Nutri-bullets: Summarizing Health Studies by Composing Segments

Darsh J Shah,[1] Lili Yu,[2] Tao Lei,[2] Regina Barzilay[1]
[1]Computer Science and Artificial Intelligence Lab, MIT
[2]ASAPP, Inc.
darsh@csail.mit.edu, liliyu@asapp.com, tao@asapp.com, regina@csail.mit.edu

Abstract

We introduce *Nutri-bullets*, a multi-document summarization task for health and nutrition. First, we present two datasets of food and health summaries from multiple scientific studies. Furthermore, we propose a novel *extract-compose* model to solve the problem in the regime of limited parallel data. We explicitly select key spans from several abstracts using a policy network, followed by composing the selected spans to present a summary via a task specific language model. Compared to state-of-the-art methods, our approach leads to more faithful, relevant and diverse summarization – properties imperative to this application. For instance, on the BreastCancer dataset our approach gets a more than 50% improvement on relevance and faithfulness.[1]

1 Introduction

Multi-document summarization is essential in domains like health and nutrition, where new studies are continuously reported (see Figure 1). Websites like *Healthline.com*[2] are critical in making food and nutrition summaries available to web users and subscribers. Such summaries provide a collective view of information – a crucial property, especially when the consumption of such knowledge leads to health related decisions. In the current environment, it is critical to provide users with the the latest health related findings. Unfortunately, summarization is time consuming and requires domain experts.

Through the introduction of *Nutri-bullets*, a multi-document summarization task, we aim to automate this summarization and expand the availability and coverage of health and nutrition information.

Recently, summarization has been predominantly solved by training sequence-to-sequence (seq2seq) models (Rush, Chopra, and Weston 2015; Vaswani et al. 2017; Hoang et al. 2019; Lewis et al. 2019). In the multi-document setting, popular seq2seq methods either concatenate all input documents as a single source or consider a hierarchical setting (Liu et al. 2018). While such methods generate fluent text, in our case, they fail to produce content faithful to the inputs (examples shown in Figure 2). Two factors make *Nutri-bullets* particularly challenging for seq2seq models: (i) The concatenation of health documents constitutes long sequences with key information being scattered making composing a good summary extremely difficult; and (ii) While a vast number of scientific abstracts are available in libraries such as Pubmed[3], a very limited number of summaries, ranging from several hundreds to a few thousands, are available to train a summarization system.

In this paper, we present a novel and practical approach which learns to *extract and compose* knowledge pieces from scientific abstracts into a summary. We side-step the scarcity of parallel data by focusing on knowledge-extraction from scientific abstracts, made possible by modules trained on such crowd-sourced annotations. Furthermore, our model learns to select a subset of knowledge pieces well-suited for a summary, from a large pool. The selection is performed through a multi-step framework, guided by reinforcement learning and distant supervision from limited available par-

> **Two recent meta-analyses reported an inverse association between dietary fiber and whole grain intake and the risk of colorectal cancer. Evidence from previous reviews is supportive of the hypothesis that whole grains may protect against various cancers. Melatonin enhances the apoptotic effects and modulates the changes in gene expression induced by docetaxel in MCF human breast cancer cells. Results from clinical trials and multiple in vivo and in vitro studies point to melatonin as a promising adjuvant molecule with many beneficial effects when concomitantly administered with chemotherapy . Oats contains selenoproteins ; it has been suggested that several selenoproteins may be protective against cancer.**

Figure 1: Our model's *extract-compose* summary describing impacts of whole grains on cancer. Segments from multiple scientific studies are composed to present the output.

- **Apples are high in protein and fiber, which may promote weight loss.**
- **Cucumber is a good source of essential nutrients, including calcium, magnesium, potassium and omega-3 fats .**
- **Sweet potatoes are antioxidant that low in protein, fiber, in fats, studies, and fat. Eating these nutrients that your body from damage.**

Figure 2: Examples of unfaithful summaries generated by seq2seq models trained on the HealthLine dataset. Such fictitious texts (red) can be extremely misleading for readers.

[1]Our code and data is submitted and will be made publicly available on acceptance.

[2]https://healthline.com

[3]https://pubmed.ncbi.nlm.nih.gov/

allel data.

While considering the extracted text as inputs to a seq2seq model already makes the learning-to-generate process easier, we further explore an alternative approach to compose the extracted text into a complete summary using a text infilling language model. Once key spans are selected from different documents, they are fused using a task specific language model (Shen et al. 2020) as illustrated in Figure 3. The fusion of separate yet related spans through generation is critical in producing meaningful and readable summaries.[4]

We conduct human and empirical evaluation to comprehensively study the applicability and quality of these approaches. While seq2seq models enjoy high fluency scores, the *extract-compose* method performs much stronger on metrics such as content, relevance, faithfulness and informativeness. Our method is particularly dominant in scenarios with scarce parallel data, since our model requires little summary data for training. For instance, on the BreastCancer dataset, humans rate the *extract-compose* model's summaries **more than 50% higher** for relevance and faithfulness than the next best baseline. Comparison with strong baselines and model ablation variants highlights the necessity of a distant supervision and reinforcement learning based multi-step approach, in selecting key ordered spans amongst several possible combinations, for text composition. Our contributions are threefold:

(i) We collect two new nutrition and health related datasets for multi-document summarization. We also collect large-scale knowledge extraction annotations, applicable to numerous tasks.

