs to the range [13-16]).

Model	#params	R-1	R-2	R-L	Flesch-Kincaid	Coleman-Liau
ORACLE†	—	53.56	25.54	49.56	14.85	16.13
BERT-base†	110M	26.60	11.11	24.59	13.44	14.40
POINTER GENERATOR†	22M	38.33	14.11	35.81	16.36	15.90
BART-base (PubMed)	139M	51.20	19.77	48.47	13.69	13.45
BART-large (PubMed)†	406M	**52.66**	**21.73**	**49.97**	13.30	14.28
EASUMM‡	8M	46.30	18.73	43.78	12.42	13.06
COGITOERGOSUMM	181M	52.23	<u>20.63</u>	49.44	14.10	13.67
- w/o RL	180M	<u>52.30</u>	20.47	<u>49.46</u>	14.06	13.64
- w/o event and RL	155M	52.13	20.42	49.30	14.02	13.69
- w/o AMR and RL	157M	52.02	20.54	49.25	13.97	13.66

Table 2: Automated evaluation on the full test set of CDSR with ROUGE (R in short) and readability metrics. Top: extractive models. Middle: abstractive models. Bottom: our semantics-augmented abstractive model. **Bold** and <u>underline</u> denote the best and second best R scores. † and ‡ results are from (Guo et al. 2021) and (Frisoni. et al. 2022), respectively. Our model significantly outperforms BART-base (Pitman's permutation test, p<0.05).

formance drops the most when removing graph encoders. Even if AMRs appear more impactful than events, the best results come from their mixture, indicating that the two types of semantic graphs complement each other in generating sounder summaries. RL effects are not appreciable by ROUGE, howbeit remarkable with a deeper analysis. Fig. 4 presents the human evaluation results[10] contrasted to automatic metrics on the same sample. The average Kendall coefficient among all evaluators' inter-rater agreement is 0.16. COGITOERGOSUMM ranks better on every quality dimension inspected, pronouncing the gap with BART (+12.46% factualness, +6.69% informativeness) and reaffirming previous deductions. The plot underlines the poor correlation between ROUGE and the desired output properties.

We validate the relative impact of our principal components (Table 3). Firstly, we scrutiny different graph encoders: (i) a GAT on Levi-transformed bipartite graphs treating nodes and edges equally, and (ii) an edge-aware GAT; we uncover that (ii) brings a substantial headroom that ascertains the value of relation types in representation learning. Secondly, we test different ways of combining event and AMR cross-attentions, documenting slightly better scores with a parallel strategy. Tests with different random seeds during fine-tuning and additional qualitative case studies are disclosed in Appendix.

	Ablation	$\tilde{R}$	Readability	BS	NN
GNN Type	EGAT	**40.74**	15.17	**14.88**	**51.55**
	GAT bipartite	40.67	**15.16**	14.76	51.18
Attn. Comb.	Parallel	**40.74**	15.17	14.82	51.55
	Sequential (AMR, event)	40.65	15.08	**14.98**	51.81
	Sequential (event, AMR)	40.69	**15.01**	14.84	**52.57**

Table 3: Ablation results averaged over three runs. Evaluation on ROUGE-1/2/L average ($\tilde{R}$), Flesch-Kincaid and Coleman-Liau average (Readability), BERTScore (BS), and average % novel n-grams w/ $n \in [1-4]$ (NN).

[10]Human judges are published for the sake of replicability; https://github.com/disi-unibo-nlp/cogito-ergo-summ/human_evaluation.xlsx

