Entity-Driven Fact-Aware Abstractive Summarization of Biomedical Literature

Amanuel Alambo
Wright State University - Main Campus

Tanvi Banerjee
Wright State University - Main Campus, tanvi.banerjee@wright.edu

Krishnaprasad Thirunarayan
Wright State University - Main Campus, t.k.prasad@wright.edu

Michael Raymer
Wright State University - Main Campus, michael.raymer@wright.edu

Follow this and additional works at: https://corescholar.libraries.wright.edu/cse

Repository Citation
Alambo, A., Banerjee, T., Thirunarayan, K., & Raymer, M. (2022). Entity-Driven Fact-Aware Abstractive Summarization of Biomedical Literature. https://corescholar.libraries.wright.edu/cse/628

This Article is brought to you for free and open access by Wright State University's CORE Scholar. It has been accepted for inclusion in Computer Science and Engineering Faculty Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
Entity-driven Fact-aware Abstractive Summarization of Biomedical Literature

Amanuel Alambo*, Tanvi Banerjee*, Krishnaprasad Thirunarayan*, Michael Raymer*
*Wright State University
{alambo.2, tanvi.banerjee, t.k.prasad, michael.raymer}@wright.edu

Abstract—As part of the large number of scientific articles being published every year, the publication rate of biomedical literature has been increasing. Consequently, there has been considerable effort to harness and summarize the massive amount of biomedical research articles. While transformer-based encoder-decoder models in a vanilla source document-to-summary setting have been extensively studied for abstractive summarization in different domains, their major limitations continue to be entity hallucination (a phenomenon where generated summaries constitute entities not related to or present in source article(s)) and factual inconsistency. This problem is exacerbated in a biomedical setting where named entities and their semantics (which can be captured through a knowledge base) constitute the essence of an article. The use of named entities and facts mined from background knowledge bases pertaining to the named entities to guide abstractive summarization has not been studied in biomedical article summarization literature. In this paper, we propose an entity-driven fact-aware framework for training end-to-end transformer-based encoder-decoder models for abstractive summarization of biomedical articles. We call the proposed approach, whose building block is a transformer-based model, EFAS, Entity-driven Fact-aware Abstractive Summarization. We conduct a set of experiments using five state-of-the-art transformer-based encoder-decoder models (two of which are specifically designed for long document summarization) and demonstrate that injecting knowledge into the training/inference phase of these models enables the models to achieve significantly better performance than the standard source document-to-summary setting in terms of entity-level factual accuracy, N-gram novelty, and semantic equivalence while performing comparably on ROUGE metrics. The proposed approach is evaluated on ICD-11-Summ-1000, a dataset we build for abstractive summarization of biomedical literature, and PubMed-50k, a segment of a large-scale benchmark dataset for abstractive summarization of biomedical literature.

Index Terms—Transformers, Named Entity Recognition, Knowledge Bases, Abstractive Summarization, ICD-11, Knowledge Retrieval, Knowledge-enhanced Natural Language Generation

I. INTRODUCTION

Neural abstractive summarization is well explored for summarization of news articles [1]-[8], and scientific articles [6]-[8]. While there are some efforts to apply neural abstractive summarization techniques for biomedical literature [6], [9], the use of named entities and background knowledge bases to guide biomedical abstractive summary generation has not been explored. On the other hand, named entity recognition/understanding in biomedical literature has been extensively studied such that named entities are known to harbor significant semantics about a biomedical article [10]-[13]. Further, linking named entities to concept definitions in background domain-specific knowledge bases boost semantic understanding of a biomedical article by improving comprehensiveness and contextualization as investigated in [13]-[16].