(ii) We demonstrate the effectiveness of our modelling approach for generating health summaries given limited parallel data. Our approach strongly outperforms all baselines and variants on human and automatic evaluation.

(iii) We conduct comprehensive (human and automatic) evaluation focusing on content relevance, faithfulness and informativeness – metrics more relevant to the task. These set new benchmarks to critically evaluate summaries in high impact domains.

2 Related Work

Multi-document Summarization Approaches in neural sequence-to-sequence learning (Rush, Chopra, and Weston 2015; Cheng and Lapata 2016; See, Liu, and Manning 2017) for document summarization have shown promise and have been adapted successfully for multi-document summarization (Zhang, Tan, and Wan 2018; Lebanoff, Song, and Liu 2018; Baumel, Eyal, and Elhadad 2018; Amplayo and Lapata 2019; Fabbri et al. 2019). Trained on large amounts of data, these methods have improved upon traditional extractive (Carbonell and Goldstein 1998; Radev and McKeown 1998; Haghighi and Vanderwende 2009) and abstractive approaches (Barzilay, McKeown, and Elhadad 1999; McKeown and Radev 1995; Ganesan, Zhai, and Han 2010). Despite producing fluent text, these techniques also tend to generate false information which is not faithful to the original inputs (Puduppully, Dong, and Lapata 2019; Kryściński et al. 2019). Side-information, such as citations in scientific domains (Qazvinian and Radev 2008; Qazvinian et al. 2013) or semantic representations (Liu et al. 2015), can be used to improve this (Sharma et al. 2019; Wenbo et al. 2019; Puduppully, Dong, and Lapata 2019; Koncel-Kedziorski et al. 2019a). However, such methods struggle in low resource scenarios. In this work, we are interested in producing faithful and fluent text in a technical domain where few parallel examples are available.

Text Fusion Traditionally, sentence fusion approaches (Barzilay and McKeown 2005) aid the concatenation of different text fragments for summarization. Recent language modeling approaches like Devlin et al. (2018); Stern et al. (2019) can also be extended for completion and fusion of partial text. These models have more flexibility than those trained on text fusion datasets (Narayan et al. 2017; Geva et al. 2019) that can combine two fragments only. In this work, we modify the Blank Language Model (Shen et al. 2020) to combine fragments coming from different source documents.

Deep Reinforcement Learning for Text Summarization The inherent discrete and sequential nature of text generation tasks has made optimizing generation models with reinforcement learning (Sutton and Barto 2018) very popular (Keneshloo et al. 2019). Typically, automatic evaluation metrics, like BLEU (Papineni et al. 2002) or ROUGE (Lin 2004), are used to provide the reward signal to train reinforcement learning algorithms in translation (Wu et al. 2018), summarization (Paulus, Xiong, and Socher 2017; Pasunuru and Bansal 2018; Chen and Bansal 2018; Xu et al. 2019) and question generation (Zhang and Bansal 2019) tasks. In this work, we are interested in using reinforcement learning to iteratively, select an explicit *ordered* subset of text phrases for downstream fusion.

3 Method

In this section, we describe the framework of our *extract-compose* solution for *nutri-bullets*. Our goal is to produce a text summary y for a food item from a pool of multiple scientific abstracts X.

3.1 Extract-compose Framework

We attain food health entity-entity relations, for both input documents X and the summary y, from entity extraction and relation classification modules trained on corresponding annotations(Table 1).

For input documents, we collect $\{(x_p, \mathcal{G}_p)_{p=1}^{N}\}$, where x_p is the sentence text and $\mathcal{G}_p$ is the set of entity-entity relations, in N sentences and $X = \{x_p\}$ is the raw text. $\mathcal{G}_p = \{(e_i^k, e_j^k, r^k)^K\}$ is composed of $K \in \{0, 1, 2, \dots\}$ tuples of two entities e_i, e_j and their relation r. r represents relations such as the effect of a nutrition entity e_i on a condition e_j (see Table 1). [5]

[4] Our analysis shows that 91% of all produced tokens are from extracted components. The generated words allow fusion and cohesion.

[5] We train an entity tagger and relation classifier to predict $\mathcal{G}$ and also for computing knowledge based evaluation scores.

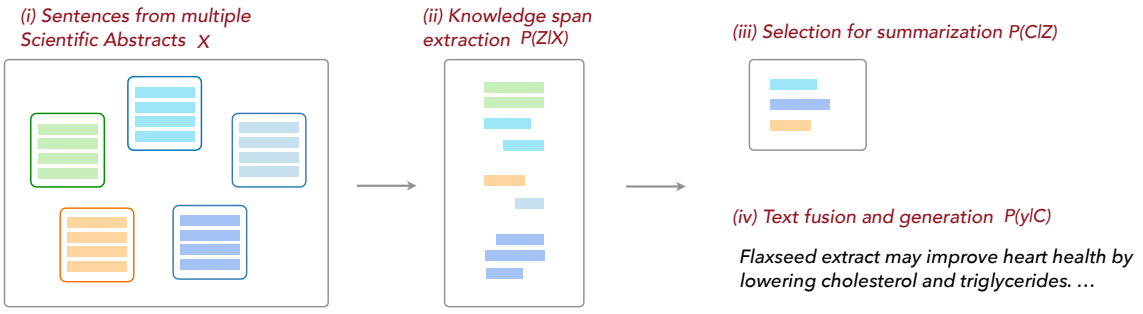

Figure 3: Illustrating the flow of our extract-compose system. (i) We begin with multiple scientific abstracts to summarize from; (ii) We extract knowledge spans as possible candidates for generation; (iii) We select key spans ("improve heart health by lowering cholestrol and triglycerides" in this case) from all possible candidates and (iv) Present a summary sentence using the spans and a domain specific language model.