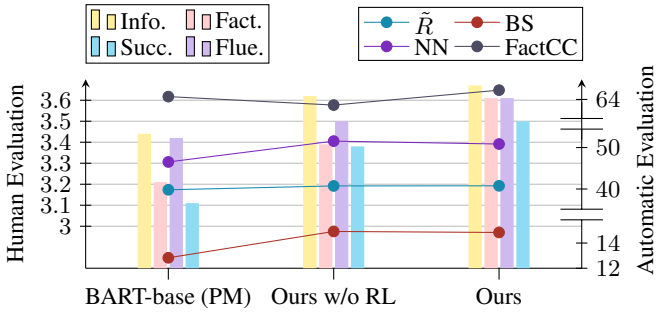

Figure 4: Human evaluation scores on **Infor**mativeness, **Fact**ualness, **Fluen**cy, **Succ**inctness, compared to the ROUGE-1/2/L average ($\tilde{R}$), BERTScore (BS), FactCC %, and average % novel n-grams w/ $n \in [1-4]$ (NN). COGITOERGOSUMM (ours) achieves significantly higher ratings than BART-base PubMed (student t-test, p=0.0305).

.

Conclusion

In this paper, we introduce a framework for infusing domain-specific and -general semantic parsing graphs—events and AMRs—into transformer-based models for biomedical abstractive summarization. We propose new decoder cross-attention modules and reward signals to generate high-quality summaries conditioned on both the source documents and their formal underlying semantics. Experiments and ablation studies on CDSR demonstrate that 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 (i.e. ROUGE). 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 (Kapanipathi et al. 2021) and neuro-logic decoding (Lu et al. 2021).

References

An, C.; Zhong, M.; Chen, Y.; Wang, D.; et al. 2021. Enhancing Scientific Papers Summarization with Citation Graph. In *AAAI*, 12498–12506. AAAI Press.

Angeli, G.; Premkumar, M. J. J.; and Manning, C. D. 2015. Leveraging Linguistic Structure For Open Domain Information Extraction. In *ACL (1)*, 344–354. ACL.

Bachlechner, T.; Majumder, B. P.; Mao, H. H.; Cottrell, G.; et al. 2021. ReZero is all you need: fast convergence at large depth. In *UAI*, volume 161 of *Proceedings of Machine Learning Research*, 1352–1361. AUAI Press.

Balachandran, V.; Pagnoni, A.; Lee, J. Y.; Rajagopal, D.; et al. 2021. StructSum: Summarization via Structured Representations. In *EACL*, 2575–2585. ACL.

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; et al. 2013. Abstract Meaning Representation for Sembanking. In *LAW@ACL*, 178–186. ACL.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT*, 610–623. ACM.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *ACL*, 5185–5198. ACL.

Bevilacqua, M.; Blloshmi, R.; and Navigli, R. 2021. One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline. In *AAAI*, 12564–12573. AAAI Press.

Cai, S.; and Knight, K. 2013. Smatch: an Evaluation Metric for Semantic Feature Structures. In *ACL (2)*, 748–752. ACL.

Cao, Z.; Wei, F.; Li, W.; and Li, S. 2018. Faithful to the Original: Fact Aware Neural Abstractive Summarization. In *AAAI*, 4784–4791. AAAI Press.

Cerroni, W.; Moro, G.; Pirini, T.; and Ramilli, M. 2013. Peer-to-Peer Data Mining Classifiers for Decentralized Detection of Network Attacks. In Wang, H.; and Zhang, R., eds., *ADC*, volume 137 of *CRPIT*, 101–108. ACS.

Chen, J.; and Yang, D. 2021. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In *NAACL-HLT*, 1380–1391. ACL.

Coleman, M.; and Liau, T. L. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60: 283–284.

DeYoung, J.; Beltagy, I.; van Zuylen, M.; Kuehl, B.; et al. 2021. MSˆ2: Multi-Document Summarization of Medical Studies. In *EMNLP*, 7494–7513. ACL.

Dohare, S.; Gupta, V.; and Karnick, H. 2018. Unsupervised Semantic Abstractive Summarization. In *ACL (3)*, 74–83. ACL.

Fan, A.; Gardent, C.; Braud, C.; and Bordes, A. 2019. Using Local Knowledge Graph Construction to Scale Seq2Seq Models to Multi-Document Inputs. In *EMNLP/IJCNLP (1)*, 4184–4194. ACL.

Fey, M.; and Lenssen, J. E. 2019. Fast Graph Representation Learning with PyTorch Geometric. *CoRR*, abs/1903.02428.