The use of named entities to guide abstractive summarization of news articles has been explored in recent studies by [17]-[20]. However, these studies do not leverage named entity-driven facts from background knowledge bases to guide abstractive summary generation. Inspired by recent advances in transformer-based encoder-decoder models [21]-[24] and knowledge and retrieval augmented natural language generation [25]-[30], we propose a technique for retrieval of entity-driven facts from biomedical knowledge bases and leveraging the retrieved facts as contextual signals for abstractive summarization of biomedical literature. Our proposed framework consists of two major components: 1) entity-driven knowledge retriever; and 2) knowledge-guided abstractive summarizer. The entity-driven knowledge retriever extracts facts from UMLS [31], ICD-10 [32], and SNOMED-CT [33] based on named entities in a biomedical article while the knowledge-guided abstractive summarizer is trained to generate summaries by attending to the source article to be summarized, the chain of named entities in the source article, and the retrieved facts from the knowledge bases.

We conduct experiments on two datasets: ICD-11-Summ-1000, and PubMed-50k. For curating ICD-11-Summ-1000, we conduct entity-driven clustering of PubMed abstracts collected per ICD-11 [34] chapter followed by entity-aware content selection to build an entity-aware pseudo extractive document. The pseudo extractive document is passed as an input source article to generate the abstractive summary during inference. The contributions of this study are summarized as follows:

• We introduce an approach for named entity-driven fact retrieval from biomedical knowledge bases using dense vector representations.
• We propose a framework based on transformer-based encoder-decoder models for knowledge-guided large-scale abstractive summarization of biomedical literature.
• We conduct extensive experiments and ablation studies to assess the efficacy of the proposed approach and show that injecting named entities and entity-aware facts mined from biomedical knowledge bases into abstractive summarization models can boost their performance in terms of entity-level factual consistency [19], [35]-[37], N-gram novelty [30], [38], and semantic equivalence.
measured using BERTScore [39].

- We develop ICD-11-Summ-1000, a dataset consisting of clusters of biomedical articles collected from PubMed for the special groups chapters in the ICD-11 catalog [1] and the abstractive summaries generated using different variations of our experimental design. Further, we plan to share our code, and data with other researchers [2].

II. RELATED WORK

While abstractive summarization is well studied for summarization of news articles with success attributed to the availability of a massive amount of training data, their applicability to scholarly articles, particularly, in the biomedical domain is limited. Further, although named entities have been extensively studied to convey the semantics of an article (news, scientific, social media) and the saliency of individual sentences [18] within an article, they have not been widely used as part of modeling abstractive summarization. [17] performed entity-aware single-document abstractive summarization using reinforcement learning for training. Their pipeline-based approach consists of an entity-aware content selection module and abstract generation module. They evaluate their approach on the CNN/Daily Mail and NYT corpora. [18] perform entity-driven multi-document abstractive summarization of news articles (WikiSum, and Multi-News) using an encoder-decoder framework augmented with Graph Attention Network (GAT). [40] proposed EntityRank, an extension of the LexRank [41] graph-based algorithm, for entity-supported summarization of biomedical abstracts.

There have been a few recent efforts towards knowledge/fact-aware abstractive summarization in different domains. [42] introduced a Fact-aware Abstractive Summarization model called FaSum for improving the factual consistency of summaries in the domain of news articles. However, their approach does not leverage named entities for fact retrieval. [20] extended a transformer-based abstractive summarization model using entities disambiguated and linked to Wikidata knowledge graph and attending to the entities for summarization of news articles. Their approach, however, does not perform named-entity based fact retrieval from the knowledge base constrained by the article to be summarized and the named entities. [43] developed an unsupervised pipeline-based approach for knowledge-infused abstractive summarization for condensing patient-to-clinician diagnostic interviews based on Multi-Sentence Compression [44] and Integer Linear Programming [45]. Nevertheless, their approach uses domain-specific lexicons as knowledge source for filtering irrelevant utterances and for retrofitting language models [46] and, does not leverage named entities or facts as part of an end-to-end training of models. [47] proposed Biomed-Summarizer, a framework for extractive summarization of biomedical literature in a multi-document setting and evaluated on PubMed abstracts. [6] built a model for abstractive summarization of long documents using a discourse-aware encoder-decoder framework and experimented on two large scale datasets including research articles collected from PubMed. To address the challenge associated with the scarcity of large-scale training data in the biomedical domain, [48] released MS2 (Multi-Document Summarization of Medical Studies). They experimented with BART [22] for abstractive summary generation on the dataset they introduced in a traditional multi-doc-to-summary setting. Though all the aforementioned studies conduct abstractive summarization of biomedical literature or the use of facts mined from knowledge bases for a different domain, they follow the well-established paradigm of source-document vs summary pairing during training/inference of models. Our approach is different in that we augment the state-of-the-art abstractive summarization models with additional contextual signals during training/inference and apply them to the biomedical domain.