Similarly, we denote the summary data as $\{(y_q, \mathcal{G}_q)_{q=1}^{M}\}$, where y_q is a concise summary and $\mathcal{G}_q$ is the set of entity-entity relation tuples, in M summaries.

Joint learning of content selection, information aggregation and text generation for multi-document summarization can be challenging. This is exacerbated in our technical domain with few parallel examples. We propose to overcome these challenges by first learning to extract and then composing the informative text pieces.

We model two intermediate variables, distilled text spans Z associated with all entity-entity relation information $\mathcal{G}$, and the key content C selected from them.

The probability of an output summary y, conditioned on the input X is

$$P(y|X) = \sum_{C,Z} P(y|C)P(C|Z)P(Z|X) \quad (1)$$

Where: (i) $P(Z|X)$ models a knowledge extraction process gathering all entity-entity relations and corresponding text spans, $\mathcal{G}$; (ii) $P(C|Z)$ models the process of selecting an important subset from all extracted text spans and (iii) $P(y|C)$ models the text fusion process that composes fluent text incorporating the extracted content.

3.2 Span Extraction $P(Z|X)$

We model $P(Z|X)$ as a span extraction task leveraging a rationale extraction system (Lei, Barzilay, and Jaakkola 2016). The extraction model picks the fewest words from the input necessary to make the correct relation label prediction for an entity pair. Let (e_i, e_j, r) be one entity relationship tuple, x be the associated sentence text and z be the considered rationale text span. $P(Z|X)$ is trained to minimize the loss:

$$\mathcal{L} = \mathcal{L}(z, r) + \mathcal{L}(z) + \mathcal{L}(z, e_i, e_j) \quad (2)$$

. Where $\mathcal{L}(z, r)$ is the cross entropy loss for predicting r with z as the extracted text. $\mathcal{L}(z)$ is a regularization term to select short and coherent text, by minimizing the span lengths and discontinuities among the spans. In addition to prior work, we introduced $\mathcal{L}(z, e_i, e_j)$ to encourage the selection of phrases that contain entities e_i and e_j. Specifically, we construct verb phrases (from constituency parse trees) containing the condition entity, and minimize the distance between the selected text span z and the verb phrase. Empirically, this loss stabilizes the span extraction and improves the quality of selected text spans, using r labels as indirect supervision as in Shah, Schuster, and Barzilay (2019). By running the extraction model on every (e_i, e_j, r) tuple, we distill the input text into a set of text spans $Z = \{z_1, z_2, \ldots, z_m\}$.

3.3 Content Selection Policy Network $P(C|Z)$

$P(C|Z)$ is a policy network, that takes a large set of text spans Z as input, and outputs C, an *ordered* set of key text spans. We model our content selection as a finite Markov decision process (MDP). Where the state is represented as $s_t = (t, \{c_1, \ldots, c_t\}, \{z_1, z_2, ..., z_{m-t}\})$ for step t, content selected so far $\{c_1, \ldots, c_t\}$ and remaining text spans $\{z_1, z_2, ..., z_{m-t}\}$. Our action space is all the remaining text spans plus one special token, $Z \cup \{STOP\}$. The number of actions is $|m - t| + 1$. We parameterize the policy $\pi_\theta(a|s_t)$ with a neural network to map the state s to a probability distribution over all available actions. At each step, the probability that the policy selects z_i as a candidate is:

$$\pi_\theta(a = z_i|s_t) = \frac{\exp(f(t, \hat{z_i}, \hat{c_i*}))}{\sum_{j=1}^{m-t+1} \exp(f(t, \hat{z_j}, \hat{c_j*}))} \quad (3)$$

where $c_i* = \arg\max_{c_j}(cos(\hat{z_i}, \hat{c_j}))$ is the selected content closest to z_i, $\hat{z_i}$ and $\hat{c_i*}$ are the encoded dense vectors, $cos(u, v) = \frac{u \cdot v}{||u|| \cdot ||v||}$ and f is a feed-forward neural network that outputs a scalar score. The selection process begins from Z. Our module iteratively samples actions from $\pi_\theta(a|s_t)$. On picking *STOP*, we end with the selected content C and a corresponding reward. Our rewards guide the selection of informative, diverse and readable phrases:

- $\mathcal{R}_e = \sum_{c \in C} cos(\hat{e_{ic}}, \hat{e_{iy}}) + cos(\hat{e_{jc}}, \hat{e_{jy}})$ is the cosine similarity of the structures of the selected content C with the structures present in the gold summary y (each gold summary structure accounted with only one c), encouraging the model to select relevant content.