Frisoni., G.; Italiani., P.; Boschi., F.; and Moro., G. 2022. Enhancing Biomedical Scientific Reviews Summarization with Graph-based Factual Evidence Extracted from Papers. In *DATA*, 168–179. INSTICC, SciTePress.

Frisoni, G.; Moro, G.; and Carbonaro, A. 2021. A Survey on Event Extraction for Natural Language Understanding: Riding the Biomedical Literature Wave. *IEEE Access*, 9: 160721–160757.

Frisoni, G.; Moro, G.; Carlassare, G.; and Carbonaro, A. 2022. Unsupervised Event Graph Representation and Similarity Learning on Biomedical Literature. *Sensors*, 22(1): 3.

Gigioli, P.; Sagar, N.; Rao, A. S.; and Voyles, J. 2018. Domain-Aware Abstractive Text Summarization for Medical Documents. In *BIBM*, 2338–2343. IEEE Comp. Society.

Guo, Y.; Qiu, W.; Wang, Y.; and Cohen, T. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. In *AAAI*, 160–168. AAAI Press.

Hardy; and Vlachos, A. 2018. Guided Neural Language Generation for Abstractive Summarization using Abstract Meaning Representation. In *EMNLP*, 768–773. ACL.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength natural language processing in python.

Huang, L.; Wu, L.; and Wang, L. 2020. Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward. In *ACL*, 5094–5107. ACL.

Ji, X.; and Zhao, W. 2021. SKGSUM: Abstractive Document Summarization with Semantic Knowledge Graphs. In *IJCNN*, 1–8. IEEE.

Kapanipathi, P.; Abdelaziz, I.; Ravishankar, S.; Roukos, S.; et al. 2021. Leveraging Abstract Meaning Representation for Knowledge Base Question Answering. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 3884–3894. ACL.

Karamanis, N. 2007. Book Reviews: Text Mining for Biology and Biomedicine, edited by Sophia Ananiadou and John McNaught. *Computational Linguistics*, 33(1): 135–140.

Kincaid, J. P.; Fishburne, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

Kryscinski, W.; McCann, B.; Xiong, C.; and Socher, R. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *EMNLP*, 9332–9346. ACL.

Kryscinski, W.; Paulus, R.; Xiong, C.; and Socher, R. 2018. Improving Abstraction in Text Summarization. In *EMNLP*, 1808–1817. ACL.

Lee, F.; Kedzie, C.; Verma, N.; and McKeown, K. R. 2021. An analysis of document graph construction methods for AMR summarization. *CoRR*, abs/2111.13993.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; et al. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880. ACL.

Li, W.; Xiao, X.; Liu, J.; Wu, H.; et al. 2020. Leveraging Graph to Improve Abstractive Multi-Document Summarization. In *ACL*, 6232–6243. ACL.

Lin, H.; and Ng, V. 2019. Abstractive Summarization: A Survey of the State of the Art. In *AAAI*, 9815–9822. AAAI Press.

Liu, Y.; and Lapata, M. 2019. Text Summarization with Pretrained Encoders. In *EMNLP/IJCNLP (1)*, 3728–3738. ACL.

Lodi, S.; Moro, G.; and Sartori, C. 2010. Distributed data clustering in multi-dimensional peer-to-peer networks. In Shen, H. T.; and Bouguettaya, A., eds., *ADC*, volume 104 of *CRPIT*, 171–178. ACS.

Lu, X.; West, P.; Zellers, R.; Bras, R. L.; et al. 2021. NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints. In *NAACL-HLT*, 4288–4299. ACL.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. T. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*, 1906–1919. ACL.

Mehdad, Y.; Carenini, G.; and Ng, R. T. 2014. Abstractive Summarization of Spoken and Written Conversations Based on Phrasal Queries. In *ACL (1)*, 1220–1230. ACL.

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *EMNLP*, 404–411. ACL.

Moradi, M.; and Ghadiri, N. 2019. Text summarization in the biomedical domain. *arXiv preprint arXiv:1908.02285*.