III. DATA PREPARATION

In addition to the benchmark PubMed-50k [6] dataset which we use to train our models, we curate ICD-11-Summ-1000 dataset using a data preparation pipeline which follows two stages: 1) ICD-11 disease lexicon curation for querying PubMed for abstracts; and 2) Entity-aware pseudo-document generation for a collection of semantically related PubMed abstracts in an ICD-11 chapter. Thus, for querying PubMed for abstracts for an ICD-11 chapter, we first query the biomedical knowledge bases for “disease” keywords using the names of each ICD-11 chapter and build a lexicon of diseases corresponding to each chapter. Figure-1 shows what a lexicon build-up for an ICD-11 chapter looks like. Once the disease related keywords are identified for an ICD-11 chapter, we use these keywords to query PubMed via the Bio Entrez parser to capture the first 1000 abstracts (PMIDs) spanning a period of last 90 days from the moment we initiated the query. We do this for each of the eight special groups ICD-11 chapters.

Fig. 1: ICD-11 based lexicon construction and querying for abstracts from PubMed using Bio Entrez parser. For illustration purpose, we show the pipeline for ICD-11 chapter Neoplasms.

Figure-2 shows our ICD-11-Summ-1000 dataset preparation pipeline. Once we have queried the 1000 PubMed abstracts for an ICD-11 chapter, we conduct named entity recognition (NER) on each of the abstracts within a chapter using the SciSpacy NER model trained using the BCS5CDR corpus [49]. Since we are interested in entity-level clustering of PubMed

---

1. https://bit.ly/3GEFWvM
2. https://github.com/AmaelFilbiomed_abstractive_summarization
3. https://biopython.org/docs/1.75/api/Bio.Entrez.html
abstracts within an ICD-11 chapter, we first conduct clustering of the named entities using agglomerative clustering as used in [19]. We use BioBERT [50] for named entity representation followed by agglomerative clustering. Once the named entities pertaining to an ICD-11 chapter are clustered into different bins, our next task is to cluster the PubMed abstracts into a bin based on how the named entities within the abstracts are related to the entities within a cluster. We use cosine similarity between named entities identified in a PubMed abstract and entities characterizing a cluster to determine an entity-aware cluster the abstract belongs to. Next, for each cluster, we perform named entity-aware salient content selection to produce an extractive pseudo-document for each cluster. This paradigm of reducing a multi-document corpus (i.e., a cluster consisting of PubMed abstracts grouped based on entity-relatedness) into an extractive pseudo-document is explored for different tasks in previous studies [51]–[53]. As part of the NER task, we use coreference resolution [54], [55] after named entities are extracted using SciSpacy to cluster the biomedical named entities and their coreferenced mentions spanning the multiple abstracts within an ICD-11 chapter.

For all tasks throughout this paper involving initializing of networks or for representation learning, we use BioBERT [50].

IV. PROPOSED METHOD

Our proposed approach is a two-stage framework consisting of 1) an entity-driven knowledge retriever, and 2) a knowledge-guided abstractive summarizer. In this section, we discuss both modules in detail.