- $\mathcal{R}_d = \mathbb{1}[\max_{i,j}(cos(\hat{c_j}, \hat{c_i})) < \delta]$ computes the similarity between pairs within selected content C, encouraging the selection of diverse spans.
- $\mathcal{R}_s = \sum_c \mathbb{1}[\sum_{q=1}^{M} cos(\hat{c}, \hat{y_q}) > \sum_{q=1}^{M} cos(\hat{z'}, \hat{y_q})]$, where z' is a text span randomly sampled from scientific abstract corpus, predicts if c is closer to a human summary than a random abstract sentence, encouraging the selection of concise spans and ignoring numerical details.
- $\mathcal{R}_m$ is the Meteor (Denkowski and Lavie 2014) score between the final selection with the gold summary, a signal which can only be provided at the completion of the episode.
- r_p is a small penalty for each action step.

The final multi-objective reward is computed as

$$\mathcal{R} = w_e\mathcal{R}_e + w_d\mathcal{R}_d + w_s\mathcal{R}_s + w_m\mathcal{R}_m - |C|r_p, \quad (4)$$

where, w_e, w_d, w_s, w_m and r_p are hyper-parameters. During training, the network is updated with these rewards. Our paradigm allows an exploration of different span combinations while incorporating delayed feedback.

3.4 Text Fusion $P(y|C)$

The text fusion module, $P(y|C)$, composes a complete summary using the text spans C selected by the policy network.

We propose to utilize the recently developed Blank Language Model (BLM) (Shen et al. 2020), which fills in the *blanks* by iteratively determining the word to place in a *blank* or adding a new *blank*, until all *blanks* are filled. The model is trained on the WikiText-103 dataset (Merity et al. 2016).

We extend this model with additional categorical *blanks* between different text spans in C, according to their relation type. This ensures control over facts and relations captured from the scientific abstracts X to present a semantically valid and fluent summary (details see Appendix).

In the Transformer variant of the model, we train a seq2seq $P(y|C)$ using the limited parallel data.

4 Data

In this section, we describe the dataset collected for our *Nutri-bullet* system.

4.1 Corpus Collection

Our Healthline[6] and BreastCancer[7] datasets consist of scientific abstracts as inputs and human written summaries as outputs.

Scientific Abstracts We collect 7750 scientific abstracts from Pubmed and the ScienceDirect, each averaging 327 words. The studies in these abstracts are cited by domain experts when writing summaries in the Healthline and BreastCancer datasets. A particular food and its associated abstracts are fed as inputs to our *Nutri-bullet* systems. We exploit the large scientific abstract corpus when gathering entity, relation and sentiment annotations (see Table 3) to overcome the challenge of limited parallel examples. Modules trained on these annotations can be applied to any food health scientific abstract.

Summaries Domain experts curate summaries for a general audience in the Healthline and BreastCancer datasets. These summaries describe nutrition and health benefits of a specific food (examples shown in Appendix). In the HealthLine dataset, each food has multiple bullet summaries, where each bullet typically talks about a different health impact (hydration, anti-diabetic etc). In BreastCancer, each food has a single summary, describing in great detail its impact on breast and other cancers.

Parallel Instances The references in the human written summaries form natural pairings with the scientific abstracts. We harness this to collect 1894 parallel (abstracts, summary) instances in HealthLine, and 141 parallel instances in BreastCancer (see Table 2). Summaries in HealthLine average 24.46 words, created using an average of 3 articles. Summaries in BreastCancer have an average length of 155.71 words referencing an average of 18 articles. Unsurprisingly, BreastCancer is the more challenging of the two.

4.2 Entity, Relation and Sentiment Annotations

Despite having a small parallel data compared to Hermann et al. (2015); Narayan, Cohen, and Lapata (2018), we conduct large-scale crowd-sourcing tasks to collect entity, relation and sentiment annotations on Amazon Mechanical Turk. The annotations (see Table 3) are designed to capture the rich technical information ingrained in such domains, alleviating the difficulty of multi-document summarization and are broadly applicable to different systems (Koncel-Kedziorski et al. 2019b).

Entity and Relation Annotations Workers identify *food*, *nutrition*, *condition* and *population* entities by highlighting the corresponding text spans.

Given the annotated entities in text, workers are asked to enumerate all the valid relation tuples (e_i, e_j, r). Table 1 lists possible combinations of e_i, e_j and r for each relation type, along with some examples.

The technical information present in our domain can make annotating challenging. To collect reliable annotations, we set up several rounds of qualification tasks [8], offer direct communication channels to answer annotators' questions and take majority vote among 3 annotators for each data point. As shown in Table 3, we collected 91K entities, 34K pairs of relations, and 7K sentiments[9]. The high value of mean Cohen's κ highlights high annotator agreement for all the tasks, despite various challenges.

Food Health Sentiment Additionally, we also collect food sentiment annotations for evaluating our system. Food health sentiment (*positive*, *negative*, *neutral*) indicates the

[6]https://www.healthline.com/nutrition

[7]https://foodforbreastcancer.com/

[8]To set up the qualification, the authors first annotate tens of examples which serve as gold answers. We leverage Mturk APIs to grade the annotation by comparing with the gold answers.

[9]For BreastCancer, we use the entity tagger and relation classifier finetuned on scientific abstracts and HealthLine datasets to extract the entities and relations.

Relation Type	e_i	e_j	r	**Example**
Containing	Food, Nutrition	Nutrition	Contain	(apple, fiber, contain)
Causing	Food, Nutrition, Condition	Condition	Increase, Decrease, Satisfy, Control	(bananas, metabolism, increase), (orange juice, hydration, satisfy)

Table 1: Details of entity-entity relationships that we study and some tuple examples.