Moro, G.; Pagliarani, A.; Pasolini, R.; and Sartori, C. 2018. Cross-domain & In-domain Sentiment Analysis with Memory-based Deep Neural Networks. In *IC3K*, volume 1, 127–138. SciTePress.

Moro, G.; and Ragazzi, L. 2022. Semantic Self-Segmentation for Abstractive Summarization of Long Documents in Low-Resource Regimes. In *AAAI*, 11085–11093. AAAI Press.

Moro, G.; Ragazzi, L.; and Valgimigli, L. 2023. Carburacy: Summarization Models Tuning and Comparison in Eco-Sustainable Regimes with a Novel Carbon-Aware Accuracy. In *AAAI*, 1–9. AAAI Press.

Moro, G.; Ragazzi, L.; Valgimigli, L.; and Freddi, D. 2022. Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature. In *ACL (1)*, 180–189. ACL.

Paulus, R.; Xiong, C.; and Socher, R. 2018. A Deep Reinforced Model for Abstractive Summarization. In *ICLR (Poster)*. OpenReview.net.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2016. Sequence Level Training with Recurrent Neural Networks. In *ICLR (Poster)*.

Ribeiro, L.; Liu, M.; Gurevych, I.; Dreyer, M.; et al. 2022. FactGraph: Evaluating Factuality in Summarization with Semantic Graph Representations. In *NAACL-HLT*, 3238–3253. ACL.

Rothe, S.; Narayan, S.; and Severyn, A. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Trans. Assoc. Comput. Linguistics*, 8: 264–280.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M. I.; et al. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *ICLR (Poster)*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; et al. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Scialom, T.; Lamprier, S.; Piwowarski, B.; and Staiano, J. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. In *EMNLP/IJCNLP (1)*, 3244–3254. ACL.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL (1)*, 1073–1083. ACL.

Sharma, E.; Huang, L.; Hu, Z.; and Wang, L. 2019. An Entity-Driven Framework for Abstractive Summarization. In *EMNLP/IJCNLP (1)*, 3278–3289. ACL.

Tan, J.; Wan, X.; and Xiao, J. 2017. Abstractive Document Summarization with a Graph-Based Attentional Neural Model. In *ACL (1)*, 1171–1181. ACL.

Trieu, H.; Tran, T. T.; Nguyen, A. D.; Nguyen, A.; et al. 2020. DeepEventMine: end-to-end neural nested event extraction from biomedical texts. *Bioinform.*, 36(19): 4910–4917.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; et al. 2018. Graph Attention Networks. In *ICLR (Poster)*. OpenReview.net.

Wan, X. 2008. An Exploration of Document Impact on Graph-Based Multi-Document Summarization. In *EMNLP*, 755–762. ACL.

Wang, D.; Liu, P.; Zheng, Y.; Qiu, X.; et al. 2020. Heterogeneous Graph Neural Networks for Extractive Document Summarization. In *ACL*, 6209–6219. ACL.

Wu, L.; Chen, Y.; Shen, K.; Guo, X.; et al. 2021a. Graph Neural Networks for Natural Language Processing: A Survey. *CoRR*, abs/2106.06090.

Wu, W.; Li, W.; Xiao, X.; Liu, J.; et al. 2021b. BASS: Boosting Abstractive Summarization with Unified Semantic Graph. In *ACL/IJCNLP (1)*, 6052–6067. ACL.

Yu, W.; Zhu, C.; Li, Z.; Hu, Z.; et al. 2020. A Survey of Knowledge-Enhanced Text Generation. *CoRR*, abs/2010.04389.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. J. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 11328–11339. PMLR.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; et al. 2020b. BERTScore: Evaluating Text Generation with BERT. In *ICLR*. OpenReview.net.

Zhu, C.; Hinthorn, W.; Xu, R.; Zeng, Q.; et al. 2021. Enhancing Factual Consistency of Abstractive Summarization. In *NAACL-HLT*, 718–733. ACL.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; et al. 2019. Fine-Tuning Language Models from Human Preferences. *CoRR*, abs/1909.08593.