A. Entity-driven Knowledge Retriever

For each extractive pseudo-document generated in the data preparation stage for ICD-11 or input article for PubMed-50k designated by $D$, we identify the named entities in the input document. The identified named entities are then used to retrieve facts from biomedical knowledge bases (UMLS, ICD-10, and SNOMED-CT). We use PyMedTermino [57] to work with the entire dump of UMLS [31] available for $\mathcal{K}$. For $m$ named entities (and their coreferenced mentions), we have a set of pairs of entities $\{(e_i, e_j) \mid 0 \leq i < j \leq m\}$ extracted from $D$, where each pair $(e_i, e_j)$ is used to query for candidate facts $F_{1}, F_{2}, F_{3}, \ldots, F_{|\mathcal{K}|}$ denoted collectively by $F_{\mathcal{D}}$ from the background knowledge bases $\mathcal{K}$ using full text search. The complete set of facts retrieved for all pairs of named entities in source document $D$ is denoted by $\mathcal{F}_{D}$.

The reason we use a pair of named entities to perform lexical query from $\mathcal{K}$ is to capture the relationship between a pair of named entities as it appears in a knowledge base since relationships among named entities harbor semantics in addition to the entities themselves and help with disambiguating relevant facts from irrelevant facts. After the candidate facts $\mathcal{F}_{D}$ are retrieved from the knowledge bases $\mathcal{K}$, we embed the candidate facts using BioBERT. Then, we perform efficient vector similarity search using Maximum Inner Product Search (MIPS) [58] implemented in the FAISS library [59] to query for the top-k facts among the candidate facts ($\mathcal{F}_{D}$) using the input document $D$ as the query. Formally, we define the similarity between fact $F_{i} \in \mathcal{F}_{D}$ and document $D$ as

$$sim(F_{i}, D) = \hat{V}(F_{i})^{T}\hat{V}(D)$$

where $\hat{V}(F_{i})$ - Vector representation of Fact $F_{i}$; $\hat{V}(D)$ - Vector representation of document $D$.

Thus, after the knowledge retrieval task, we have 1) the input document $D$ which is obtained during the data preparation phase for ICD-11 and readily available for PubMed-50k; 2) the named entity chain (i.e., chain of named entities extracted from the pseudo-doc) $\mathcal{E}$ [19]; and 3) top-k facts $F_{1}, F_{2}, F_{3}, \ldots, F_{|\mathcal{K}|}$ retrieved from the background knowledge bases collectively represented as $F_{\mathcal{K}} \subseteq \mathcal{F}_{D}$. We set the value of $K$ to 3 following the study by [19]. We experiment with different values of $K$ as detailed in the ablation studies section. The combination of these contextual signals will be used to guide the summarization model at training/inference time. The rationale for using maximum inner product search for knowledge retrieval

\[^{a}\text{https://bit.ly/3E0z1lI}
\[^{b}\text{https://github.com/facebookresearch/faiss}\]
is inspired by the works of [26–28], [59], [60], albeit they used it mainly for open domain question answering [61], [62].

Table I shows sample facts, as they appear and retrieved from the SciSpacy NER model trained on the BCSCDR corpus [49].

The decoder component, which is trained using teacher forcing [66] at training time, consists of two cross-attention sub-layers to attend to: 1) the input source article; and 2) the affine transformed concatenation of facts and named entity chain’s codings. The following formulations show the transformations in the decoder component where the ground truth output sequence $y$ is passed to the sequence of transformer decoder layers and is used to initialize the first hidden layer $h_0$ of the decoder network. Note that we have a Masked Multi-head Self-Attention in the decoder network denoted by $MMHAtt$.

\[
\hat{h}_i^l := \text{LayerNorm}(h_i^{l-1} + M MHAtt(h_i^{l-1}))
\]

\[
h_i^l := \text{LayerNorm}(h_i^l + FFN(h_i^l))
\]

V. MODEL TRAINING

All models are trained with a cross-entropy loss using back-propagation:

\[
\mathcal{L}_\theta = -\frac{1}{n} \sum_{k=1}^{n} P(t_k | l_k, X, \mathcal{E}, F_K; \theta)
\]

Where $X$ - the input sequence to be summarized; $\mathcal{E}$ - the named entities chain in the input sequence $X$; and $F_K$ - top-k facts extracted from biomedical KB; $\theta$ - model parameters.