Data	**Train**	**Dev**	**Test**
Input Scientific Abstracts	6110	750	866
Average words	327.7	323.3	332.3
HealthLine summaries	1522	179	193
Average words	24.7	23.3	23.9
Abstracts Per Instance	3.19	2.82	3.43
BreastCancer summaries	104	18	19
Average words	161.4	131.6	148.1
Abstracts Per Instance	18.90	17.21	17.67

Table 2: Statistics for scientific abstracts, HealthLine and BreastCancer datasets.

Data	**Task**	**# annotations**	**mean κ**
Scientific Abstracts	entity	83543	0.75
	relation	28088	0.79, 0.81
	sentiment	5000	0.65
HealthLine	entity	7860	0.86
	relation	5974	0.73, 0.90
	sentiment	2000	0.89

Table 3: Entity, relation and sentiment annotation statistics. Each annotation is from three annotators. Mean κ is the mean pairwise Cohen's κ score.

kind of impact the food has on human health. The annotation is performed at a sentence level, and modules trained on this data are used to assess the contrastiveness in a food's summary bullets.

Annotation interfaces, instructions as well as more data details can be found in Appendix.

5 Experimental Setup

Datasets We randomly split both HealthLine and BreastCancer datasets into training, development and testing sets(see Table 2).

Evaluation The subjective nature of summarization demands human judgements to comprehensively evaluate model performance. We have human annotators score our models on faithfulness, relevance, fluency and informativeness. Given input scientific abstracts, *faithfulness* characterizes if the output text is consistent with the input. Given a reference summary, *relevance* indicates if the output text shares similar information. *Fluency* represents if the output text is grammatically correct and written in well-formed English. *Informativeness* characterizes the degree of health related knowledge conveyed by the summaries.[10]

Annotators rate faithfulness, relevance and fluency on a 1-4 scale likert score, which is commonly used as a standard psychometric scale to measure responses in social science and attitude research projects (Croasmun and Ostrom 2011; Li 2013; Sullivan and Artino Jr 2013). For rating informativeness, we perform a one-vs-one evaluation between our model and the strong Transformer baseline (Hoang et al. 2019). We have 3 annotators score every data point and take an average across the scores.

We further evaluate our systems using the following automatic metrics. *Meteor* is an automatic metric used to compare the model output with the gold reference. *KG(G)* computes the number of entity-entity pairs with a relation in the gold reference, that are present in the output.[11] This captures relevance in context of the reference. *KG(I)*, similarly, computes the number of entity-entity pairs in the output that are present in the input scientific abstracts. This measures faithfulness with respect to the input documents. *Diversity* calculates the proportion of unique trigrams in the outputs (Li et al. 2016; Rao and Daumé III 2019). *Contrastiveness* calculates the proportion of article summaries belonging to the same food that contain both positive and negative/neutral sentiment about a food's health impact.

Baselines We compare our method against several state-of-the-art methods.

- Copy-gen: See, Liu, and Manning (2017) is a top performing technique for summarization, which can copy words from the input or generate new words.
- Transformer: Hoang et al. (2019) is a system that utilizes a pretrained Transformer for summarization.
- GraphWriter: Koncel-Kedziorski et al. (2019b) is a graph transformer based model, which generates text using a seed title and a knowledge graph. This model utilizes all the extraction capabilities used by our $P(Z|X)$ module.

In addition, we study variants of our model (1) replacing BLM with a transformer text fusing model – Ours (Transformer), (2) replacing the policy gradient selector with a fully supervised selector trained on knowledge structure information (Select w/ Sup.), and (3) different reward combinations. Further, we compare our system with a BERT-based Extractive Summarization system.

Implementation Details We adapt the implementation of span extraction from Bao et al. (2018). Our policy network

[10]On the HealthLine datasets, each food article contains multiple bullet summaries. We group these bullet summaries per food for annotator comparison. For BreastCancer, a single summary output is a food's complete summary.

[11]We run entity tagging plus relation classification on top of the model output and gold summaries. We match the gold (e_i^g, e_j^g, r^g) tuples using word embedding based cosine similarity with the corresponding entities in the output structures (e_i^o, e_j^o, r^o). If the cosine score exceeds a threshold of 0.7, a match is found.

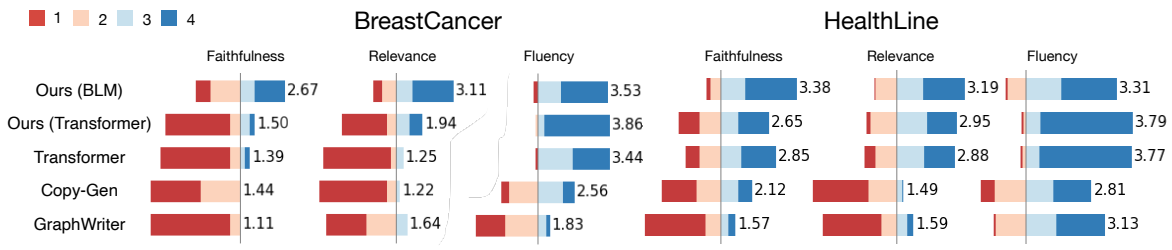

Figure 4: Human evaluation Gantt charts showing ratings on faithfulness, relevance and fluency for both BreastCancer and HealthLine datasets, where Ours (BLM) – *extract-compose* model strongly outperforms all others. Each example is rated by three judges and the mean is reported.

	BreastCancer	HealthLine
Better	94 %	75 %
Worse	6 %	25 %

Table 4: Human evaluation on informativeness of Ours (BLM) when comparing to Transformer.

is a three layer feedforward neural network. We set w_e, w_d and w_s to 1 and w_m to 0.75. r_p is 0.02 and δ is 0.99. We use a large, 6 layer BLM (Shen et al. 2020) for fusion. Additional experimentation details can be found in the Appendix .

We train a Neural CRF tagger (Yang and Zhang 2018) for the food, condition and nutrition entity tagging. We use BERT (Devlin et al. 2018) text classifiers to predict the relation between two entities, and to predict food sentiments of a summary, trained using annotations from Table 3.

6 Results

In this section, we describe the performance of *Nutri-bullet* systems (baselines and models) on the HealthLine and BreastCancer datasets. We also report ablation studies, highlighting the importance of our policy network's ability to explore different span combinations, while incorporating long term rewards from distant supervision.

Human Evaluation Figure 4 and Table 4 report the human evaluation results of our model and baselines on the test set of both datasets.

Our method outperforms all other models by a large margin on faithfulness, owing to the extraction and controlled text generation process. Our method is also rated the highest on relevance (3.11 & 3.19), demonstrating its ability to select the important and relevant content consistently, even with little supervision. Our model produces fluent text.

On the contrary, despite being fluent, transformer models fail to generate faithful content. They achieve decent averaged relevance scores (≈2.9) on the HealthLine dataset. However, annotators rate them a *score 1* far more often,[12] due to hallucinated content and factually incorrect outputs.

Copy-gen and GraphWriter struggles to score high on either of our metrics, showing the challenges of learning meaningful summaries in such a low-resource and knowledge-rich domain.

Finally, we evaluate if our output text can benefit the reader and provide useful nutritional information about the food being described. In our human evaluation of informativeness, we outperform 94%-6% and 75%-25% against the Transformer baseline on BreastCancer and HealthLine respectively.

Automatic Evaluation Table 6 presents the automatic evaluation results on BreastCancer and HealthLine datasets.

High KG(I) (52% & 40%) and KG(G) (96% & 88%) scores for our method highlight that our produced text is faithful and relevant, consistent with human evaluation. Additionally, high diversity (91% & 90%) and contrastiveness (94% & 83%) scores indicate that our model is also able to present distinct (across sentiment and content) information for the same food.

In contrast, the sequence-to-sequence based methods tend to get a higher Meteor score with a lower diversity, suggesting that they repeatedly generate similar outputs, regardless of the input text. Low KG scores show that they fail to capture the relevant or faithful text, which is crucial in this domain. Among these, our transformer variation performs strong, especially on the BreastCancer dataset. This was also observed in Liu and Lapata (2019), where abstractive summarization fed with only a subset of the input sentences outperformed vanilla abstractive summarization.

Case Study Table 5 shows examples of various system outputs on HealthLine. The food name and titles of corresponding scientific abstracts are also presented. We observe that Transformer, GraphWriter and Copy-gen fail to generate meaningful summaries that are faithful to the original input text. With carefully selected knowledge, our (transformer) and our (BLM) both produce relevant and useful information. Ours (transformer) generates concise and easier language. Ours (BLM) composes faithful text with rich details, for example, it captures "anti-aging properties", which is relevant to "free radicals" in the gold summary. Detailed analysis shows that 91% of all produced tokens by our model are from the extracted segments – critical for faithful summarization. The words generated in between allow cohesive fusion. Examples from the BreastCancer dataset can be found in the Appendix. Additional model outputs with human evaluation scores can be found in Appendix.

$P(C|Z)$ **Variants** To further understand our model, we implement an alternative content selection method, using

[12] Annotators rate transformer with *score 1* 6 times and ours(transformer) 3 times more often than BLM model.

Food	banana
Pubmed Titles	(1) High content of dopamine a strong antioxidant, in Cavendish banana; (2) Flavonoid intake and risk of CVD: a systematic review and meta-analysis of prospective cohort studies; (3) Antioxidants in health, disease and aging; (4) Banana(Musa spp) from peel to pulp: ethnopharmacology, source of bioactive compounds and its relevance for human health.
Gold	*Bananas are high in several antioxidants, which may help reduce damage from free radicals and lower your risk of some diseases.*
Copy-gen	Seaweed contain several compounds that can reduce feelings of fullness and prevent fullness hormones, that are needed to health health benefits.
GraphWriter	The type of insoluble fiber found in banana banana may help reduce blood sugar levels and help lower blood pressure.
Transformer	The antioxidants in walnuts may help lower blood pressure in people with type 2 diabetes.
Ours (Trans.)	Bananas are a rich source of antioxidants, which may help reduce the risk of many chronic diseases.
Ours (BLM)	Bananas are used to help overcome or prevent a substantial number of illnesses , as depression and bananas containing antioxidants may lower the incidence of disease , such as certain cancers , cardiovascular and neurodegenerative diseases , DNA damage , or even have anti-aging properties .

Table 5: Example outputs for a HealthLine input.