We train each model using cross entropy loss to generate the ground truth summaries for the PubMed-50k dataset. We use the following hyperparameters setting: number of epochs to 5, fixed learning rate of 5e-5 with adam optimizer [67], batch size to 8, beam size of 5 with a length penalty [68] $\alpha$ between the range of 0.6 and 1 [5] at inference time; to deal with long-document summarization using the traditional transformer encoder-decoder models, we split the source article into
chucks of a maximum of 512 tokens and independently encode each chunk, after which we concatenate and project back to 768 dimension using a linear layer. The approach of splitting the long input sequence into smaller chunks of 512 tokens and then embedding independently is motivated by the recent work by [48]. For Longformer-Encoder-Decoder (LED) [69] and BigBird [70], however, we set the maximum length of the input sequence to 8192 tokens since they can deal with long input sequences without having to truncate; maximum output sequence length is set to 210 tokens following the experiments by [6]. To mitigate redundancy in the generated summaries, we enable trigram blocking [71] during inference. For each backbone model, we use its base variant with 12 encoder and 12 decoder layers. The train/validation/test sizes for PubMed-50k is 50,000/5,000/5,000 and each model is trained using early stopping. A checkpoint of the model that is saved to the file system. All models are built and trained using PyTorch on NVIDIA Tesla T4 GPU. We perform model training experiments with different input guidance settings: input document only, input document + named entities chain, input document + named entities chain + knowledge facts. For our base summarization document + named entities chain, input document + named entities chain, input document + named entities chain + knowledge facts. At inference time, we experiment with five transformer-based encoder-decoder models and show that our entity-driven knowledge-aware approach enables us to achieve the best performance in entity-level factual consistency, N-gram novelty, and semantic equivalence while performing comparably on the commonly used ROUGE metrics. At inference time, we experiment with two settings (w/o named entities, and w/ named entities).

VI. EXPERIMENTS AND RESULTS

While all models are trained using the PubMed-50k corpus, they are evaluated using a hold-out test set from the original PubMed dataset as well as the ICD-11-Summ-1000 corpus we curate. The experimental results are shown in Table II through Table IX. Results of evaluation w.r.t source articles reported are average results for both the ICD-11-Summ-1000 and PubMed corpora since the ICD-11 pseudo-extractive documents do not have a ground truth summary. For lexical (ROUGE) evaluation, we report ROUGE F1 scores [72]. Similarly, here for evaluation conducted w.r.t source articles, the results reported are average results across the PubMed and ICD-11-Summ-1000 corpora. Entity-level factual accuracy [37] is measured in terms of precision, and recall w.r.t ground truth summary (for PubMed), and w.r.t source articles (for both PubMed and ICD-11-Summ-1000). Entity-level precision and recall w.r.t ground truth summaries are denoted with precision-source and recall-source, respectively. The F1 score is the harmonic mean of the precision and recall for either case. For measuring semantic equivalence between generated summaries and ground truth summaries, we leverage BERTScore as proposed by [39]; specifically, we use BioBERT for representing each token in a generated summary and the ground truth summary after which we perform pairwise cosine similarity as proposed in [39]. All experimental results are reported in percentages. The average full text length of input source articles in PubMed-50k is 3,224 words and the average abstract length is 218 words, while for ICD-11-Summ-1000, the average length of an extractive pseudo-doc (i.e., input source article) is 4816 words.

VII. ABSTRACTION STUDIES

To assess the impact of facts mined on the quality of summaries generated, we conduct an ablation study where we experiment with different values of $K$ in top-k for the backbone models. Figure-4 shows results of ablation to assess precision-source and recall-target. Since we want to minimize entity hallucination which is measured in terms of precision-source and want to maximize the number of entities in the ground truth summary that are retrieved in the generated summary as measured by recall-target, we report the impact of different values of $K$ for these two metrics. As shown in the two plots, precision-source and recall-target consistently improve as we increase $K$.

| Backbone Model | Training Config (k=n) | B1 | B2 | B3 |
|----------------|-----------------------|----|----|----|
| T5             | T5 w/o named entities (Baseline) | 26.078 | 26.385 | 26.935 |
|                | T5 w/ named entities (Baseline) | 26.591 | 26.892 | 27.492 |
| BART           | BART w/o named entities (Baseline) | 26.577 | 26.874 | 27.474 |
|                | BART w/ named entities (Baseline) | 27.241 | 27.549 | 28.149 |
| Pegasus        | Pegasus Vanilla (Baseline) | 26.551 | 26.852 | 27.452 |
|                | Pegasus w/ named entities (Baseline) | 27.207 | 27.515 | 28.115 |
| BigBird        | BigBird Vanilla (Baseline) | 30.872 | 31.283 | 31.883 |
|                | BigBird w/ named entities (Baseline) | 31.426 | 31.838 | 32.438 |
| LED            | LED Vanilla (Baseline) | 29.196 | 29.518 | 30.118 |
|                | LED w/ named entities (Baseline) | 30.129 | 30.521 | 31.121 |

TABLE II: Lexical (ROUGE) Evaluation w.r.t Ground Truth Summary (vanilla input @ inference time). The input in this experimental setting is the raw input article to be summarized (i.e., w/o named entity chain)

| Backbone Model | Training Config (k=n) | B1 | B2 | B3 |
|----------------|-----------------------|----|----|----|
| T5             | T5 w/o named entities (Baseline) | 9.583 | 7.743 | 11.228 |
|                | T5 w/ named entities (Baseline) | 8.801 | 7.232 | 9.848 |
| BART           | BART w/o named entities (Baseline) | 9.583 | 7.743 | 11.228 |
|                | BART w/ named entities (Baseline) | 8.801 | 7.232 | 9.848 |
| Pegasus        | Pegasus Vanilla (Baseline) | 4.739 | 8.263 | 11.581 |
|                | Pegasus w/ named entities (Baseline) | 6.275 | 9.976 | 13.288 |
| BigBird        | BigBird Vanilla (Baseline) | 6.821 | 34.288 | 6.275 |
|                | BigBird w/ named entities (Baseline) | 8.263 | 11.581 | 6.275 |
| LED            | LED Vanilla (Baseline) | 11.733 | 32.537 | 34.214 |
|                | LED w/ named entities (Baseline) | 11.352 | 33.734 | 35.426 |

TABLE III: Entity-level Factual Consistency Evaluation w.r.t Ground Truth Summary (vanilla input @ inference time). The input in this experimental setting is the raw input article to be summarized (i.e., w/o named entity chain)

| Backbone Model | Training Config (k=n) | B1 | B2 | B3 |
|----------------|-----------------------|----|----|----|
| T5             | T5 w/o named entities (Baseline) | 48.387 | 46.757 | 51.593 |
|                | T5 w/ named entities (Baseline) | 46.851 | 45.265 | 50.452 |
| BART           | BART w/o named entities (Baseline) | 28.255 | 28.986 | 29.715 |
|                | BART w/ named entities (Baseline) | 27.241 | 27.972 | 28.703 |
| Pegasus        | Pegasus Vanilla (Baseline) | 27.207 | 27.972 | 28.703 |
|                | Pegasus w/ named entities (Baseline) | 26.827 | 26.601 | 27.377 |
| BigBird        | BigBird Vanilla (Baseline) | 16.533 | 17.388 | 18.243 |
|                | BigBird w/ named entities (Baseline) | 15.934 | 16.789 | 17.644 |
| LED            | LED Vanilla (Baseline) | 26.198 | 25.957 | 26.857 |
|                | LED w/ named entities (Baseline) | 25.938 | 25.697 | 26.557 |