	BreastCancer					HealthLine				
Model	Me	KG(G)	KG(I)	Di	Co	Me	KG(G)	KG(I)	Di	Co
Copy-gen	14.0	21	23	70	0†	7.4	21	51	82	43
GraphWriter	10.1	0 †	0†	16	0†	7.6	3	69	31	25
Transformer	13.0	31	11	76	66	10.2	21	67	53	28
Ours (Tranformer)	**15.0**	49	13	81	50	**10.3**	23	64	55	28
Ours (BLM)	13.8	**52**	**96**	**91**	**94**	8.7	**40**	**88**	**90**	**83**

Table 6: Meteor score (Me), KG in gold(G), KG in input(I), Diversity (Di) and Contrastiveness (Co) in our models and various baselines, on both BreastCancer and HealthLine datasets. †denotes cases which model generates meaningless results due to small training size. The best results are in bold and Ours(BLM) – *extract-compose* is the most dominant.

a supervised classification module (implementation details described in Appendix . Table 7 reports the results. Being an extract-compose variant, the supervised model (first row) produces faithful summaries (KG(I)). However, our Policy Network's joint selection and ability to explore span combinations with guidance from gold structure rewards and the Meteor score, lead to an improved performance on KG(G), Diversity and Meteor. Additionally, on human evaluated relevance, the Policy Network approach scores a higher 3.19 while the supervised extraction variant scores 2.5. We also observe the importance of $\mathcal{R}_m$ for KG(I) and Meteor, and $\mathcal{R}_d$ for Diversity.

Model	Me	KG(G)	KG(I)	Di	Co
Select w/ Sup.	7.6	28	84	75	**88**
Policy (w/o R_m, R_d)	8.4	36	83	70	**88**
Policy (w/o R_m)	8.2	33	**89**	**91**	**88**
Policy (full $\mathcal{R}$)	**8.7**	**40**	88	90	83

Table 7: Automatic evaluation of extract-compose model's variants on HealthLine. The best results are in **bold**.

Comparison with Extractive Summarization To understand the benefits of fine-span extraction followed by fusion, we compare our model with an extractive summarization system (implementation details in Appendix . Our model achieves 40% KG(G), while the extractive summarization system achieves 26%. Our *extract-compose* approach performs strongly since: (1) We extract knowledge pieces more precisely by selecting key spans instead of complex complete sentences from scientific abstracts; (2) RL model's content selection jointly through multi-steps, and (3) The Text Fusion module consolidates knowledge pieces, which may otherwise remain incomplete due to linguistic phenomena, such as coreference and discourse.

Even unsupervised methods like TextRank (Mihalcea and Tarau 2004) are not particularly applicable when we need to select key spans amongst multiple candidates. Instead our model is able to capture relevant, readable and coherent pieces of text by utilizing guidance from distant supervision and the use of a domain specific language model for fusion.

7 Conclusion

High impact datasets, content selection, faithful decoding and evaluation are open challenges in building multi-document health summarization systems. First, we propose two new datasets for *Nutri-bullets*. Next, we tackle this problem by exploiting annotations on the source side and formulating an extraction and composition method. Comprehensive human evaluation demonstrates the efficacy of our method in producing faithful, informative and relevant summaries.

References

Amplayo, R. K.; and Lapata, M. 2019. Informative and controllable opinion summarization. *arXiv preprint arXiv:1909.02322* .

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367* .

Barzilay, R.; McKeown, K.; and Elhadad, M. 1999. Information fusion in the context of multi-document summarization. In *ACL*, 550–557.

Barzilay, R.; and McKeown, K. R. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics* 31(3): 297–328.

Baumel, T.; Eyal, M.; and Elhadad, M. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704* .

Carbonell, J.; and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335–336.

Chen, Y.-C.; and Bansal, M. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *ACL*, 675–686. Melbourne, Australia: ACL. doi:10.18653/v1/P18-1063.

Cheng, J.; and Lapata, M. 2016. Neural Summarization by Extracting Sentences and Words. In *ACL*, 484–494. Berlin, Germany: ACL. doi:10.18653/v1/P16-1046.

Croasmun, J. T.; and Ostrom, L. 2011. Using Likert-Type Scales in the Social Sciences. *Journal of Adult Education* 40(1): 19–22.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Fabbri, A.; Li, I.; She, T.; Li, S.; and Radev, D. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. 1074–1084. Florence, Italy: ACL. doi:10.18653/v1/P19-1102.

Ganesan, K.; Zhai, C.; and Han, J. 2010. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 340–348. Beijing, China: Coling 2010 Organizing Committee. URL https://www.aclweb.org/anthology/C10-1039.

Geva, M.; Malmi, E.; Szpektor, I.; and Berant, J. 2019. DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion. *NAACL HLT* 3443–3455. doi:10.18653/v1/N19-1348.

Haghighi, A.; and Vanderwende, L. 2009. Exploring content models for multi-document summarization. In *NAACL HLT*, 362–370.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, 1693–1701.

Hoang, A.; Bosselut, A.; Celikyilmaz, A.; and Choi, Y. 2019. Efficient adaptation of pretrained transformers for abstractive summarization. *arXiv preprint arXiv:1906.00138* .

Keneshloo, Y.; Shi, T.; Ramakrishnan, N.; and Reddy, C. K. 2019. Deep reinforcement learning for sequence-to-sequence models. *IEEE Transactions on Neural Networks and Learning Systems* .