TABLE IV: Entity-level Factual Consistency Evaluation w.r.t Ground Truth Summary (vanilla input @ inference time). The input in this experimental setting is the raw input article to be summarized (i.e., w/o named entity chain).
TABLE V: Lexical (ROUGE) Evaluation w.r.t Ground Truth Summary (input article + named entity chain @ inference time), i.e., the input in this experimental setting is the raw input article to be summarized with the named entities (i.e., w/ named entity chain).

| Backbone Model | Training Config (w/o) | BLEU-0.5 | BLEU-1.5 | BLEU-2.5 |
|----------------|-----------------------|----------|----------|----------|
| TS             | TS Vanilla (Baseline) | 26.802   | 25.472   | 24.386   |
| BART           | BART Vanilla (Baseline) | 28.342   | 27.135   | 26.173   |
| Pegasus        | Pegasus Vanilla (Baseline) | 27.592   | 26.382   | 25.367   |
| BigBird        | BigBird Vanilla (Baseline) | 25.615   | 24.402   | 23.415   |
| LED            | LED Vanilla (Baseline) | 30.800   | 29.714   | 28.726   |

TABLE VI: Entity-level Factual Consistency w.r.t Ground Truth Summary. The input in this experimental setting is the raw input article to be summarized with the named entities (i.e., w/ named entity chain) @ inference time.

| Backbone Model | Training Config (w/o) | F1 score-source | F1 score-target |
|----------------|-----------------------|-----------------|----------------|
| TS             | TS Vanilla (Baseline) | 63.722           | 65.742          |
| BART           | BART Vanilla (Baseline) | 57.360           | 59.372          |
| Pegasus        | Pegasus Vanilla (Baseline) | 57.282           | 59.294          |
| BigBird        | BigBird Vanilla (Baseline) | 57.195           | 59.207          |
| LED            | LED Vanilla (Baseline) | 58.316           | 60.328          |

TABLE VII: Entity-level Factual Consistency w.r.t source article (input article + named entity chain @ inference time), i.e., the input in this experimental setting is the raw input article to be summarized with the named entities (i.e., w/ named entity chain).

| Backbone Model | Training Config (w/o) | F1 score-source | F1 score-target |
|----------------|-----------------------|-----------------|----------------|
| TS             | TS Vanilla (Baseline) | 63.722           | 65.742          |
| BART           | BART Vanilla (Baseline) | 57.360           | 59.372          |
| Pegasus        | Pegasus Vanilla (Baseline) | 57.282           | 59.294          |
| BigBird        | BigBird Vanilla (Baseline) | 57.195           | 59.207          |
| LED            | LED Vanilla (Baseline) | 58.316           | 60.328          |

TABLE VIII: N-gram Novelty w.r.t source articles w/o and w/ named entity chain during inference.

| Backbone Model | Training Config (w/o) | N-gram Novelty |
|----------------|-----------------------|----------------|
| TS             | TS Vanilla (Baseline) | 34.728          |
| BART           | BART Vanilla (Baseline) | 33.842          |
| Pegasus        | Pegasus Vanilla (Baseline) | 33.731          |
| BigBird        | BigBird Vanilla (Baseline) | 33.629          |
| LED            | LED Vanilla (Baseline) | 34.863          |

IV. RESULTS

As can be seen, retrieve more relevant facts from the biomedical knowledge bases and train our models.

VIII. DISCUSSION OF RESULTS

From the results reported in the previous section, we generally see entity-level factual consistency (particularly, precision-source, and recall-target) improve when a model is trained with named entities and/or facts included as an additional signal in the training with the same objective of generating the ground truth summary using cross-entropy loss. The addition of more facts further improves entity-level factual consistency as shown in Figure-4. Further, we notice N-gram novelty improves with our proposed framework for the five backbone models. Semantic equivalence generally improves when named entities and/or facts are included during training for all models. Thus, the corresponding entries for the various models and training configurations show improvement in semantic based scores. The ROUGE scores, however, drop slightly from when there is no additional context at training or inference time. The drop in ROUGE is a result of augmenting the models with facts from background knowledge bases which in turn leads to higher N-gram novelty. Thus, the proposed framework enables us to achieve better abstractive scores in terms of entity-level factual consistency, paraphrasing and semantic equivalence.