Koncel-Kedziorski, R.; Bekal, D.; Luan, Y.; Lapata, M.; and Hajishirzi, H. 2019a. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342* .

Koncel-Kedziorski, R.; Bekal, D.; Luan, Y.; Lapata, M.; and Hajishirzi, H. 2019b. Text Generation from Knowledge Graphs with Graph Transformers. In *NAACL HLT*, 2284–2293. Minneapolis, Minnesota: ACL. doi:10.18653/v1/N19-1238.

Kryściński, W.; McCann, B.; Xiong, C.; and Socher, R. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv* arXiv–1910.

Lebanoff, L.; Song, K.; and Liu, F. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. *arXiv preprint arXiv:1808.06218* .

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *EMNLP*, 107–117. Austin, Texas: ACL. doi:10.18653/v1/D16-1011.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* .

Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541* .

Li, Q. 2013. A novel Likert scale based on fuzzy sets theory. *Expert Systems with Applications* 40(5): 1609–1618.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* .

Liu, F.; Flanigan, J.; Thomson, S.; Sadeh, N.; and Smith, N. A. 2015. Toward Abstractive Summarization Using Semantic Representations. In *NAACL HLT*, 1077–1086. Denver, Colorado: ACL. doi:10.3115/v1/N15-1114.

Liu, P. J.; Saleh, M.; Pot, E.; Goodrich, B.; Sepassi, R.; Kaiser, L.; and Shazeer, N. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198* .

Liu, Y.; and Lapata, M. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* .

McKeown, K.; and Radev, D. R. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 74–82.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843* .

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

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745* .

Narayan, S.; Gardent, C.; Cohen, S. B.; and Shimorina, A. 2017. Split and Rephrase. *EMNLP* 606–616. doi:10.18653/v1/D17-1064.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 311–318. ACL.

Pasunuru, R.; and Bansal, M. 2018. Multi-Reward Reinforced Summarization with Saliency and Entailment. In *NAACL HLT*, 646–653. New Orleans, Louisiana: ACL. doi:10.18653/v1/N18-2102.

Paulus, R.; Xiong, C.; and Socher, R. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* .

Puduppully, R.; Dong, L.; and Lapata, M. 2019. Data-to-text generation with entity modeling. *arXiv preprint arXiv:1906.03221* .

Qazvinian, V.; and Radev, D. R. 2008. Scientific paper summarization using citation summary networks. *arXiv preprint arXiv:0807.1560* .

Qazvinian, V.; Radev, D. R.; Mohammad, S. M.; Dorr, B.; Zajic, D.; Whidby, M.; and Moon, T. 2013. Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research* 46: 165–201.

Radev, D. R.; and McKeown, K. R. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics* 24(3): 469–500.

Rao, S.; and Daumé III, H. 2019. Answer-based adversarial training for generating clarification questions. *arXiv preprint arXiv:1904.02281* .

Rush, A. M.; Chopra, S.; and Weston, J. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *EMNLP*, 379–389. Lisbon, Portugal: ACL. doi:10.18653/v1/D15-1044.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*, 1073–1083. Vancouver, Canada: ACL. doi:10.18653/v1/P17-1099.

Shah, D. J.; Schuster, T.; and Barzilay, R. 2019. Automatic Fact-guided Sentence Modification. *arXiv preprint arXiv:1909.13838* .

Sharma, E.; Huang, L.; Hu, Z.; and Wang, L. 2019. An Entity-Driven Framework for Abstractive Summarization. In *EMNLP-IJCNLP*, 3280–3291. Hong Kong, China: ACL. doi:10.18653/v1/D19-1323.

Shen, T.; Quach, V.; Barzilay, R.; and Jaakkola, T. 2020. Blank Language Models. *arXiv preprint arXiv:2002.03079* .

Stern, M.; Chan, W.; Kiros, J.; and Uszkoreit, J. 2019. Insertion transformer: Flexible sequence generation via insertion operations. *arXiv preprint arXiv:1902.03249* .

Sullivan, G. M.; and Artino Jr, A. R. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education* 5(4): 541–542.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wenbo, W.; Yang, G.; Heyan, H.; and Yuxiang, Z. 2019. Concept Pointer Network for Abstractive Summarization. *arXiv preprint arXiv:1910.08486* .

Wu, L.; Tian, F.; Qin, T.; Lai, J.; and Liu, T.-Y. 2018. A study of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1808.08866* .

Xu, P.; Wu, C.-S.; Madotto, A.; and Fung, P. 2019. Clickbait? Sensational Headline Generation with Auto-tuned Reinforcement Learning. *arXiv preprint arXiv:1909.03582* .

Yang, J.; and Zhang, Y. 2018. Ncrf++: An open-source neural sequence labeling toolkit. *arXiv preprint arXiv:1806.05626* .

Zhang, J.; Tan, J.; and Wan, X. 2018. Adapting Neural Single-Document Summarization Model for Abstractive Multi-Document Summarization: A Pilot Study. In *Proceedings of the 11th International Conference on Natural Language Generation*, 381–390. Tilburg University, The Netherlands: ACL. doi:10.18653/v1/W18-6545.

Zhang, S.; and Bansal, M. 2019. Addressing Semantic Drift in Question Generation for Semi-Supervised Question Answering. *arXiv preprint arXiv:1909.06356* .