IX. CONCLUSIONS AND FUTURE WORK

In this study, we proposed a framework to integrate named entities in a source article and facts extracted from biomedical knowledge bases pertaining to the named entities in transformer-based encoder-decoder models and applied to the task of abstractive summarization of biomedical literature. Through extensive experiments, we showed the proposed approach improves the semantics of generated summaries in terms of entity-level factual consistency and semantic equivalence while generating novel words. For future steps, we plan to jointly train the knowledge-retriever and the knowledge-guided abstractive summarizer in an end-to-end fashion. While our current architecture optimizes a single cost function given different input signals, we plan to augment the existing framework using multi-objective optimization to further enhance the factual accuracy and semantic equivalence of generated summaries using different cost functions during training.
G. Erkan and D. R. Radev, “Lexrank: Graph-based lexical centrality as salience in text summarization,” Journal of artificial intelligence research, vol. 22, pp. 457–479, 2004.

C. Zhu, W. Hinthorn, R. Xu, Q. Zeng, M. Zeng, X. Huang, and M. Jiang, “Enhancing factual consistency of abstractive summarization,” arXiv preprint arXiv:2003.08612, 2020.

G. Manas, V. Aribandi, U. Kursuncu, A. Alambo, V. L. Shalin, K. Filippova, “Multi-sentence compression: Finding shortest paths in word graphs,” in Proceedings of the 23rd international conference on computational linguistics (Coling 2010), 2010, pp. 322–330.

A. Schrijver, Theory of linear and integer programming. John Wiley & Sons, 1998.

M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, “Clinical context–theory of linear and integer programming,” arXiv preprint arXiv:1411.4166, 2014.

M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, “AllenNLP: A deep semantic natural language processing platform,” arXiv preprint arXiv:1803.07640, 2018.

S. Banerjee, P. Mitra, and K. Sugiyama, “Multi-document abstractive summarization using ilp based multi-sentence compression,” in Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

L. Jean-Baptiste, “Using medical terminologies with pymedtermino and umls,” in Ontologies with Python. Springer, 2021, pp. 207–239.

A. Shrivastava and P. Li, “Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips),” arXiv preprint arXiv:1405.5869, 2014.

G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” arXiv preprint arXiv:2007.01282, 2020.

D. Singh, S. Reddy, W. Hamilton, C. Dyer, and D. Yogatama, “End-to-end training of multi-document reader and retriever for open-domain question answering,” Advances in Neural Information Processing Systems, vol. 34, 2021.

J. M. Prager, “Open-domain question-answering.” Found. Trends Inf. Retr., vol. 1, no. 2, pp. 91–231, 2006.

J. Zhang, J. Tan, and X. Wan, “Towards a neural network approach to abstractive multi-document summarization,” arXiv preprint arXiv:1804.09610, 2018.

J. Zhang, J. Tan, and X. Wan, “Towards a neural network approach to abstractive multi-document summarization,” arXiv preprint arXiv:1804.09610, 2018.

Y. Gao, W. Zhao, and S. Eger, “Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization,” arXiv preprint arXiv:2005.03724, 2020.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer, “End-to-end neural coreference resolution,” arXiv preprint arXiv:1707.07045, 2017.

M. Gardiner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, “AllenNLP: A deep semantic natural language processing platform,” arXiv preprint arXiv:1803.07640, 2018.

S. Banerjee, P. Mitra, and K. Sugiyama, “Multi-document abstractive summarization using ilp based multi-sentence compression,” in Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.